MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Mestrado no Programa de Pós-Graduação em Biologia Computacional e Sistemas

PREDIÇÃO COMPUTACIONAL DAS INTERAÇÕES PROTEÍNA-PROTEÍNA EM ESPÉCIES DO GÊNERO *CRYPTOCOCCUS* SPP.

ELVIRA CYNTHIA ALVES HORÁCIO

Rio de Janeiro
Agosto de 2016

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

Elvira Cynthia Alves Horácio

PREDIÇÃO COMPUTACIONAL DAS INTERAÇÕES PROTEÍNA-PROTEÍNA EM ESPÉCIES DO GÊNERO *CRYPTOCOCCUS* SPP.

Dissertação apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Mestre em Ciências.

Orientador (es): Prof. Dr. Jeronimo C. Ruiz
Prof. Dra. Daniela M. Resende

RIO DE JANEIRO
Agosto de 2016

Ficha catalográfica elaborada pela
Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

H811 Horácio, Elvira Cynthia Alves

Predição computacional das interações proteína-proteína em espécies do gênero *Cryptococcus* spp. / Elvira Cynthia Alves Horácio. – Rio de Janeiro, 2016.
xvi, 137 f. : il. ; 30 cm.

Dissertação (Mestrado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2016.
Bibliografia: f. 63-72

1. Interações proteína-proteína. 2. *Cryptococcus* spp. 3. Redes. 4. Redes de interação. I. Título.

CDD 572.696

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

ELVIRA CYNTHIA ALVES HORÁCIO

PREDIÇÃO COMPUTACIONAL DAS INTERAÇÕES PROTEÍNA-PROTEÍNA EM ESPÉCIES DO GÊNERO *CRYPTOCOCCUS* SPP.

ORIENTADOR (ES): **Prof. Dr. Jeronimo C. Ruiz**
Prof. Dra. Daniela M. Resende

Aprovada em: _____/_____/_____

EXAMINADORES:

Prof. Dr. Ernesto Raul Caffarena **- Presidente** (IOC)
Profa. Dra. Cristiana Ferreiara Alves Brito (CPQRR)
Prof. Dr. Douglas Eduardo Valente Pires (CPQRR)
Prof. Dr. Gabriel da Rocha Fernandes (CPQRR)
Prof. Dr. Antonio Basilio de Miranda (IOC)

Rio de Janeiro, 29 de Agosto de 2016

Dedico a Deus, aos meus pais, familiares e amigos.

AGRADECIMENTOS

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES pelo auxílio financeiro.

Agradeço imensamente a Deus, por me dar proporcionar a sabedoria nos momentos de fraqueza, ser meu ater ego para me ajudar a refletir e procurar respotas, ao contrários de outros seres humanos que a resposta era quase sempre procure no google (como se eu já não tivesse perguntado e encontrado mais de uma resposta diferente).

Agradeço aos meus pais pelo suporte, aos meus familíares e amigos por compreenderem a minha ausência inclusive nos finais de semanas e feriados (trabalhar com grande jornada sem recesso de final de semana poderia parecer má vontade de particitar da família ou do ciclo de amizade), mas compreenderam que a minha ausencia era para o meu crescimento, meu muito obrigada.

Muito obrigada aos meus amigos "computeiros" que me auxiliaram nessa longa jornada, que as 3 horas da amanhã me orientava e respondia, aos domingos ou as noites que me disponibilizaram conversar via internet ou ate mesmo sedendo um tempo pessoalmente para que me ajudasse a compreender algo que eu não pude ter disciplinas ou alguem que pudesse me dar uma aula ou ceder mais que meia hora para me explicar de forma simplificada a questão.

Aos meus amigos e colegas, que encontrei durante essa jornada, meu muito obrigada, por me suportar em momentos tão complexos.

Aos meus diversos médicos, enfermeiros e auxiliares, que foram muitos ao longo desse tempo, mais do que eu tive a minha vida inteira, gostaria de agradecer por não errar tantas veias, por quando abrir o computador para trabalhar ameaçar retirar da tomada, por proporcionar um alivio de dor momentaneo, pelo "sabia que ia ver você aqui de novo" (como se fosse agradavel essa frase para uma pessoa que tem pavor de lugares hospitales). Vocês foram importante para relembrar a todo momento o quanto somos insignificantes perante esse universo, e que por isso, apesar de termos problemas, o nosso corpo e a nossa alma pode padecer, mas a forma como agimos e tratamos nossos semeslhantes não. As pessoas não tem culpa de como gostariamos de ser ou produzir.

Agradeço ao programa de Pós Graduação em Biologia Computacional e Sistemas por todo apoio e zelo desmostrados.

Por fim aos meus oritadores por tentar me oritentar e me mostrar a melhor e a pior versão de mim mesma.

Stand on the shoulders of giants.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

PREDIÇÃO COMPUTACIONAL DAS INTERAÇÕES PROTEÍNA-PROTEÍNA EM ESPÉCIES DO GÊNERO *CRYPTOCOCCUS* SPP.

RESUMO

DISSERTAÇÃO DE MESTRADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

ELVIRA CYNTHIA ALVES HORÁCIO

A biologia de sistemas procura modelar sistemas complexos, usufruindo de conteúdos interdisciplinares e possibilitando a integração da informação biológica disponível sobre determinado organismo. Nesse contexto, uma das abordagens integradas é a de redes, que pode ser definida como um conjunto de entidades ou nós que se conectam e que viabilizam a integração de dados. Um exemplo são as redes de interação proteína-proteína (PPI). Os organismos modelos utilizados neste estudo pertencem ao gênero *Cryptococcus* spp. Aproximadamente um milhão de casos de Criptococose ocorrem no mundo e estima-se a ocorrência de 400 mil mortes anualmente. O presente trabalho teve como objetivo o emprego dessa metodologia em espécies do gênero *Cryptococcus* spp. depositadas em bancos de dados de domínio público, visando à análise comparativa de redes associadas à patogenia e à virulência desses organismos.Um fator crucial nas análises genômicas comparativas é a anotação funcional dos genomas estudados. Por isso, este trabalho inclui também uma etapa inicial de reanotação e anotação desses genomas. Nosso recurso básico para a geração de redes PPI são os genes preditos e nesse contexto utilizamos 12 genomas de *Cryptococcus* spp. Dentre as técnicas computacionais para a predição de PPI utilizamos a abordagem de coevolução. Aplicamos duas estratégias distintas que envolvem a referida abordagem. Na primeira estratégia, realiza-se o alinhamento múltiplo global das proteínas a serem comparadas e avalia-se o valor de PI. Na segunda estratégia, realiza-se uma análise combinatória de sequências par-a-par oriundas do agrupamento original. Essa etapa é seguida pela execução de um novo alinhamento múltiplo global e a posterior avaliação do valor de PI. A partir daí, torna-se possível inferir as interações proteína-proteína do organismo em estudo. Como resultado, obtiveram-se 24 redes. Calculamos as estatísticas associadas às redes, além da localização dos principais *hubs*. O menor e o maior número de nós obtidos das redes do grupo utilizando o agrupamento variaram de 63 a 118 e as conexões (arestas) correspondentes foram de 393 a 476. No grupo de pares do agrupamento a variação foi de 149 a 2.465 e as conexões correspondentes variaram de 1.516 a 46.468. Construímos um banco relacional em MySQL com proteínas associadas à virulência e à patogenicidade de onde foram identificadas 226.065 proteínas associadas aos termos de busca. Na primeira estratégia utilizando o valor PI do agrupamento, encontramos poucas proteínas à virulência e à patogenicidade, mas na segunda estratégia, obtivemos de nem uma proteína associadas à virulência e à patogenicidade em um organismo que não atinge humanos a 91 proteínas. Dentre elas proteínas de *heat shock* e proteína de resistência a múltiplas drogas 1 (MDR1).

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

COMPUTATIONAL PREDICTION OF PROTEIN-PROTEIN INTERACTIONS IN THE SPECIES OF THE GENUS *CRYPTOCOCCUS* SPP

ABSTRACT

MASTER DISSERTATION IN COMPUTATIONAL BIOLOGY AND SYSTEMS

ELVIRA CYNTHIA ALVES HORÁCIO

The biology of systems search for complex models, taking advantage of interdisciplinary contents and enabling the integration of available biological information on a given organism. In this context, one of the integrated approaches is networks, which can be defined as a set of entities or nodes that connect and enable integration of data. An example is how protein-protein interaction networks (PPI). The models of the organisms used in this study belong to the genus *Cryptococcus* spp. Approximately one million cases of Cryptococcosis occur worldwide and an estimated 400,000 deaths are estimated annually. The present work had as objective the use of the methodology in species of the genus *Cryptococcus* spp. Deposited in public domain databases, aiming the comparative analysis of networks associated with pathogenesis and virulence of organisms. The crucial factor for comparative genomic analysis is a functional annotation of the studied genomes. Therefore, this work also includes an initial stage of reassessment and annotation of genomes. We used the predicted gene sequences to obtain the networks, for this purpose we sought 12 genomes of Cryptococcus spp. Among the computational techniques to predict PPI we used the coevolution approach. We developed and applied two different strategies involving this approach. In the first strategy, we carried out global multiple sequence alignments of proteins and evaluated the PI value. In the second strategy it was performed a pairwise combinatory analysis of sequences from the original group. This step was followed by the implementation of a new global multiple alignment and the subsequent assessment of the PI value. From there, it becomes possible to infer how protein-protein interactions of the organism under study. As a result, 24 networks were obtained. Calculations such as statistics associated with networks, as well as the location of the main hubs. The lowest and the highest number of nodes obtained in the cluster with the varied clustering from 63 to 118 and the corresponding connections (edges) were from 393 to 476. There is no comparison group of variables from 149 to 2465 and the corresponding connections varied from 1,516 to 46,468. We have a MySQL relational database with proteins associated with virulence and pathogenicity from which 226,065 proteins associated with search terms were identified. In the first use of value-added measures, there are few proteins for virulence and pathogenicity, they have a second strategy, they obtain a protein associated with virulence and they are pathogenic in a non-human body. Among them, thermal shock and multiple drug resistance protein 1 (MDR1).

ÍNDICE

ÍNDICE DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

NCBI	*National Center for Biotechnology Information*
CDD	*Conserved Domains Database*
nr	banco não redundante
nrS1	dataset1 banco não redundante
nrS2	dataset2 banco não redundante
DFVF	*Database of fungal virulence factors*
MLST	*Multi Locus Sequence Typing*
Ecdf	função da distribuição cumulativa empirica
GFF	*General feature format*
Embl	*European Molecular Biology Laboratory*
PERL	*Practical Extraction and Reporting Language*

1 INTRODUÇÃO

1.1 Biologia de Sistemas

Vivemos na atualidade um contexto de explosão da quantidade de dados biológicos gerados e disponíveis em bancos de dados de domínio público. Tendo como ponto alto o Projeto Genoma Humano, essa geração de dados em larga escala também foi catalisada pelos avanços técnicos e tecnológicos, que estão ligados às abordagens de sequenciamento e à bioinformática, vinculada às análises necessárias ao entendimento e contextualização da informação biológica contida nas sequências geradas.

Com o intuito de sistematizar e analisar esses dados, tornou-se essencial o desenvolvimento de metodologias de processamento e modelagem de sistemas biológicos visando a integração das informações geradas. Devido à essa crescente quantidade de dados, a utilização do termo *systems biology* (Biologia de Sistemas) em artigos aumentou substancialmente (PUJOL, MOSCA *et al*., 2010).

A ciência normalmente utiliza estratégias para compreender parte do sistema biológico. Os métodos mais comumente utilizados envolvem a separação das partes que compõem o objeto de interesse para estudá-las individualmente como parte do sistema global. Nos sistemas *bottom-up* usamos a abordagem reducionista e estudamos os componentes básicos e integramos esses para encontrarmos padrões e funções. Por exemplo, esses sistemas são utilizados na biologia para examinar os mecanismos por meio dos quais as propriedades funcionais surgem nas interações de componentes conhecidos e integrar as informações a cada passo, o sistema original se torna maior (PUJOL, MOSCA *et al*., 2010; KITANO, 2002).

Essa teoria, chamada reducionista, foi por muito tempo o principal mecanismo utilizado pela ciência para a explicação de fenômenos biológicos. Um exemplo que podemos citar é a compreensão da função de uma proteína dentro do sistema (SAUER, HEINEMANN *et al*., 2007) .Para entender os sistemas biológicos complexos, uma possível abordagem para uma integração da investigação experimental e computacional, em outras palavras, uma abordagem de biologia de sistemas (KITANO, 2002).

A Biologia de sistemas é uma área que atua em um campo interdisciplinar do conhecimento, associando diversos conteúdos e integrando muitas disciplinas científicas, como biologia, ciência da computação, engenharia,

bioinformática, física e outras. Esse estudo utiliza abordagens holísticas, isto é, que considera o todo, levando em conta as partes e suas inter-relações, e é aplicada para pesquisas biológicas e biomédicas.

Essa área estuda sistemas biológicos de algumas formas: perturbando-os sistematicamente (biologicamente, geneticamente, ou quimicamente); monitorando o gene, a proteína e as respostas das vias metabólicas; integrando os dados; e, finalmente, formulando modelos matemáticos que descrevem a estrutura do sistema e a sua resposta a perturbações individuais (IDEKER *et al*, 2001). Essa vertente da biologia identifica uma rede de interação com base no comportamento molecular, correlacionado-as com estudos genômicos.

Uma das possibilidades para se integrar dados em sistemas biológicos é por meio da utilização do método de redes, pois ele permite visualizar relações, principalmente em larga escala dinâmicas (XENARIOS, 2000).

1.2 Redes

Uma rede é um sinônimo de grafo, grafos são representados essencialmente por duas entidades, vértices e arestas. Os vértices, também conhecidos como nós, são elementos representativos dos componentes da rede. Por outro lado, as arestas representam as interações entre os vértices, como por exemplo conexões entre cidades e largura de banda da conexão entre dois computadores de uma rede (EASLEY E KLEINBERG, 2010).

Existem diversos tipos de rede, dentre elas podemos destacar as redes livres de escala que são redes complexas cujo grau de distribuição segue a lei de potência. Redes livre de escala tem a maioria dos vértices apresentando poucas ligações (arestas), contrastando com a existência de alguns nós que apresentam um elevado número de ligações. Uma das propriedades desse tipo de rede é chamada conexão preferencial que resumidamente pode ser descrita como a tendência de um novo vértice se conectar a um vértice da rede que tem um grau elevado, ou seja, muitas conexões (EASLEY E KLEINBERG, 2010; BARABASI E ALBERT, 1999).

As propriedades estatísticas das redes são importantes e devem ser avaliadas. Entre elas podemos destacar: a) Média do grau do número de grau, que representa uma média aritmética do número de arestas dividido pelo número de nós; e b) Mediana do grau é o ponto central que divide esse conjunto em dois subconjuntos com o mesmo número de elementos. Outras características

importantes a serem avaliadas são: a) os *hubs*, que são nós altamente conectados (com muitas arestas); e b) as comunidades (que são grupos de nós altamente conectadas entre si) podem ser detectadas através da utilização de diversos algoritmos que em sua maioria utiliza a informação de densidade de ligações entre os nós (EASLEY E KLEINBERG, 2010).

A célula de um organismo é um sistema complexo cujas características são definidas através da atividade de muitos componentes, que interagem uns com os outros através de interações. Um desses componente são as proteínas que por sua vez podem ser representadas por uma série de nós, os quais são conectados por ligações, que representam as interações. Para compreender melhor o funcionamento da rede algumas análises que podem ser utilizadas, dentre elas o *betweenness* e o *walktrap community* (PONS e LATAPY, 2005).

Betweenness (intermediação) considera um ator como meio para alcançar outros atores, visto que ele está posicionado nos caminhos geodésicos entre outros pares de atores na rede. Ele é um indicador de um determinado vértice pertencer ao centro do grafo que é obtida pela contagem de caminhos mínimos que passam por um determinado vértice, isto é, a medida da centralidade de um determinado nó. Um nó com alta centralidade tem uma grande influência sobre a transferência de itens através da rede, sob a suposição de que as transferências dos pontos seguem os caminhos mais curtos (EASLEY E KLEINBERG, 2010). Comunidades são grupos de nós que são densamente conectados entre si e esparsamente conectados com outros agrupamentos.

Dentre as análises que podem ser empregadas para verificar de que forma as redes se agrupam destaca-se a *walktrap community*, que é uma abordagem baseada em caminhadas aleatórias.

A ideia da abordagem geral da *walktrap community* é de que ao se realizar passeios aleatórios no grafo de forma iterativa, os caminhos têm a propensão de permanecer dentro da mesma comunidade, pois há poucas arestas que levam para fora dela. Deste modo, é possível observar redes que não possuem uma conexão com a rede central ou sub-redes.

Uma das formas de se evidenciar a estrutura é relacionar quão central um nó pode ser, ou a quantidade de conexões que o nó possui. Isto é feito através da utilização da função ecdf que é a função de probabilidade cumulativa empírica, em que a cada valor da lista será associada uma fração da quantidade de valores menores que o mesmo.

1.3 Interações proteína–proteína

Proteínas são polímeros de aminoácido ou macromoléculas, são altamente variadas e constituem uma grande parte da massa orgânica de cada forma de vida. As proteínas são necessárias à dieta de todos os animais e outros organismos não fotossíntetizantes. Tais macromoléculas podem ser compostas por até 20 ou mais tipos de aminoácidos diferentes que por sua vez são ligados numa sequência linear geneticamente controlada em uma ou mais cadeias polipeptídicas (BERG, 2008);.

As proteínas possuem funções especializadas, como: o colágeno, de tecido de suporte; a hemoglobina, para o transporte de gases; os anticorpos, para a defesa imune; e as enzimas, para o metabolismo. É importante salientar que essas moléculas são determinadas pela sequência genética de um gene, podendo sofrer alterações por fatores pré ou pós-traducionais (BERG, 2008; HARRINGTON, JENSEN *et al*. 2008).

As proteínas interagem entre si, o conjunto conhecido de interações proteína-proteína (PPI) para uma dada célula ou organismo constitui o interatoma, que por sua vez pode ser representado através de uma rede. Um aspecto relevante associado à construção das redes está ligado à fonte de origem dos dados, que pode ser teórica ou experimental. As abordagens experimentais são usualmente laboriosas, demandando tempo e recursos financeiros, e usualmente empregam técnicas como a co-imunoprecipitação, a cromatografia de afinidade, o duplo híbrido e a espectrometria de massas, sendo as duas últimas responsáveis pela detecção em larga escala de interações proteína-proteína (CHIEN *et al*., 1991; YOUNG, 1998; SAYEON CHO, 2004). Apesar da grande contribuição científica dessas abordagens, um aspecto que tem gerado enormes discussões é a taxa de falso-positivos que pode ser bastante elevada, gerando erros sistemáticos na detecção das interações (CHIEN *et al*., 1991, YOUNG, 1998).

Considerando os aspectos mencionados acima, uma alternativa viável em termos de diminuição de tempo e custos é a utilização das abordagens computacionais para a construção de redes de interação.

1.4 Predição Computacional de interações proteína-proteína

Existem várias abordagens de predição computacional de redes de interação proteína-proteína, dentre elas destacam-se:

A) Fusão gênica: método que identifica eventos de fusão de genes em genomas completos, baseado unicamente na comparação de sequências. Ele considera que deve haver pressão seletiva para certos genes se fundirem ao longo de evolução. Esta metodologia explora o fato de certas famílias de proteínas, numa determinada espécie, possuírem domínios fundidos que geralmente correspondem a única proteína em outras espécies, indicando que é possível predizer interações, desde que se tenha a sequência (ENRIGHT *et al*., 1999).

B) Interação filogenética: Perfis de sequência semelhantes mostram um padrão correlacionado de herança e, por implicação, uma ligação funcional. Este método prevê que as funções das proteínas são descaracterizadas susceptíveis de ser semelhante a proteínas caracterizadas dentro de um *cluster.* A técnica apoia na ideia de que proteínas com perfis semelhantes são susceptíveis de ser funcionalmente ligado. A técnica utiliza o agrupamento das sequências e correlaciona com suas funções, indicando que é possível relacionar as proteínas através de um perfil filogenético, trazendo para as interações das proteínas (HUYNEN E BORK, 1998; PELLEGRINI *et al*., 1999; REZENDE *et al*, 2012).

C) Vizinhança genômica: pode ser vista como uma extensão da abordagem do perfil filogenético, pois, além da procura de determinados genes co-herdados, ela busca a tendência desses genes permanecerem agrupados no genoma. Dessa forma, pode haver uma pressão seletiva para esses genes ficarem próximos no genoma, indicando uma relação funcional através da qual as interações podem ser preditas (HARRINGTON, JENSEN *et al*., 2008, REZENDE, 2012).

D) Coevolução: assume que proteínas que interagem possuem árvores filogenéticas moleculares similares, pois existe uma coevolução mantida pela interação. Neste caso, as taxas evolutivas das proteínas são comparadas para selecionar os pares de interação (SATO,YAMANISHI *et al*., 2005). A coevolução considera que as proteínas evoluem a uma taxa próxima, possuindo dessa forma uma alta similaridade. Essa foi abordagem de predição de redes adotada no presente trabalho.

E) Homologia de sequência contra banco de dados de interação: as sequências de interesse são comparadas contra interações proteína-proteína dos bancos de dados de interações (MATTHEWS, VAGLIO *et al*., 2001).

1.5 *Cryptococcus* spp.

O gênero *Cryptococcus* inclui mais de 50 espécies de leveduras encapsuladas do Filo Basidiomycota que apresentam tamanho entre 5 a 10 µm. Dentre as espécies, *Cryptococcus neoformans* e *Cryptococcus gattii* são importantes agentes etiológicos da Criptococose, uma micose sistêmica que ocorre no homem e em animais, hígidos ou imunossuprimidos (CASADEVALL, 1998).

O trato respiratório superior representa a provável porta de entrada do fungo, que é alcançado pela inalação dos propágulos infectantes (GRIFFITHS, KRETSCHMER *et al*, 2012). Apesar de acometer tanto animais como humanos, a transmissão direta do micro-organismo entre humanos ou de animais para humanos ainda não foi documentada (CHAYAKULKEEREE E PERFECT, 2006).

Atualmente, as espécies de *Cryptococcus* spp. são classificadas em cinco diferentes sorotipos. A classificação é baseada nas reações de aglutinação dos antígenos polissacarídeos presentes em sua cápsula proteica. Nos sorotipos A, D e híbrido AD estão incluídas as cepas da espécie *C. neoformans*; enquanto para as cepas de *C. gattii* são atribuídos os sorotipos B e C (SEVERO *et al*., 2009).

As descrições referentes à virulência e à patogenicidade no gênero normalmente provêm das espécies *C. neoformans* e *C. gattii*. Essas espécies compartilham boa parte dos fatores de virulência, tornando-as ainda mais interessantes, pelo fato de o perfil dos indivíduos infectados serem diferentes. *C. neoformans* geralmente afeta indivíduos imunocomprometidos, enquanto *C. gattii* está mais associado a infecções em indivíduos hígidos (KROSNTADET *et al*., 2011, CALVO, 2010; PAPPAS, 2001).

Um fator de virulência altamente descrito na literatura é a cápsula polissacarídica, onde as principais funções a mesma são o aumento da capacidade de invasão, além de proteger a levedura do seu ambiente externo e do hospedeiro quando o infecta (CASADEVAL *et al*., 2009; GULLO, 2013; FERNANDES, 2008).

2 JUSTIFICATIVA

As proteínas participam de diversas funções biológicas, podendo atuar como enzimas catalizadoras de reações químicas, proteínas transportadoras e de armazenamento, proteínas contráteis, proteínas estruturais, proteínas reguladoras e outras funções. Sabe-se que uma proteína pode interagir com outra, desempenhando funções diferentes. Nesse sentido, estudar as redes de interação proteína-proteína possibilita a compreensão das interações do organismo por meio de uma visão sistêmica dos processos biológicos que ocorrem dentro do organismo.

O estudo integrado de um sistema biológico proporciona os meios que permitem avaliar o papel sistêmico uma determinada proteína. Entre as caracterictas que podem ser estudadas destaca-se: a sua função e atuação e a possível identificação de potenciais alvos de drogas ou vacinas relacionados a ela.

Como exemplo, pode-se destacar o tratamento de uma determinada patogenia e a identificação de biomarcadores. De fato, atualmente os biomarcadores vêm sendo utilizados e considerados como fatores de grande importância no tratamento do câncer (PUJANA, HAN *et al*, 2007).

A predição computacional de redes de interação proteína-proteína representa uma abordagem atual, rápida e de baixo custo para o estudo de interações em um dado organismo.

Neste trabalho, utilizou-se, como organismo modelo, espécies do gênero *Cryptococcus*. Ressalta-se que aproximadamente um milhão de casos de criptococose ocorrem no mundo, causando grande quantidade de óbitos (WHO, 2009), principalmente em indivíduos imunosuprimidos, sendo um dos principais responsáveis pela mortalidade de pacientes portadores do vírus HIV. Nesse sentido, o objetivo primordial deste trabalho envolveu a análise comparativa das redes de interação proteína-proteína vinculadas às espécies de *Cryptococcus* e suas diferenças patogênicas.

3 OBJETIVOS

3.1 Objetivo Geral

Predizer computacionalmente as redes de interação proteína-proteína para espécies do gênero *Cryptococcus* spp. depositadas em bancos de dados de domínio público, visando à análise comparativa de *clusters* associados à patogenia e virulência desses organismos.

3.2 Objetivo específicos

3.2.1 Realizar a predição de redes de interação proteína-proteína para 12 genomas do gênero *Cryptococcus* spp. depositados em bancos de dados de domínio público;
3.2.2 Empregar a abordagem preditiva associada à identificação de interação proteína-proteína por meio da metodologia de coevolução;
3.2.3 Comparar as redes identificando proteínas descritas na literatura relacionadas à patogenicidade nas espécies de *Cryptococcus* spp.;
3.2.4 Identificar nas redes preditas *clusters* (subredes) as proteínas associadas ao processo de virulência e patogenicidade do organismo.

4 MATERIAL E MÉTODOS

Os aspectos técnicos e metodológicos do presente trabalho serão descritos no decorrer desse captítulo. Os tópicos aresentados passam desde a anotação do genoma, ao banco de interação proteína-proteína ate a construção das redes e a procura da proteínas associadas a virulência e a patogenicidade.

4.1 Genomas

No presente trabalho analisamos os 12 genomas abaixo detalhados (tabela 1):

Tabela 1: Nome do *Cryptococcus* utilizado, sua linhagem e o repositório de origem.
A tabela mostra a espécie do genoma a linhagem e o nome do banco onde foi retirado a sequência.

Tabela 1: Nome e linhagens das espécies de *Cryptococcus*

Espécie	**Linhagem**	**Repositório**
C. bestiolae	CBS 10118	NCBI
C. dejecticola	CBS 10117	NCBI
C. flavescens	NRRL Y-50378	NCBI
C. gattii	CBS 7750	NCBI
C. gattii	R265	NCBI
C. gattii	WM276	NCBI
C. neoformans* var. *grubii	H99	NCBI
C. neoformans* var. *neoformans	B-3501ª	NCBI
C. neoformans* var. *neoformans	JEC21	NCBI
C. pinus	CBS 10737	NCBI
C. heveanensis	BCC8398	Broad
C. heveanensis	CBS569	Broad

4.2 Normalização dos formatos das sequências genômicas

Um dos problemas cotidianos quando se trabalha com diferentes bancos de dados de domínio público relaciona-se à existência de diferentes formatos de apresentação de sequências biológicas.

Durante o desenvolvimento deste projeto os dados provenientes de diferentes bancos (BROAD e NCBI) apresentaram diferentes formatos de anotação.

O resultando foi a necessidade de padronizar um formato que viabilizasse a anotação e a reanotação genômica, assim como as análises comparativas entre os dados.

O formato de escolha foi EMBL (ftp://ftp.ebi.ac.uk/pub/databases/embl/doc/usrman.txt) e assim sendo utilizou-se o programa readseq.jar (http://www.ebi.ac.uk/Tools/sfc/readseq/) para a conversão dos formatos originais (gbk, fasta, gff e etc) disponibilizados pelos bancos de dados no referido formato.
Como o organismo modelo possui em média 14 cromossomos e, consequentemente, para cada genoma seriam encontrados 14 arquivos no formato EMBL, utilizamos um *script* (*psu_union.pl*) desenvolvido pelo grupo para a união desses cromossomos em um único arquivo que, a partir desse momento, será chamado de pseudomolécula genômica.

4.3 Anotação dos genomas

A anotação dos genomas é uma etapa crucial para o desenvolvimento do trabalho, uma vez que os dados obtidos por meio da anotação forneceram as informações sobre os genes e, consequentemente, as proteínas contidas nos genomas, dando um significado biológico para os dados obtidos. Abaixo, encontra-se um fluxograma (figura 1) referente ao processo de anotação dos genomas que foi realizado junto com a aluna Grace Santos Tavares.

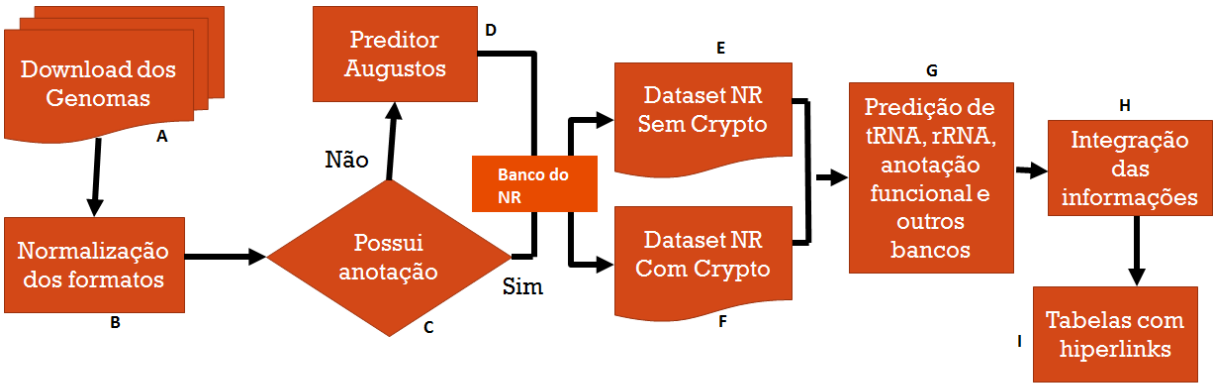

Figura 1: Fluxograma do processo de anotação referente aos genomas do gênero *Cryptococcus*. a) Obtenção dos dados genômicos do NCBI e Broad; b) conversão dos formatos encontrados para o formato EMBL através do programa readseq.jar; c) Verificar de existe anotação gênica ou não; d) Caso o genoma não tenha anotação, realiza-se a anotação gênica através do preditor Augustus (STANKE *et al*, 2004); e) Realização da análise de similaridade de sequência com o BLAST contra o *dataset* do NR sem *Cryptococcus*; f) Realização da análise de similaridade de sequência com o BLAST contra o *dataset* do NR com *Cryptococcus*; g) Predição de tRNA, rRNA, anotação funcional com o Interproscan e a realização do BLAST contra outros bancos; h) Tabelas com hiperlinks.

4.4 Predição gênica *ab initio* nos genomas

Para os genomas que possuíam anotação estrutural foi realizada a predição gênica *ab initio* para todos, porém utilizou-se somente as dos 8 genomas que não continham nem uma informação gênica. Para tanto utilizamos o programa Augustus. Dentre o conjunto de dados utilizados para o treinamento do programa existem dois organismos do gênero *Cryptococcus*: o *Cryptococcus neoformans gattii* e o *Cryptococcus neoformans neoformans* (STANKE *et al*, 2004).

O formato de saída do programa é GFF (General Feature Format) e para convertê-lo ao formato EMBL desenvolvemos um *script* na linguagem de programação PERL.

Posteriormente, verificou-se o *start codon* e *stop codon* dos genes obtidos por meio da predição. Essa etapa foi realizada pelo programa Artemis (RUTHERFORD *et al*., 2000, CAVER *et al*, 2005), que possui como uma de suas funções a capacidade de selecionar para a metionina (*start codon*) inicial e estender até o próximo *stop codon*.

4.5 Anotação por similaridade de sequências

Os genes preditos pelo programa Augustus foram posteriormente submetidos à anotação funcional através de buscas por similaridade de sequências.

O programa utilizado para busca por similaridade de sequências foi o algoritmo BLAST (ALTSCHUL *et al*, 1990). Utilizamos quatro diferentes bancos de dados locais. Resumidamente: a) *subset* do NR (banco de dados não redundante do NCBI), somente com os organismos do gênero *Cryptococcus*; b) o *subset* do NR sem os organismos do gênero *Cryptococcus*; c) Swisprot (http://www.ebi.ac.uk/uniprot) e d) CDD (Conserved Domains Database) (http://www.ncbi.nlm.nih.gov/cdd); Os resultados gerados foram formatados de tal forma a serem integrados na anotação dos genomas analisados.

O processo de anotação gênica por similaridade de sequências envolve a comparação dos genes preditos na etapa *ab initio* contra quatro bancos de dados específicos, sendo eles: swissprot, CDD, nrS1 (subset do banco nr contendo apenas as sequências de *Cryptococcus*) e nrS2 (subset do banco nr em que as sequências de *Cryptococcus* foram removidas).

A utilização do banco swissprot deve-se ao fato de que esse banco possui uma curadoria maior comparada ao nr por exemplo. Trata-se de um banco de dados de sequência de proteína curadas que se esforça para fornecer um alto nível de anotação (tal como a descrição da função de uma proteína, a sua estrutura de domínio, modificações pós-traducionais, variantes, etc), um nível mínimo de redundância e um elevado nível de integração com outras bases de dados.

O CDD é um recurso de anotação proteína que consiste de uma coleção de bem anotadas vários modelos de alinhamento de sequências para domínios antigos e proteínas de comprimento total (MARCHLER-BAUER *et al*, 2003).

Utilizamos *scripts* em PERL criados pelo Grupo Informática de Biossistemas; ou modificou-se outros *scripts* visando a integração dos arquivos gerados através do programa Artemis (RUTHERFORD *et al*., 2000)..

4.6 Predição de RNA transportador e RNA ribossomal

A predição de RNA transportador (tRNA) foi realizada com o algoritmo tRNAscan-SE (LOWE, EDDY,1999) (http://selab.janelia.org/tRNAscan-SE/). Tal

algoritmo identifica genes de RNA de transportador na sequência do DNA, utilizando um sistema baseado em regras hierárquicas, nas quais cada tRNA potencial deve exceder determinados limiares de similaridade por dois promotores intragênicos.

Para a predição de RNA ribossomal (rRNA), utilizou-se o algoritmo RNAmmer que prediz as regiões 5/8S, 16/18S, 23/28S rRNA ribossomal em sequências genômicas completas. O programa utiliza modelos ocultos de Markov para encontrar sequências que são suficientemente semelhantes a outros rRNAs dentro desse reino taxonômico. Ambos os programas foram executados localmente. Para a integração das predições realizadas pelos programas *scripts* PERL foram desenvolvidos ou adaptados.

4.7 Anotação funcional

O processo de anotação dos genomas também incluiu a integração das assinaturas de domínios funcionais conservados nos genes mapeados. Nessa etapa utilizamos o programa Interproscan (https://www.ebi.ac.uk/interpro/search/sequence-search).

Os resultados de anotação funcional gerados para cada um dos genes dos genomas analisados foram formata do através do *script ipro2tab.pl* que tem como objetivo transpor todos os resultados de uma forma que seja possível para uma posterior integração na anotação genômica.

4.8 Visualização e curadoria manual de genomas

Utilizamos como ferramenta de anotação o programa Artemis (http://www.sanger.ac.uk/Software/Artemis) (RUTHERFORD *et al*., 2000).

Todos os resultados gerados nas diferentes etapas de anotação foram integrados à pseudomolécula dos genomas através de *scripts* PERL e analisados usando o referido programa.

4.9 Seleção de sequências associadas à Patogenicidade e Virulência

A seleção de sequências associadas à patogenicidade e virulência do organismo modelo de estudo foi realizada utilizando duas formas de buscas, a primeira em bancos de sequências que estão associados à patogenicidade e à

virulência, e a segunda utilizando buscas lexicas no NCBI (National Center for Biotechnology Information) (http://www.ncbi.nlm.nih.gov/).

Especificamente utilizamos os seguintes bancos: a) MLST ("Multi-Locus Sequence Typing") (http://mlst.mycologylab.org/), usa análise de sequência parcial de sete a 10 genes, para uma abordagem de uma investigações epidemiológicas de microrganismos. Os dados gerados podem ser usados, para determinar se os isolados de fungos são clonais ou sofreram recombinação. Mas utilizamos esse banco pois existe sequências relacionadas à virulência e à patogenicidade. Por esse motivo existem sequências. b) FungiDB (http://FungiDB.org) que é um recurso da genômica funcional para genomas pan-fúngicos que foi desenvolvido em parceria com o centro de recursos eucarióticas Pathogen Bioinformatic (http://EuPathDB.org). FungiDB usa a mesma interface de utilizador e a infra-estrutura como EuPathDB, que permite pesquisas sofisticadas e integradas para ser realizadas utilizando um sistema gráfico intuitivo. A versão atual do FungiDB contém sequência do genoma e anotação de 18 espécies, abrangendo diversas classes de fungos, incluindo as classes Ascomycota, eurotiomycetes, sordariomycetes, Saccharomycetes e as ordens Basidiomycota, pucciniomycetes e tremellomycetes, e linhagem basais 'Zygomycete' Mucormycotina. c) A intenção do DFVD, database of fungal virulence factors (http://sysbio.unl.edu/DFVF/) é construir um banco de dados abrangente com todos os fatores de virulência de fungos conhecidos e desenvolver novos algoritmo para prever fatores de virulência putativos para um determinado patógeno. O projeto proposto visa preencher esta lacuna, criando uma plataforma central que irá estimular e facilitar estudos futuros no banco por fungos patogênicos. O banco de dados de sequência tem a pretensão de estimular fortemente e facilitar novos estudos em fungos patogênicos; ambos os biólogos experimentais e biólogos computacionais podem usar o banco de dados e/ou os fatores de virulência previstos para orientar a sua busca por novos fatores de virulência e/ou descoberta de novos mecanismos de interação patógeno-hospedeiro em fungos.

Os termos utilizados para a pesquisa no buscador do NCBI foram *Pathogenic and Cryptococcus* categoria gene; *Pathogenic and Cryptococcus categoria protein; Pathogen and Cryptococcus* categoria *gene; Pathogenic and Cryptococcus* categoria *protein; Pathogenesis and Cryptococcus* categoria *gene; Pathogenesis and Cryptococcus* categoria *protein; Pathogenicity and Cryptococcus categoria gene; Pathogenicity and Cryptococcus* categoria *protein; Pathogenes and*

Cryptococcus categoria gene; Pathogenes and Cryptococcus categoria *protein; Pathogenesises and Cryptococcus* categoria *protein; Virulence and Cryptococcus*.

Posterior a essa seleção realizamos uma procura por similaridade de sequência utilizando programa BLAST, contra cada genoma de interesse (nesse caso 12 genomas), com o intuito de descobrir quais proteínas estão associadas com virulência ou patogenicidade e que foram preditas nas redes utilizando a metodologia de coevolução por agrupamento ou par a par. Possibilitando então a identificação da mesma dentro da rede. Além disso, outro processo de agrupamento de sequências utilizando o programa ORTHOMCL, foi realizado com as sequências encontradas, a fim de verificar quais dessas sequências poderiam ser as mesmas ou que poderiam ser agrupadas.

4.10 Predição computacional de interações proteína–proteína: coevolução

O processo geral de predição computacional das interações proteína-proteína utilizando a coevolução, empregado nesse estudo, está descrito na figura 2.

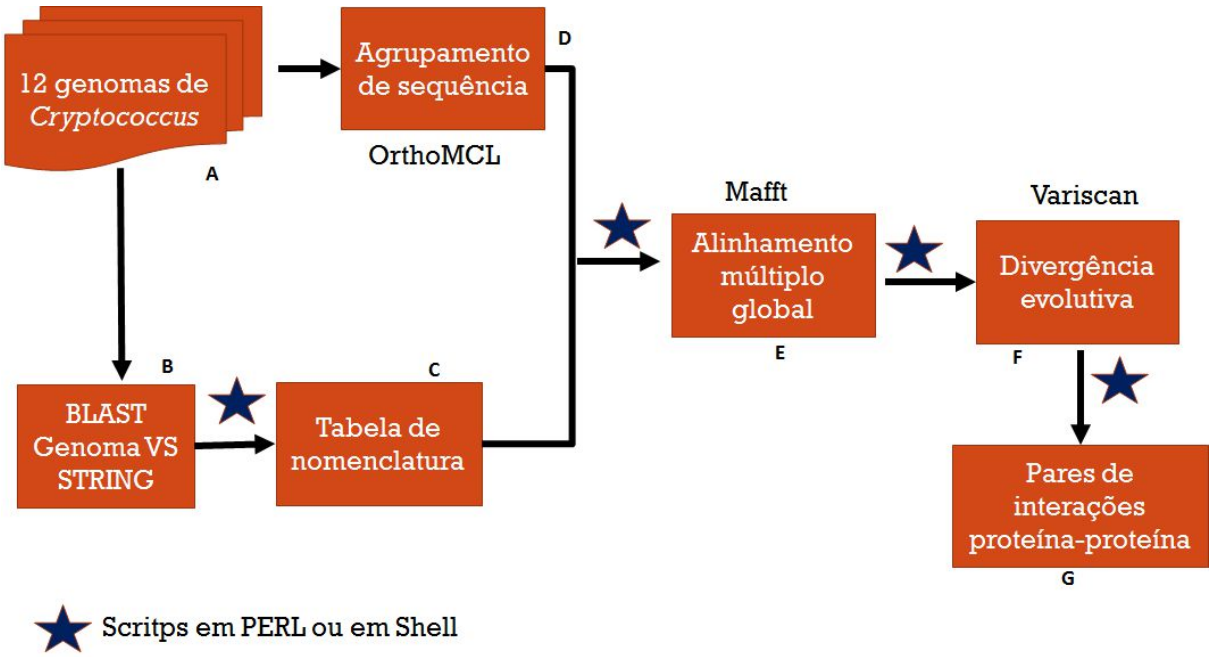

Figura 2: Fluxograma do processo de predição computacional de interações proteína–proteína por coevolução. A) Obtenção dos genes de 12 genomas de *Cryptococcus*; B) Análise de similaridade de sequência utilizando o BLAST utilizando o STRING. C) Através de diversos *scritps* cria-se uma tabela de nomenclatura onde contém os pares de interações; D) O agrupamento das sequências através do ORTHOMCL; E) A realização do Alinhamento múltiplo global através do programa Mafft; F) Cálculo da taxa de divergência evolutiva utilizando o Variscan. G) Obtenção dos pares de evolução por metodologia de coevolução.

O processo de predição de interações proteína-proteína se inicia através da análise de agrupamento das sequências. Para isto, utilizou-se programa OrthoMCL (LI, *et al*., 2003). Este realiza um BLAST de todos os genes dos genomas contra todos genes dos genomas. Como resultado o programa fornece um arquivo onde agrupa os genes, baseado na similaridade de sequência formando um agrupamento, uma vez que genes ortólogos são derivados de um ancestral comum, através de eventos de especiação (GABALDÓN *et al*., 2009).

Na próxima etapa, utilizou-se um algoritmo que dispunha do alinhamento múltiplo global, que nesse caso o programa escolhido foi Mafft (KATOH *et al*., 2005), que, dentre suas atribuições, possui essa característica. Os formatos de saída que o Mafft oferece não são compatíveis com o utilizado pelo Variscan (VILELLA, BLANCO-GARCIA *et al*., 2005). Por esse motivo, tornou-se necessária a conversão dos formatos.

No programa Variscan que de modo geral é um pacote de *software* para a análise de polimorfismos de sequências de DNA em escala genômica, um dos seus parâmetros fornece o cálculo de PI das sequências de entrada referente ao alinhamento, onde PI pode ser compreendido como uma medida sobre quantidade de variação nucleotídica das sequências.

Em todo o processo de análise, entre cada etapa aqui descrita, se fez necessário desenvolver *scripts* em PERL ou em Shell, devido à alta demanda de processamento de informações, assim como a execução de diversas entradas em um programa.

4.11 Banco de dados de interações proteína-proteína

O banco de dados de interações proteína-proteína utilizado nesse estudo foi o STRING (VON MERING, JAEGGI *et al*., 2003, JESEN *et al*, 2009). O STRING é um banco de interações que descreve associações diretas (físicas) e indiretas (funcionais), que por sua vez são derivadas de dados genômicos, experimentos de alto rendimento, coexpressão (conservação) e mineração de dados.

O banco possui três organismos correspondentes ao gênero de estudo, são eles: a) *Cryptococcus neoformans* var. *neoformans* JEC21; b) *Cryptococcus neoformans* var. *neoformans* B-3501A; e c) *Cryptococcus gattii*.

Desses dados selecionamos apenas as interações que continham evidência experimental das interações.

4.12 Construção das Tabelas de nomenclatura de interação

O termo tabela de nomenclatura refere-se aos pares de interação proteína-proteína do genoma de estudo obtidos através de um *hit* bidirecional com proteínas do banco do STRING. Esse processo será detalhado a seguir.

A predição computacional de interações proteína-proteína passa por uma etapa de similaridade de sequência entre as proteínas. Um programa amplamente utilizado para essa função e que tem como característica um processamento de alta vazão é o BLAST. No presente estudo utilizou-se o banco de dado STRING como repositório de informações associadas às interações proteína-proteína.

A identificação das redes foi realizada através de busca por similaridade de sequências e os critérios de corte empregados foram: e-value 0.00005, os 50 primeiros *hits*.

A quantidade de dados gerados pelo resultado do BLAST pode-se tornar enorme, de acordo com os parâmetros que são disponibilizados pelo programa. Devido a isto estabelecemos parâmetros de cortes após o resultado, no intuito de obter uma maneira de eliminar dados que não fossem de interesse, estabelecemos que dos resultados gerados seriam eliminadas as sequências menores que 100 aminoácidos e menores que 80% de identidade de sequência, com a intenção de obter uma melhor cobertura das proteínas encontradas.

Outra informação relevante a ser ressaltada é que o banco STRING não contém somente predições computacionais, existem também informações experimentais comprovando o contato físico entre duas proteínas através dos diversos métodos de detecção. Por esse motivo foi necessário filtrar a tabela onde estão correlacionadas as interações proteína-proteína do organismo de *Cryptococcus*, para conter somente dados referentes a métodos experimentais.

Através dos resultados da comparação do genoma de interesse com o banco de dados do STRING, obtidos pelo BLAST, desenvolveu-se um *script* que cria uma tabela inicial relacionando o identificador utilizado pelo STRING e o identificador utilizado pelo BLAST, na qual intitulou-se de inicial de tabela de nomenclatura. Essa tabela apresenta a relação da nomenclatura dos pares de

interações encontrado no STRING, com a nomenclatura das proteínas encontradas do nosso genoma, ela é demonstrada na tabela 2, onde observamos os pares e seus correspondentes, sendo os espaços vazios, devido aos fatos de não encontrar a proteína correspondente.

Tabela 2: Parte de uma tabela de nomenclatura de interação para *C. flavenses*.
Os campos vazios denotam situações onde não foi possível a identificação dos pares de interação proteína-proteína. As colunas 1 e 2 correspondem aos pares de interação depositados no STRINGdb e as colunas 3 e 4 identificam as proteínas encontradas no genoma em questão.

Identificador A do gene - STRINGdb	Identificador B do gene - STRINGdb	Identificador A do gene do Genoma 1	Identificador B do gene do Genoma 1
214684.XP_567629.1	214684.XP_570686.1	cflav.003990	
214684.XP_567629.1	214684.XP_570916.1	cflav.003990	
214684.XP_567629.1	214684.XP_572266.1	cflav.003990	
214684.XP_567656.1	214684.XP_567862.1		cflav.076960
214684.XP_567673.1	214684.XP_570420.1	cflav.071530	
214684.XP_567824.1	214684.XP_567049.1		cflav.004570
214684.XP_567862.1	214684.XP_567656.1	cflav.076960	
214684.XP_567862.1	214684.XP_568353.1	cflav.076960	
214684.XP_567862.1	214684.XP_571804.1	cflav.076960	
214684.XP_567862.1	214684.XP_572160.1	cflav.076960	
214684.XP_567862.1	214684.XP_572598.1	cflav.076960	
214684.XP_567939.1	214684.XP_570895.1		cflav.044900
214684.XP_567964.1	214684.XP_568462.1		cflav.004190
214684.XP_568101.1	214684.XP_569650.1		cflav.048540
214684.XP_566981.1	214684.XP_569167.1	cflav.018640	cflav.043150
214684.XP_566981.1	214684.XP_569827.1	cflav.018640	cflav.032020
214684.XP_566981.1	214684.XP_572717.1	cflav.018640	cflav.005630
214684.XP_567364.1	214684.XP_566981.1	cflav.068760	cflav.018640
214684.XP_567364.1	214684.XP_569827.1	cflav.068760	cflav.032020

O *hit* unidirecional do BLAST não é suficiente para definir o par de interação. É necessária a existência de *hit* bidirecional. Esse é o pre-requisito para definição de par de interação. Para resolver o problema associado à correta identificação dos pares de interação, fizemos um pipeline descrito na figura 3.

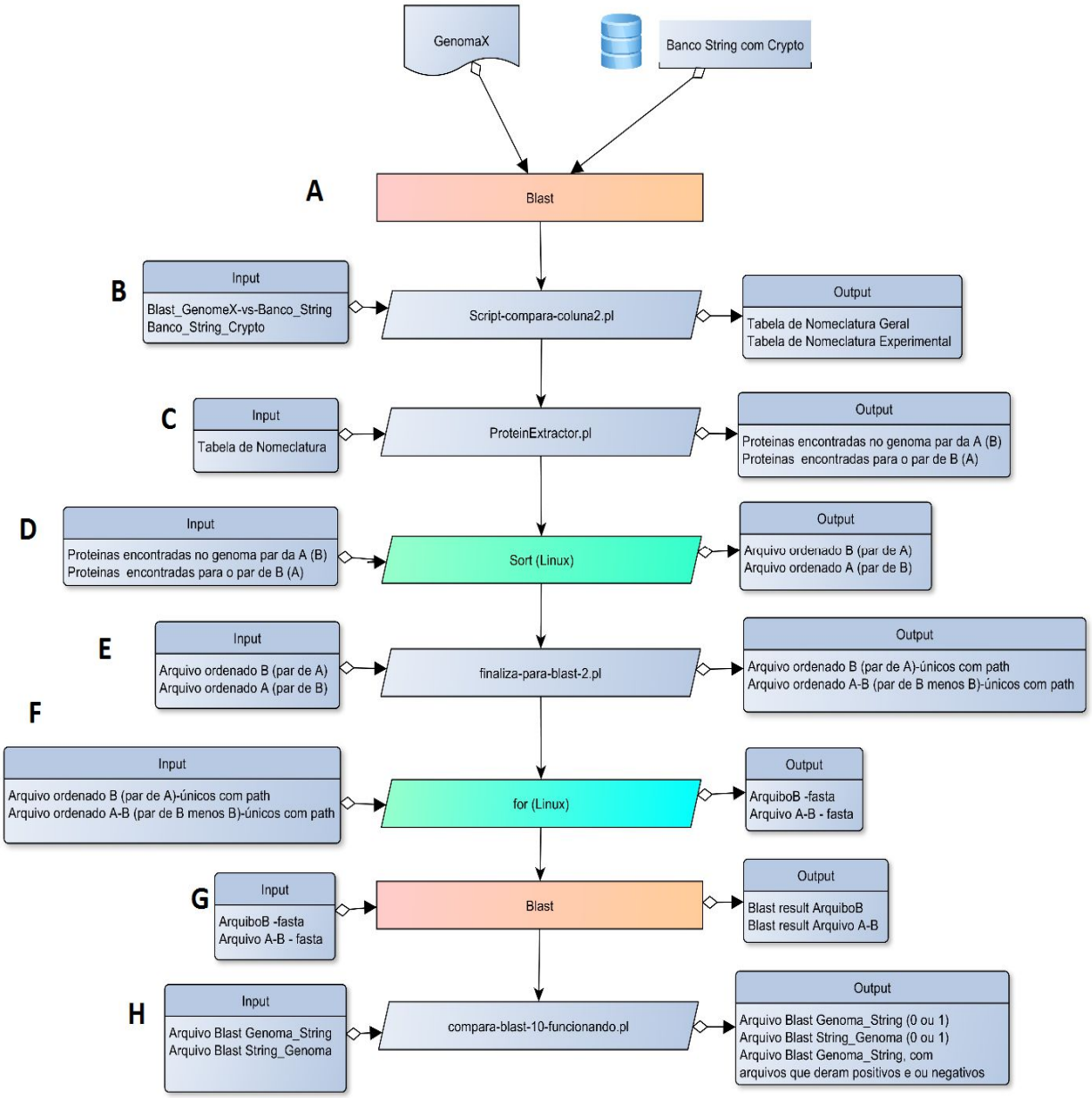

Figura 3: Fluxograma com as etapas para verificar se é um *hit* bidirecional dos genes. A) Executação do BLAST do genoma utilizando o STRING como banco de dados. B) Comparação dos resultados obtidos do BLAST com a Tabela de interação proteína-proteían do STRING; C) Extração de todas as proteínas da tabela contendo as relações dos pares ; D) O *Sort* facilita processamento por ser em ordem de pesquisa, otimizando o tempo; E) Busca do path do arquivo fasta de acordo com a proteína indicada na tabela; F) Retira essas sequências do diretório e concatena em dois arquivos; G) Realizaa o BLAST reciproco das sequências obtidas do STRING utilizando como banco as sequências do genoma; H) Obtem uma tabela com o *hit* bidirecional além de dar um arquivo com 0 ou 1, para indicar se o *hit* foi bidirecional 0 ou não 1.

No fluxograma é possível observar as etapas gerais. O primeiro elemento de cada nível de execução do fluxograma indica o arquivo de entrada (*input*), o elemento entre as caixas azuis é o *script* ou programa para execução e o último elemento é o arquivo de saída (*output*). Desta forma realizou-se o processo que passa pela etapa de criação da Tabela de Nomenclatura, com o *script Proteinextractor.pl*, ele extrai os nomes das proteínas e gera uma lista, retirando elementos que estão repetidos. Após este processo utilizou-se um programa de uma linha em Shell com intuito de ordenar a coluna referente a nomenclatura, através disso é possível ganhar tempo na consulta destas colunas.

Na próxima etapa criamos um *script* para obter o caminho do diretório onde se encontra o arquivo em formato fasta separado. Com um programa de uma linha em Shell, extraiu-se os arquivos correspondentes ao caminho, estes arquivos fastas foram obtidos do STRING utilizando como o arquivo de entrada para o BLAST. Desta forma obteve-se todos os arquivos para o BLAST. Posteriormente um *script* para comparar as duas tabelas e gerar uma nova tabela, dando a relação de reciprocidade com 0 como positivo ou 1 como negativo para a relação. Isto é, se o identificador se encontrava em ambas as tabelas 0, caso contrário 1, além disso gera a tabela de nomenclatura que corresponde as proteínas que estão no genoma de interesse que corresponde ás proteínas encontradas no STRING .

A intenção de empregar essa forma se dá pelo motivo da possibilidade de agregar as interações de outros organismos, pois elas podem ter uma alta identidade contra a sequência de interesse, inferindo uma possível interação que não foi descrita para o genoma de trabalho, e que não possuem evidências experimentais descritas para *Cryptocoocus* . Pois o que observa-se na literatura é um problema na geração de dados experimentais, devido ao fato da análise experimental demandar uma grande quantidade de tempo e recurso (ENRIGHT, 1999; PUJANA, 2007; KROETZ, 2009).

Seguindo esta sequência de etapas pode-se incluir como perspectiva em nossas análises a viabilidade de proteínas ainda preservadas, descritas em outros grupos, que estejam contidas em alguns dos nossos genomas e que possuem uma grande identidade dentre outras características. Esta análise se torna inviável da forma tradicional, que seria fazer um BLAST de forma reciproca, uma vez que atualmente os dados provenientes do banco possuam mais de 9.500.000 sequências. Estas etapas nos permitem encontrar sequências especificas dentro do conjunto de sequências, possibilitando então agregar interações, e avaliando se

esse ganho de interações é viável, pois foi observado em alguns casos que o melhor *hit* do BLAST das proteínas contra o banco geral, foi um alinhamento que não correspondeu ao organismo de *Cryptococcus* spp.

4.13 Validação dos pares de interação (Verificação de Melhor *hit*)

O processo de validação dos pares de interação envolvem a verificação do melhor *hit* recíproco entre as duas proteínas que compõem o par. O processo envolve a identificação dos genes nos dois genomas analisados que apresentam maior similaridade, identidade e *E-value* (figura 4).

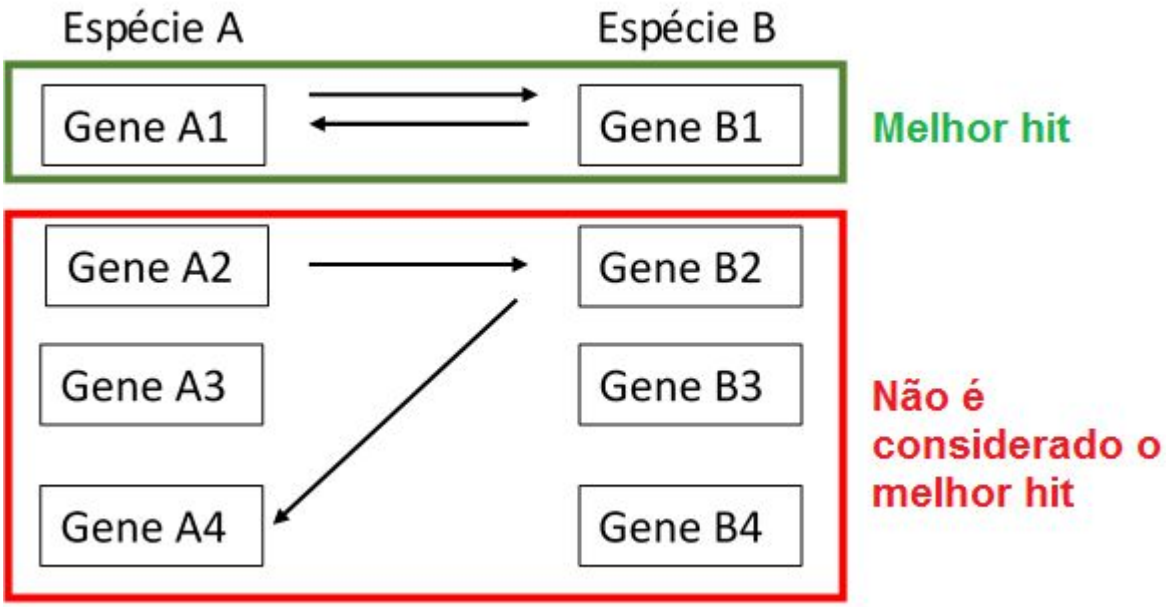

Figura 4: Validação dos pares de interação pela estratégia de melhor hit bidrecional. Em destaque (caixa verde) vem representado o melhor *hit* bidirecional. O GeneA1 da espécie A é considerada homologo ao Gene B1 da espécie B, se o Gene A1 for o melhor alinhamento para o GeneB1, e o Gene B1 representar o melhor alinhamento para o Gene A1. Em destaque (caixa vermelha) vem representado um exemplo de *hit* não bidirecional. Um *hit* não bidirecional é representado pela busca do Gene A2 da espécieA dentro do banco de dados formado pelos genes da espécie B, onde o correspondente do Gene A2 não foi reciproco na busca por similaridade de sequência. Na figura é representado como o Gene A2 tendo o melhor *hit* o Gene B2, e na busca contraria o melhor *hit* do Gene B2 foi o Gene A4.

4.14 Cálculo de Variação nucleotídica

No presente trabalho utilizou-se o programa Variscan para o cálculo de PI . O PI pode ser entendido como uma métrica para medir a quantidade de variação de nucleotídeos diferentes entre pares de sequência, e essas diferenças recebem pesos diferentes de acordo com a frequência (NEI E LI, 1979).

A fórmula para o cálculo do PI utilizada foi:

$$\left[\mathbf{PI} = \sum_{ij} p_i p_j \mathbf{PI}_{ij} \right] \quad (1)$$

1- Equação que mede a variação nucleotídeca entre os pares das sequências.
onde *pi* e *pj* representam a frequência da sequência *i* e *j* e *PIij* é a proporção de nucleotídeos que diferem quando as sequências *i* e *j* são comparadas.

4.15 Alinhamento das sequências e realização do cálculo da taxa de divergência evolutiva

No processo de predição computacional algumas etapas para obtenção dos pares de interações proteína-proteína, devem ser feitas após o agrupamento das sequências referentes a todos os genomas, entre elas: a) obtenção das sequências em um arquivo multifasta; b) o alinhamento múltiplo global; e c) o cálculo propriamente dito da taxa de divergência evolutiva (vide figura 5).

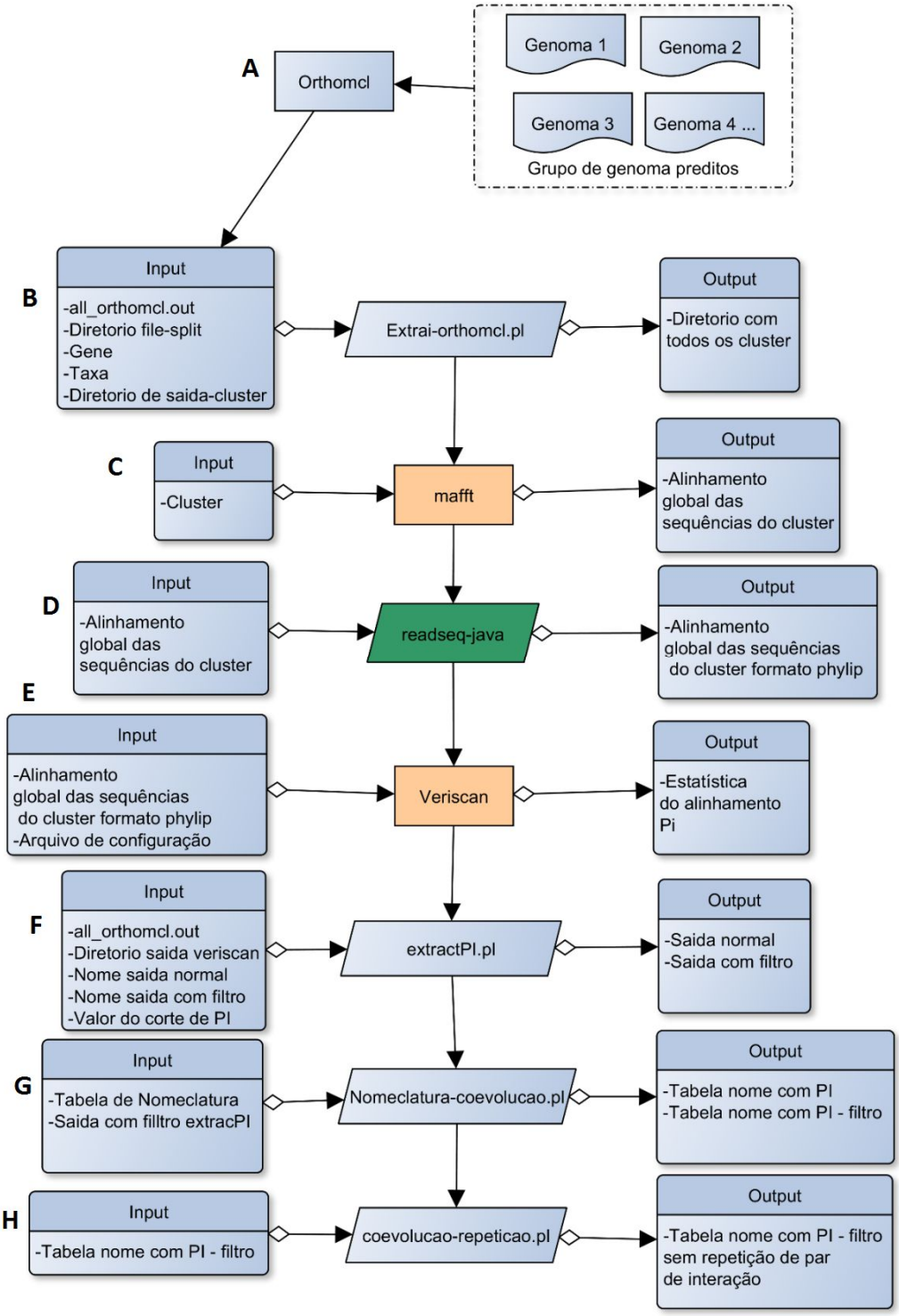

Figura 5: Fluxograma para predição de interações proteína – proteína por coevolução. A etapa inicial passa pelo agrupamento de sequências, utilizando nesse caso o programa OrthoMCL, em busca de genes ortólogos que são derivados de um ancestral comum. Após o resultado utilizou um *script* chamada extrai OrthoMCL, que visa produzir arquivos multifasta com as sequências agrupadas, dessa forma obtém as sequências para realizar o alinhamento múltiplo global, utilizando o Mafft, ao obtermos os alinhamentos, transformamos a saída em um formato adequado para a entrada, para a execução do Variscan, ele calcula-se o valor de PI, que mede a variação nucleotídica. Após isso os *scripts* posteriores visam reconhecer quais proteínas foram encontradas e que possuem pares de interação e o valor do PI encontrado para cada uma.

Posteriormente um *script* chamado Extrai-orthomcl.pl foi utilzado, onde as entradas desse *script* são o número de genes e de taxa, o arquivo de saída do OrthoMCL, o caminho contendo todos os arquivos de todos os genomas separados com o seu nome e o diretório de saída. Como saída o *script* gera um arquivo com o nome do agrupamento do OrthoMCL e dentro do arquivo as sequências encontradas na linha (Figura 10). Então cada arquivo contém as sequências respectivas a sua linha do OrthoMCL.

Depois de obter os arquivos fastas correspondentes às sequências que foram agrupadas realizou-se um alinhamento global das sequências. Para isso utilizou-se o Mafft. Executou-se os diversos arquivos multifastas no programa Mafft através de um programa de uma linha, dessa forma foi possível realizar o processo do alinhamento em fila, sem precisar da execução individual dos agrupamentos. Um exemplo na linha 7855 da saída do programa OrthoMCL, que é o *cluster* ORTHOMCL7854, esse *cluster* possui dois genes e duas taxas, o arquivo multifasta gerado, contém as duas sequências.

Como resultado do Mafft, obteve-se o alinhamento das sequências, porém esses não se encontram no formato de entrada para a próxima etapa. Foi executado um *script* Shell*,* para utilizando o *readseq.jar,* convertendo todos os arquivos para um formato que seja compatível com o necessário para a execução do próximo programa. Sendo esse formato de saída o Phyllip.

Para inferirmos que a proteína em um organismo possui a mesma interação em outro organismo, é necessário calcular a taxa de divergência evolutiva, uma das possibilidades para isso é calcular o PI do alinhamento das sequências em questão (NEI E LI, 1979; REZENDE, 2012).

O programa Variscan em um dos seus resultados gera o valor de PI do alinhamento das sequências. Desenvolveu-se outro *script* que coloca no arquivo, o valor de PI associado a linha do número do agrupamento da saída do OrthoMCL, pois o PI refere-se ao alinhamento das sequências, onde o PI pode ser compreendido como uma medida sobre quantidade de variação nucleotídica das sequências.

O resultado do Variscan contém o valor de PI do alinhamento múltiplo global das sequências, neste caso pertenciam aos agrupamentos. Com o resultado de PI foi possível inferir que a taxa evolutiva foi pequena, quando PI é menor ou igual a 0.03, pois quanto mais próximo de zero, menor é a diferença entre as

sequências. Essa diferença é a considerada ideal, porém existe outros valores de cortes que não são restringentes quanto este.

Esse resultado é associado com a tabela de nomenclatura, gerando então uma tabela de interações proteína–proteína por coevolução de sequência, contendo o valor de PI das sequências descritas. Esse processo é descrito no fluxograma da figura 5.

A tabela de nomenclatura utilizada neste caso foi a de *Cryptococcus neoformans* var. *neoformans* JEC21, e *Cryptococcus gattii R265,* esse procedimento foi realizado para os dois outros organismos depositados no banco STRING, afim de gerar futuramente uma tabela de interações proteína – proteína única para cada organismo.

Na tabela final é possível nota que nem todas a proteínas possuem um par de interação proteína-proteína. Existem alguns fatores importantes para isso ter acontecido, o primeiro seria pelo fato de serem espécies diferentes, por isso existem genes que não serão encontrados em ambos; a segunda possibilidade é que esta proteína existe no nosso genoma, porém é diferente da encontrada no banco de dado do STRING, não estando contida dentro dos nossos parâmetros de cortes para gerar a tabela de nomenclatura. Na terceira corresponde à proteína se encontra no genoma, mas o valor de PI calculado estar fora do valor de corte, excluindo a proteína da análise; a quarta possibilidade vem do fator de PI ser dado através do alinhamento do *cluster*, por isso uma ou mais sequências dentro do *cluster* podem ser diferentes das demais e com isso ter elevado o valor de PI do agrupamento inteiro, indicando que as outras proteínas podem ter associação de coevolução, mas alguma ou algumas proteínas dentro da sequência teriam uma divergência, que alterou o valor.

Para corrigir este quarto problema, realiza-se o fluxograma abaixo, que é similar com o fluxograma anterior, com a diferença que ele utiliza os pares de sequências encontrados no resultado do agrupamento, ao contrário desse que utiliza todas as sequências do agrupamento.

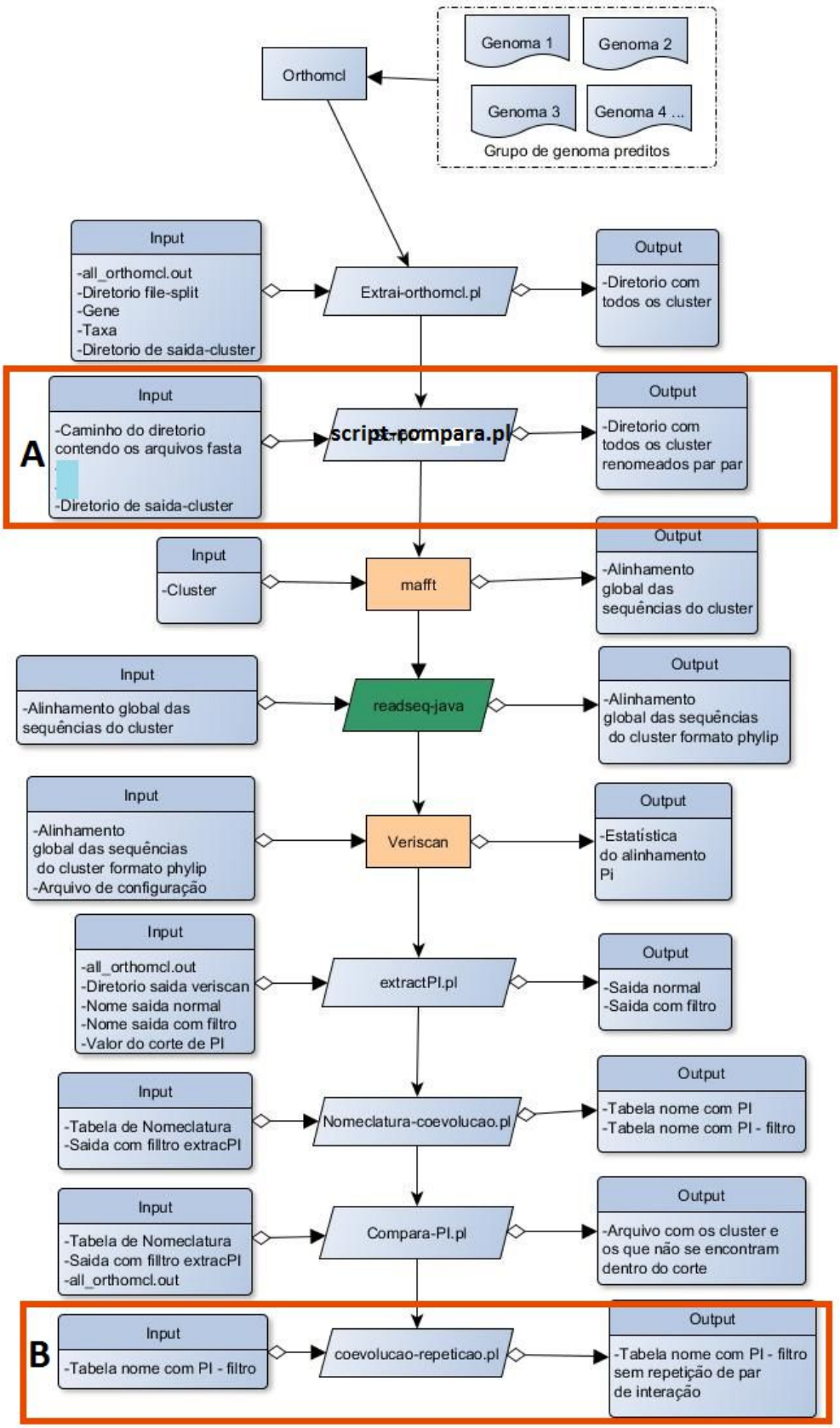

Figura 6: Variante de fluxograma empregado para predição de interações proteína – proteína por coevolução. Somente os blocos destacados diferem do fluxograma apresentado na figura 6. A) Nessa etapa realiza-se uma análise combinatória de pares das sequências contidas no agrupamento; B) realiza a diferença do valor PI adequado para aquela sequência de acordo com o resultado obtido na análise combinatória.

O processo realizado nessa etapa difere pelo fato de que a análise não é realizada com o agrupamento, mas com os pares de sequências. Os pares são gerados através de uma análise combinatória de cada agrupamento, isso é um agrupamento com as sequências A, B e C, no método anterior ele é considerado uma sequência, na análise combinatória a sequências são A e B, A e C e B e C. E todas elas passam pelas etapas posteriores. O Cálculo de análise combinatória utilizado se encontra abaixo.

Na etapa de obtenção das sequencias para o alinhamento representado no fluxograma na figura 6 empregamos a seguinte estratégia para calcular as combinações possíveis entre as sequências.

$$\left[C_{n,p} = \frac{n!}{p!\cdot(n-p)!} \right] \quad (2)$$

2- Fórmula referente a análise combinátoria para se obter os pares de sequência do agrupamento.

C = Combinação

n = Elementos.

p = Agrupamento

4.16 Análise estatística das redes no software R

No presente trabalho utilizou-se a linguagem de programação em R disponível gratuitamente (https://cran.r-project.org/index.html). Particularmente utilizamos uma interface gráfica conhecida como o R Studio, que possibilita além da janela de execução, a janela de históricos e visualização de gráficos, facilitando dessa forma a criação de *scripts* no ambiente R. *Scripts* específicos foram criados para obter o grau, a média, a mediana relacioando ao grau, a localização dos *hubs*, a análise de comunidade (figura 7).

Através dele criou-se um *script* para a análise específica de acordo com o conjunto de dados, gerando as análises estatísticas e os gráficos relacionados a ela, como os histogramas com a frequência do grau cada nó, estatísticas como a quantidade de nós e graus e a análise dos vizinhos, dentre outras análises de redes, o formato do arquivo de entrada das redes era cvs ou pajec.

Para as estatísticas gerais da rede, utilizou-se a linguagem e as bibliotecas específicas do R, uma das principais nomeada de Igraph (http://igraph.org/). Essa biblioteca possui modelos e estatísticas que possam ser utilizadas com redes, possibilitando diversas ferramentas para dados estatísticos para serem realizados em redes.

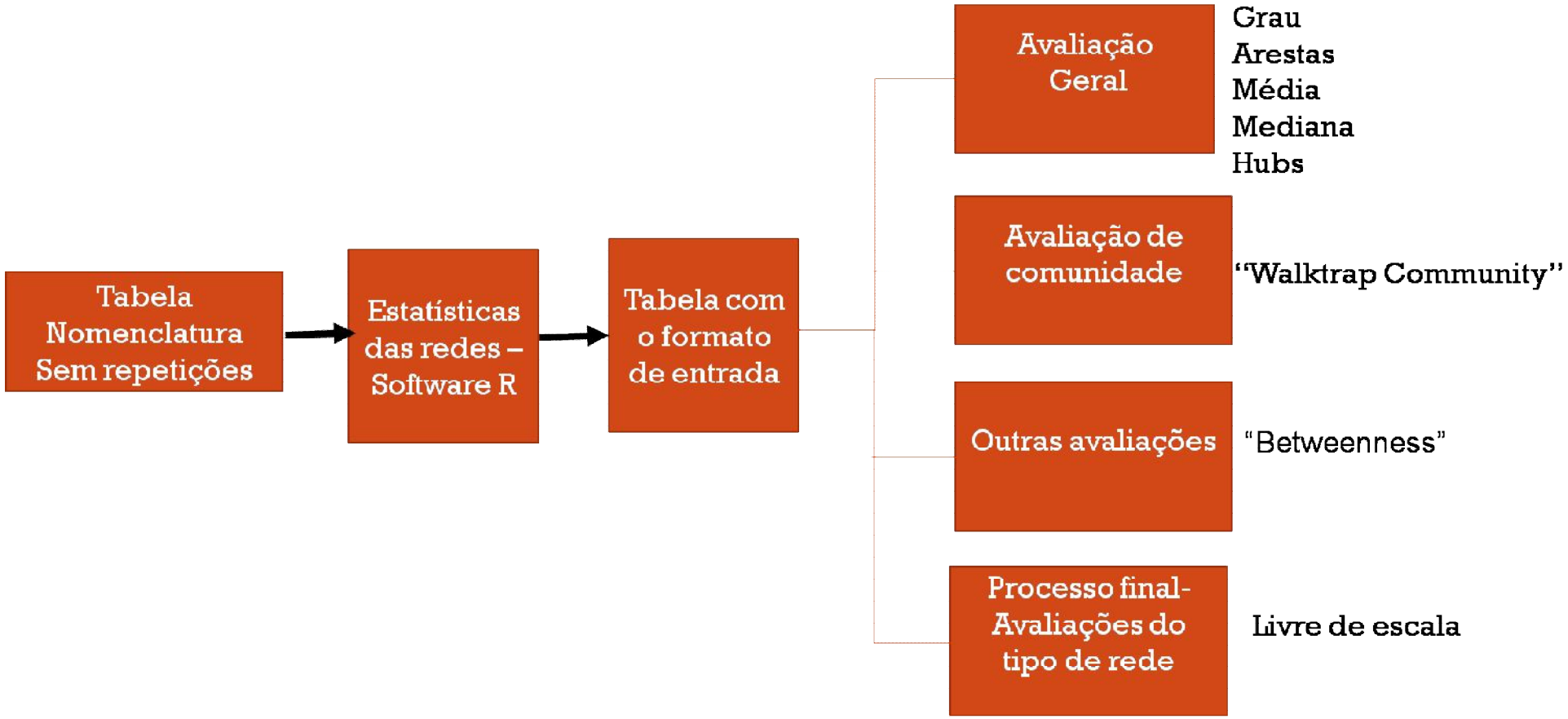

Figura 7: Análises estatísticas e propriedades calculadas das redes geradas. O fluxograma mostra como as estatísticas das redes foram distribuídas de forma geral, criando grupos de avaliação geral, como o grau, as arestas, qual análise de comunidade foi utilizada.

4.17 Visualização das Redes

O software Cytoscape (SHANNON *et al*, 2003) é uma plataforma de software que possui código aberto, sendo amplamente utilizado para a visualização de redes de interação molecular e vias biológicas, além de integrar estas redes com anotações, perfis de expressão gênica e outros dados. Uma característica adicional importante trata-se da possibilidade do usuário criar *plugins* (que são aplicativos) que integram informações e análises adicionais à rede. Esses *apps* estão disponíveis no próprio site do Cytoscape, no Cytoscape App Store ou pelo próprio programa, sendo a maioria disponibilizada gratuitamente, possibilitando então sua a distribuição e utilização de modo global.

4. 18 Scripts PERL e SHELL

Devido à grande quantidade de dados gerados nos diversos processos de análises, torna-se necessária a implementação e a utilização de linguagens de programação específicas para catalisar a geração dos resultados. As linguagens utilizadas foram a PERL e a *Shell script*. Para as diversas etapas adaptou-se e criou-se novos *scripts*, para suprirem a necessidade de extração de informação e análise.

4.19 Banco em MySQL para relacionar as proteínas virulentas e patogênicas encontradas nas redes

O banco relacional criado em MySQL está representado na figura 8 simplificado da estrutura do banco, pois uma figura relacional seria muito grande, devido ao grande número de tabelas.

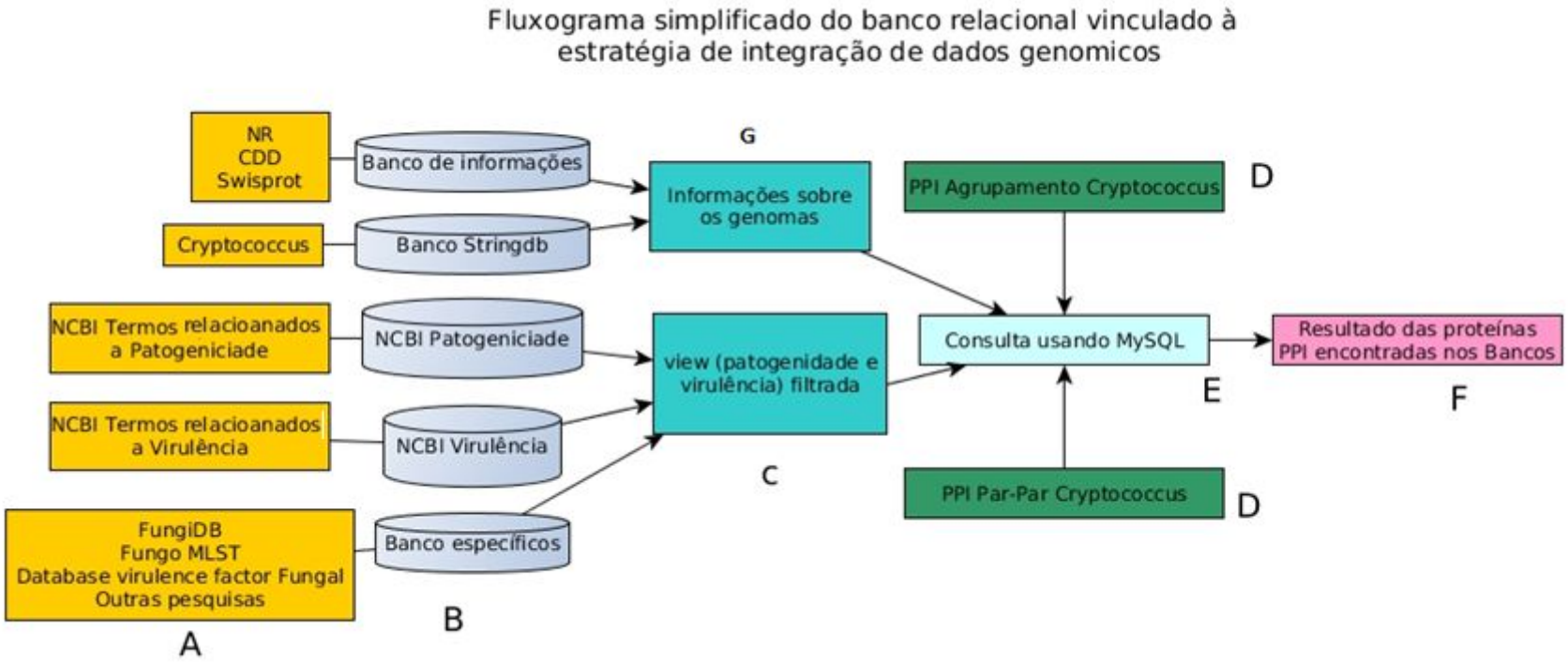

Figura 8: Fluxograma simplificado do banco relacional vinculado à virulência estratégia de integração dos dados genômicos. A) Representa a fonte das tabelas, essas tabelas são obtidas através do BLAST, e além dessa tabela existe uma tabela de correspondência para cada uma para identificar o ID identificado no banco; B) As tabelas pertencem a determinados grupos dentro do banco, sendo esses os cinco grupos utilizados; C) Das tabelas provenientes de dados relacionados à virulência e à patogenicidade criamos uma *VIEW* com filtro para eliminar dados que poderiam não ser correlacionados com as proteínas dos genomas; D) São as rede de interação proteína-proteína de *Cryptococcus,* utilizando as duas estratégias; E) Consulta relacionado as proteínas das redes de interação proteína-proteína, com as proteínas associadas à virulência e à patogenicidade. F) Resultado das proteínas associadas à virulência e à patogenicidade encontrada nas redes. G) Busca de informações sobre proteínas que possam ser de interesse.

5 RESULTADOS E DISCUSSÃO

Os resultados e discussões estão no mesmo capítulo para que a comprensão das ideias e dos resutlados sejam continuas.

5.1 Genomas

Os genomas utilizados no presente trabalho, foram obtidos de dois repositórios, o NCBI (http://www.ncbi.nlm.nih.gov/) e o Broad (https://www.broadinstitute.org). Ao todo foram analisados 12 genomas completos (vide tabela 3).

Tabela 3: Lista dos genomas analisados do gênero *Cryptococcus* spp com o repositório utilizado. Na tabela encontram-se o nome das espécies com suas respectivas linhagens, a informação sobre anotação genômica, o número de contigs presentes no assembly e o agrupamento das sequências em cromossomos ou não.

Espécie	**Linhagem**	**Repositório**	**Anotação genômica no DB de origem**	**nº contigs**	**Agrupamento em cromossomos/quantidade**
C. bestiolae	CBS 10118	NCBI	NÃO	42	não
C. dejecticola	CBS 10117	NCBI	NÃO	69	não
C. flavescens	NRRL Y-50378	NCBI	NÃO	712	não
C. gattii	CBS 7750	NCBI	NÃO	2938	não
C. gattii	R265	NCBI	NÃO	--	não
C. gattii	WM276	NCBI	SIM	14	Sim/14
C. neoformans var. grubii	H99	NCBI	SIM	15	Sim
C. neoformans var. neoformans	B-3501ª	NCBI	SIM	14	Sim/14
C. neoformans var. neoformans	JEC21	NCBI	SIM	70	Sim
C. pinus	CBS 10737	NCBI	NÃO	42	não
C. heveanensis	BCC8398	Broad	NÃO	--	não
C. heveanensis	CBS569	Broad	NÃO	--	não

Optou-se por genomas depositados no NCBI uma vez que esses estavam com informações referentes às anotações gênicas, com isso apenas os

organismos de *Cryptococcus heveanensis* foram retirados do banco de sequências do Broad.

5.2 Anotação e Reanotação dos genomas

O processo de anotação gênica por similaridade de sequências envolve a comparação dos genes preditos na etapa *ab initio* contra três bancos de dados específicos, sendo eles: swissprot, CDD, nrS1 (subset do banco nr contendo apenas as sequências de *Cryptococcus*) e nrS2 (subset do banco nr em que as sequências de *Cryptococcus* foram removidas).

A utilização do banco swissprot deve-se ao fato de que esse banco possui uma curadoria maior comparada ao nr por exemplo. Trata-se de um banco de dados de sequência de proteína curadas que fornece um alto nível de anotação (tal como a descrição da função de uma proteína, a sua estrutura de domínio, modificações pós-traducionais, variantes, etc), um nível mínimo de redundância e um elevado nível de integração com outras bases de dados.

O Banco CDD possui uma anotação de proteína que consiste em uma coleção de bem anotada vários modelos de alinhamento de sequências de domínios e proteínas de comprimento total (MARCHLER-BAUER *et al*, 2003). O nr é um banco amplamente utilizado para obter uma anotação, diminuindo a capacidade de erro por anotação. Utiliza-se, um banco com e sem o organismo.

Os resultados do processo de predição *ab initio* com o programa Augustus resultaram em média em 7109 genes para os oito genomas analisados que não possuíam anotação prévia. O detalhamento do processo vem explicitado na tabela 4.

Tabela 4: Lista de genomas do gênero *Cryptococcus* spp. e resultado da predição gênica *ab initio.* Número de genes na anotação original depositada = (AO); Número de genes na predição gênica *ab initio* (AA).

Espécie	Anotação genômica	nº contigs	Predição Augustus	Modelo de treinamento	nº CDS
C_bestiolae_CBS_10118	não	42	Sim	c_neoformans_neoformans_	6617-AA
C_dejecticola_CBS_10117	não	69	Sim	c_neoformans_neoformans	7423-AA
C_flavescens_NRRL_Y-50378	não	712	Sim	c_neoformans_neoformans	8714-AA
C_gattii_CBS_7750	não	2938	sim	cryptococcus_neoformans_gattii	7332-AA
C_gattii_R265	não		sim	cryptococcus_neoformans_gattii	6209-AA
C_gattii_WM276	sim	14	não	cryptococcus_neoformans_gattii	6460-AA/6575/AO
C_neoformans_var_grubii_H99	sim	15	não	c_neoformans_neoformans	6525-AA/7798/AO
C_neoformans_var_JEC21_uid10698	sim	14	não	c_neoformans_neoformans	6808-AA / 6475/AO
C_neoformans_var_neoformans_B-3501ª	sim	70	não	c_neoformans_neoformans	6572-AA/ 6578-AO
C_pinus_CBS_10737	não	42	sim	c_neoformans_neoformans	5160-AA
C_heveanensis_BCC8398	não	--	sim	c_neoformans_neoformans	8128-AA
C_heveanensis_CBS569	não	--	sim	c_neoformans_neoformans	7295-AA

5.3 Anotação por similaridade de sequência

Com o formato tabular proveniente dos *scripts* como *anota-genome.pl e anota-auto.pl.* As saídas desses programas foram utilizadas como entrada no programa Artemis, tornando possível incluir as informações provenientes do banco de sequências do NR, Swisprot e CDD na anotação do genoma através do

programa Artemis. Através dessa plataforma de integração de informações estruturais e funcionais foi possível integrar dados obtidos da anotação feita com banco NR do NCBI, o CDD e o Swissprot (vide figura 9). Na figura 9 é apresentado um exemplo da anotação realizada.

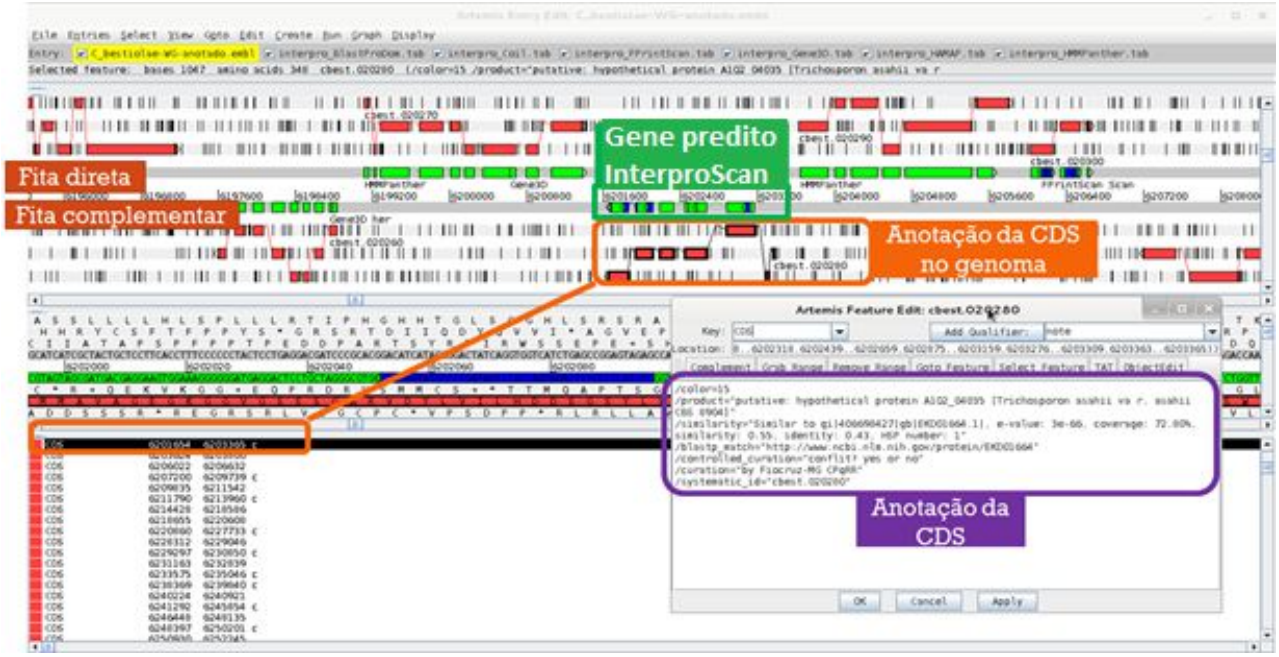

Figura 9: Visualização do genoma de *Cryptococcus bestiolae* CBS 10118, através do programa Artemis. Em destaque a anotação estrutural e funcional de umas das CDS. Vários termos anotadores são utilizados, mas não se limitam à: product, sistemtic_id, similarity, blastp, curation e etc;

Para facilitar o processo de análise dos dados, utilizamos uma abordagem de integração de informações que utiliza como base uma planilha que contêm *hiperlinks* viabilizando o acesso direto à várias informações, entre elas: a) acesso direto às sequência fasta dos genes; b) acesso direto ao resultado das buscas por similaridade de sequência; c) acesso direto ao número de acesso do banco que originou a anotação; d) acesso as informações de similaridade, e-value, cobertura e etc.

5.4 Agrupamento das sequências

As sequências biológicas são utilizadas para inferir a história evolutiva dos organismos ou de determinado gene, pois um alto grau de conservação pode implicar em uma função bioquímica conservada (ARVESTAD *et al*., 2004). Para a realização de inferências filogenéticas torna-se necessário em alguns casos as predições e as análises de homologia sejam precisas, tal como encontrar um gene

e/ou grupos de genes em organismos modelos que estão relacionados a doenças humanas, e com isso, inferir a função de um gene que foi sequênciado recentemente (GABALDÓN *et al*., 2009; ALTENHOFF, 2012).

Para inferir que a conservação gênica entre os genomas foi utilizada um programa que realiza o agrupamento (*clusters*) das sequências. Adotamos o programa OrthoMCL (LI, *et al*., 2003) que resumidadmente realiza um BLAST recíproco de todos os genes dos genomas. Como resultado o programa fornece um arquivo que relata o agrupamento dos genes baseado na similaridade de sequência (GABALDÓN *et al*., 2009). Vide na figura 10 um exemplo do resultado gerado pelo programa.

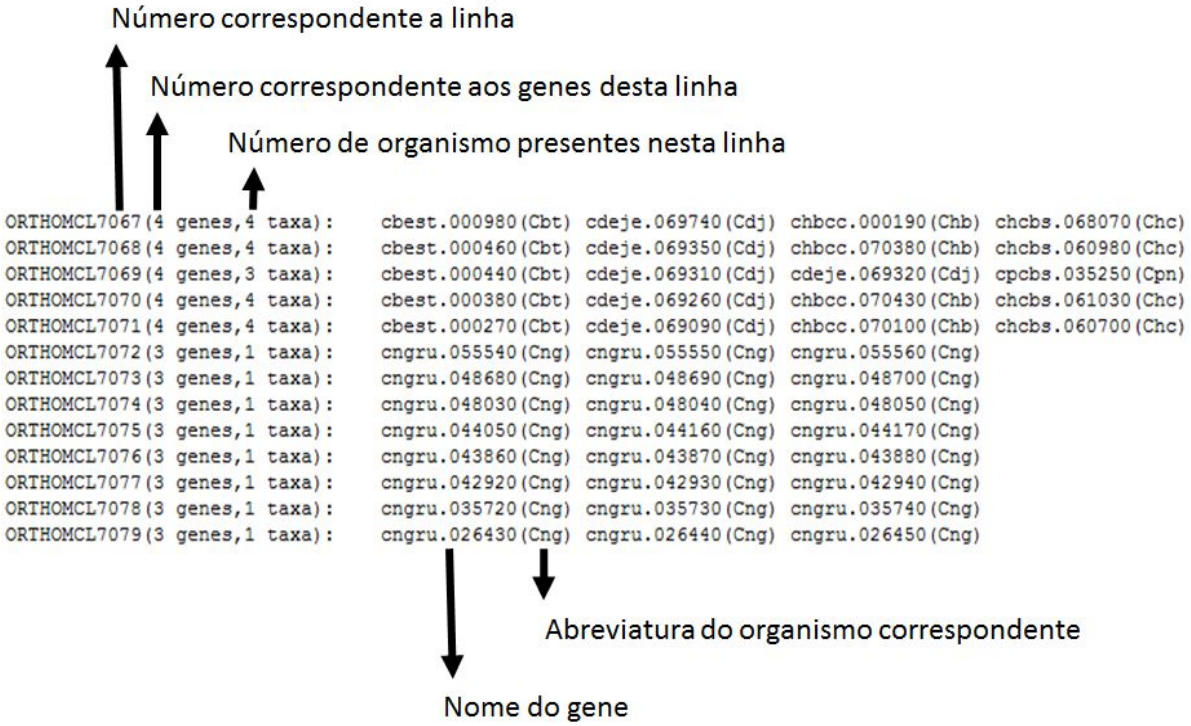

Figura 10: A figura mostra a visualização de um dos arquivos de saida gerados pelo programa OrthoMCL. Cada uma das linhas indica um *cluster* mapeado pelo programa. As linhas descrevem o número de genes encontrados, o número de taxa, ou seja, o número de espécies/cepas contidas no agrupamento. Os organismos são representados através de uma abreviatura de três letras contidas entre os parenteses (no exemplo Cbt = *C. bestiolae*).

Ao obter esse resultado, realizou-se a extração dos grupos que possuíssem a mesma quantidade de genes e a mesma quantidade de taxas, por exemplo 4 genes,4 taxas, ou seja, aquele gene foi agrupado para os genomas uma única vez. Realizou-se estes procedimentos respectivamente para os outros grupos iguais de genes e taxa.

5.5 Seleção de sequências associadas à Patogenicidade e Virulência

A seleção de sequências associadas à patogenicidade e à virulência para o gênero *Cryptococcus* foi importante para buscar sequências que se relacionavam aos termos de busca detalhados no 4.9 da seção de materiais e métodos, uma vez que esse fungo não possui um banco especifico contendo todas as sequências que possam estar relacionadas aos termos.

Para isso buscamos duas formas de se obter os dados referentes à seleção de sequências, à primeira foi através da utilização de repositórios específicos e a segunda forma foi buscar termos em pares como descrito no item 4.10 de materiais e métodos.

Na busca de bancos onde seria possível encontrar sequências associadas à virulência e à patogenicidade encontramos três, o MLST, O FungiDB e o DFVF.

Para cada locus estudado, diferentes sequências genéticas presentes dentro de uma espécie são atribuídas como alelos distintos. Esse banco possui sequências que estavam associadas aos termos.

As outras sequências foram obtidas através do buscador do NCBI utilizando as combinações descritas em materiais e métodos. Como termos específicos de busca utilizamos combinações das palavras: a) patogenicity; b) *virulence*; e c) *Cryptococcus*.

Os resultados dessas buscas são apresentados na tabela 5.

Tabela 5: Resultados das buscas de sequências relacionadas à virulência e à patogenicidade em bancos de dados.

Foi utilizada a busca por sequências contendo termos anotadores relacionados à virulência e à patogenicidade (*virulence* or *pathogenicity*).

Nome do banco	Nº de sequências encontradas
DFVF: database of fungal virulence factors	75
FungiDB	6
Fungal MSLT database	25
Total	106

Na busca de sequências relacionadas à virulência (*virulence*) e *Cryptococcus* no banco do NCBI, as quantidades de sequências encontradas estão descritas abaixo (tabela 6):

Tabela 6: Resultados das buscas de sequências relacionadas ao par *virulence* and *Cryptococcus* no buscador do NCBI

Forma de procura	Nº de sequências encontradas
Termo principal " Virulence and Cryptococcus"	15878
Descrição " Virulence" dentro do termo principal	49
Descrição " Pathogen" dentro do termo principal	10
Busca ativa	2
Outras formas de "virulence"	20
Total	15959

As sequências dessa tabela são as sequências encontradas através do buscador utilizando esses termos, além disso o número de sequências que continham diretamente em sua descrição o termo "*virulence*" ou algo relacionado a patogenicidade, e no caso de busca ativa e outras formas, artigos ou notas relacionadas a virulência que em seu conteúdo disponibilizava uma sequência. A tabela representa o número de sequências obtidas após a busca realizada no site do NCBI com os termos presentes e o número de sequências encontradas em cada termo.

Os termos relacionados à patogenicidade apresentaram grandes derivações ortográficas. Os termos e o número de sequências encontrados se encontram abaixo (tabela 7):

Tabela 7: Resultados das buscas de sequências relacionadas ao termo de patogenicidade no buscador do NCBI.

A tabela contendo aos termos utilizados na busca por sequências associadas a patogenicidade e a quantidade de sequências encontradas em cada par de termo.

Termo de procura	Nº de sequências encontradas
Pathogenic Cryptococcus – gene	19
Pathogenic Cryptococcus – protein	854
Pathogen Cryptococcus – gene	15
Pathogen Cryptococcus – protein	32331
Pathogenesis Cryptococcus – gene	47
Pathogenesis Cryptococcus – protein	123
Pathogenicity Cryptococcus – gene	6
Pathogenicity Cryptococcus – protein	71
Pathogenes Cryptococcus – gene	14
Pathogenes Cryptococcus – protein	71
Pathogenesises Cryptococcus – protein	176449
Total	210000

O total de proteínas selecionadas associadas à virulência e à patogenicidade foram de 226.065 proteínas. Para a vremoção da redundância de sequências utilizamos o programa de clusterização ORTHOMCL que como resultado gerou-se 6666 agrupamentos. Esse grande número acontece ao fato de que muitas proteínas estão repetidas dentro do banco, mas com nome diferente, além disso variações pequenas de sequências podem agrupar uma mesma proteína em outro grupo.

5.6 Estatísticas das redes

Na tabela de interações proteína-proteína, houve uma verificação de redundância dos pares de interações. Essas repetições prejudicam a contagem dos *hubs* dentro da rede. Os *hubs* são nós altamente conectados, e possuem uma importância estrutural na manutenção da rede uma vez que se retirar esses nós, a rede se desestrutura, sendo um ponto importante para a conexão das sub redes, indicando então que a proteína correspondente ao *hub* tem uma grande probabilidade na manutenção da homeostase do organismo, podendo ou não participar de funções vitais (ENRIGHT, 1999; REZENDE, 2012).

Complementarmente, outros parâmetros importantes são calculados, entre eles: a) média dos graus; b) mediana dos graus; c) *betweenness*; d) grau das proteínas (número de pares de interações); e) número de nós (número de proteínas). As compreensões desses fatores são importantes para entender e possibilitar a compreensão da forma estrutural da rede (REZENDE, 2012).

As análises estatísticas descritas acima foram aplicadas aos dois grupos utilizados para predição de redes de interação proteína-proteína por coevolução. Esses grupos foram descritos detalhadamente em materiais e métodos (4.10). Vale a pena destacar que foi retirada a redundância dos resultados obtidos (tabela 8).

Tabela 8: Tabela referente ao número de pares de interações encontrados na predição computacional utilizando a metodologia de coevolução através do agrupamento.
Na tabela estão representados os organismos do presente estudo com o número das proteínas e o número dos pares de interações (arestas) encontrados.

Organismo	Nº de pares de interação	Nº de proteínas
C. bestiolae CBS	446	89
C. dejecticola CBS	433	87
C. flavescens NRRL	353	63
C. pinus CBS_10737	425	79
C. heveanensis BCC8398	393	84
C. heveanensis CBS569	415	81
C. gattii CBS_7750	473	113
C. gattii R265	480	114
C. gattii WM276	476	118
C. neoformans var neoformans B-3501A	482	115
C. neoformans var JEC21	482	115
C. neoformans var grubii H99	454	111
Média	450	100
Variância	1666.96	354.26
Desvio Padrão	40.83	18.82

Nos resultados obtidos através dos agrupamentos das sequências percebe-se que acontece um número de pares de interação entre 353 e 482, tendo um desvio-padrão de 40,83. Por outro lado, nota-se na tabela que representa o resultado dos pares de interação proteína-proteína obtidos utilizando os pares do agrupamento que o número de pares variou de 1.516 a 46.468, tendo um desvio padrão de mais de 18.000 (tabela 9).

Tabela 9 : Representação do número de pares de interações encontrados na predição computacional utilizando a metodologia de coevolução usando os pares dos agrupamentos encontrados, e o número de proteínas encontradas nas respectivas redes.

Organismo	Nº de pares de interação	Nº de proteínas
C. bestiolae CBS	5863	397
C. dejecticola CBS	5463	475
C. flavescens NRRL	1516	149
C. pinus CBS_10737	6019	450
C. heveanensis BCC8398	6284	493
C. heveanensis CBS569	6345	495
C. gattii CBS_7750	38778	2151
C. gattii R265	42007	2229
C. gattii WM276	38117	2161
C. neoformans var neoformans B-3501A	46062	2465
C. neoformans var JEC21	46468	2465
C. neoformans var grubii H99	25665	1702
Média	16005	1098.5
Variância	348346742.8	913376.42
Desvio Padrão	18664.04	955.71

A variação nucleotídica é um cálculo que permite avaliar a taxa de divergência evolutiva das sequências. Ao comparar os genes do genoma com um banco de interações proteína-proteína e verificar uma alta similaridade de sequência entre pares de proteínas é possível induzir que pelo fato de terem um ancestral comum a função em uma, pode se manter conservada em outra, mas para isso é necessário verificar a taxa de divergência evolutiva.

Os resultados indicam que a metodologia utilizando a estratégia combinatória dos pares do agrupamento, obtivemos um número muito maior de proteínas e interações, isso se deve ao fato de que na outra análise (utilizando somente o agrupamento) algumas sequências divergentes do grupo poderiam interferir na obtenção do valor de PI. Os organismos virulentos *C. gattii* CBS_7750, *C. gattii* R265*, C. gattii* WM276, *C. neoformans* var *grubii* H99, *C. neoformans* var *neoformans* JEC21, C *neoformans* var *neoformans* B-3501A, obtiveram um número maior de proteínas e interações preditas para os mesmos. Isso se deve ao fato de que os organismos depositados no StringDB pertencem a essas espécies, tornando então pertinente a diferença encontrada.

Em uma análise comparativa, os resultados obtidos através da metodologia de pares de agrupamento obteve de três a 95 vezes a mais pares de interação proteína-proteína quando comparado à metodologia de agrupamento.

Em um trabalho realizado com *Leishmania braziliensis* (REZENDE, 2012) foram encontrados 1818 proteínas e 39420 conexões. Já em outro trabalho utilizou-se a metodologia de coevolução onde obteve-se 1.377 pares de interação envolvendo 621 proteínas no gênero de *Saccharomyces* (FRASER, 2004), porém o número de proteínas do organismo é muito maior, e em um trabalho do mesmo gênero foram encontrados 1548 nós com 2358 arestas (SCHWIKOWSKI, 2000), dentro de uma espécie e gênero altamente estudada pelo grande impacto comercial.

O banco utilizado para a obtenção das informações dos pares de interação foi STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) é um banco de dados e web, recurso biológico de conhecidos e previsíveis interações proteína-proteína. O banco de dados STRING contém informações de várias fontes, incluindo dados experimentais, métodos de previsão computacionais e coleções de textos públicos. O recurso online também serve para destacar enriquecimentos funcionais em listas fornecidas pelo usuário de proteínas, usando um número de sistemas de classificação funcionais, tais como GO, Pfam e KEGG. A última versão contém informações sobre cerca de 9,6 milhões de proteínas a partir de mais de 2000 organismos.

5.6.1. Estratégia utilizando o Agrupamento

As análises e gráficos mostrados a seguir foram gerados para todos os 12 genomas em estudo. Como exemplo escolhemos os resultados gerados para a especie *Cryptococcus neoformans* var. *neoformans* JEC21 sendo os demais resultados incluídos no anexo B.

As redes geradas através do agrupamento das sequências possuem de maneira geral um nó com alta conectividade, ou seja, um nó com um alto grau. Através da análise dos histogramas apresentados na figura 11 observamos a existência de um nó altamente conectado (grau 80) para todos os 12 genomas analisados.

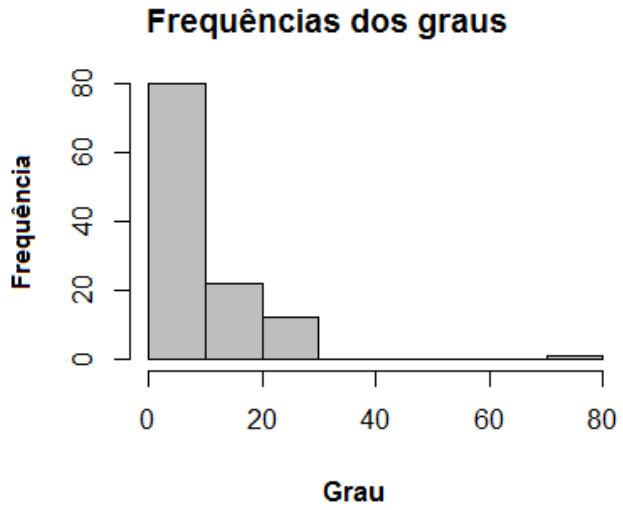

Figura 11: Histograma da distribuição da frequência de grau em *Cryptococcus neoformans* var. *neoformans* JEC21. O histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus neoformans* var. *neoformans* JEC21, utilizando a metodologia de predição de interações proteína-proteína por coevolução através do agrupamento.

O *boxplot* é formado pelo primeiro e terceiro quartil e pela mediana do grau das proteínas. Na figura 12 podemos ver que no painel da esquerda um *outlier*, isso é um nó que não se encontra dentro do grupo, pois o seu valor está muito acima dos outros nós que fazem parte do conjunto de nós preditos para essa espécie. O painel da direita refere-se ao *boxplot* sem o *outlier*. A alteração foi realizada para que possamos ver melhor a distribuição do grau dos nós nessa rede onde a mediana é bem abaixo de 10.

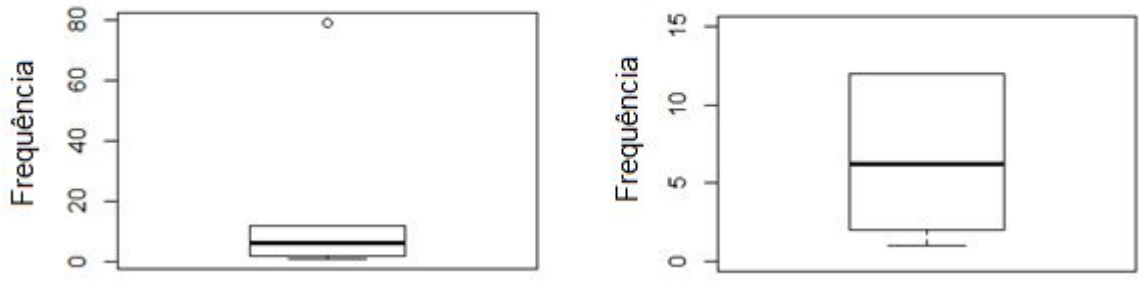

Figura 12: *Boxplot* representando a distribuição dos graus da rede de *Cryptococcus neoformans* JEC21: O painel da esquerda mostra o conjunto de nós com o *outlier* e o painel da direita mostra o conjunto de nós sem o *outilier.* Nos dois casos utilizou-se a metodologia de predição de interações proteína-proteína por coevolução através do agrupamento. Eixo x representa o grau dos nós.

A função ecdf é a função de distribuição cumulativa em que a cada valor da lista será associada uma fração da quantidade de valores menores que o mesmo. Observou-se que na função ecdf do grau é possível visualizar a acumulação dos nós em determinado grau e que o grau de cada nó de forma geral se encontrava bem abaixo de 20. O resultado indica que somente um nó obteve um grau acima de 40 (Figura 13).

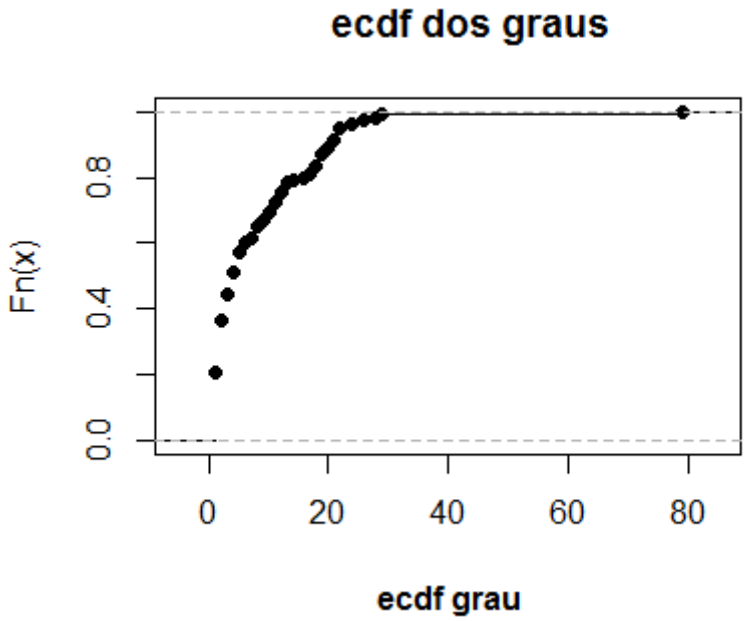

Figura 13: Gráfico referente ao ecdf da frequência dos graus. O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede de *Cryptococcus neoformans* var. *neoformans* JEC21, utilizando a metodologia de predição de interações proteína-proteína por coevolução através do agrupamento.

Com relação à análise de *betweenness* (figura14), observou-se no painel A a existência de poucos nós que podem ser centrais para a rede e no painel B, reafirmando esta constatação, temos o resultado da função ecdf.

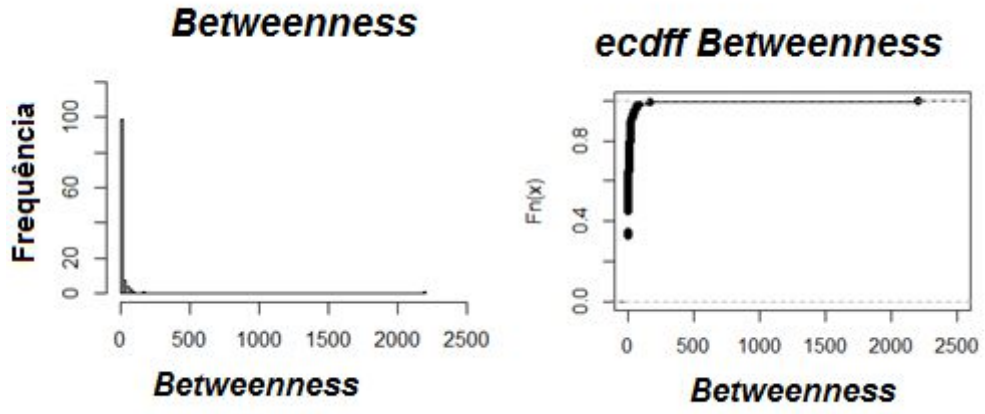

Figura 14: Análise de *Betweenness* da rede de *Cryptococcus neoformans* var. *neoformans* JEC21 utilizando a metodologia de predição de interações proteína-proteína por coevolução através do agrupamento. Painel a esquerda Histograma da frequência do grau utilizando o *betweenness*; painel a direita Distribuição do ecdf do *betweenness*.

Walktrap Community baseia-se em uma caminhada aleatória entre os nós, os caminhos tendem a ser percorridos com uma maior frequência (quantidade de vezes que o percurso passa pela mesma aresta devido à iteração utilizada pela técnica) na parte da rede que correspondente a uma comunidade. É importante compreender a comunidade para compreender os *hubs* das mesmas e as suas funções. Desta forma conseguimos distinguir os grupos (sub-redes) que acontecem dentro da própria rede. Na figura 15 visualizamos que existem 13 comunidades gerais sendo quase todas, exceto uma, com número máximo de quatro nós. A maior comunidade observada na figura possui quatro sub-comunidades.

Figura 15: Rede de *Cryptococcus neoformans* var. *neoformans* JEC21 gerada através de *Walktrap Community*, que é uma abordagem baseada em caminhadas aleatórias. Onde os caminhos são mais propensos a permanecer dentro da mesma comunidade.

Após a predição computacional das interações proteína-proteína por coevolução utilizando os pares dos agrupamentos, as análises nas redes geradas, considerando grau e *betweenness*, foram similares em relação à curva de distribuição nos histogramas e distribuição ecdf. Nos dois casos, há uma maior frequência de nós com menor grau ou *betweenness*.

5.6.2. Estratégia utilizando análise combinatória por par do Agrupamento

As redes geradas através da análise combinatória por par provenientes do agrupamento das sequências possuem de maneira geral um nó com alta conectividade, ou seja, um nó com um alto grau. Através da análise dos

histogramas apresentados na figura 16 observamos a existência de um nó altamente conectado em comparação com os outros nós.

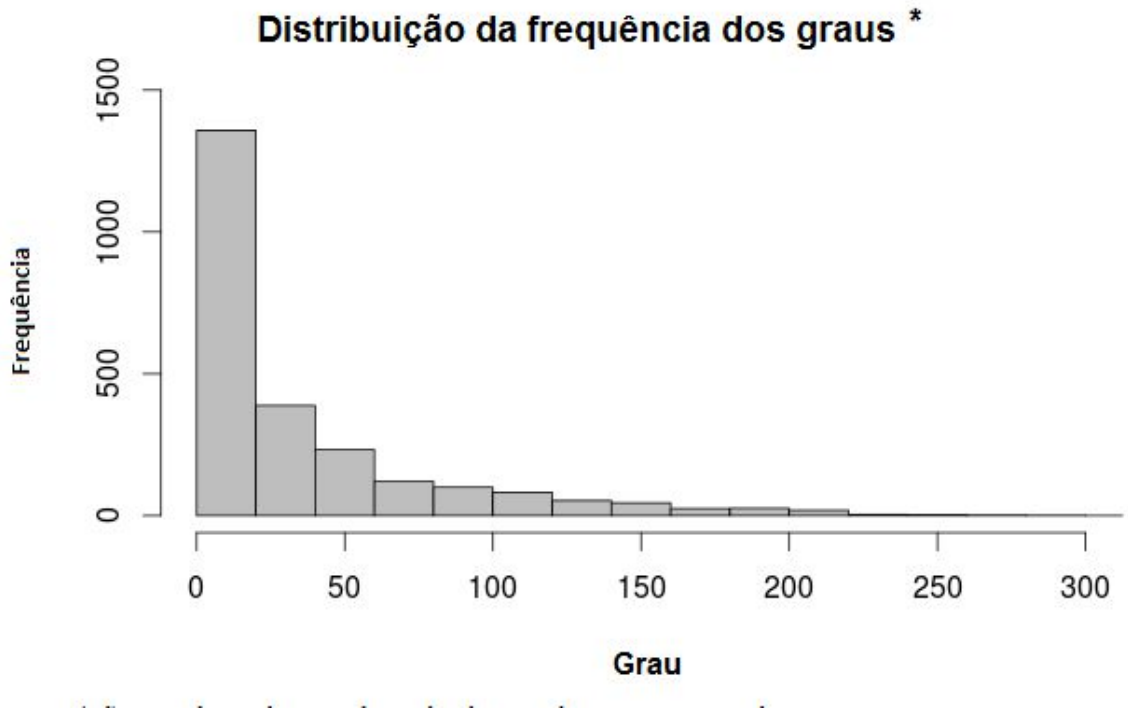

Figura 16: Histograma da distribuição da frequência dos graus dos nós em *Cryptococcus neoformans* var. *neoformans* JEC21. O histograma mostra a distribuição do grau dos nós em relação à frequência na espécie *Cryptococcus neoformans* var. *neoformans* JEC21 utilizando a metodologia de predição de interações proteína-proteína por coevolução através dos pares combinatórios dentro do agrupamento. O Histograma não contém os nós acima de 300, pois eram poucos, e estavam dificultando a visualização da distribuição dos graus.

Em seguida, foi feita a verificação de *outlier* em função do grau no conjunto de dados. Essa informação é importante para saber a forma de estruturação da rede.

Nessa análise do *boxplot* da distribuição dos graus em *Cryptococcus neoformans* var *neoformans* JEC21 (figura 17) observou-se uma proteína com alto grau.

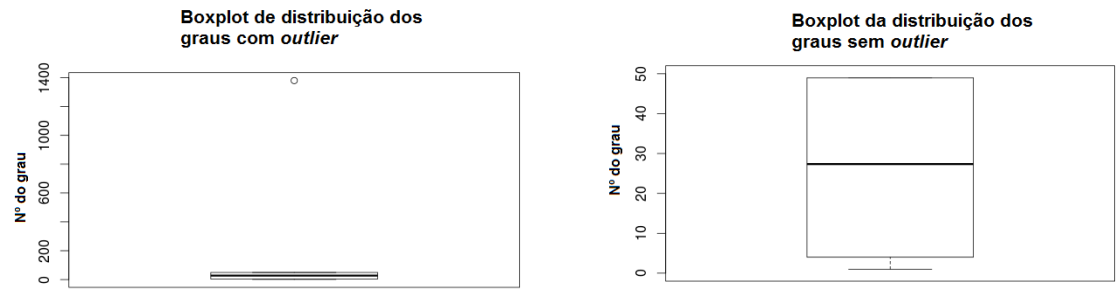

Figura 17: *Boxplot* representando a distribuição dos graus da rede de *Cryptococcus neoformans* var. *neoformans* JEC21 através da metodologia de predição de interações proteína-proteína por coevolução utilizando a análise de pares do agrupamento. No painel à esquerda podemos observar um *outlier*; no painel à direita, o *boxplot* se encontra sem o *outlier* para que possamos observar a distribuição da mediana e a distribuição do número do grau das proteínas nesse conjunto de nós.

A partir do gráfico de distribuição cumulativa (ecdf) dos pares de interação de *Cryptococcus neoformans* var. *neoformans* JEC21 (figura 18), observamos que existe uma proteína com um alto grau, divergindo do conjunto do grupo. Além disso, outras duas proteínas também possuem um alto grau, mas menor em relação à primeira.

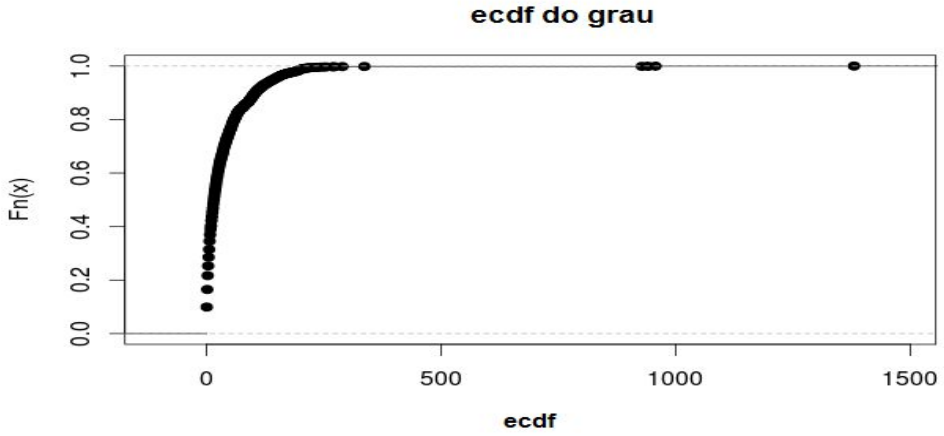

Figura 18: Gráfico referente ao ecdf da frequência dos graus da rede de *Cryptococcus neoformans* var. *neoformans* JEC21 gerada através da metodologia de predição de interações proteína-proteína por coevolução utilizando a análise de pares do agrupamento.

Betweenness (intermediação) é um indicador de que o vértice pode pertencer ao centro do grafo. Dessa forma, o *betweenness* é igual ao número de caminhos mais curtos (caminho mínimo) a partir de todos os vértices para todos os

outros que passam por esse nó. Através dessa métrica podemos compreender qual nó possui uma maior centralidade, ou seja, possui uma maior importância na estruturação da rede. Essa análise foi importante, uma vez que as redes geradas utilizando os pares de agrupamento possuem um número muito maior de nós e arestas quando comparado às redes geradas com os agrupamentos, dificultando a detecção visual dessas proteínas. Na figura 19 observamos que poucos nós tem um *betweenness* alto.

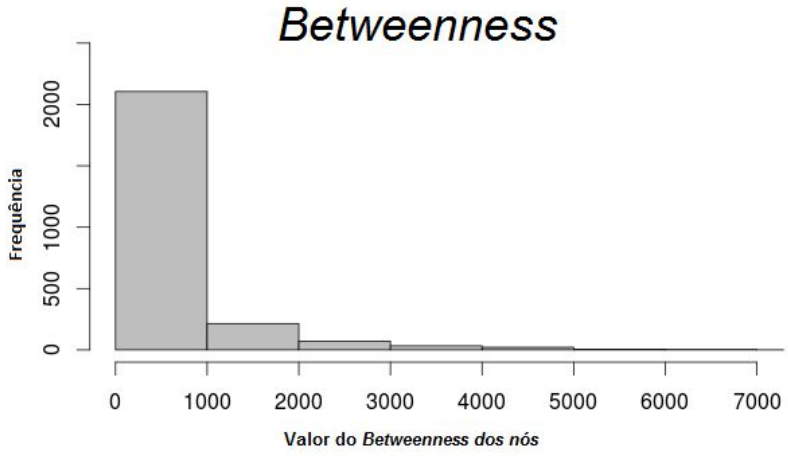

Figura 19: Gráfico representando o histograma do *betweenness.* O histograma indica que poucos nós podem pertencer ao centro do grafo.

Devido à grande quantidade de proteínas encontradas através da metodologia de coevolução através dos pares combinatórios provenientes do agrupamento, não é possível realizar a análise de *Walktrap Community* visual, pois o número de iterações necessário para o processo torna o método inviável, uma vez que o tempo para o cálculo seria muito grande.

5.7 Visualização das Redes

Trabalhar com uma grande quantidade de dados dificulta a apresentação e a integração das informações de forma a extrair informações biológicas, por isso a visualização da rede faz parte da compreensão dos sistemas.

A visualização gráfica demonstra a posição de determinadas proteínas dentro da rede e a forma como elas interagem, facilitando a compreensão dos

dados. A análise visual possibilita a observação de eventos individuais relacionados a cada proteína, além de integrar informações e mapear proteínas que podem ser importantes no funcionamento do sistema.

Para a visualização de redes utilizamos o programa Cytoscape. Na visualização da rede de agrupamento de *Cryptococcus neoformans* var. *neoformans* JEC21 observou-se uma maior quantidade de comunidades. No total são 13 conjuntos de interações e somente um nó possui um alto grau. Na rede gerada existem 115 nós (proteínas) e 482 interações encontradas entre elas.

As funções encontradas no *hub* principal de todas as redes aqui trabalhadas (vide anexo B), referentes ao resultado do *dataset* do NR sem *Cryptococcus* e com *Cryptococcus,* foram: “ATP-dependent protein binding protein” e “ubiquitin”.

A ubiquitina é uma pequena proteína regulatória de 8,5 kDa que tem sido encontrada em quase todos os tecidos de organismos eucariotas. A adição de ubiquitina a uma proteína é chamado ubiquitinação. A ubiquitinação pode afetar proteínas de muitas maneiras: ele pode sinalizar para a sua degradação através do proteassoma, alterar a sua localização celular, afetar a sua atividade, e promover ou prevenir interações proteicas, indicando então a importância na rede e justificando seu alto grau.

O termo do GO (Gene Ontology) “ATP-dependent protein binding protein” define a função de interagir seletivamente e de forma não covalente com qualquer proteína ou complexo de proteínas utilizando energia a partir da hidrólise de ATP, que é outra função relevante e que acontece em todo o organismo, sendo importante para a manutenção do mesmo.

Dessa forma evidenciou-se que o nó descrito com alto grau possui uma função central dentro da rede, devido a sua função, uma vez que pode interagir com diversas proteínas e participar de vários processos (figura 20).

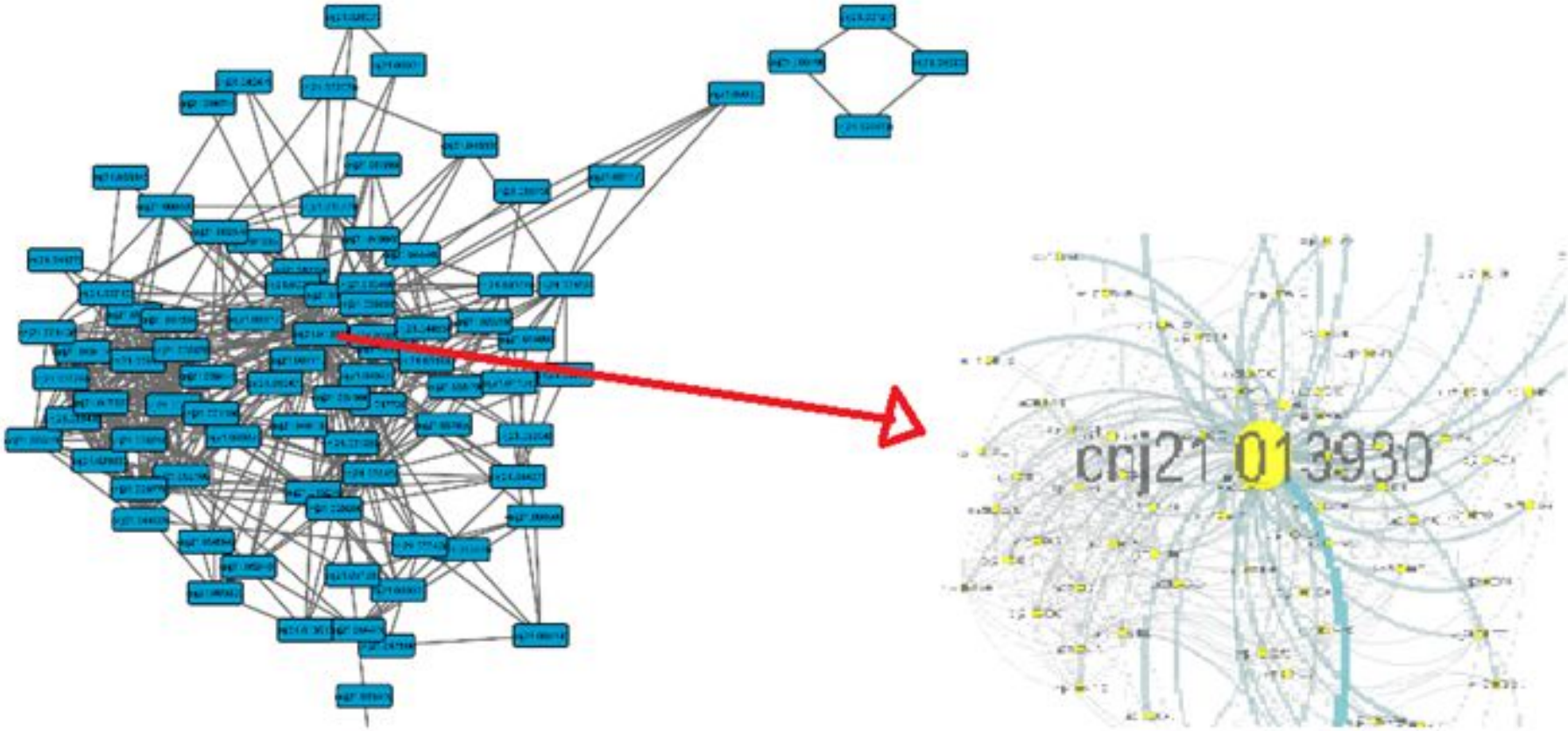

Figura 20: Rede de interações proteína-proteína gerada por coevolução através do agrupamento do organismo *Cryptococcus neoformans* var. *neoformans* JEC21. Os nós altamente conectados são os *hubs*. A seta evidencia o nó que possui um maior grau dentro dessa rede.

Nas redes com os pares de proteínas obtidos através da predição de interações proteína-proteína por coevolução utilizando a análise de pares do agrupamento, observamos que houve um maior número de interações comparando com a metodologia que utilizava o agrupamento. Podemos ver isso na figura 21 que visualiza a rede de *Cryptococcus gattii* WM276.

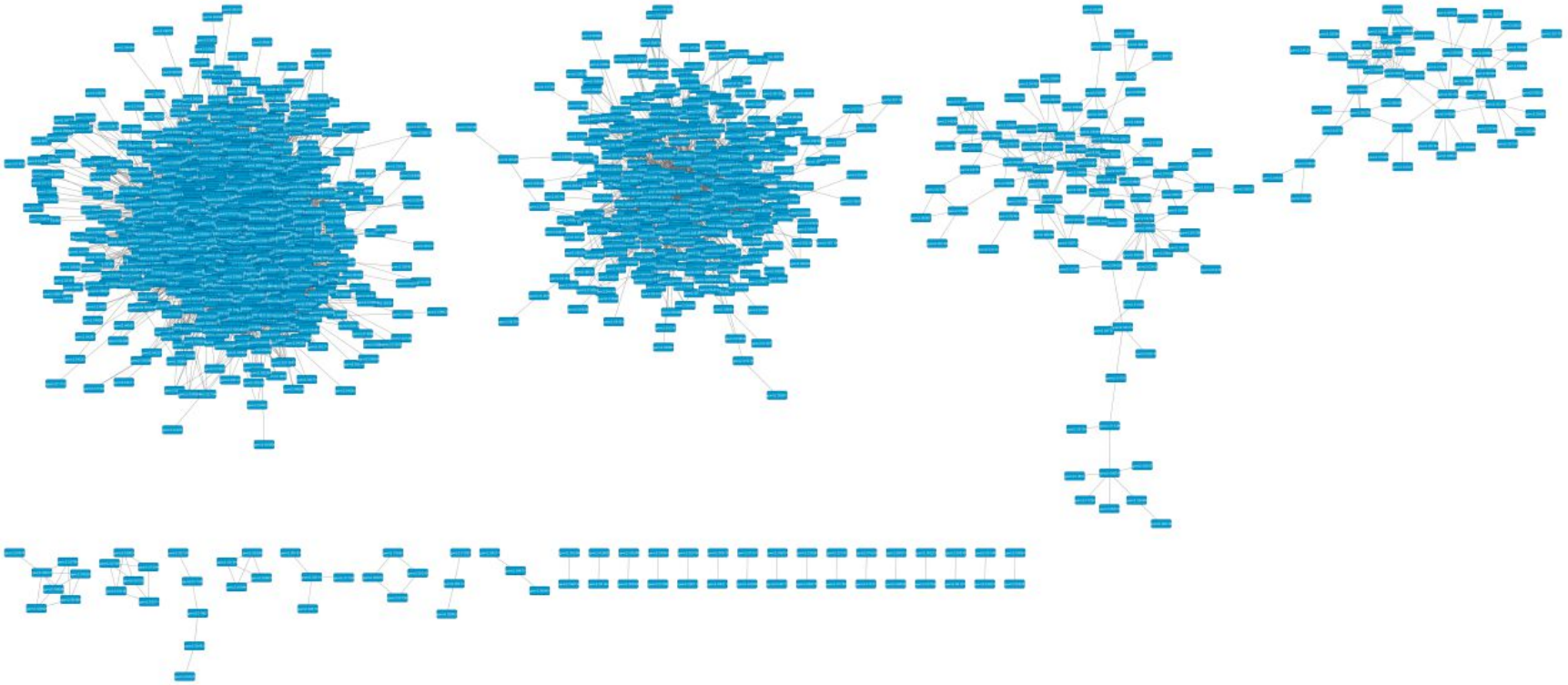

Figura 21: Conjunto das redes de interações proteína-proteína por coevolução utilizando a análise de pares do agrupamento do organismo *Cryptococcus gattii* WM276. Visualização global das redes geradas. Apesar de não ser possível visualizar os nomes das proteínas envolvidas, destacam-se os seguintes pontos: o conjunto de pares de interação preditos para esse organismo mostra que existem pares ou redes não conectadas à rede principal, obtivemos 28 redes que não possuem interações entre si. Outro fato é que existe sempre uma rede principal que engloba a maior parte das proteínas. Isso acontece para todas as redes trabalhadas.

Na visualização da rede de agrupamento de *Cryptococcus gattii* WM276 *Cryptococcus gattii* WM276, observou-se uma maior quantidade de conjuntos de redes ou pares de interações proteína- proteína. No total são 28 sub-redes de interações que não se conectam. O conjunto gerado possui 2161 nós (proteínas) e 38117 interações preditas entre elas.

O nó com o maior grau foi o cgwm2.055390 com um grau de 906 e o resultado do BLAST utilizando o *dataset* do NR sem *Cryptocoocus* foi de “ATP-dependent protein binding protein, ubiquitin [Dacryopinax sp. DJM-731 SS1]”, utilizando o dataset do NR sem *Cryptocoocus* “ATP-dependent protein binding protein [Cryptococcus gattii WM276] ATP-dependent protein binding protein, putative [Cryptococcus gattii WM276], ubiquitin protein 1 [Cryptococcus gattii R265]”. Exatamente o mesmo gene encontrado na análise utilizando o agrupamento.

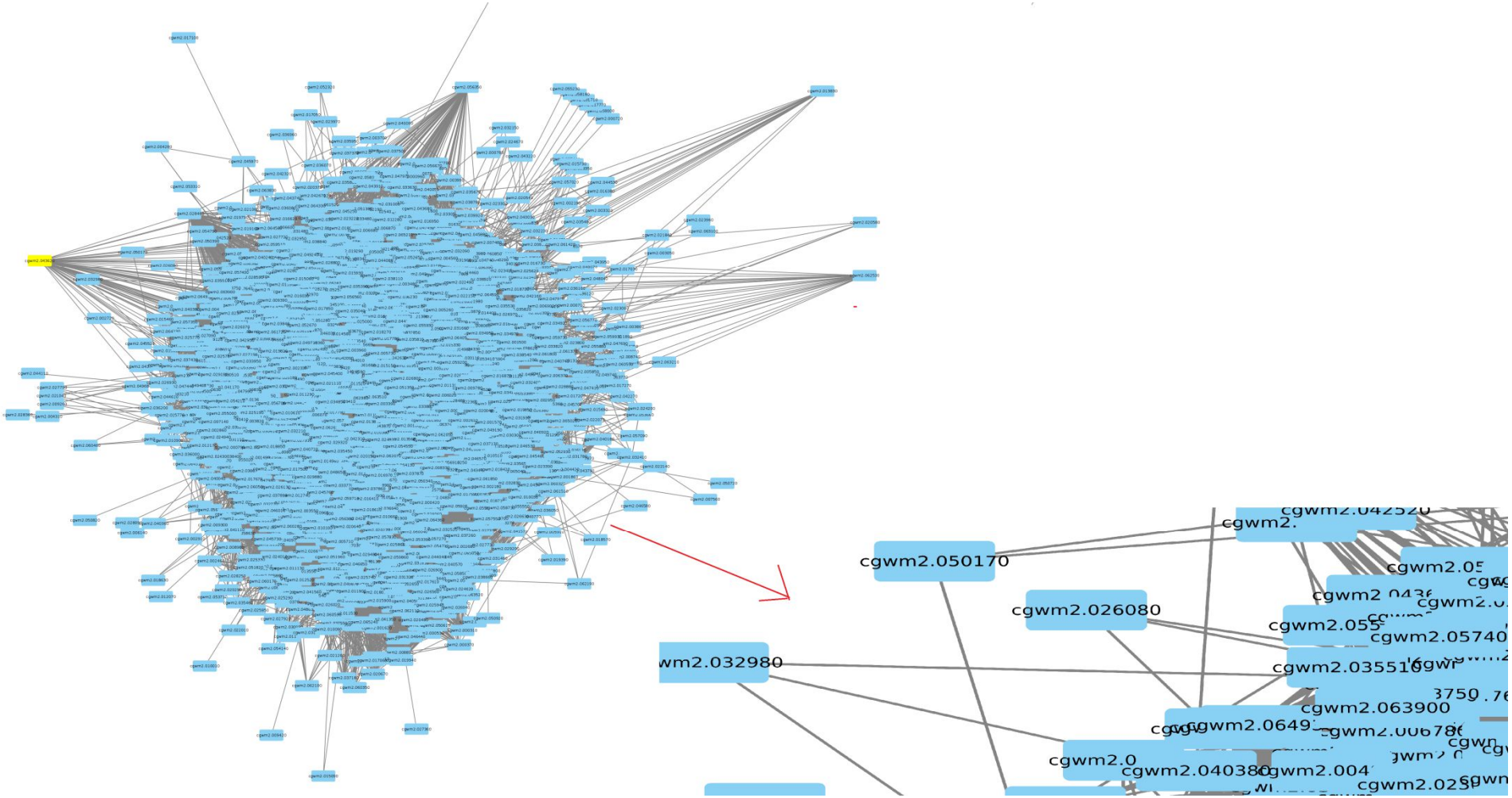

Figura 22: Ampliação da maior rede de interações proteína-proteína por coevolução (vide Figura 21) utilizando a metodologia de predição de interações proteína-proteína por coevolução utilizando a análise de pares do agrupamento do organismo *Cryptococcus gattii* WM276. Apesar de não ser possível indenficar as proteínas nas redes, observamos que nós altamente conectados são os *hubs*. A seta em vermelho mostra a ampliação para visualizar o nome da proteína.

Na figura 22 observou-se a maior rede predita para *Cryptococcus gattii* WM276. Essa rede possui 35.923 conexões e 1.522 nós (proteínas), englobando a maior parte (94,24%) de conexões e nós do conjunto de rede.

5.8 Proteínas relacionadas à virulência e à patogenicidade no Banco

A grande quantidade de proteínas e a variedade de identificadores utilizados assim como as diferentes anotações funcionais vinculadas a cada proteína evidenciam dois pontos antagônicos: o primeiro tem uma conotação positiva, uma vez que traz diversidade e abrangência às informações associadas a cada proteína e o segundo tem uma conotação negativa que está vinculada à redundância de dados gerada.

Por isso é importante na utilização no banco MySQL o resultado do agrupamento de todas as sequências associadas à virulência e à patogenicidade gerado pelo programa ORTHOMCL. Como a quantidade de proteínas é grande 226.025 as buscas no banco são usualmente demoradas dependendo do que se deseja consultar.

Além disso, toda a estrutura de obtenção de dados referente às buscas dos pares de interação e o banco de dados de virulência e patogenicidade foi estruturada para que *desktops* pudessem gerar os resultados, isso é, sem a utilização de grandes servidores. Com isso a metodologia utilizada tanto nas estatísticas como na obtenção dos pares de interação pode ser empregada em escala genômica.

Como o objetivo de selecionarmos proteínas associadas à virulência e à patogenicidade entre as proteínas presentes na rede de interação com a maior confiança possível estabelecemos como parâmetro de corte um valor de E-value 0.000001.

Para *Cryptococcus neoformans* var. *neoformans* JEC21 foi encontrada somente uma proteína associada à virulência e à patogenicidade presente na rede gerada através da análise de pares do agrupamento. O nome sistemático dessa proteína é cnj21.046130, e ela possui um grau de 47. Essa proteína foi encontrada após a busca que utilizou quatro diferentes termos de busca no banco de dados de proteínas do NCBI: "*Pathogen and Cryptococcus*"; "*Pathogens and Cryptococcus*"; "*Pathogenesises and Cryptococcus*"; "*Pathogenesis and Cryptococcus*". Ao

buscarmos esses termos no banco de dados, percebemos uma grande redundância nos resultados que foi removida. O resultado obtido antes da remoção de redundância associa a referida protéina 188 vezes aos termos utilizados e sempre com a mesma anotação: "*fatty-acid synthase complex protein [Cryptococcus neoformans var. neoformans JEC21]*".

Através do banco em MySQL foi possível procurar proteínas dentro da rede que foram associadas à virulência e à patogenicidade. O resultado dessa busca foi encontrar o nó com o maior grau o gene do complexo proteico de síntese de ácido graxo. Onde na literatura mostrou uma importância desse gene na manutenção da homeostase da levedura de *Cryptococcus* como demonstrado em um trabalho (LONG NAM NGUYEN *et al,* 2009) que as células de levedura Fas2 (uma das subunidades do complexo) são profundamente susceptíveis à glicose, que levou a propor que as células de levedura sem ácidos gordos apresentam metabolismo não controlado em resposta à glicose. Além disso a inibição da síntese de ácido graxo mostou-se ser prejudicial é essa homeostase (MAHMOUD, 1996). Outro trabalho apresentou que uma dose sub-inibitórias de fluconazol mostrou atividade anticryptococcal in vitro na presença de cerulenina (outro antibiótico), um inibidor da sintaxe dos ácidos graxos.

Ao procurar proteínas utilizando a estratégia do valor PI obtido através do alinhamento par a par obtivemos um número muito maior de proteínas relacionadas à virulência e à patogenicidadede acordo com o banco, isso avaliando a diferença funcional obtida. Isso se deve ao fato dessa estratégia propiciar um número muito maior de proteínas e interações entre as mesmas. O resultado obtido se encontra na tabela 10 . O número de proteínas relacionados à virulência e à patogenicidade em *C. gattii,* uma espécie que atinge humanos, obteve o maior número de proteínas. A espécie de *C. neoformans*, que também é uma espécie que afeta humanos obteve um número muito menor. A Supresa foi o C. pinus, uma organismo que atinge plantas ter um alto número de proteínas encontradas.

Tabela 10: Tabela que relaciona os genomas utilizados com as proteínas associadas à virulência e à patogenicidade que foram obtidas atavés da estratégia do PI através do alinhamento dos pares do agrupamento. Os que estão em negrito, são os genomas virulêntos, os que estão em verde, são os genomas que continham anotação gênica no banco de origem.

Nome das Sequências	**Nº Proteínas**
C. bestiolae CBS 10118	9
C. dejecticola CBS 10117	1
C. flavescens NRRL Y-50378	1
C. pinus CBS 10737	31
C. heveanensis BCC8398	0
C. heveanensis CBS569	1
***C. gattii* CBS 7750**	**21**
***C. gattii* R265**	**37**
***C. gattii* WM276**	**93**
***C. neoformans* var *grubii* H99**	**2**
***C. neoformans* var *neoformans* JEC21 uid10698**	**1**
***C. neoformans* var *neoformans* B-3501**	**94**

Nas proteínas que foram encontradas nessa procura, as que mais chamaram atenção foram em *C. neoformans grubii*: piruvato carboxilase é uma enzima da classe ligase que catalisa piruvato para oxaloacetato. Como descrito por Schär e *et al* (2010), o piruvato carboxilase faz parte da Gliconeogenese e caso tenha algum problema na produção do mesmo, o organismo morre, por esse motivo é considerado uma proteína patogenica.

Em *C. gattii WM276* foi encontrada a proteína de resistência a múltiplas drogas 1 (MDR1), que é uma importante proteína da membrana celular que bombeia muitas substâncias estranhas para fora das células. Ela se torna importante para o organismo para resistir a drogas (HUBER *et al*, 2010).

Em todas as cepas de *C. gatti* foi encontrada a proteína de splicing pré-mRNA, importante na via de secreção pelo retículo endoplasmático e complexo de Golgi. O splicing alternativo é um processo celular que aumenta a capacidade de codificação de uma célula a partir de um conjunto limitado de genes (GRÜTZMANN *et al*, 2014).

Em *C. pinus* a proteína "virulence-associated DEAD Box protein" que são enzimas envolvidas em muitos aspectos do metabolismo do RNA. Essa proteína foi descrita na literatura como uma proteína que pode ser encontrada em *Cryptococcus neoformans,*e que ela é um importante fator de virulência para o desenvolvimento da criptococose (HEUNG *et al*, 2005). Porém *C. pinus* é uma

espécie descrita que atinge a árvore do Pinus, por isso o nome, algo bem interessante, uma vez que de acordo com nosso corte dentro do banco, essa proteína não foi encontrada em *Cryptococcus neoformans.*

Em *C. flavences* foi encontado CDC4 (proteína de controle da dividão celular 4), que atua como um mediador de transferência de ubiquitina, levando à sua degradação subsequente através da ubiquitina via proteassoma. Na litatura encotramos que o fenótipo da CDC4 Ca mutante (ausência dessa proteína) sugere que a degradação de proteínas mediadas por ubiquitina é envolvida na regulação do interruptor de dimorphic *C. albicans* (ATIR-LANDE *et al*, 2005).

No organismo de *C. dejecticola* foi encontrada a proteína "major karyopherin", que pertence a um grupo de proteínas envolvidas no transporte de moléculas entre o citoplasma e o núcleo de uma célula eucariótica;

C. bestiolae foi encontrado as duas proteínas de C. dejecticola e C. Flavences, além de uma proteína "heat shock". Essa proteína é descrita para desempenhar um papel fundamental na homeostase térmica, ajustando os níveis de chaperones essenciais para mudanças na temperatura de crescimento, por exemplo em pacientes febris em fungo (BROWN *et al*, 2010).

Com esse resultado percebeu-se que algumas proteínas que estão associadas à virulência e à patogenicidade desse fungo, se encontra em organismos que não afetam humanos, porém o fato de ter encontrado não significa que ela é transcrita, mas é levantamento bem interessante para o trabalho experimental.

6 PERSPECTIVAS

Como perspectivas deste trabalho, podemos destacar:

- Utilização do pipeline para redes de outros organismos com intuito de procurar proteínas *hubs* e não *hubs* que possuem associadas com virulência e à patogenicidade.
- Localizar dentre esses hubs, proteínas que podem ser de interesse como alvo farmacologico ou vacinal.
- Possibilitar a integração de informações dentro de uma rede, como dados provenientes de RNAseq.

7 CONCLUSÕES

Como parte crucial do desenvolvimento do projeto realizamos a anotação de 12 genomas de *Cryptococcus*. No total foram anotados funcionalmente 84.304 genes. Destes, 56.878/84304 (67,46%) representam genes até então não identificados nos genomas através de predição gênica. O processo de anotação e/ou reanotação é importante pois viabiliza a análise comparativa dos genomas e consequentemente a predição computacional das redes é realizada de forma mais acurada.

A partir desse *dataset* inicial construimos redes de interação para os 12 genomas. Para tanto, um *pipeline* computacional foi desenvolvido empregando duas diferentes abordagens. Essas estratégias utilizam como cerne a metodologia de predição computacional de interações proteína-proteína utilizando a abordagem de coevolução. A diferença está relacionada à obtenção do valor de PI (taxa de divergência nucleotídica). Para um grupo este valor foi obtido através do agrupamento de sequências, e para o outro grupo o valor de PI foi calculado através dos pares obtidos através de uma análise combinatória das sequências do agrupamento. Observamos que na segunda metodologia obtivemos um conjunto de pares de interação muito maior (três a 95 vezes maior) quando comparado a primeira abordagem descrita. Os resultados obtidos pela segunda abordagem em relação ao número de nós são corroborados pela literatura. No total foram preditos 24 conjuntos de pares de interação proteína-proteína.

Podemos destacar como pontos interessantes dessas redes que poucos nós possuem um alto grau e *betweenness* indicando que poucos nós são centrais na rede e que em todas as redes houve um nó classificado como *outlier*. Em todas as redes preditas esse *outlier* foi anotado como tendo a função “ATP-dependent protein binding protein” e “ubiquitin” .

Na busca de sequências associadas à virulência e à patogenicidade obtivemos 226.025 proteínas. Em *C. neoformans* var *neoformans* JEC21 encontramos uma proteína classificada como pertencente ao complexo proteico de síntese de ácido graxo. A inibição dessa proteína prejudica a homeostase da levedura, uma vez que ela não controla o metabolismo em resposta a glicose.

A metodologia empregada se torna viável para obtenção das interações proteína-proteína em escala genômica. O banco relacional utilizando MySQL se mostrou útil na obtenção das proteínas associadas à virulência e à

patogenicidade, encontrando as proteínas nas redes que pudesse ser indicas como virulentas e patogênicas.

As proteínas associadas à virulência e à patogenicidade para o organismo de *C. gattii* foram encontradas em um número muito maior, comparado as outras espécies. E que pelo paramentro de corte utilizado nem sempre uma proteínas descrita para organismos virulêntos em humanos como proteínas *heat shock* vão ser encontras de acordo com o corte utilizados e o banco produzido.

Concluiu-se que algumas proteínas descritas como associadas à virulência e à patogenicidade fazem parte do ciclo normal do seu metabolismo, isso acontece pelo fato que proteínas extremamente específicas à virulência e à patogenicidade passam por essas vias, sendo um processo em cadeira.

8 REFERÊNCIAS BIBLIOGRÁFICAS

Altenhoff AM, Dessimoz C. Inferring orthology and paralogy. Evolutionary Genomics: Springer; 2012. p. 259-79.

Altschul SF, Gish W, Miller W, Myers Altenhoff AM, Dessimoz C. Inferring orthology and paralogy. Evolutionary Genomics: Springer; 2012. p. 259-79.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.

Arvestad L, Berglund A-C, Lagergren J, Sennblad B, editors. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. Proceedings of the eighth annual international conference on Resaerch in computational molecular biology; 2004: ACM.

Atir-Lande, Avigail. Gildor, Tsvia. Kornitzer, Daniel. Role for the SCFCDC4 Ubiquitin Ligase in Candida albicans Morphogenesis. Molecular Biology of the Cell. Vol. 16, 2772–2785, June 2005.

Barabasi,A.L.;Albert,R.Emergence of scaling in random networks. Science, v. 286, n. 5439, p. 509--512, 1999. ISSN 0036--8075. Disponívelem:<<GotoISI>://WOS:000083121200054>.

Berg, J.M.; Tymoczko, J.L; Stryer, L. Bioquímica. 6 ed. Rio de Janeiro: Guanabara Koogan, 2008.

Brown, Alistair JP. Leach, Michelle D. Nicholls, Susan. The relevance of heat shock regulation in fungal pathogens of humans. Virulence. 2010 Jul-Aug; 1(4): 330–332. doi: 10.4161/viru.1.4.12364.

Calvo, E. et al. Antifungal therapy in a murine model of disseminated infection by Cryptococcus gattii. Antimicrob Agents Chemother, v. 54, n. 10, p. 4074-7, Oct 2010.

Carver TJ RK, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. Bioinformatics. 21(16):3422-3, 2005

Casadevall, A. Cryptococcus neoformans. 1st. Washington, DC, USA: American Society for Microbiology Press, 1998.
Casadevall, A. et al. Vesicular transport across the fungal cell wall. Trends Microbiol, v. 17, n. 4, p. 158-62, Apr 2009.

Chayakulkeeree, M.; Perfect, J. R. Cryptococcosis. Infect Dis Clin North Am, v. 20, n. 3, p. 507-44, v-vi, Sep 2006.

Chien, C.-T., P.L. Bartel, R. Sternglanz, and S. Fields. The two-hybrid system: A method to identify and clone genes for proteins that interact with a protein of interest. Proc. Natl. Acad. Sci. 88: 9578-9582, 1991.

Easley, David. Kleineberg, Jon. Networks, Crowds, and Markets:Reasoning About a Highly Connected World. 2010.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86-90, 1999.

Fraser, Hunter B. Hirsh, Aaron E. Wall, Dennis P. Eisen, Michael B. Coevolution of gene expression among interacting proteins. PNAS June 15, 2004 vol. 101 no. 24 .

Gabaldon T, Dessimoz C, Huxley-Jones J, Vilella AJ, Sonnhammer EL, Lewis S. Joining forces in the quest for orthologs. Genome biol. 2009;10(9):403.

Griffiths, E. J.; Kretschmer, M.; Kronstad, J. W. Aimless mutants of Cryptococcus neoformans: Failure to disseminate. Fungal Biology Reviews, p. 1-12, 2012.

Gru¨tzmann, Konrad. Szafranski, Karol. Pohl, Martin. Voigt, Kerstin, Petzold, Andreas. Schuster, STEFAN . Fungal Alternative Splicing is Associated with Multicellular Complexity and Virulence: A Genome-Wide Multi-Species Study. DNA RESEARCH 21, 27–39, 2014. doi:10.1093/dnares/dst038.

Gullo FP, Rossi S a, Sardi JDCO, Teodoro VLI, Mendes-Giannini MJS, Fusco-Almeida a M. Cryptococcosis: epidemiology, fungal resistance, and new alternatives for treatment. Eur J Clin Microbiol Infect Dis. Nov; 32(11): 1377–91, 2013.

Harrington ED,Jensen LJ, Bork P. Predicting biological networks from genomic data. Febs Letters 582:1251--1258, 2008.

Heung LJ, Kaiser AE, Luberto C, Del Poeta M (2005) The role and mechanism of diacylglycerol-protein kinase C1 signaling in melanogenesis by Cryptococcus neoformans. J Biol Chem 280: 28547–28555.

Huber, Paula C. Maruiama, Cintia H. Almeida, Wanda P. GLICOPROTEÍNA-P, RESISTÊNCIA A MÚLTIPLAS DROGAS (MDR) E RELAÇÃO ESTRUTURA-ATIVIDADE DE MODULADORES. Quim. Nova, Vol. 33, No. 10, 2148-2154, 2010.

Huynen, M. A.; Bork, P. Measuring genome evolution. Proceedings of the National Academy of Sciences of the United States of America,v.95,n.11,p.5849---5856,May1998. ISSN 0027--8424.

Ideker, T., Galitski T. and Hood, L. A new approach to decoding life: Systems biology. Annu. Rev. Genomics Hum. Genet. 2, 343–372, 2001.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37 (Database issue): D412–6, 2009.

Katoh, K. et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, v. 30, n. 14, p.8, 2002.

Kitano H1. Computational systems biology. Nature. 2002 Nov 14;420(6912):206-10

Kroetz, M.B, Dan Su, Hochstrasser, Mark. Identification of SUMO-Interacting Proteins by Yeast Two-Hybrid Analysis. Molecular Biology of the Cell.Methods Mol Biol. 497, 107–120, 2009 .

Kronstad J1, Saikia S, Nielson ED, Kretschmer M, Jung W, Hu G, Geddes JM, Griffiths EJ, Choi J, Cadieux B, Caza M, Attarian R. Adaptation of Cryptococcus neoformans to mammalian hosts: integrated regulation of metabolism and virulence. Eukaryot Cell. 2012 Feb;11(2):109-18. doi: 10.1128/EC.05273-11. Epub 2011 Dec 2.
Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research. 2003;13(9):2178-89.
Lowe TM1, Eddy SR. A computational screen for methylation guide snoRNAs in yeast. Science. 1999 Feb 19;283(5405):1168-71.

Mahmoud YA1, Abu el Souod SM, Niehaus WG. Purification and characterization of fatty acid synthetase from Cryptococcus neoformans. Mycopathologia. 1996-1997;136(2):75-84.

Marchler-Bauer, A.; Zheng, C.; Chitsaz, F.; Derbyshire, M. K.; Geer, L. Y.; Geer, R. C.; Gonzales, N. R.; Gwadz, M.; Hurwitz, D. I.; Lanczycki, C. J.; Lu, F.; Lu, S.; Marchler, G. H.; Song, J. S.; Thanki, N.; Yamashita, R. A.; Zhang, D.; Bryant, S. H. (2012). "CDD: Conserved domains and protein three-dimensional structure". Nucleic Acids Research. 41 (Database issue): D348–D352. doi:10.1093/nar/gks1243. PMC 3531192.

Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, et al. .Identification of potential interaction networks using sequence-¬-based searches for conserved protein-¬-protein interactions or “interologs”.Genome research 11:2120-2126, 2001.

NEI, M.; LI, W. H. Mathematical--model for studying genetic--variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of theU nited States of America,v.76,n.10,p.5269--5273,1979. ISSN 0027--8424.

Nguyen LN, Trofa D, Nosanchuk JD (2009) Fatty Acid Synthase Impacts the Pathobiology of Candida parapsilosis In Vitro and during Mammalian Infection. PLoS ONE 4(12): e8421. doi:10.1371/journal.pone.0008421.

Pappas, P. G. et al. Cryptococcosis in human immunodeficiency virus-negative patients in the era of effective azole therapy. Clin Infect Dis, v. 33, n. 5, p. 690-9, Sep 2001.

Paulo J, Fernandes A, David F, Sousa N, Lage RA, Izael MDA, Gurgel, A. Criptococose - uma revisão bibliográfica. Acta Vet Bras. 2,2, 32–8. 2008.

Pellegrini, M. et al. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proceedings of The National Academy of Sciences of the United States of America, v. 96, n. 8, p. 4285-4288, 1999.

Pons, Pascal. Latapy, Matthieu. Computing Communities in Large Networks Using Random Walks. Journal of Graph Algorithms and Applications. http://jgaa.info/ vol. 10, no. 2, pp. 191–218 ,2006.

Pujana, M. A. et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nature Genetics, v. 39, n. 11, p. 1338-1349, Nov 2007.
Pujol, A. et al. Unveiling the role of network and systems biology in drug discovery.Trends in Pharmacological Sciences,v.31,n.3,p.115-¬123, 2010.

Resende, A. M. Modelagem de Redes de Interação de Proteínas em Genomas de Parasito Tese de doutorado. Programa de Pos Graduação em Bioinformatica. Belo Horizonte, 2012.

Rezende, Antonio M. ; Folador, Edson l. ; Resende, Daniela de M. ; Ruiz, Jeronimo C. . Computational Prediction of Protein-Protein Interactions in Leishmania Predicted Proteomes. Plos One, v. 7, p. e51304, 2012.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, et al. Artemis: sequence visualization and annotation. Bioinformatics.16(10):944-5, 2000.

Sato T, Yamanishi Y, Kanehisa M,Toh H. The inference of protein- protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21: 3482-3489, 2005.

Sauer,U.;heinemann,M.;zamboni,N.Genetics-Getting closer to the whole picture. Science,v.316,n.5824,p.550--551,2007.ISSN0036--8075.

Sayeon Cho, Sung Goo Park, Do Hee Lee and Byoung Chul Park. Protein-protein Interaction Networks: from Interactions to Networks. Journal of Biochemistry and Molecular Biology, v. 37, 1, January 20, 45-52, 2004.
Schwikowski B1, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol. 2000 Dec;18(12):1257-61.

Severo C, Gazzoni A, Severo L. Criptococose pulmonar. J Bras Pneumol.; 35(11): 1136–44, 2009.

Shannon P, Markiel A, Ozier O, et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". Genome Res. 13 (11): 2498–504, 2003.

Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Research 32: W309–W312, 2004.

Vilella, A. J. et al. VariScan: Analysis of evolutionary patterns from large--scale DNA sequence polymorphismdata.Bioinformatics,v.21,n.11, p. 2791--2793, Jun 1 2005.

Von Mering, C and Huynen, M and Jaeggi, D and Schmidt, S and Bork, P and Snel, B. "STRING: a database of predicted functional associations between proteins". Nucleic Acids Res 31 (1): 258–261, 2003

Xenarios, I. et al. DIP: the Database of Interacting Proteins. Nucleic Acids Research, v.28, n.1, Jan 2000.

Young K . Yeast two-hybrid: so many interactions, (in) so little time. Biol Reprod 58, 1998.

EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.

Arvestad L, Berglund A-C, Lagergren J, Sennblad B, editors. Gene tree reconstruction and orthology analysis based on an integrated model for duplications

and sequence evolution. Proceedings of the eighth annual international conference on Resaerch in computational molecular biology; 2004: ACM.

Berg, J.M.; Tymoczko, J.L; Stryer, L. Bioquímica. 6 ed. Rio de Janeiro: Guanabara Koogan, 2008.

Calvo, E. et al. Antifungal therapy in a murine model of disseminated infection by Cryptococcus gattii. Antimicrob Agents Chemother, v. 54, n. 10, p. 4074-7, Oct 2010.

Carver TJ RK, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. Bioinformatics. 21(16):3422-3, 2005

Casadevall, A. et al. Vesicular transport across the fungal cell wall. Trends Microbiol, v. 17, n. 4, p. 158-62, Apr 2009.

Casadevall, A. Cryptococcus neoformans. 1st. Washington, DC, USA: American Society for Microbiology Press, 1998.

Chayakulkeeree, M.; Perfect, J. R. Cryptococcosis. Infect Dis Clin North Am, v. 20, n. 3, p. 507-44, v-vi, Sep 2006.

Chien, C.-T., P.L. Bartel, R. Sternglanz, and S. Fields. The two-hybrid system: A method to identify and clone genes for proteins that interact with a protein of interest. Proc. Natl. Acad. Sci. 88: 9578-9582, 1991.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86-90, 1999

Gabaldon T, Dessimoz C, Huxley-Jones J, Vilella AJ, Sonnhammer EL, Lewis S. Joining forces in the quest for orthologs. Genome biol. 2009;10(9):403.

Griffiths, E. J.; Kretschmer, M.; Kronstad, J. W. Aimless mutants of Cryptococcus neoformans: Failure to disseminate. Fungal Biology Reviews, p. 1-12, 2012

Gullo FP, Rossi S a, Sardi JDCO, Teodoro VLI, Mendes-Giannini MJS, Fusco-Almeida a M. Cryptococcosis: epidemiology, fungal resistance, and new alternatives for treatment. Eur J Clin Microbiol Infect Dis. Nov; 32(11): 1377–91, 2013.

Harrington ED,Jensen LJ, Bork P. Predicting biological networks from genomic data. Febs Letters 582:1251--1258, 2008.

Ideker, T., Galitski T. and Hood, L. A new approach to decoding life: Systems biology. Annu. Rev. Genomics Hum. Genet. 2, 343–372, 2001.
Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37 (Database issue): D412–6, 2009.

Kroetz, M.B, Dan Su, Hochstrasser, Mark. Identification of SUMO-Interacting Proteins by Yeast Two-Hybrid Analysis. Molecular Biology of the Cell.Methods Mol Biol. 497, 107–120, 2009 .

Lakhani, K.R.; von Hippel, E.. "How Open Source Software Works: Free User to User Assistance". Research Policy June 32, 6: 923–943, 2003.

Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research. 2003;13(9):2178-89.
Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, et al. .Identification of potential interaction networks using sequence-¬-based searches for conserved protein-¬-protein interactions or “interologs”.Genome research 11:2120-2126, 2001.

Mitchell P. Levesque and Philip N. Benfey 2004 Current Biology Vol 14 No 5.

Montojo J, Zuberi K, Rodriguez H, et al.: GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics. ; 26(22): 2927–2928, 2010

Pappas, P. G. et al. Cryptococcosis in human immunodeficiency virus-negative patients in the era of effective azole therapy. Clin Infect Dis, v. 33, n. 5, p. 690-9, Sep 2001.

Paulo J, Fernandes A, David F, Sousa N, Lage RA, Izael MDA, Gurgel, A. Criptococose - uma revisão bibliográfica. Acta Vet Bras. 2,2, 32–8. 2008.

Pellegrini, M. et al. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proceedings of The National Academy of Sciences of the United States of America, v. 96, n. 8, p. 4285-4288, 1999.

Pujana, M. A. et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nature Genetics, v. 39, n. 11, p. 1338-1349, Nov 2007.

Pujol, A. et al. Unveiling the role of network and systems biology in drug discovery.Trends in Pharmacological Sciences,v.31,n.3,p.115-¬123, 2010.

Resende, A. M. Modelagem de Redes de Interação de Proteínas em Genomas de Parasito Tese de doutorado. Programa de Pos Graduação em Bioinformatica. Belo Horizonte, 2012.

Rezende, Antonio M. ; Folador, Edson l. ; Resende, Daniela de M. ; Ruiz, Jeronimo C. . Computational Prediction of Protein-Protein Interactions in Leishmania Predicted Proteomes. Plos One, v. 7, p. e51304, 2012.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, et al. Artemis: sequence visualization and annotation. Bioinformatics.16(10):944-5, 2000.

Sato T, Yamanishi Y, Kanehisa M,Toh H. The inference of protein- protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21: 3482-3489, 2005.

Sayeon Cho, Sung Goo Park, Do Hee Lee and Byoung Chul Park. Protein-protein Interaction Networks: from Interactions to Networks. Journal of Biochemistry and Molecular Biology, v. 37, 1, January 20, 45-52, 2004.

Severo C, Gazzoni A, Severo L. Criptococose pulmonar. J Bras Pneumol.; 35(11): 1136–44, 2009.

Shannon P, Markiel A, Ozier O, et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". Genome Res. 13 (11): 2498–504, 2003.

Skrabana R, Skrabanova M, Csokova N, Sevcik J, Novak M. Intrinsically disordered tau protein in Alzheimer's tangles: a coincidence or a rule? Bratisl Lek List. ;107(9-10):354–358, 2006.

Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Research 32: W309–W312, 2004.

Von Mering, C and Huynen, M and Jaeggi, D and Schmidt, S and Bork, P and Snel, B. "STRING: a database of predicted functional associations between proteins". Nucleic Acids Res 31 (1): 258–261, 2003

Xenarios, I. et al. DIP: the Database of Interacting Proteins. Nucleic Acids Research, v.28, n.1, Jan 2000.

Young K . Yeast two-hybrid: so many interactions, (in) so little time. Biol Reprod 58

9 ANEXOS

Os anexos contêm os dados referentes a algumas tabelas, figuras e gráficos referentes as redes. As informações aqui contidas não estão inclusas no relatório principal devido ao tamanho e principalmente a quantidade de informação aqui contida, uma vez que são 12 genomas, e para cada um gerou-se diversos gráficos, e figuras.

	Query name	size(pb)	Subject name	Link to Subject Accession_number	Subject Description
74	cbest.000730	798	gi|540384503|gb|AFR96265.2|	AFR96265	hypothetical protein CNAG_07687 [Cryptococcus neoformans var. grubii H99]
75	cbest.000740	666	gi|540384421|gb|AFR96095.2|	AFR96095	phosphatidylinositol glycan, class S [Cryptococcus neoformans var. grubii H99]
76	cbest.000750	468	gi|58268182|ref|XP_571247.1|	XP_571247	hypothetical protein [Cryptococcus neoformans var. neoformans JEC21] expressed protein [Cryptococcus ne
77	cbest.000760	553	gi|134113478|ref|XP_774764.1|	XP_774764	hypothetical protein CNBF4430 [Cryptococcus neoformans var. neoformans B-3501A] hypothetical protein Cl
78	cbest.000770	153	gi|321260294|ref|XP_003194867.1|	XP_003194867	hypothetical protein CGB_F4030C [Cryptococcus gattii WM276] hypothetical protein CNBF2670 [Cryptococcu
79	cbest.000780	1154	gi|134113955|ref|XP_774225.1|	XP_774225	hypothetical protein CNBG2070 [Cryptococcus neoformans var. neoformans B-3501A] hypothetical protein C
80	cbest.000790	803	gi|321260296|ref|XP_003194868.1|	XP_003194868	F-box protein; Met30p [Cryptococcus gattii WM276] F-box protein, putative; Met30p [Cryptococcus gattii WM
81	cbest.000800	771	gi|540384419|gb|AFR96092.2|	AFR96092	hypothetical protein CNAG_05772 [Cryptococcus neoformans var. grubii H99]
82	cbest.000810	1081	gi|321260398|ref|XP_003194919.1|	XP_003194919	importin beta-4 subunit [Cryptococcus gattii WM276] Importin beta-4 subunit, putative [Cryptococcus gattii
83	cbest.000820	371	gi|686624101|gb|KGB75929.1|	KGB75929	hypothetical protein CNBG_1767 [Cryptococcus gattii R265]
84	cbest.000830	168	No hits found		
85	cbest.000840	4589	gi|58268752|ref|XP_571532.1|	XP_571532	motor [Cryptococcus neoformans var. neoformans JEC21] motor, putative [Cryptococcus neoformans var. ne
86	cbest.000850	307	gi|321260380|ref|XP_003194910.1|	XP_003194910	hypothetical protein CGB_F5530C [Cryptococcus gattii WM276] conserved hypothetical protein [Cryptococcu
87	cbest.000860	319	gi|58268178|ref|XP_571245.1|	XP_571245	pyrroline-5-carboxylate reductase [Cryptococcus neoformans var. neoformans JEC21] hypothetical protein C
88	cbest.000870	1040	gi|134113382|ref|XP_774716.1|	XP_774716	hypothetical protein CNBF3950 [Cryptococcus neoformans var. neoformans B-3501A] hypothetical protein Cl
89	cbest.000880	162	gi|405121437|gb|AFR96206.1|	AFR96206	ubiquitin-conjugating enzyme E2 [Cryptococcus neoformans var. grubii H99]
90	cbest.000890	444	gi|321260342|ref|XP_003194891.1|	XP_003194891	pre-mRNA splicing factor [Cryptococcus gattii WM276] pre-mRNA splicing factor, putative [Cryptococcus gatt
91	cbest.000900	578	gi|405121504|gb|AFR96273.1|	AFR96273	nicotinate (nicotinamide) nucleotide adenylyltransferase [Cryptococcus neoformans var. grubii H99] nicotina
92	cbest.000910	181	No hits found		
93	cbest.000920	359	gi|405121444|gb|AFR96213.1|	AFR96213	hypothetical protein CNAG_05893 [Cryptococcus neoformans var. grubii H99]

Figura 1: Visualização de parte da tabela com os resultados do BLAST e com o hiperlink.

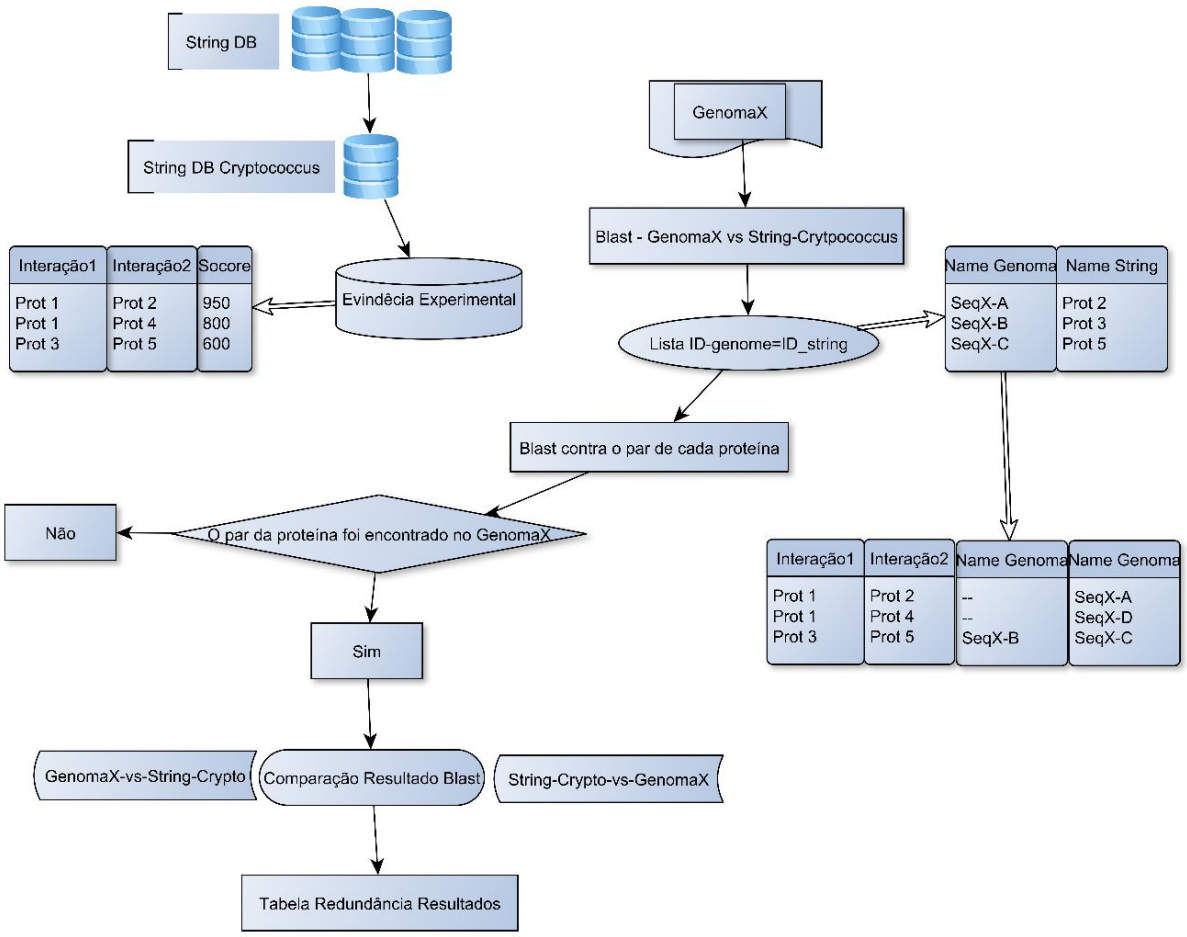

Figura 2: Fluxograma de criação da Tabela de Nomenclatura e conferência dos resultados através do BLAST.

Tabela 1: Tabela correspondente ao número de linhas, onde encontramos um ou dos pares da interação proteína-proteína (Nº linhas sem filtro), ou o par da interação proteína-proteína (Nº linhas com filtro), encontrados nos genomas de *Cryptococcus bestiolae, Cryptococcus dejecticola, Cryptococcus flavescens.*

Organismo	*Cryptococcus bestiolae*		*Cryptococcus dejecticola*		*Cryptococcus flavescens*	
"Cluster"	Nº linhas sem filtro	Nº linhas com filtro	Nº linhas sem filtro	Nº linhas com filtro	Nº linhas sem filtro	Nº linhas com filtro
2 genes 2 taxas	0	0	0	0	0	0
3 genes 3 taxas	1536	0	1536	0	0	0
4 genes 4 taxas	32	0	0	0	0	0
5 genes 5 taxas	0	0	0	0	0	0
6 genes 6 taxas	0	0	0	0	0	0
7 genes 7 taxas	0	0	0	0	0	0
8 genes 8 taxas	760	0	0	0	762	0
9 genes 9 taxas	2540	4	3490	10	260	0
10 genes 10 taxas	1904	6	1298	2	706	0
11 genes 11 taxas	1124	2	2202	4	1934	2
12 genes 12 taxas	30578	972	29582	894	26678	748
Total	38474	984/446	38108	910/433	30340	750/353

Tabela 2: Tabela correspondente ao número de linhas, onde encontramos um ou dos pares da interação proteína-proteína (Nº linhas sem filtro), ou o par da interação proteína-proteína (Nº linhas com filtro), encontrados nos genomas de *Cryptococcus neoformans var. neoformans JEC21, Cryptococcus neoformans var. grubii H99, Cryptococcus neoformans var. neoformans B-3501.*

Organismo	*Cryptococcus neoformans var. neoformans JEC21*		*Cryptococcus neoformans var. grubii H99*		*Cryptococcus neoformans var. neoformans B-3501*	
"Cluster"	Nº linhas sem filtro	Nº linhas com filtro	Nº linhas sem filtro	Nº linhas com filtro	Nº linhas sem filtro	Nº linhas com filtro
2 genes 2 taxas	990	2	0	0	926	2
3 genes 3 taxas	1386	2	1174	2	1386	2
4 genes 4 taxas	1162	2	1130	2	1162	2
5 genes 5 taxas	438	0	0	0	438	0
6 genes 6 taxas	3632	8	3760	8	3766	8
7 genes 7 taxas	2790	12	2790	12	2790	12
8 genes 8 taxas	1968	0	1968	0	1574	0
9 genes 9 taxas	4316	12	4316	12	4316	12

10 genes 10 taxas	2702	6	2598	6	2688	6
11 genes 11 taxas	2624	6	2624	6	2624	6
12 genes 12 taxas	31112	958	30610	902	31112	958
Total	53120	1008/482	50970	950/454	52782	1008/482

Tabela 3: Tabela correspondente ao número de linhas, onde encontramos um ou dos pares da interação proteína-proteína (Nº linhas sem filtro), ou o par da interação proteína-proteína (Nº linhas com filtro), encontrados nos genomas de *Cryptococcus gattii WM276, Cryptococcus gattii CBS 7750, Cryptococcus gattii R265.*

Organismo	*Cryptococcus gattii WM276*		*Cryptococcus gattii CBS 7750*		*Cryptococcus gattii R265*	
"Cluster"	Nº linhas sem filtro	Nº linhas com filtro	Nº linhas sem filtro	Nº linhas com filtro	Nº linhas sem filtro	Nº linhas com filtro
2 genes 2 taxas	0	0	3378	6	3670	6
3 genes 3 taxas	624	0	0	0	412	0
4 genes 4 taxas	1236	2	0	0	0	0
5 genes 5 taxas	6	0	6	0	6	0
6 genes 6 taxas	4036	6	3622	8	3756	8
7 genes 7 taxas	3832	20	2528	12	2528	12
8 genes 8 taxas	2458	2	1968	0	1968	0
9 genes 9 taxas	4608	14	4816	16	4816	16
10 genes 10 taxas	2702	6	2292	2	2292	2
11 genes 11 taxas	2624	6	2106	6	2624	6
12 genes 12 taxas	28966	980	31022	944	3112	958
Total	51092	1036/476	51738	994/473	25184	1008/480

Tabela 4: Tabela correspondente ao número de linhas, onde encontramos um ou dos pares da interação proteína-proteína (Nº linhas sem filtro), ou o par da interação proteína-proteína (Nº linhas com filtro), encontrados nos genomas de *Cryptococcus pinus, Cryptococcus heveanensis BCC8398, Cryptococcus heveanensis CBS569.*

Organismo	*Cryptococcus pinus*		*Cryptococcus heveanensis BCC8398*		*Cryptococcus heveanensis CBS569*	
"Cluster"	Nº linhas sem filtro	Nº linhas com filtro	Nº linhas sem filtro	Nº linhas com filtro	Nº linhas sem filtro	Nº linhas com filtro
2 genes 2 taxas	0	0	900	0	900	0
3 genes 3 taxas	2514	0	0	0	0	0
4 genes 4 taxas	0	0	0	0	0	0
5 genes 5 taxas	0	0	0	0	0	0
6 genes 6 taxas	0	0	0	0	0	0
7 genes 7 taxas	1170	0	216	0	262	0
8 genes 8 taxas	382	0	52	0	814	0
9 genes 9 taxas	0	0	1476	0	926	2

10 genes 10 taxas	468	0	1578	6	1004	2
11 genes 11 taxas	1594	2	1990	4	2038	4
12 genes 12 taxas	29290	892	28186	818	28694	866
Total	35418	894/495	34398	828/393	34638	874/415

Tabela 5: Tabela correspondente ao número de nós e arestas que cada genoma possui.

Organismo	Nº nós ou vértices	Nº arestas ou conexões
C. bestiolae	89	446
C. dejecticola	87	433
C. flavescens	63	353
C. neoformans var. neoformans JEC21	115	482
C. neoformans var. grubii H99	111	454
C. neoformans var. neoformans B-3501	115	482
C. gattii WM276	118	476
C. gattii CBS 7750	113	473
C. gattii R265	114	480
C. pinus	79	425
C. heveanensis BCC8398	81	393
C. heveanensis CBS569	84	415

Tabela 6: Análise estatística das redes de interações proteína– proteína de 6 genomas

Intens	*C. bestiolae*	*C. dejecticola*	*C. flavescens*	*C. neoformans var. neoformans JEC21*	*C. neoformans var. grubii H99*	*C. neoformans var. neoformans B-3501*
Grau mínimo	1	1	1	1	1	1
1st Qu	3	3	3	2	2	2
Mediana	6	7	9	4	4	4
Média	10.02	9.954	11.21	8.383	8.18	8.383
3rd Qu	17	17.5	18	12	12	12
Grau máximo	76	75	60	79	77	79

Tabela 7: Análise estatística das redes de interações proteína– proteína dos outros 6 genomas

Intens	*C. gattii WM276*	*C. gattii CBS 7750*	*C. gattii R265*	*C. pinus*	*C. heveanensis BCC8398*	*C. heveanensis CBS569*
Grau mínimo	1	1	1	1	1	1
1st Qu	2	2	2	3	3	2
Mediana	4	4	4	8	7	7
Média	8.068	8.372	8.421	10.76	9.704	9.881

3rd Qu	12	12	12	18	17	17.250
Grau máximo	75	78	79	75	70	72

Estatística de interações proteína – proteína por coevolução em *Cryptococcus bestiolae*

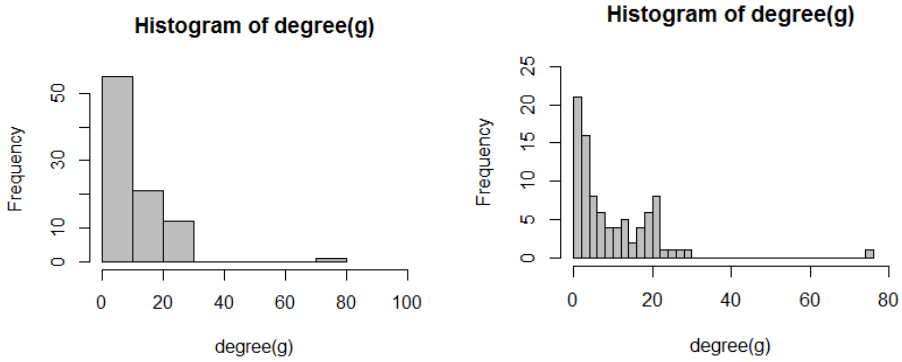

Figura 3: Histogramas mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus bestiolae.* Eixo x grau do nó, Eixo Y frequência.

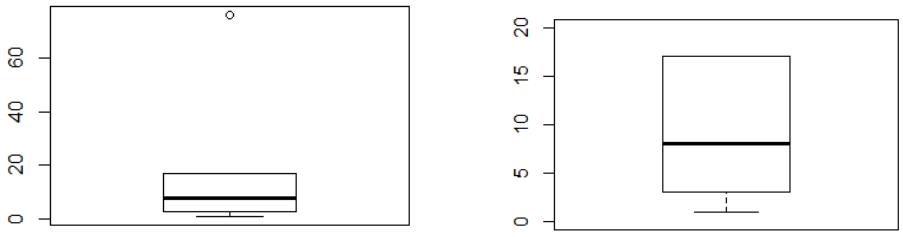

Figura 4: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus bestiolae.* Eixo x grau dos nós.

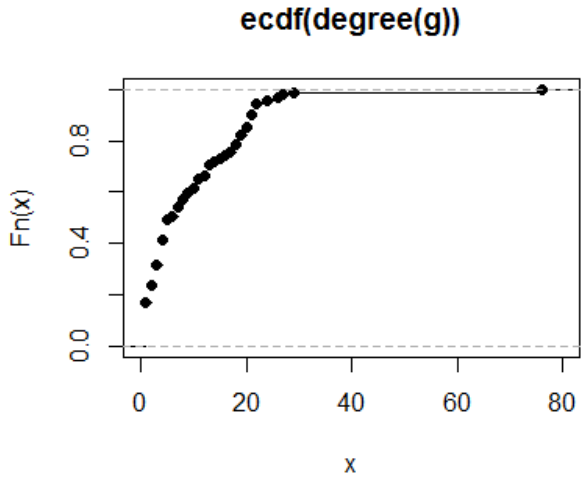

Figura 5: O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus bestiolae*.

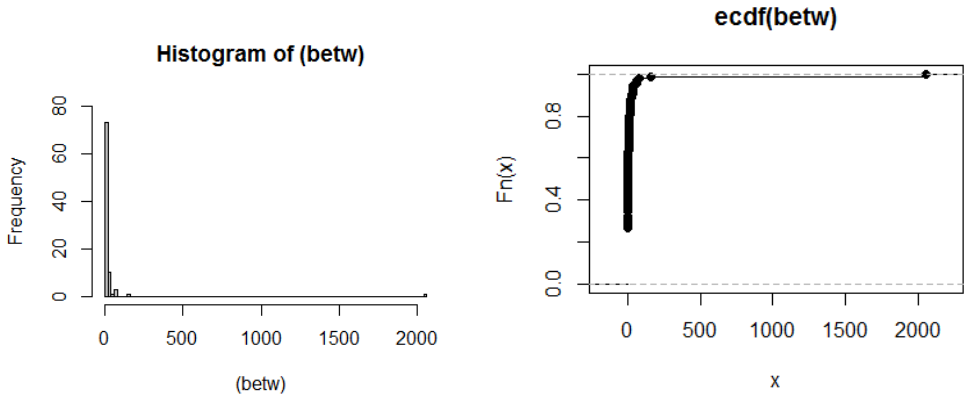

Figura 6 : Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus bestiolae.*

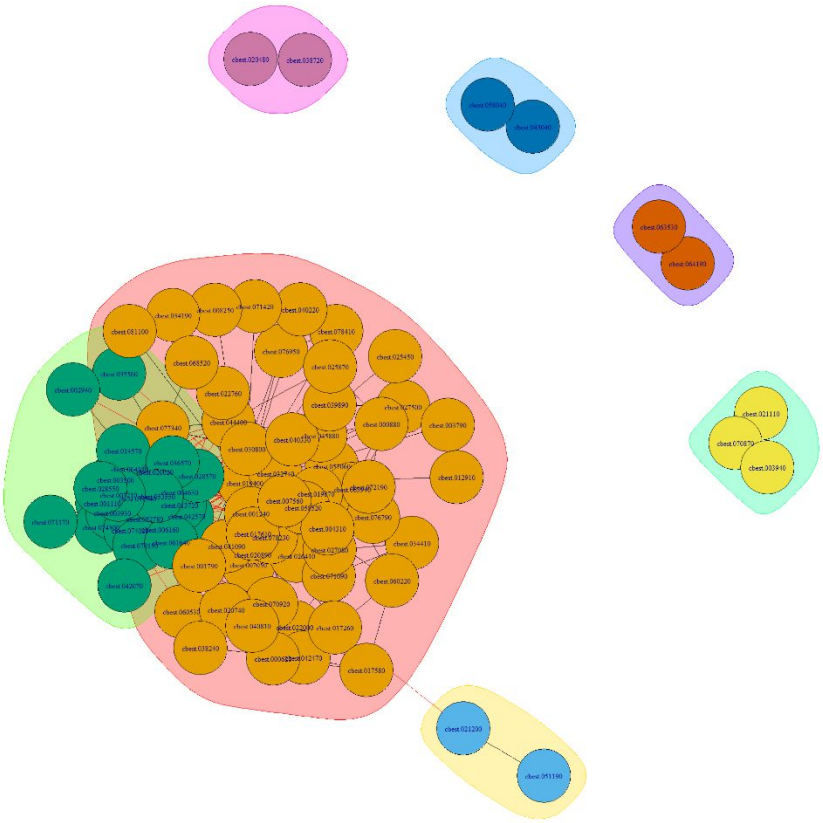

Figura 7; Rede gerada através de “Walktrap Community” que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie de *Cryptococcus bestiolae.*

3.1.1 Estatística de interações proteína – proteína por coevolução em *Cryptococcus dejecticola*

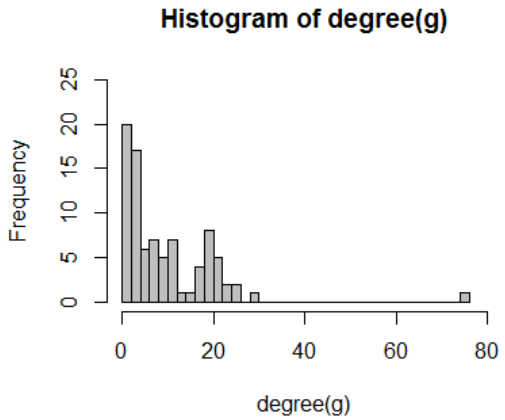

Figura 8: Histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus dejecticola.* Eixo x grau do nó, Eixo Y frequência.

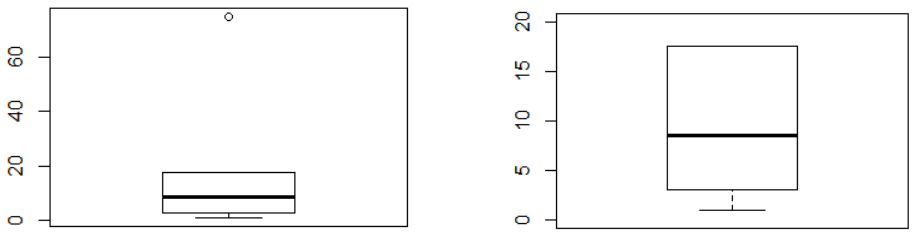

Figura 9: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus dejecticola.* Eixo x grau dos nós.

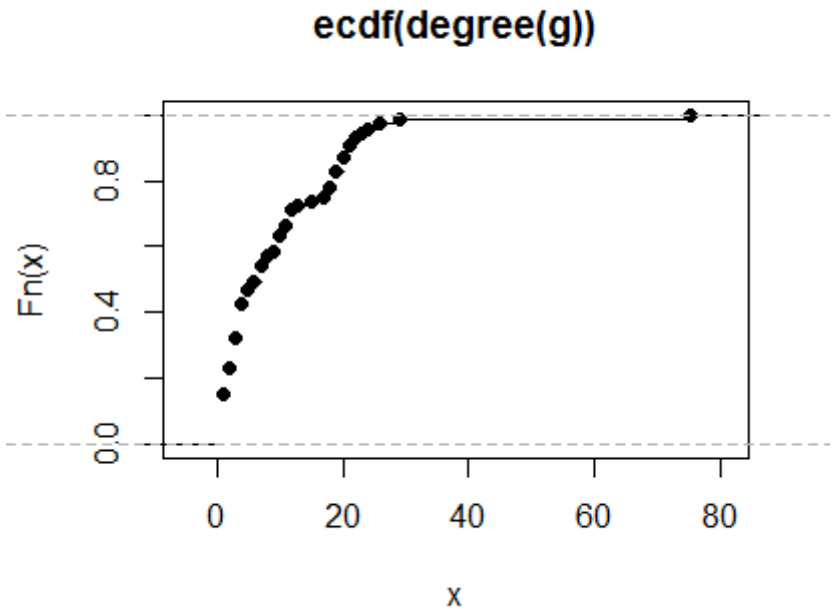

Figura 10: O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus dejecticola*.

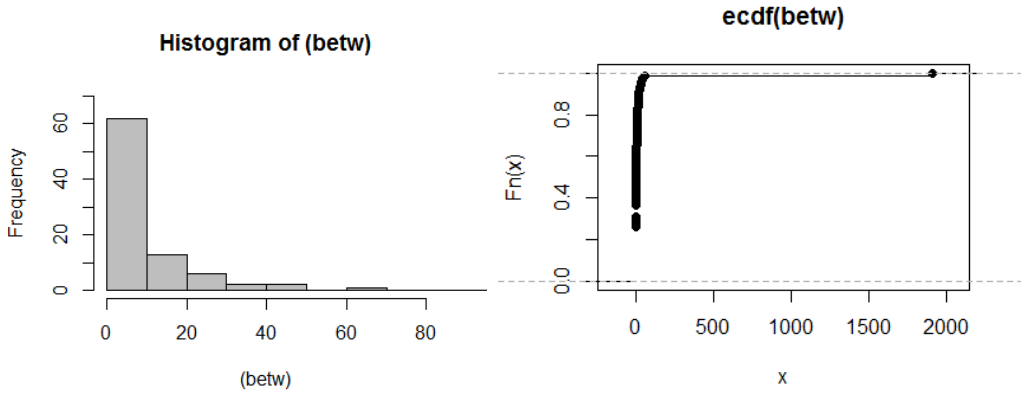

Figura 11: Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus dejecticola.*

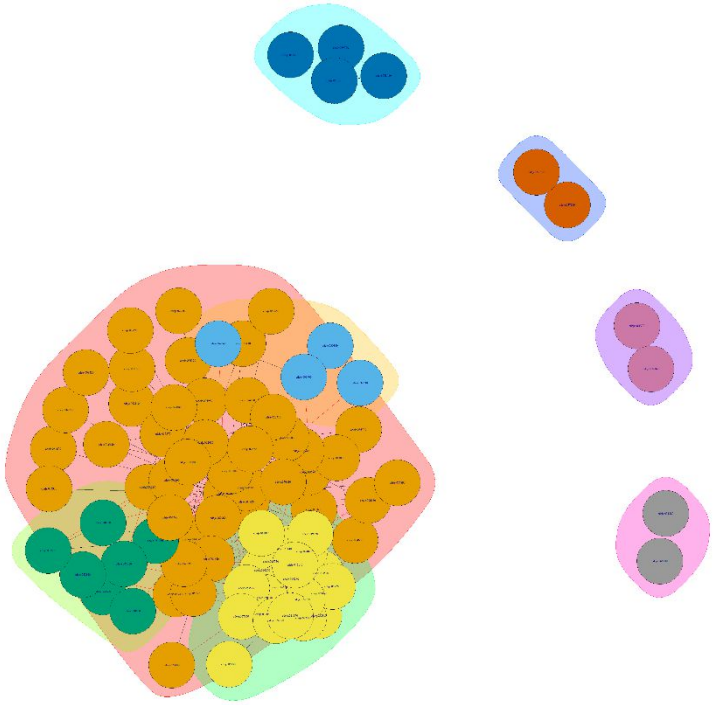

Figura 12: Rede gerada através de “Walktrap Community” que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie de *Cryptococcus dejecticola.*

Estatística de interações proteína – proteína por coevolução em *Cryptococcus flavescens* NRRL Y-50378

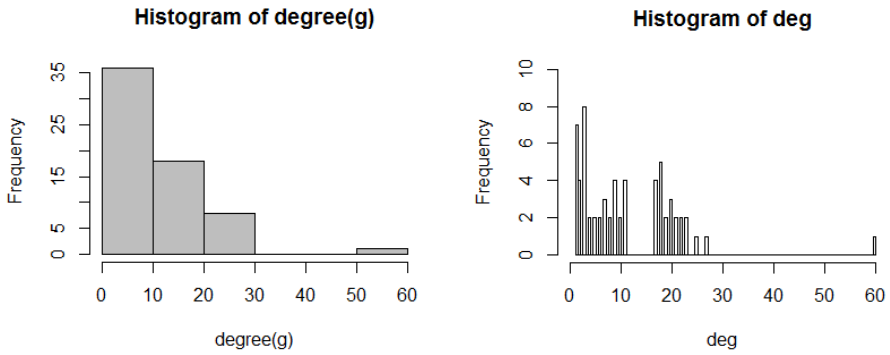

Figura 13: Histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus flavescens.* Eixo x grau do nó, Eixo Y frequência.

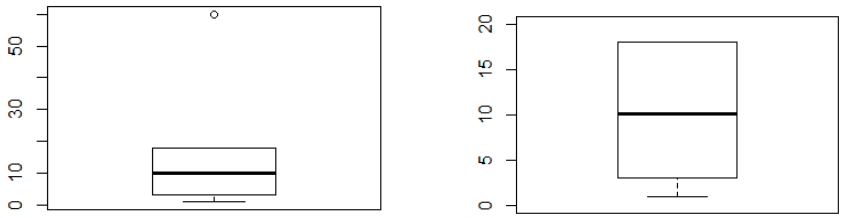

Figura 14: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus flavescens.* Eixo x grau dos nós.

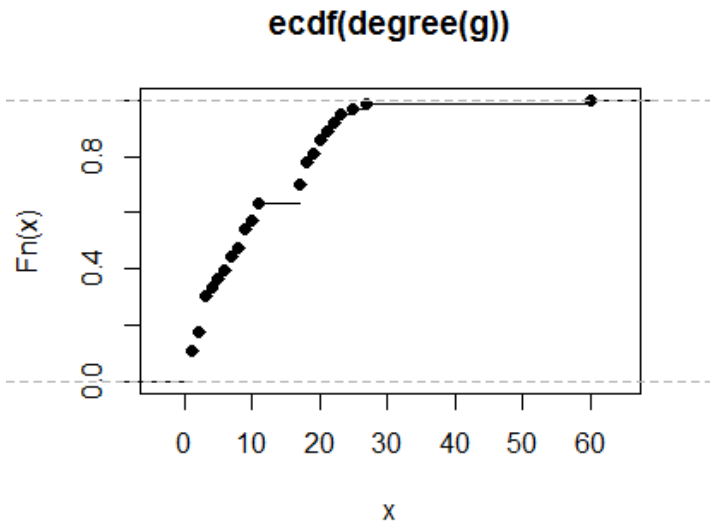

Figura 15 O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus flavescens*

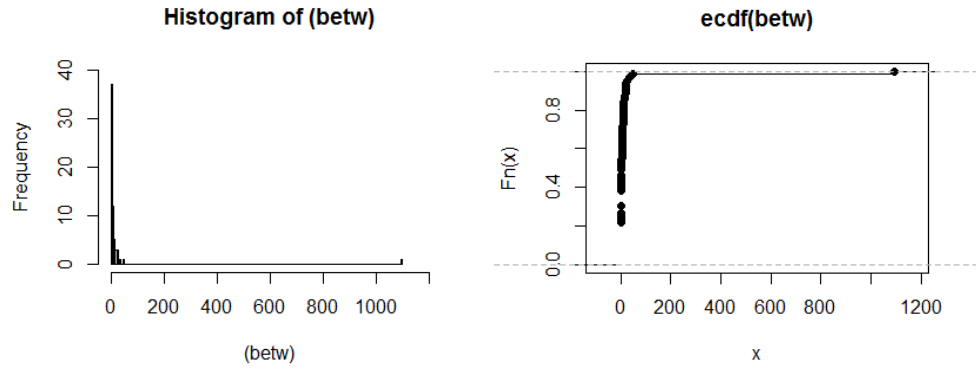

Figura 16: Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus flavescens.*

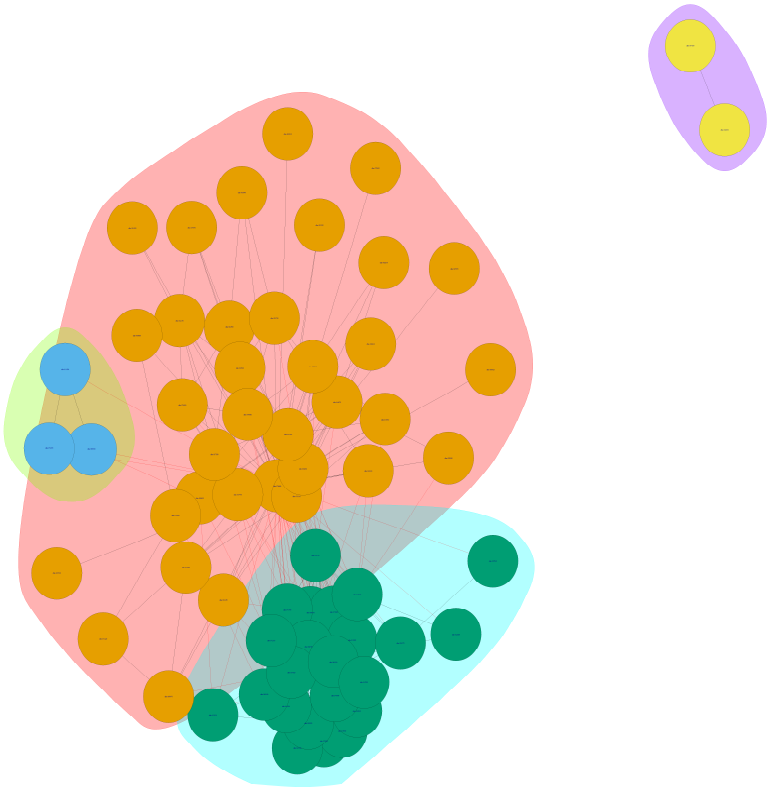

Figura 17; Rede gerada através de "Walktrap Community" que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie de *Cryptococcus flavescens.*

Estatística de interações proteína – proteína por coevolução *Cryptococcus neoformans* var. *neoformans* JEC21

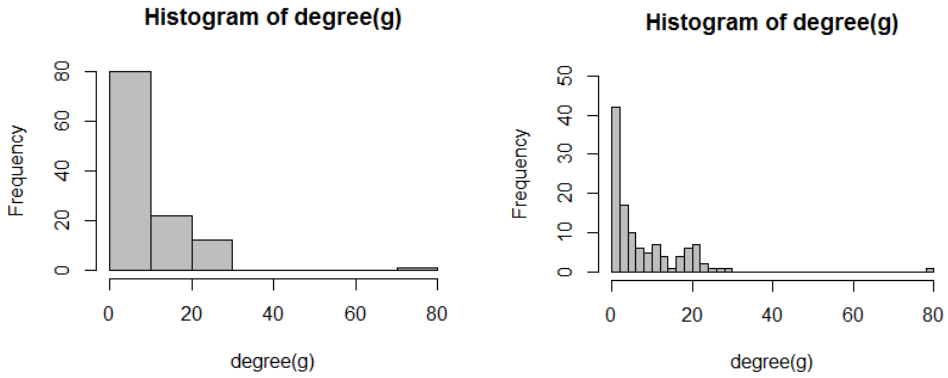

Figura 18: Histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus neoformans* var. *neoformans* JEC21. Eixo x grau do nó, Eixo Y frequência.

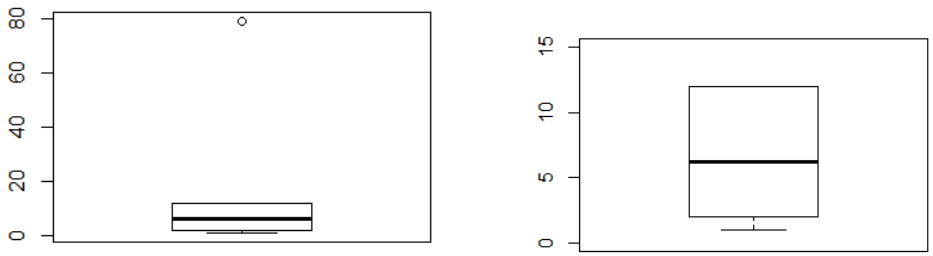

Figura 19: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus neoformans* var. *neoformans* JEC21. Eixo x grau dos nós.

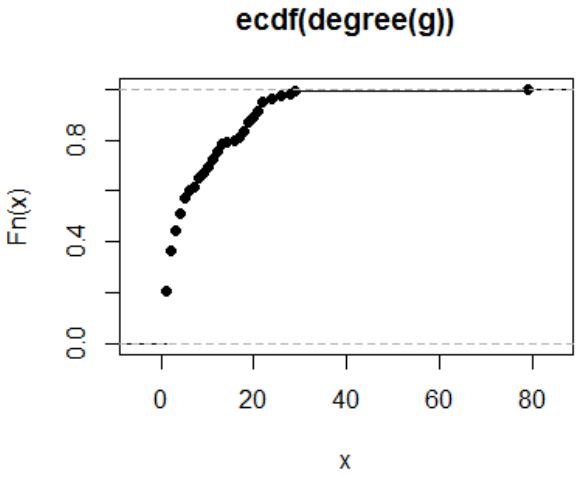

Figura 20: O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus neoformans* var. *neoformans* JEC21.

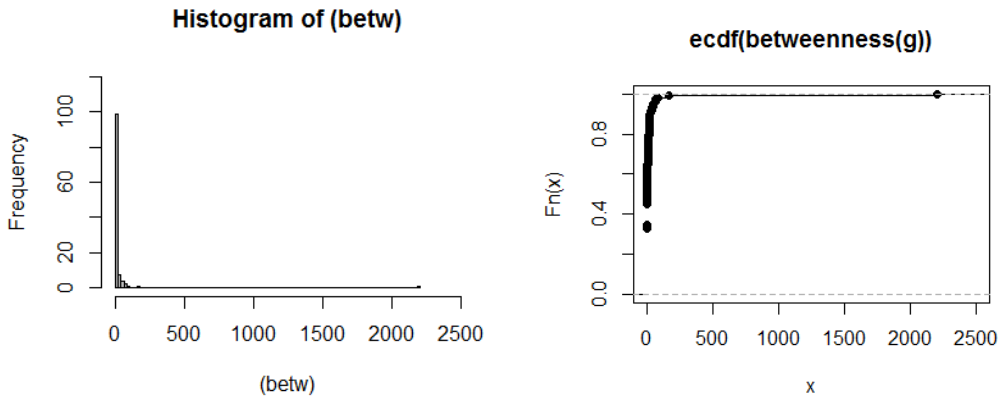

Figura 21: Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus neoformans* var. *neoformans* JEC21.

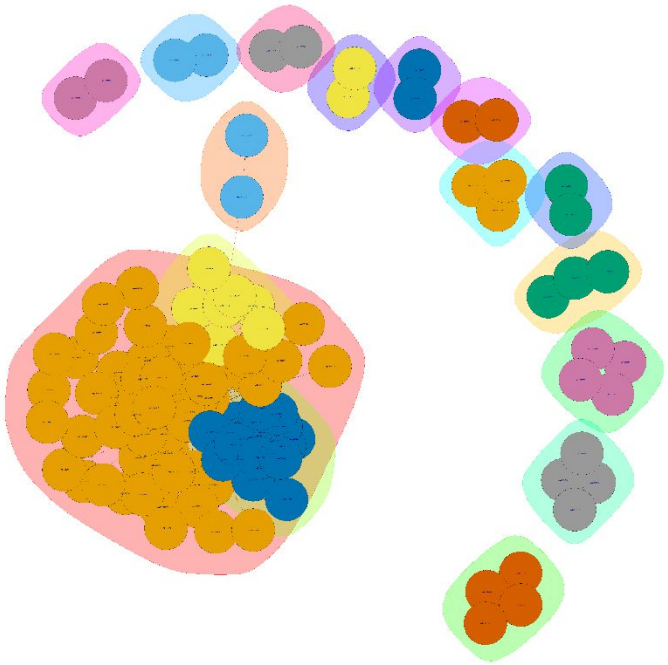

Figura 22; Rede gerada através de “Walktrap Community” que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie de *Cryptococcus neoformans* var. *neoformans* JEC21.

Estatística de interações proteína – proteína por coevolução *Cryptococcus neoformans* var. *grubii* H99

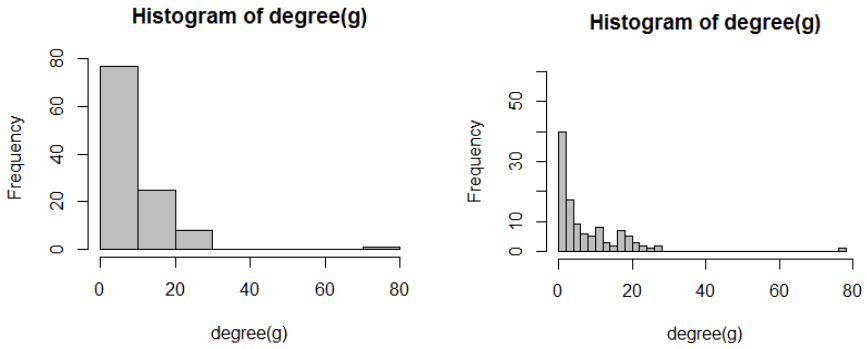

Figura 23: Histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus neoformans* var. *grubii* H99. Eixo x grau do nó, Eixo Y frequência.

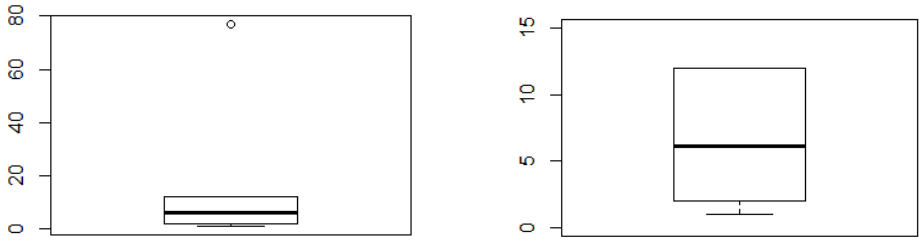

Figura 24: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus neoformans* var. *grubii*. Eixo x grau dos nós.

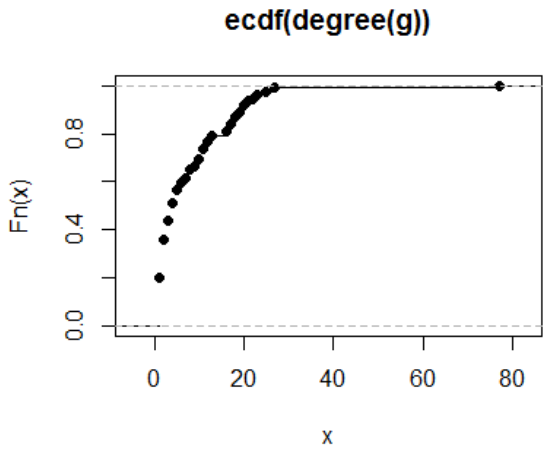

Figura 25: O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus neoformans* var. *grubii*.

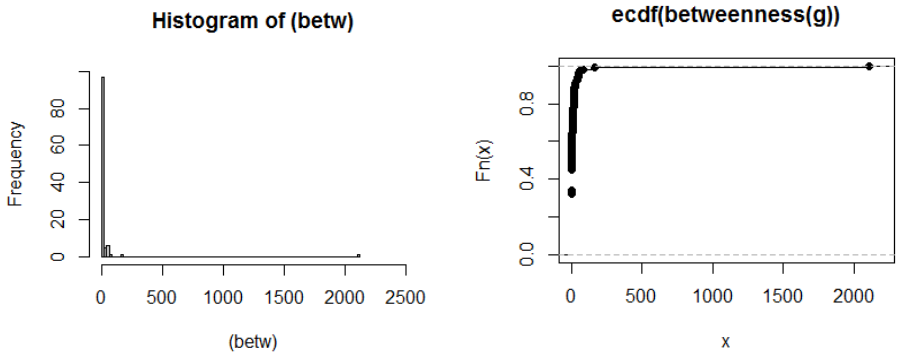

Figura 26: Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus neoformans* var. *grubii*.

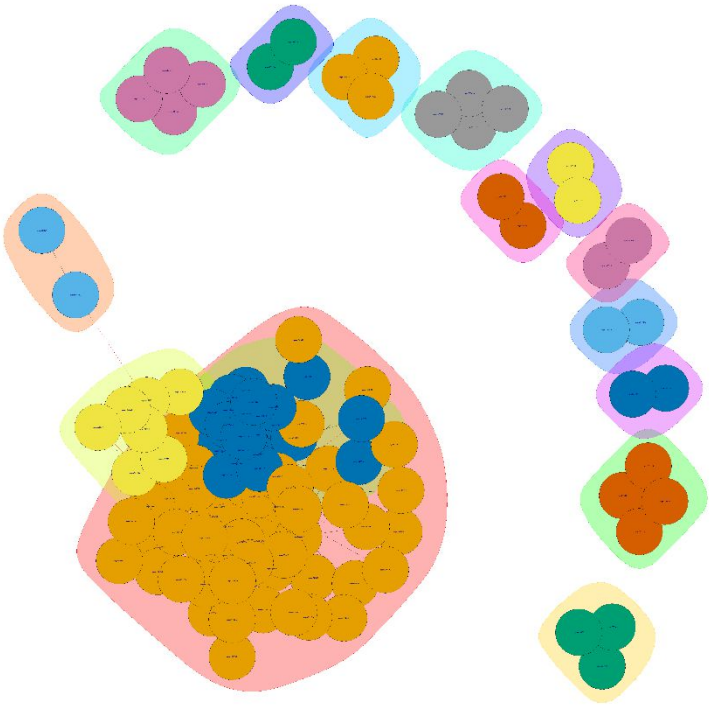

Figura 27; Rede gerada através de "Walktrap Community" que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie de *Cryptococcus neoformans* var. *grubii.*

Estatística de interações proteína – proteína por coevolução *Cryptococcus neoformans* var. *neoformans* B-3501

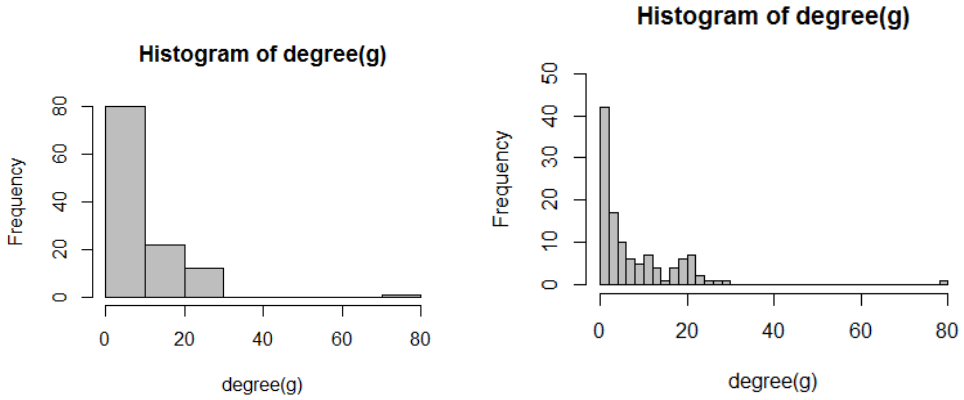

Figura 28: Histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus neoformans* var. *neoformans* B-3501 . Eixo x grau do nó, Eixo Y frequência.

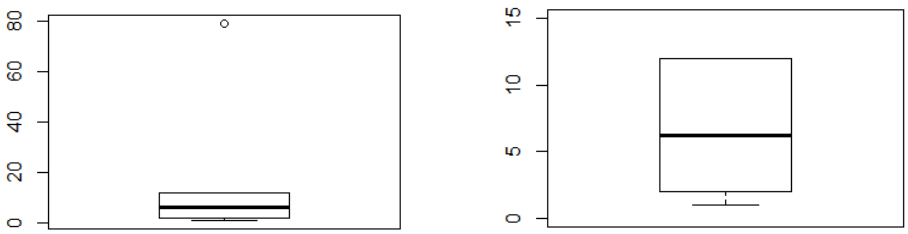

Figura 29: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus neoformans* var. *neoformans* B-3501. Eixo x grau dos nós.

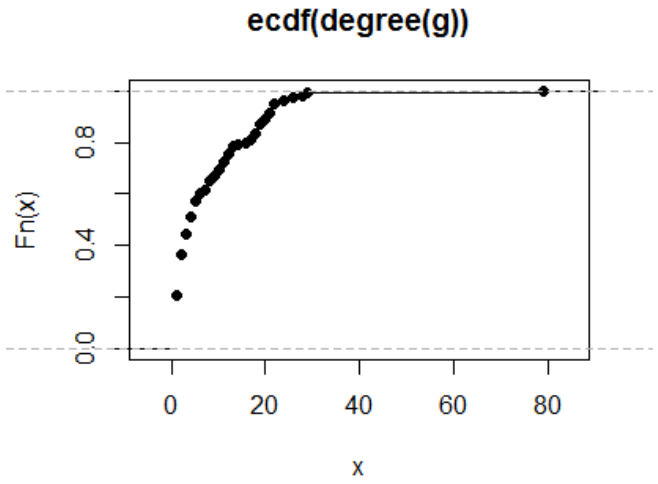

Figura 30: O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus neoformans* var. *neoformans* B-350.

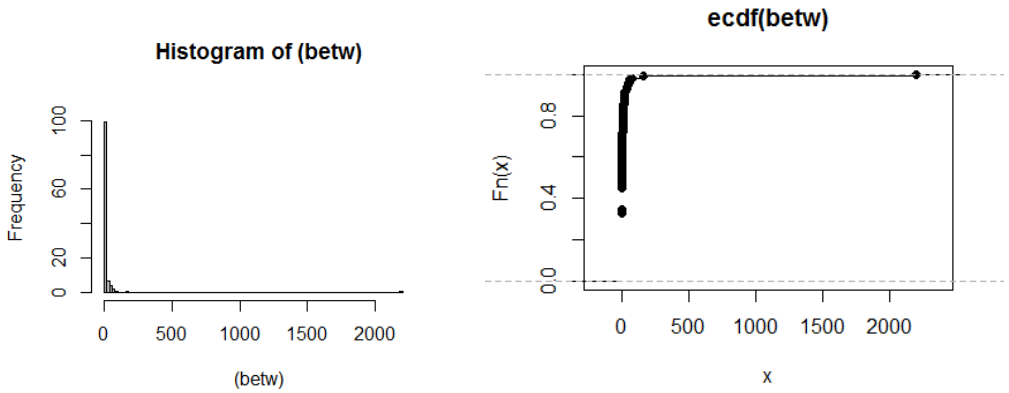

Figura 31: Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus neoformans* var. *neoformans* B-350.

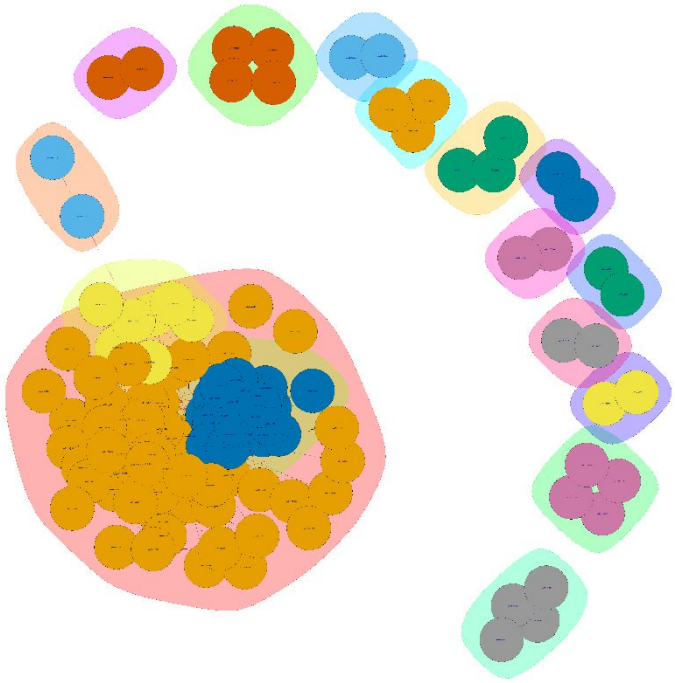

Figura 32; Rede gerada através de “Walktrap Community” que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie de *Cryptococcus neoformans* var. *neoformans* B-350.

Estatística de interações proteína – proteína por coevolução *Cryptococcus gattii* WM276

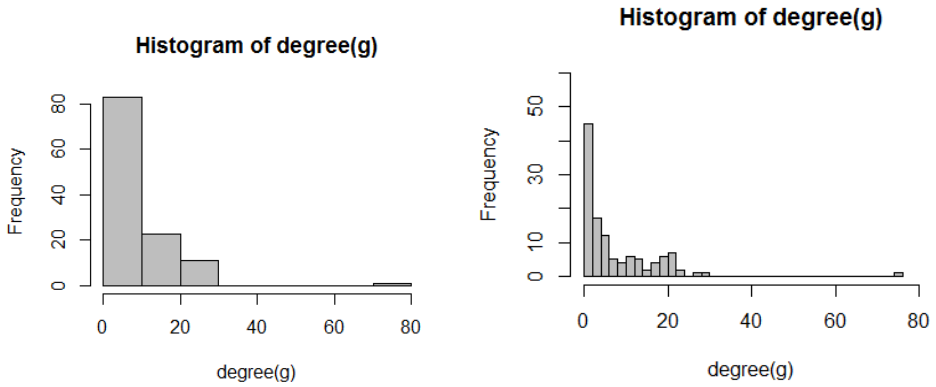

Figura 33: Histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus gattii* WM276 . Eixo x grau do nó, Eixo Y frequência.

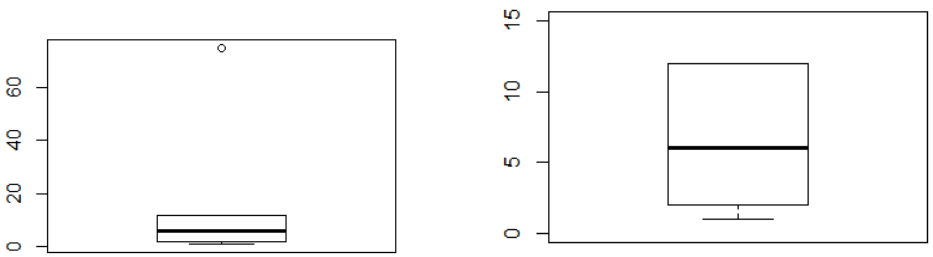

Figura 34: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus gattii* WM276. Eixo x grau dos nós.

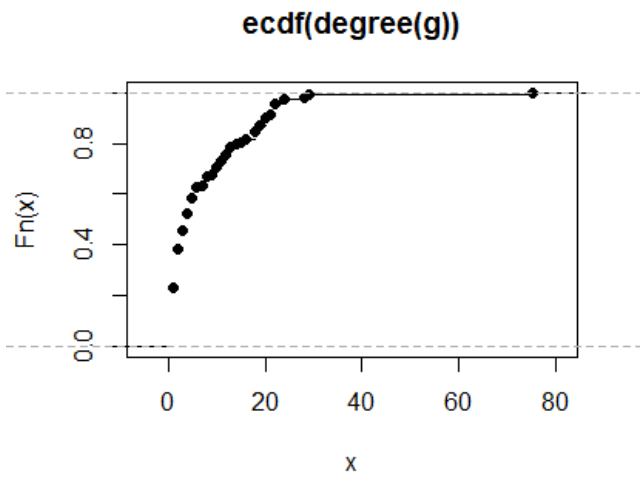

Figura 35: O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus gattii* WM276.

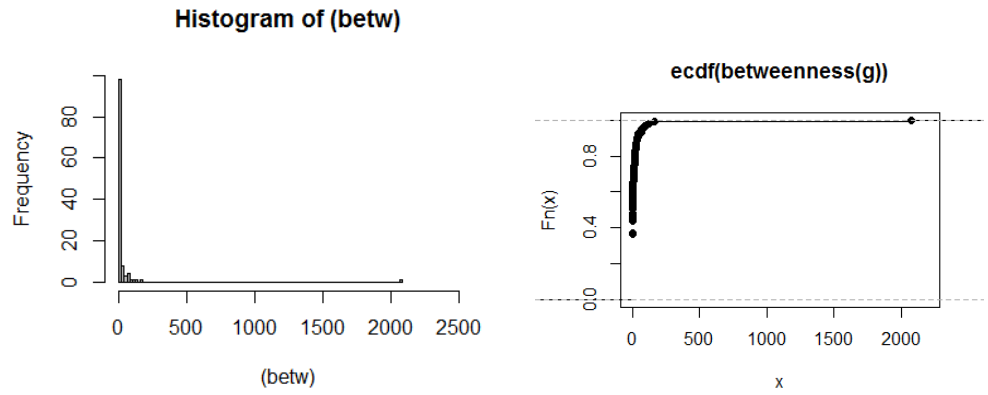

Figura 36: Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus gattii* WM276.

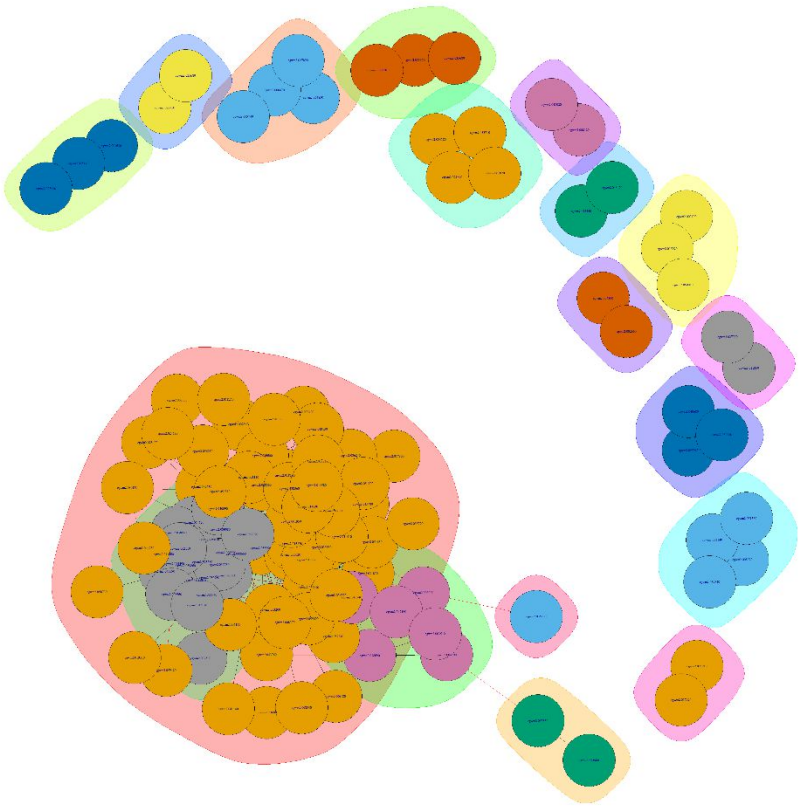

Figura 37; Rede gerada através de "Walktrap Community" que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie de *Cryptococcus gattii* WM276.

Estatística de interações proteína – proteína por coevolução *Cryptococcus gattii* CBS 7750

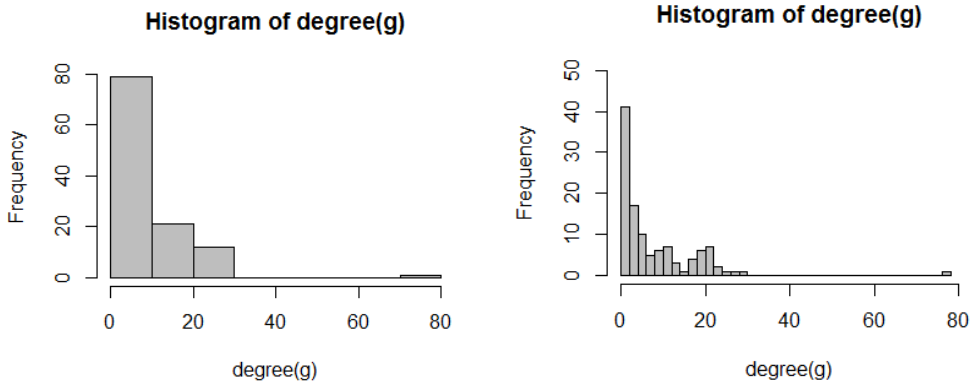

Figura 38: Histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus gattii* CBS 7750. Eixo x grau do nó, Eixo Y frequência.

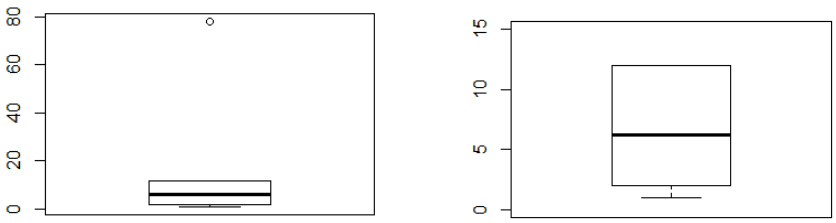

Figura 39: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus gattii* CBS 7750. Eixo x grau dos nós.

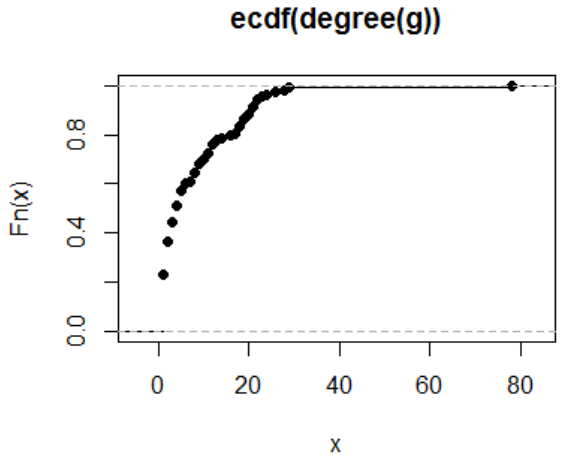

Figura 40 O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus gattii* CBS 7750.

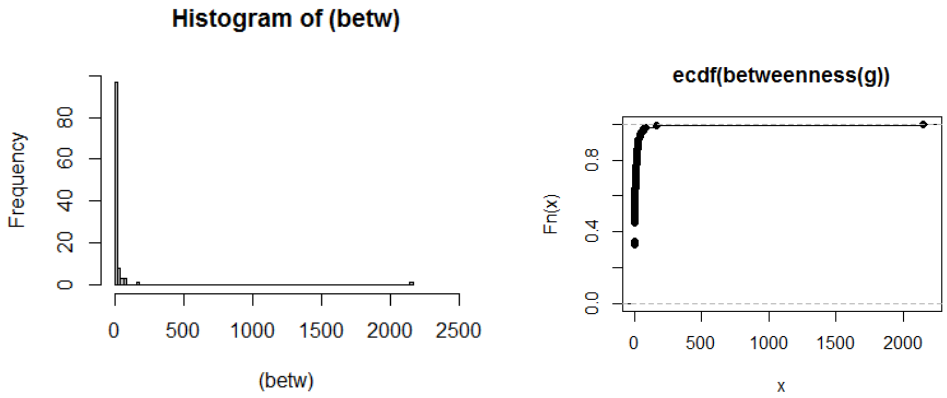

Figura 41: Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus gattii* CBS 7750.

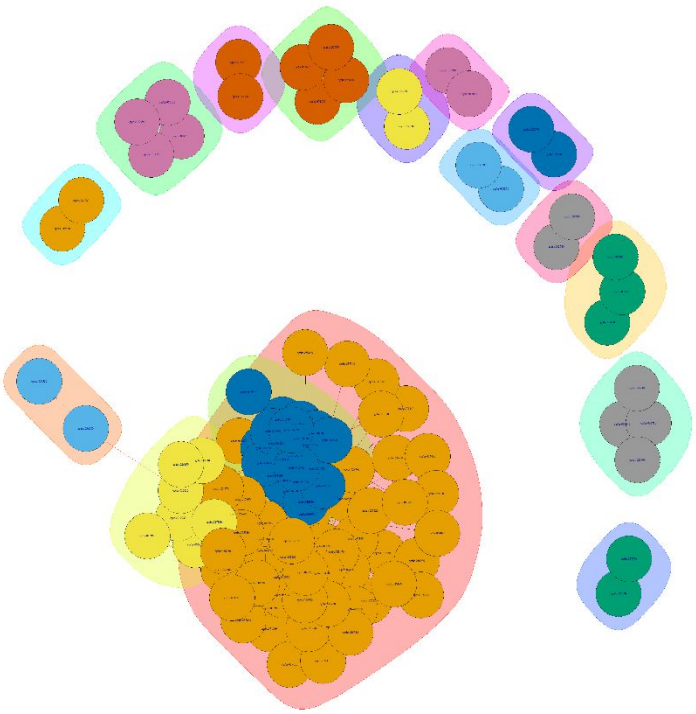

Figura 42; Rede gerada através de “Walktrap Community” que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie *Cryptococcus gattii* CBS 7750.

Estatística de interações proteína – proteína por coevolução *Cryptococcus gattii* R265

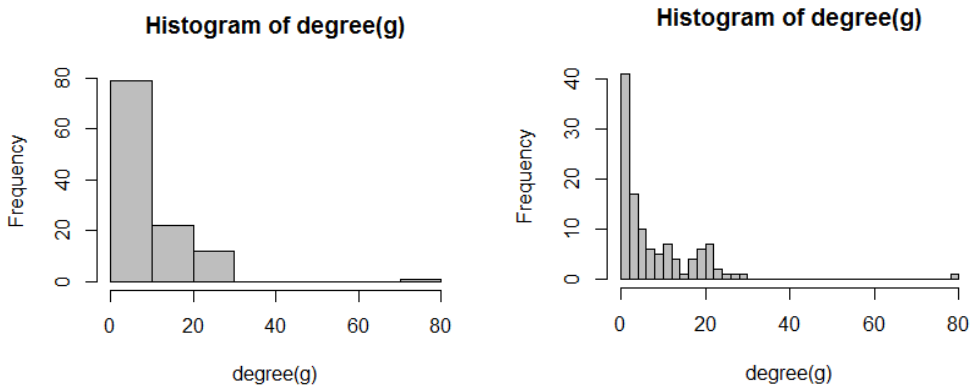

Figura 43: Histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus gattii* R265. Eixo x grau do nó, Eixo Y frequência.

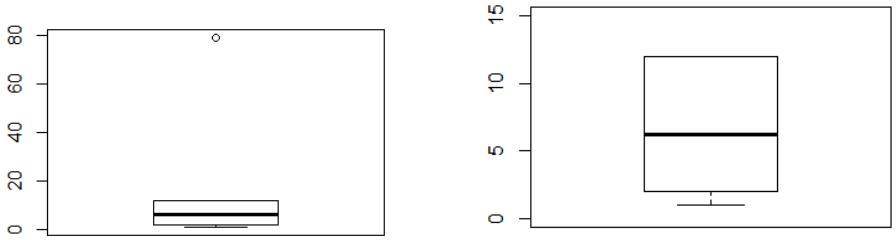

Figura 44: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus gattii* R265. Eixo x grau dos nós.

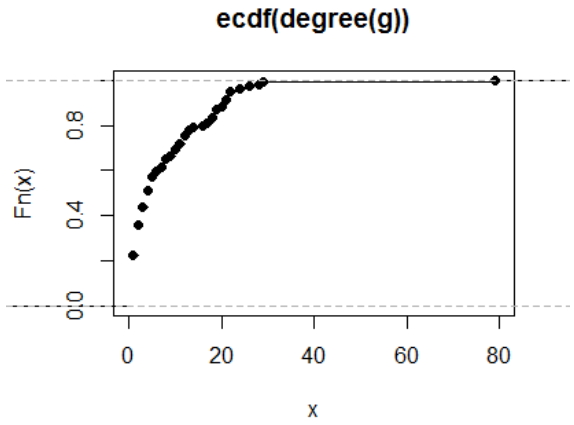

Figura 45: O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus gattii* R265.

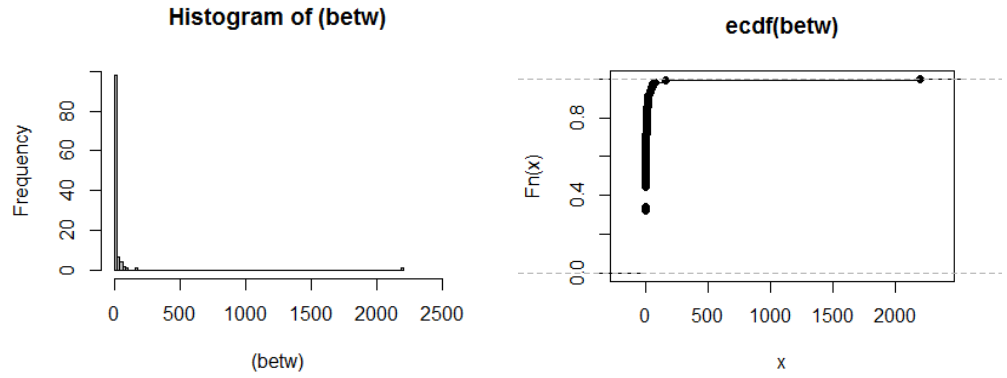

Figura 46: Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus gattii* R265.

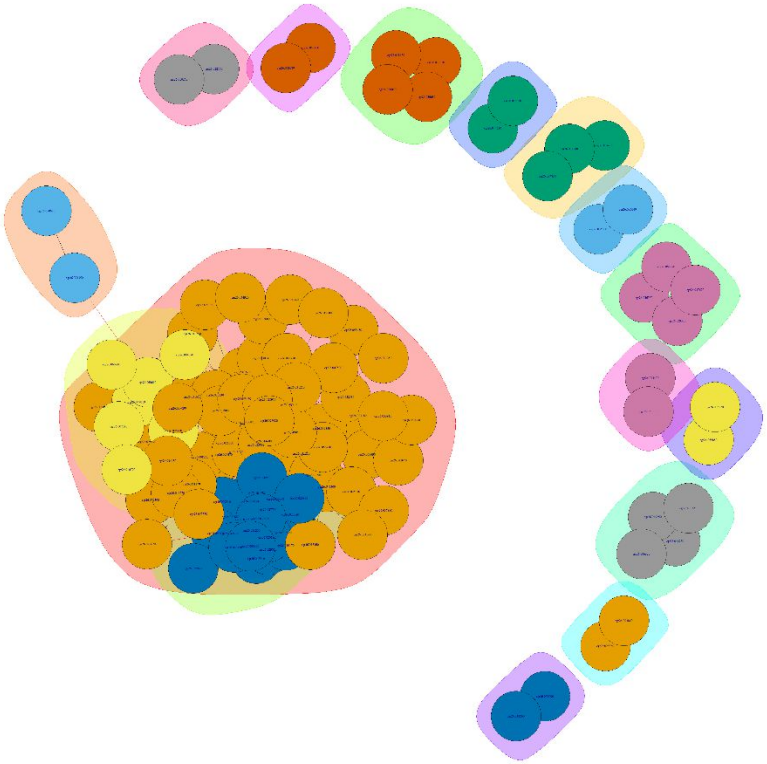

Figura 47; Rede gerada através de "Walktrap Community" que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie *Cryptococcus gattii* R265.

3.1.2 Estatística de interações proteína – proteína por coevolução *Cryptococcus pinus* CBS 10737

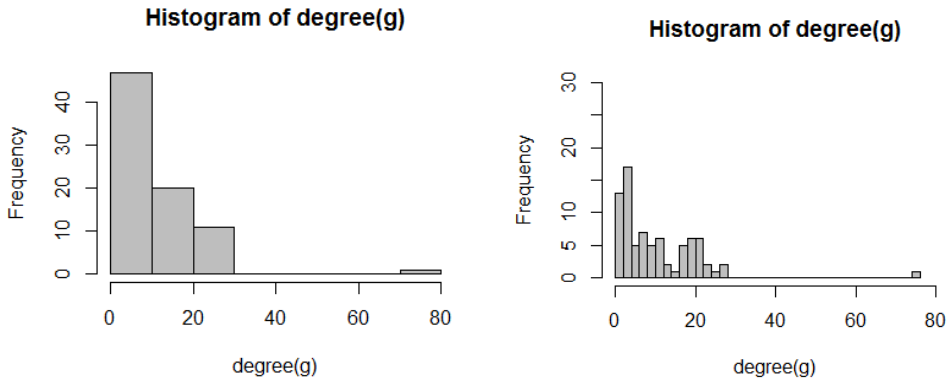

Figura 56: Histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus pinus* CBS 10737. Eixo x grau do nó, Eixo Y frequência.

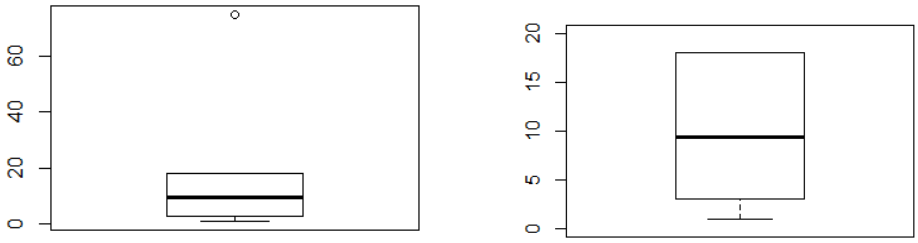

Figura 48: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus pinus* CBS 10737. Eixo x grau dos nós.

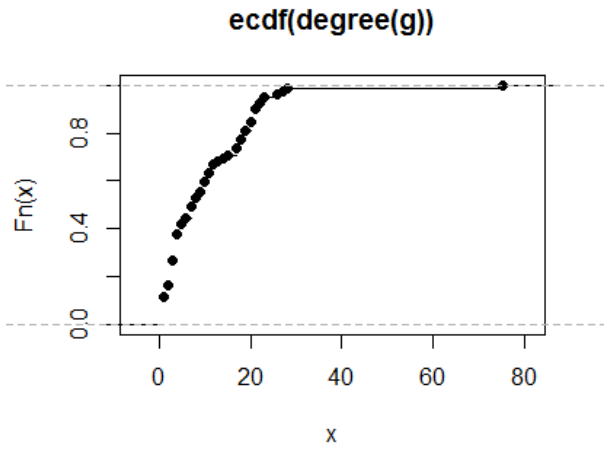

Figura 49: O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus pinus* CBS 10737.

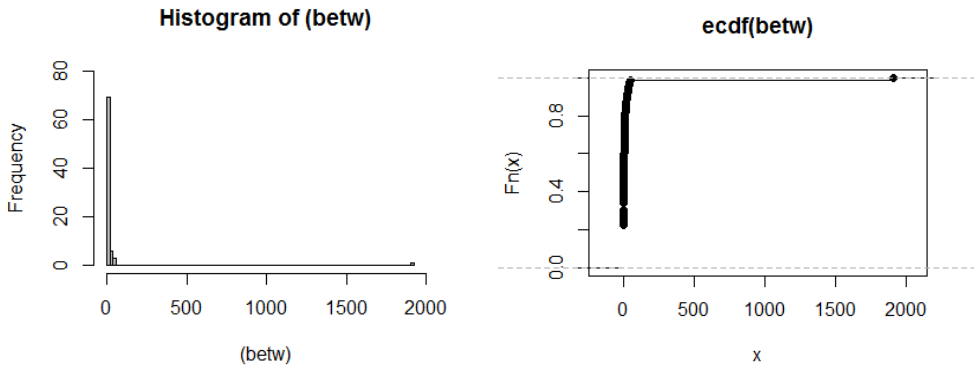

Figura 50: Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus pinus* CBS 10737.

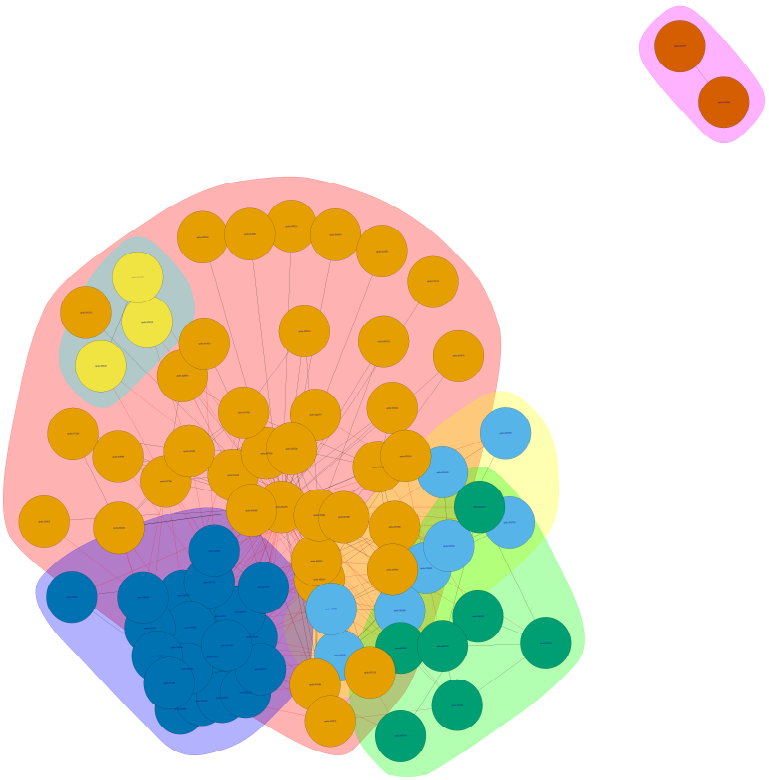

Figura 51; Rede gerada através de “Walktrap Community” que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie *Cryptococcus pinus* CBS 10737

3.1.3 Estatística de interações proteína – proteína por coevolução *Cryptococcus heveanensis* BCC8398

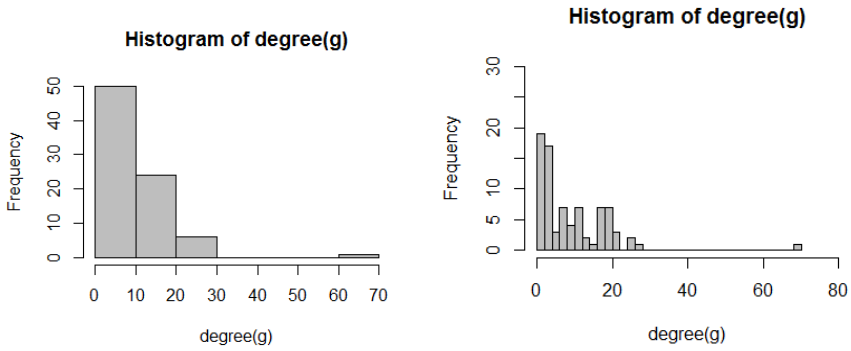

Figura 52: Histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus heveanensis* BCC8398. Eixo x grau do nó, Eixo Y frequência.

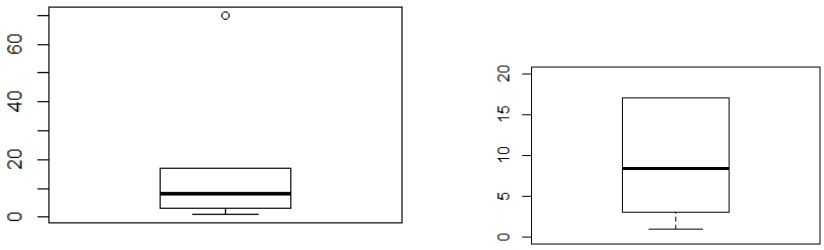

Figura 53: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus heveanensis* BCC8398. Eixo x grau dos nós.

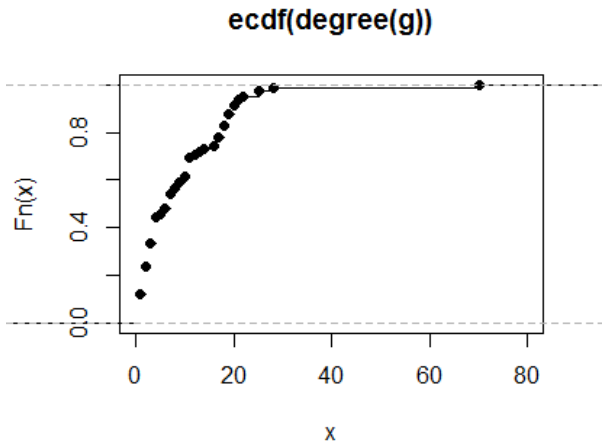

Figura 54: O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus heveanensis* BCC8398.

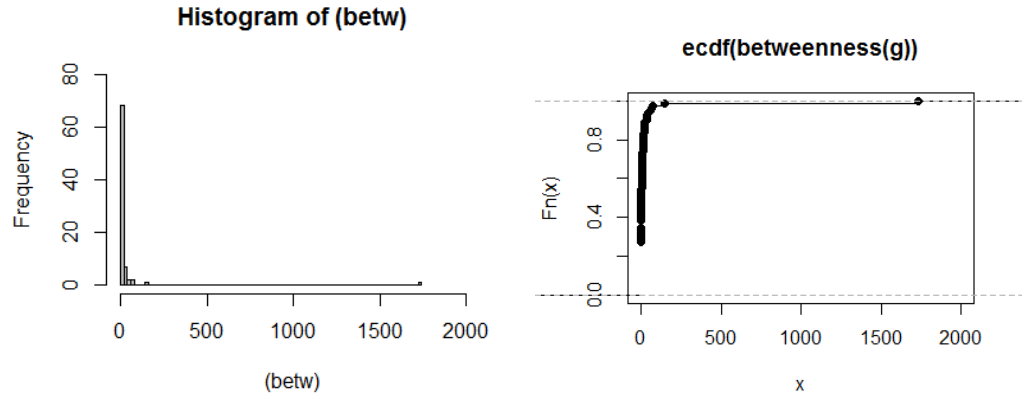

Figura 55: Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus heveanensis* BCC8398.

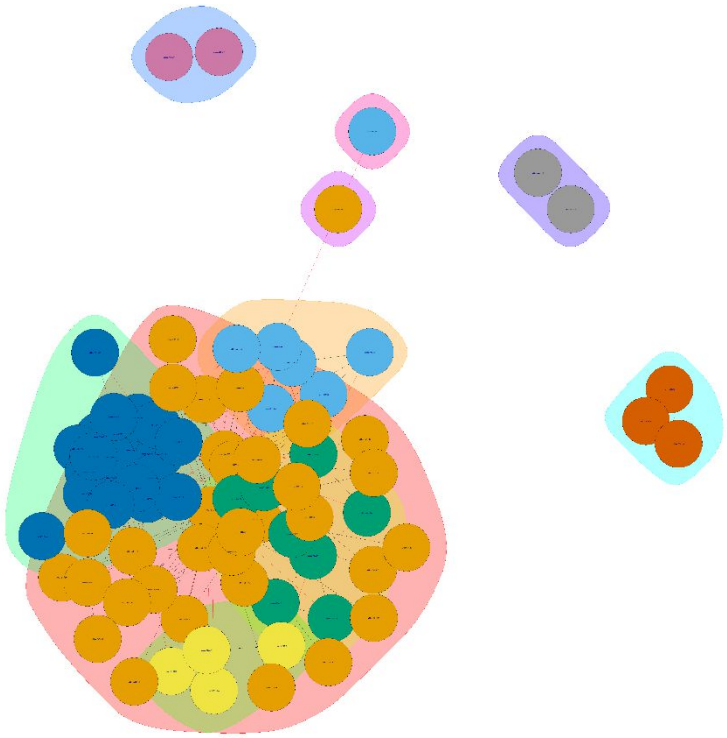

Figura 56; Rede gerada através de "Walktrap Community" que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie *Cryptococcus heveanensis* BCC8398.

Estatística de interações proteína – proteína por coevolução *Cryptococcus heveanensis* CBS569

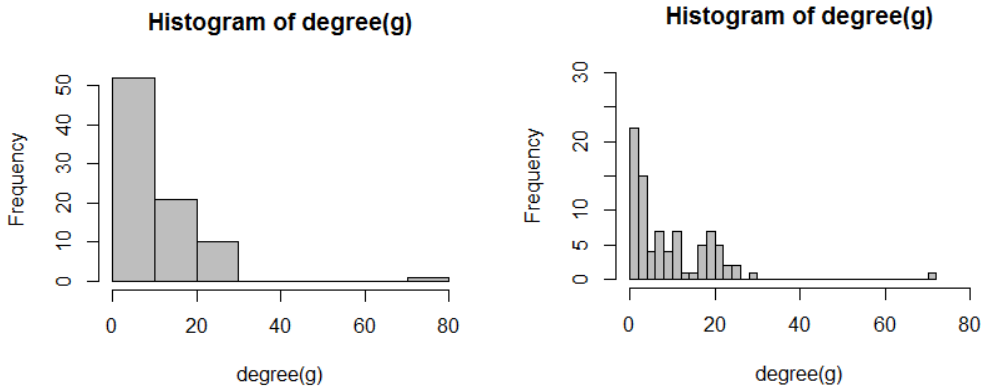

Figura 57: Histograma mostra a distribuição do grau dos nós em relação a frequência na espécie de *Cryptococcus heveanensis* CBS569. Eixo x grau do nó, Eixo Y frequência.

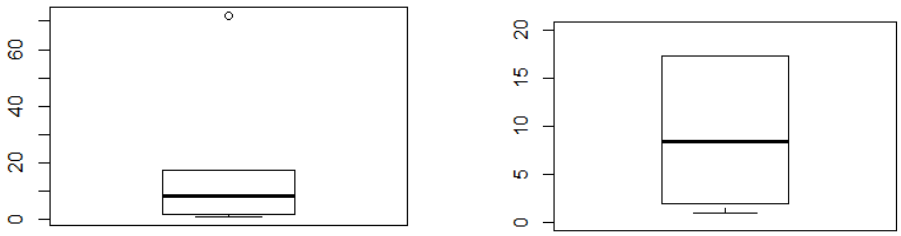

Figura 58: "boxplot" da esquerda mostra um *out line*, isso é um dos nós que não se encaixa no grau, dentro do conjunto de nós da rede; a esquerda vemos a caixa de "boxplot",sem *out line*, para que possamos observar a distribuição da mediana, e a distribuição assimétrica, isso na espécie de *Cryptococcus heveanensis* CBS569. Eixo x grau dos nós.

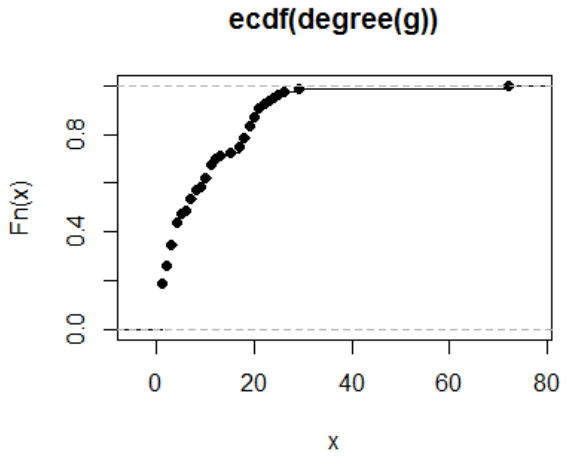

Figura 59: O ecdf é a função de probabilidade cumulativa dos graus, no conjunto de nós da rede *Cryptococcus heveanensis* CBS569

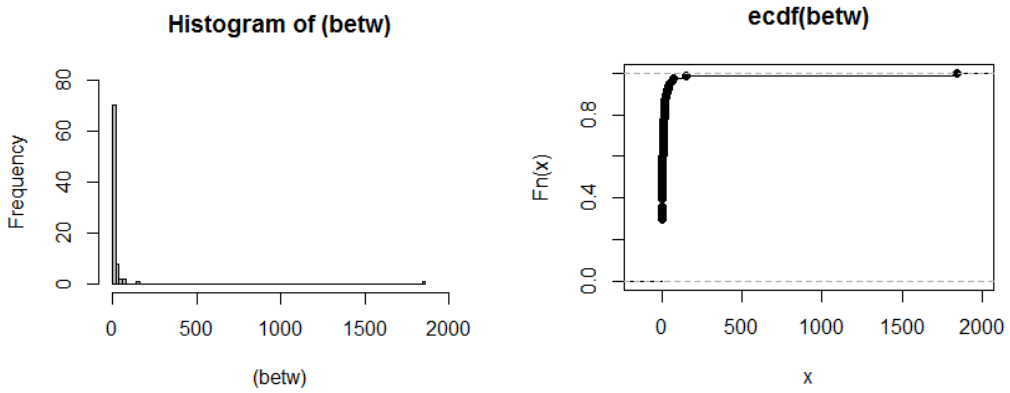

Figura 60: Histograma do lado esquerdo mostra a distribuição de "betweenness" é a probabilidade de um determinado vértice pertencer ao meio do grafo, indicando então que apenas um nó possui alto nível e alguns outros possui um valor muito menor, mas que aparecem. No gráfico da direita podemos perceber essa mesma relação, mas com o ecdf do "betweenness" que é a função acumulativa. Ambos os gráficos são referentes ao genoma de *Cryptococcus heveanensis* CBS569.

.

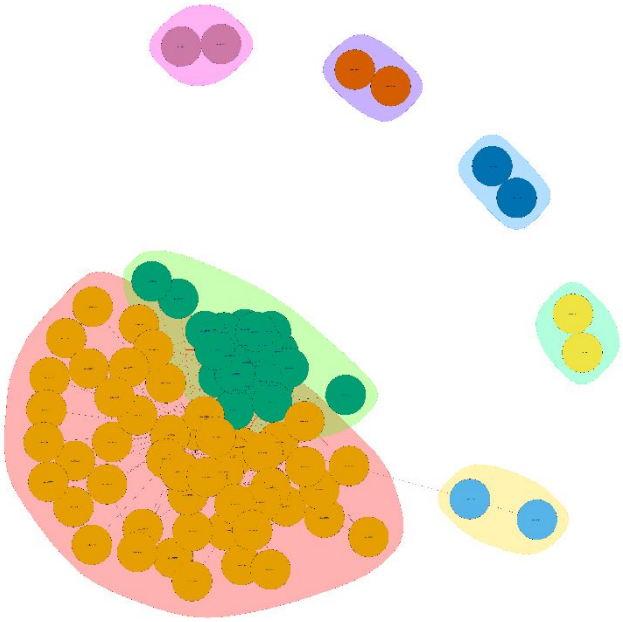

Figura 61; Rede gerada através de “Walktrap Community” que é uma abordagem baseada em caminhadas aleatórias, onde os caminhos são mais propensos a permanecer dentro da mesma comunidade. A rede foi da espécie *Cryptococcus heveanensis* CBS569.

Rede de *Cryptococcus bestiolae*

Nesse primeiro organismo, aproveitamos para discutir a forma de apresentação de uma rede. Uma rede pode ser organizada de diversas formas, porém independente da forma ela deve ser importante para visualizar ou extrair alguma informação referente aos dados que rede apresentada. Um exemplo seria a rede da figura 62, onde ela seria interessante para visualizar o grau dos nós, e verificar os com maiores conexões, mas para compreender de que forma essas interações estão dispostas dentro do sistema, a sua utilização não seria viável.

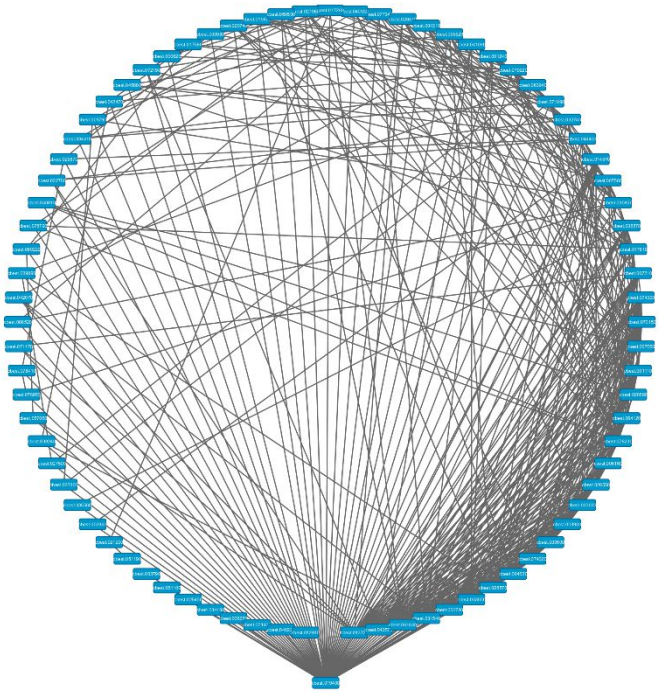

Figura 62: Mostra a rede gerada me forma circular e crescente nos números de ligações da esquerda para direita, possibilitando a visualização que a maior parte dos nós possuem poucas ligações, sendo o nó deslocado que o que possui maior grau.

Outra rede que é interessante é a rede onde organizamos de forma hierárquica. Como podemos ver na figura 63, quando possuímos um grande número de nós, a compreensão da distribuição da rede, assim como a localização de um nó, no caso uma proteína, se torna complexa, não sendo útil então na premissa de que uma rede facilitaria e agilizaria a compressão do sistema. Porém um uma escala reduzida de uma sub rede ou tendo em mente alguns nós específicos, essa disposição poderia se tornar interessante.

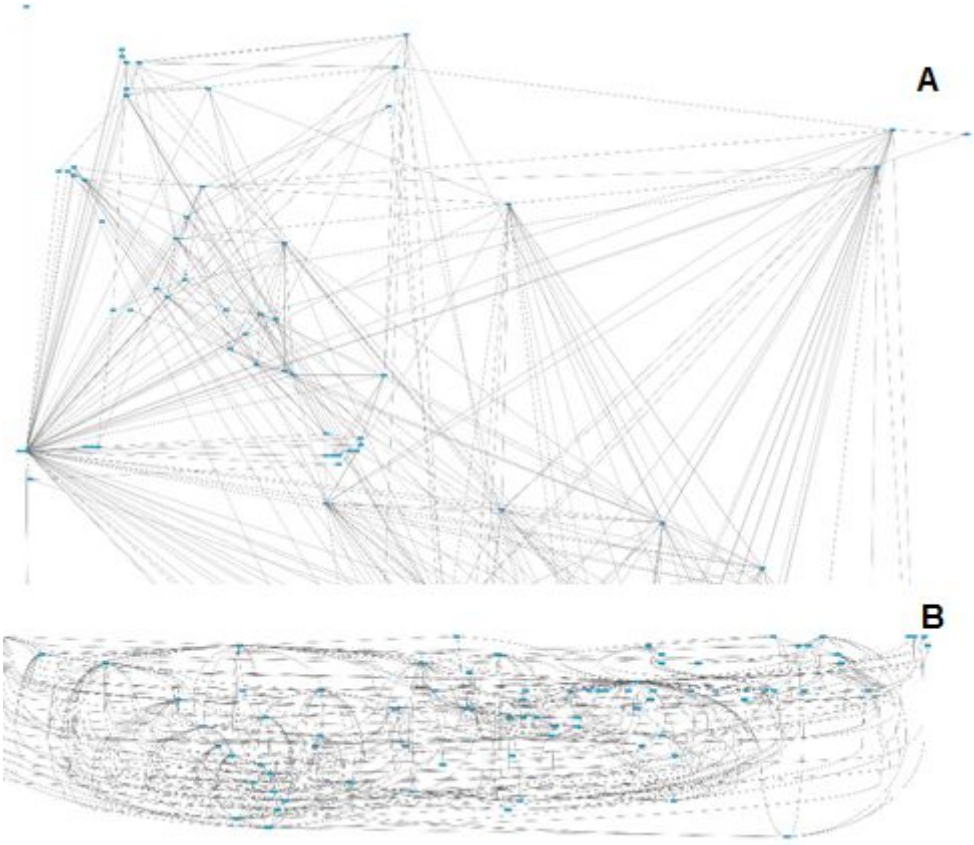

Figura 63: Rede em forma hierárquica, sendo A uma ampliação do no com maior grau e B a rede de forma geral. Isso evidencia que alguns tipos de visualização não são uteis quando se trabalha em uma escala maior. Porém quando se trabalha em menor escala, facilita a compreensão de alguns mecanismos.

Após essa breve discussão sobre a disposição das redes, iremos mostrar como se encontra a rede de *Cryptococcus bestiolae*. As figuras 64 e 65 mostram a disposição da rede, diferentes distribuições apesar de semelhantes são importantes para localizar e observar a rede ou algum nó por isso se encontra de duas formas diferentes. Nessa rede visualizamos que existe 5 conjuntos de redes, sendo 4 pequenas, e uma rede global que emprega a maior parte dos nós. Isso é importante para a construção do conhecimento referente ao sistema do organismo.

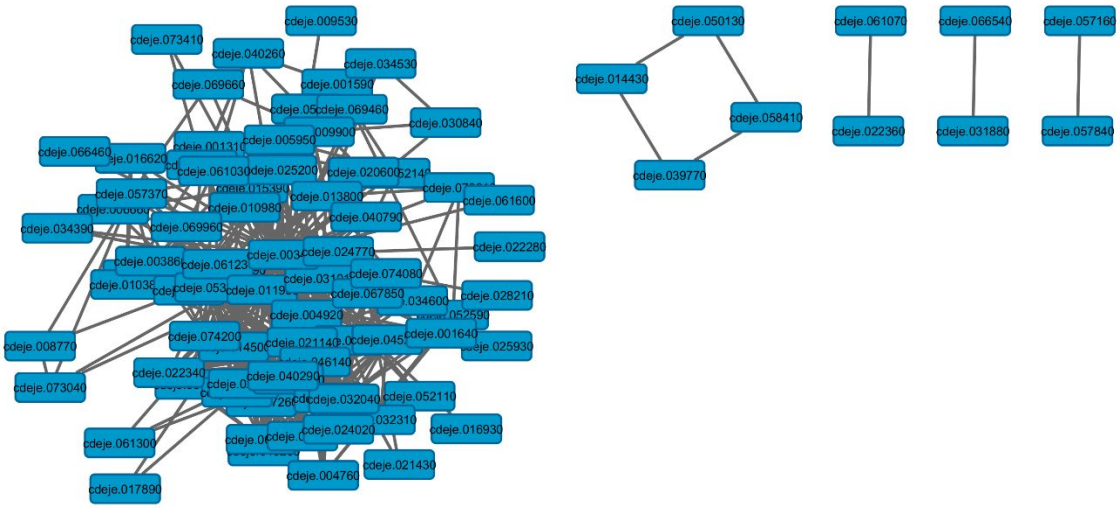

Figura 64 Mostra a Rede completa, incluído os nós que foram conectados fora da rede maior

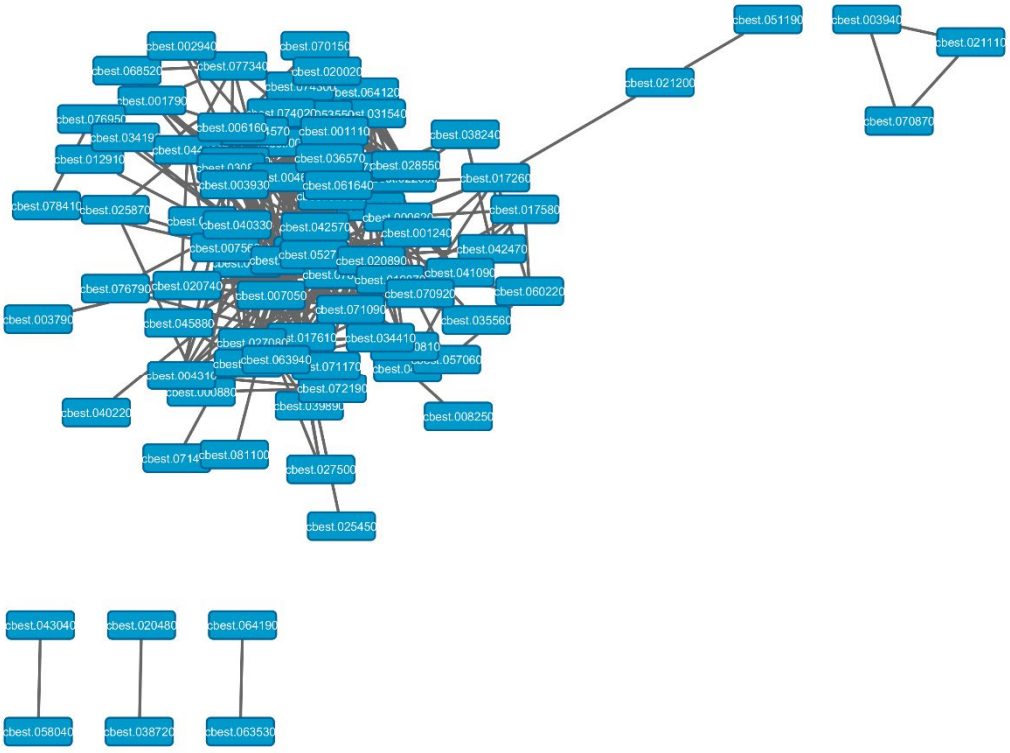

Figura 65: A rede de interações proteína-proteína visualizada com outra forma de organização dos seus nós

O nó que possui maior grau ou conexão é o nó cbest.019400, a sua função no sistema segundo o BLAST foi ATP-dependent protein binding protein [Cryptococcus gattii WM276], o organismo dentro do gênero de *Cryptococcus* e sendo importante dentro de uma rede por ser uma proteína de ligação ATP dependente. Podemos ver através da figura X a sua alta representação, onde ele possui um grau de 76, isso é ele possui

76 arestas, proteínas que interagem experimentalmente com o nó. O outro resultado do blast mostra como hypothetical protein V492_04946 em Pseudogymnoascus pannorum VKM F-4246, que seu identificado é gi|682296803|gb|KFY10578.1|.

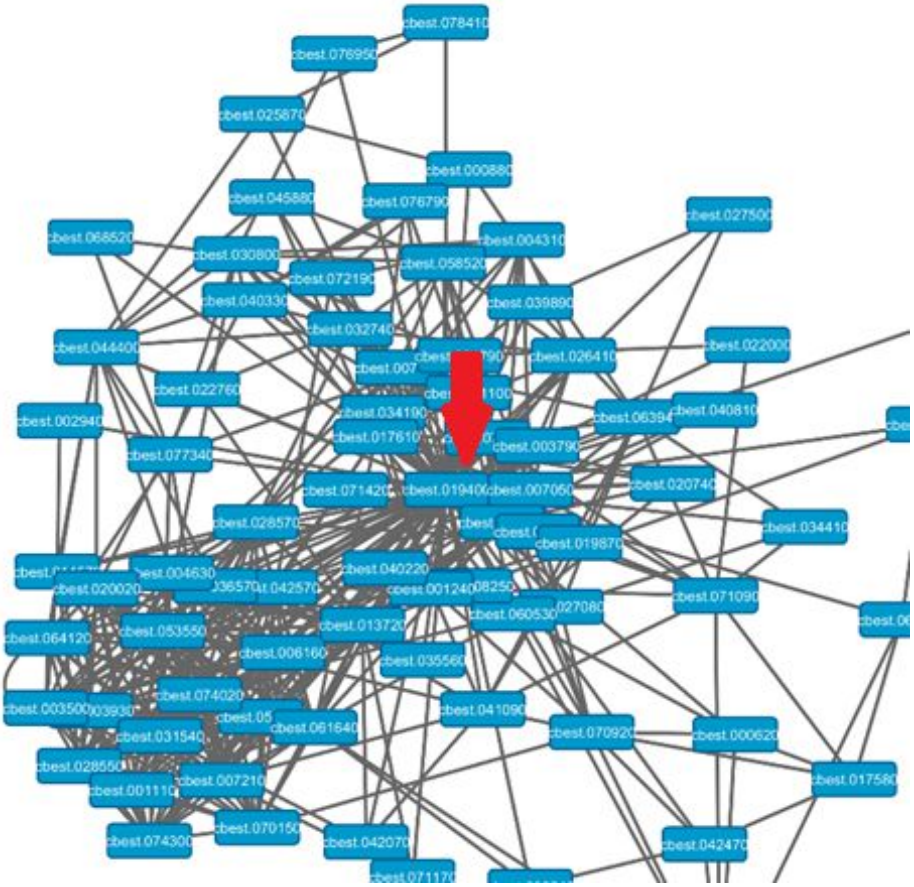

Figura 66: Uma ampliação da rede para observar a nomenclatura e evidenciar os o nó com maior ligação dentro dessa rede, indicado pela seta vermelha.

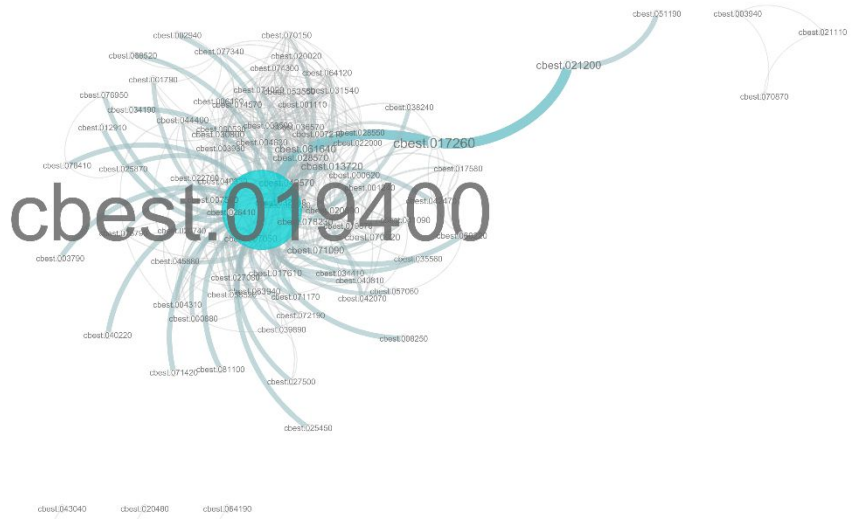

Figura 67 : Visualização priorizando os nós com maior grau, nesse caso evidenciando o gene cbest.019400

Rede de *Cryptococcus dejecticola* CBS 10117

Na rede de de *Cryptococcus dejecticola* observamos também que existe 5 grupos de redes sendo a rede central com o maior número de nós. A rede de forma geral apresenta 87 nós com 433 arestas.

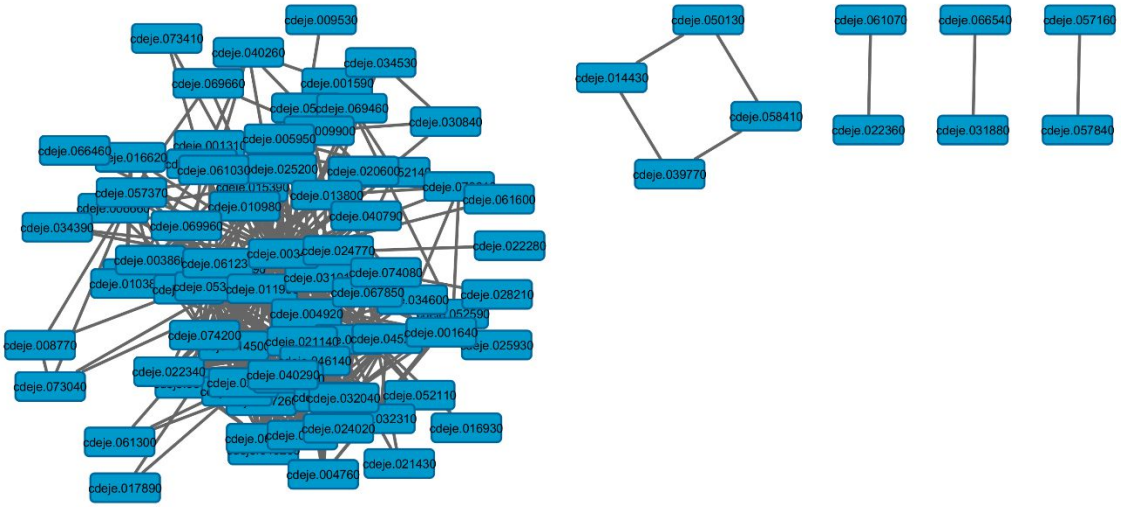

Figura 68: Visualização da rede de interações proteína-proteína geral

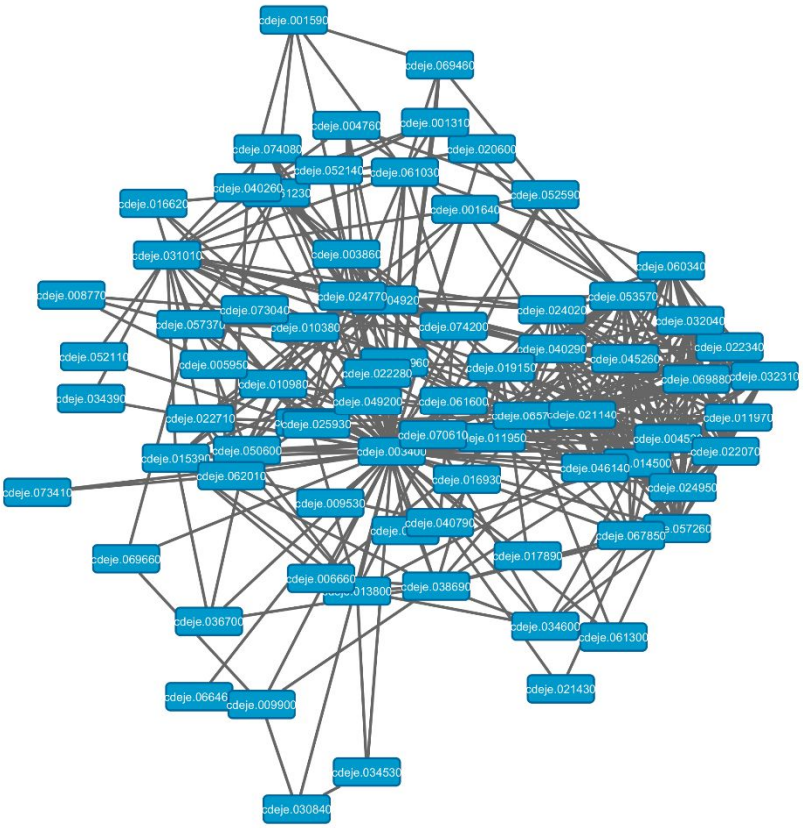

Figura 69: Visualização da rede principal

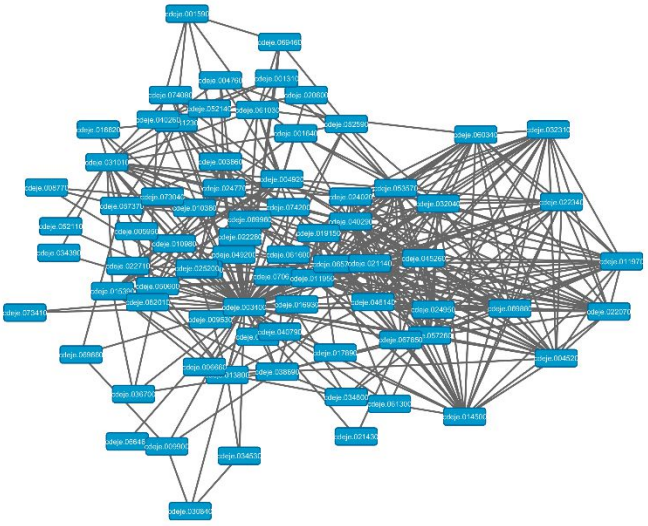

Figura 70: Visualização da rede principal com outra forma de organização

Através dessa visualização podemos perceber um comportamento onde os nós altamente conectados se ligam entre si. Isso é algo a ser avaliado posteriormente para compreender a finalidade desse comportamento.

O nó que observamos com maior número de arestas foi o cdeject.003400, esse nó possui 75 conexões, a função descrita no subset do NR sem *Cryptococcus* é | hypothetical protein V492_04946 [Pseudogymnoascus pannorum VKM F-4246]

(gi|682296803|gb|KFY10578.1), e no subset somento com *Cryptococcus* foi ATP-dependent protein binding protein, putative [Cryptococcusgattii WM276] ubiquitin protein 1 [Cryptococcus gattii R265] (gi|321263887|ref|XP_003196661.1). Sendo uma função similar a encontra no organismo anterior.

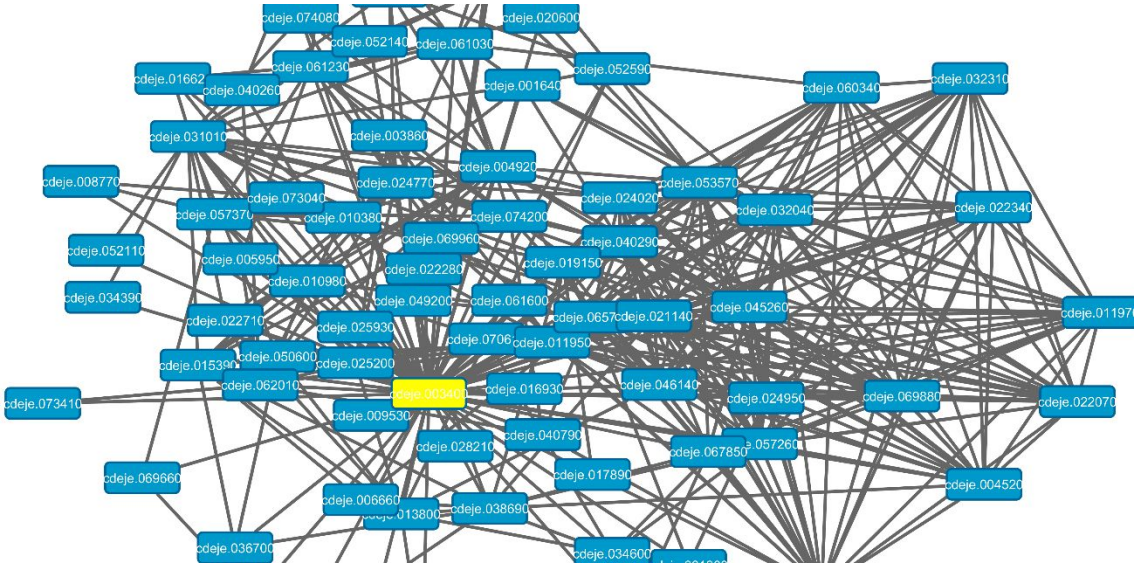

Figura 71: Visualização de parte da rede, evidenciando o nó com maior grau.

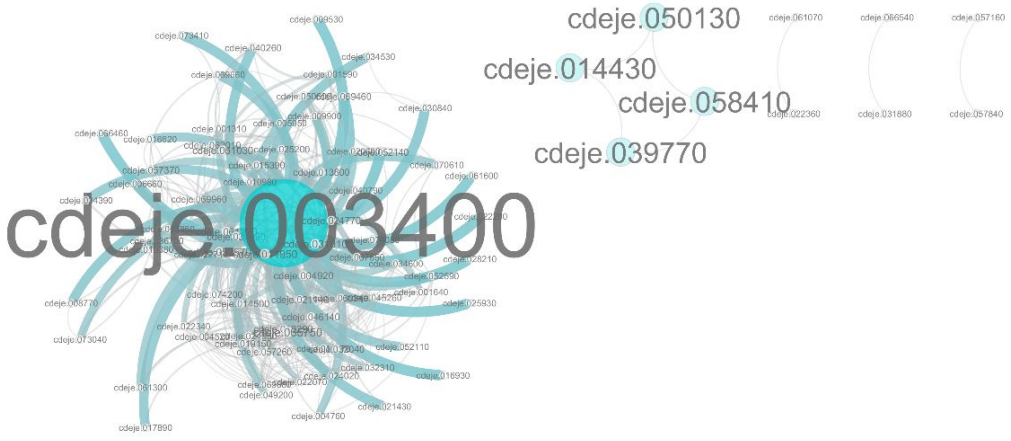

Figura 72: Visualização da rede evidenciando os nós com maior grau dentro dos sub grupos.

Rede de *Cryptococcus flavescens* NRRL Y-50378

Na rede que construímos para *Cryptococcus flavescens* observamos dois conjuntos de redes, na verdade um dos conjuntos é somente um par de interação. O outro abrange todos os outros nós e arestas. No conjunto total são 63 nós e 353 interações.

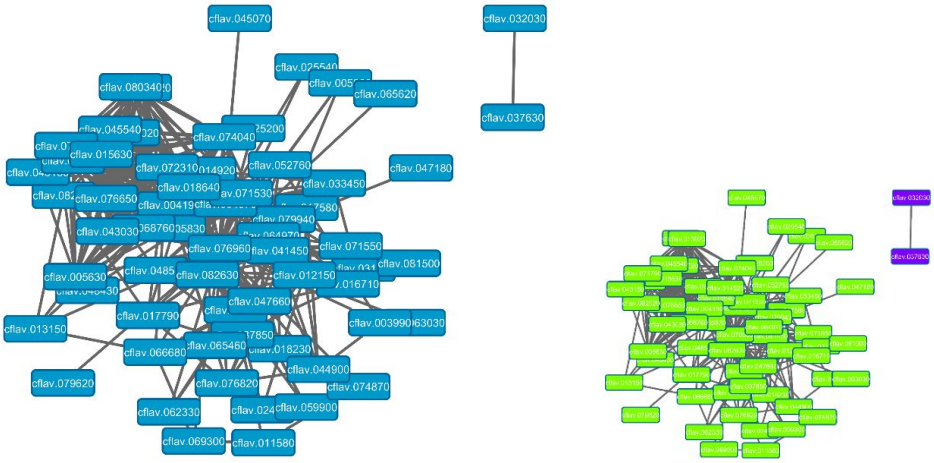

Figura 74: Visualização da rede de interações proteína-proteína geral, onde segunda mostra os dois grupos.

Observamos nessa rede o nó com maior grau foi o cflav.076960, cuja a função descrita no subset do NR sem *Cryptococcus foi* ubiquitin *[Metarhizium acridum CQMa 102]* ubiquitin *[Metarhizium acridum CQMa 102] (gi|629687373|ref|XP_007808982.1|).* E o subset do NR com somente *Cryptococcus foi* ATP-dependent protein binding protein [Cryptococcus neoformans var. neoformans JEC21] hypothetical protein CNBK0920 [Cryptococcus neoformans var.neoformans B-3501A] hypothetical protein CNBK0920 [Cryptococcus neoformans var. neoformans B-3501A] ATP-dependent protein binding protein, putative [Cryptococcus neoformans var. neoformans JEC21] polyubiquitin [Drechslerella stenobrocha 248] (*gi|58260904|ref|XP_567862.1|*).Dessa forma percebemos o papel vital desse nó na relação da interação da rede.

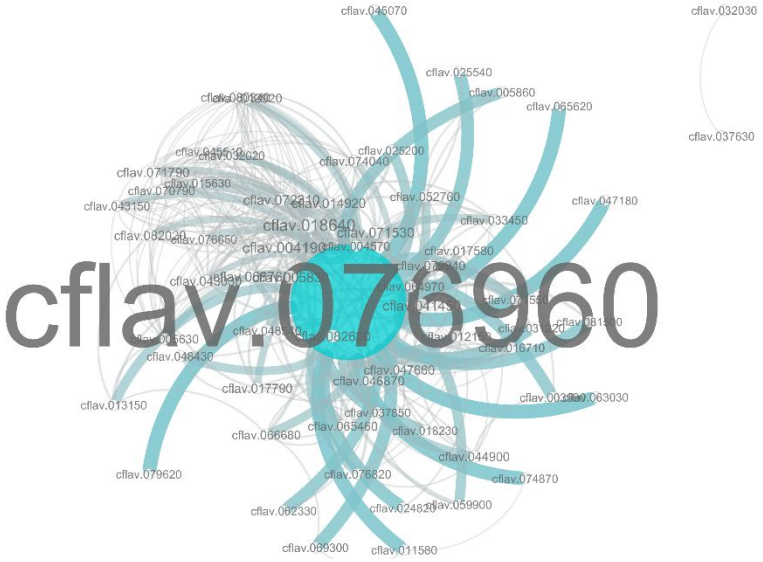

Figura 75: Visualização da rede evidenciando os nós com maior grau dentro dos subgrupos.

Rede de *Cryptococcus neoformans* var. *neoformans* JEC21

Na rede que construímos para *Cryptococcus neoformans* var. *neoformans* JEC21, observamos uma maior quantidade de conjuntos de redes ou pares de interações proteína- proteína São 12 conjuntos de interações, porém como nós anteriores somente um possui um grande número de nós. O conjunto gerado possui 115 nós e 482 interações descritas.

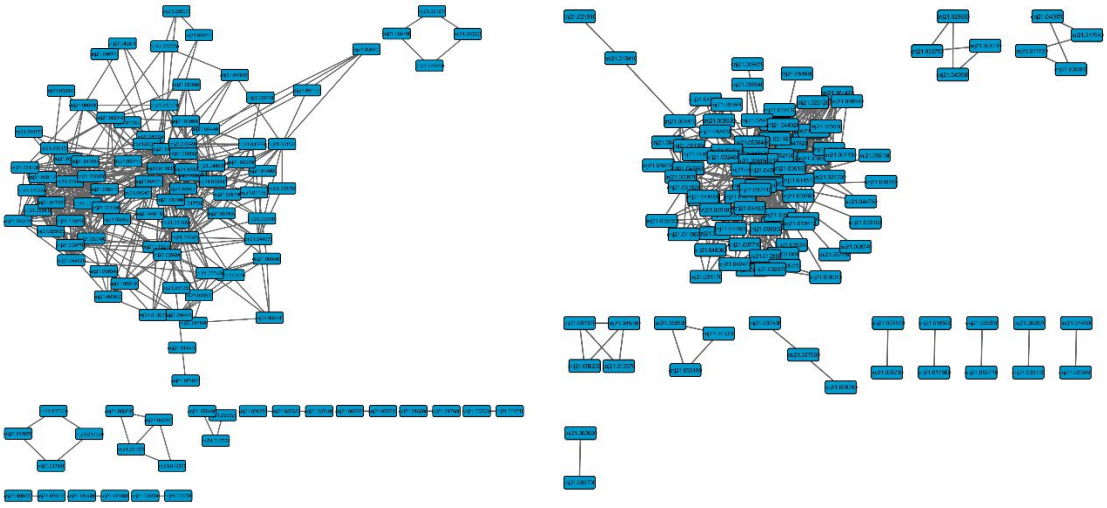

Figura 76: Visualização da rede de interações proteína-proteína de forma geral, onde mostra duas formas de organização para visualziação da rede.

Observamos na rede o nó com um maior grau foi o cnj21.013930, com 78 arestas, cuja a função descrita no subset do NR sem *Cryptococcus* foi Polyubiquitin-A [Triticum urartu] (gi|474093556|gb|EMS54972.1|). No subset somente com *Cryptococcus* ATP-dependent protein binding protein [Cryptococcus neoformans var. neoformans JEC21], hypothetical protein CNBK0920 [Cryptococcus neoformans var.neoformans B-3501A], hypothetical protein CNBK0920 [Cryptococcus neoformans var.neoformans B-3501A], ATP-dependent protein binding protein, putative [Cryptococcus neoformans var. neoformans JEC21], polyubiquitin [Drechslerella stenobrocha 248](gi|58260904|ref|XP_567862.1).

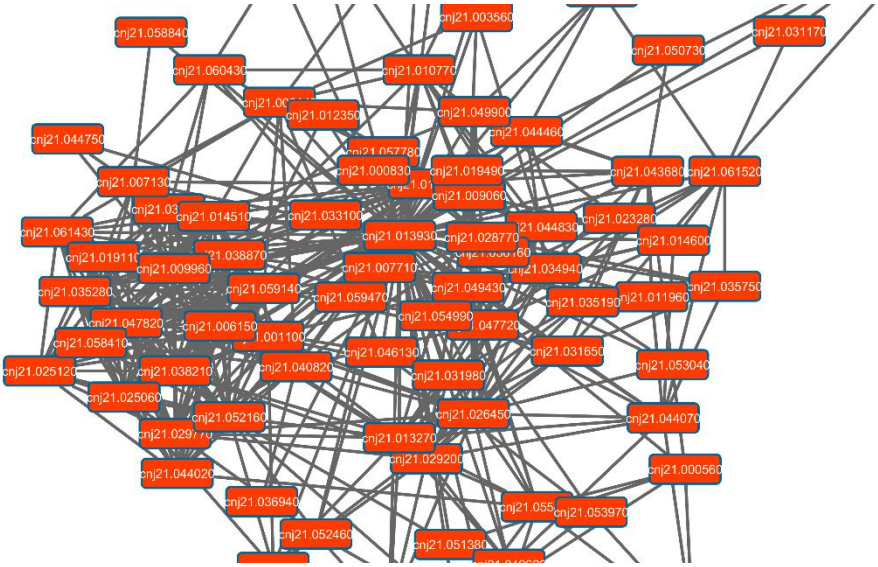

Figura 77: Visualização de parte da rede, evidenciando o nó com maior grau.

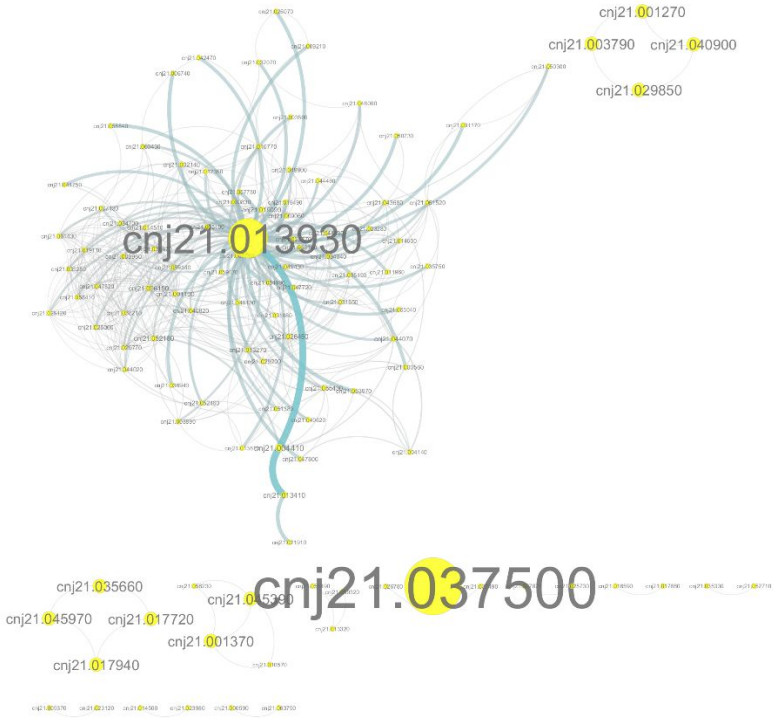

Figura 78: Visualização da rede evidenciando os nós com maior grau dentro dos subgrupos.

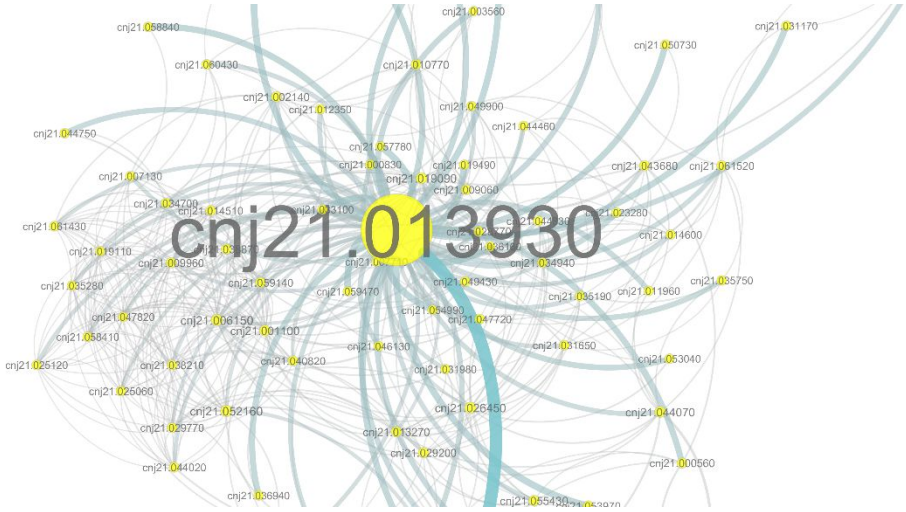

Figura 79: Visualização da rede evidenciando os nós com maior grau dentro da principal rede.

Rede de *Cryptococcus neoformans* var. *grubii* H99

Na rede que construímos para *Cryptococcus neoformans* var. *grubii* H99, observamos uma maior quantidade de conjuntos de redes ou pares de interações proteína-

proteína São 11 conjuntos de interações, porém como nós anteriores somente um possui um grande número de nós. O conjunto gerado possui 111 nós e 454 interações descritas.

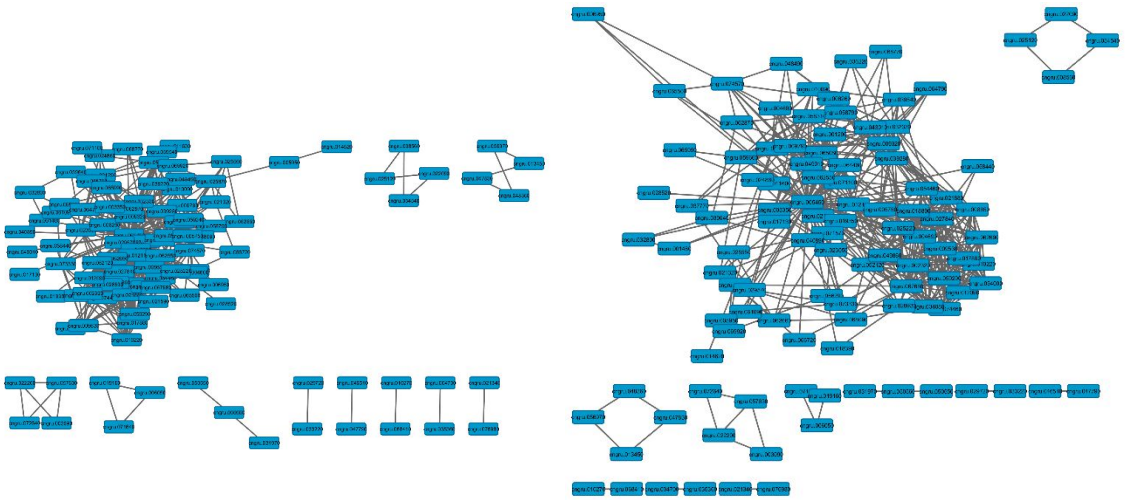

Figura 80: Visualização da rede de interações proteína-proteína de forma geral, onde mostra duas formas de organização para visualziação da rede.

Observamos na rede o nó com um maior grau foi o cngru.005450, com 77 arestas, cuja a função descrita no subset do NR sem *Cryptococcus* foi hypothetical protein JAAARDRAFT_29850 [Jaapia argillacea MUCL 33604] (gi|646399668|gb|KDQ63803.1|), no subset do NR somente com *Cryptococcus* foi ATP-dependent protein binding protein [Cryptococcus neoformans var. neoformans JEC21], hypothetical protein CNBK0920 [Cryptococcus neoformans var.neoformans B-3501A], hypothetical protein CNBK0920 [Cryptococcus neoformans var. neoformans B-3501A], ATP-dependent protein binding protein, putative [Cryptococcus neoformans var. neoformans JEC21] polyubiquitin [Drechslerella stenobrocha 248] (gi|58260904|ref|XP_567862.1|)

Figura 81: Visualização de parte da rede, evidenciando o nó com maior grau.

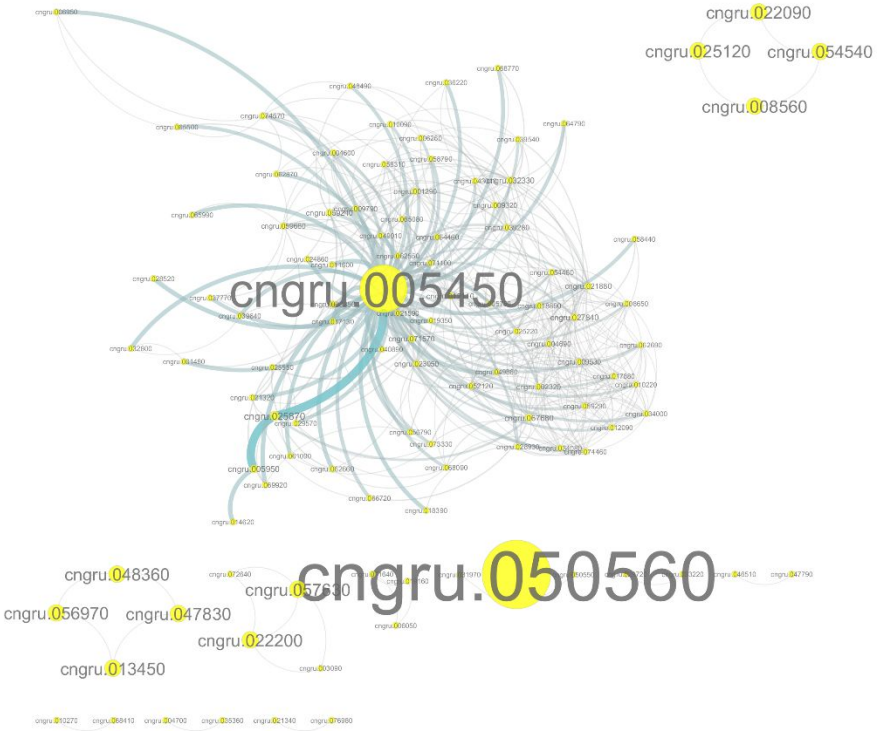

Figura 82: Visualização da rede evidenciando os nós com maior grau dentro dos subgrupos.

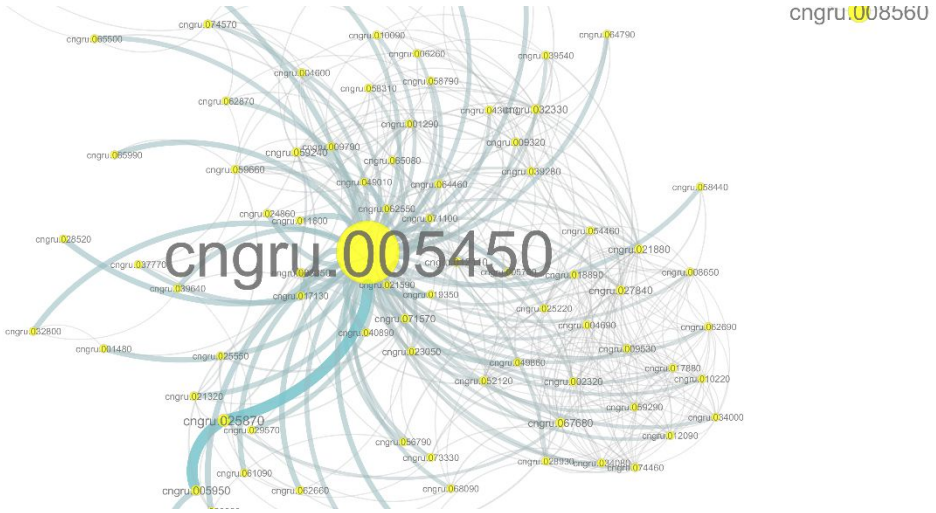

Figura 83: Visualização da rede evidenciando os nós com maior grau dentro da principal rede.

Rede de *Cryptococcus neoformans* var. *neoformans* B-3501

Na rede que construímos para *Cryptococcus neoformans* var. *neoformans* B-3501, observamos uma maior quantidade de conjuntos de redes ou pares de interações proteína- proteína São 12 conjuntos de interações, porém como nós anteriores somente um possui um grande número de nós. O conjunto gerado possui 115 nós e 482 interações descritas.

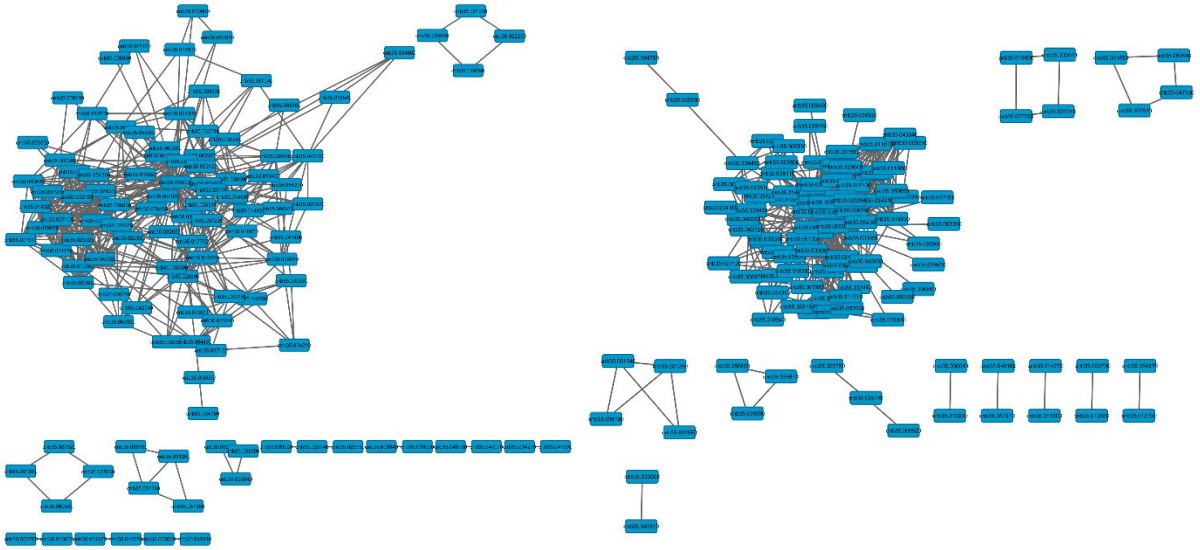

Figura84: Visualização da rede de interações proteína-proteína de forma geral, onde mostra duas formas de organização para visualziação da rede.

Observamos na rede o nó com um maior grau foi o cnb35.055030, com 79 arestas, cuja a função descrita no subset do NR sem *Cryptococcus* foi Polyubiquitin-A [Triticum urartu](gi|474093556|gb|EMS54972.1|), no subset do NR somente com *Cryptococcus* foi | ATP-dependent protein binding protein [Cryptococcus neoformans var. neoformans JEC21], hypothetical protein CNBK0920 [Cryptococcus neoformans var. neoformans B-3501A], hypothetical protein CNBK0920 [Cryptococcus neoformans var. neoformans B-3501A], ATP-dependent protein binding protein, putative [Cryptococcus neoformans var. neoformans JEC21], polyubiquitin [Drechslerella stenobrocha 248] (gi|58260904|ref|XP_567862.1).

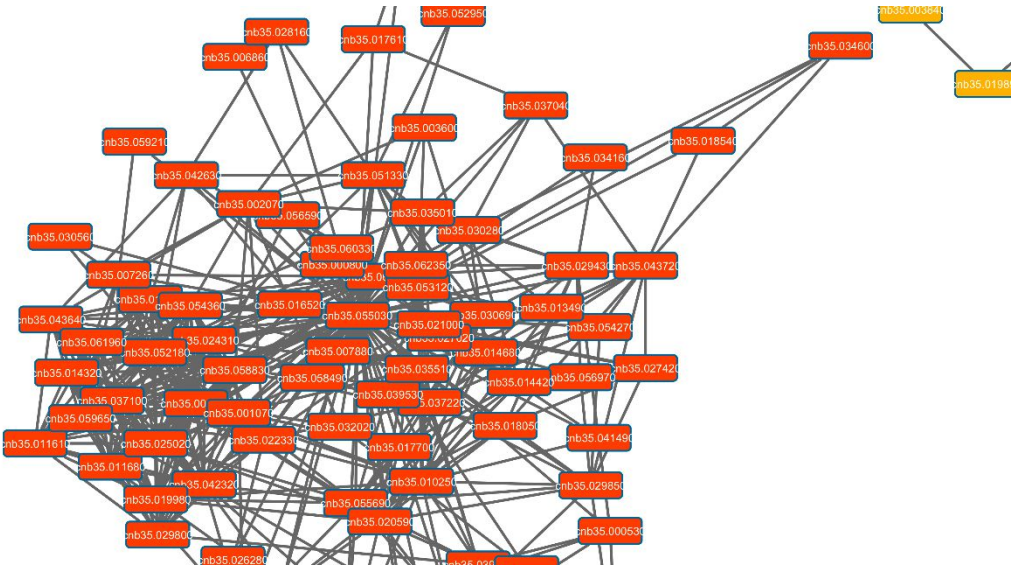

Figura 85: Visualização de parte da rede, evidenciando o nó com maior grau.

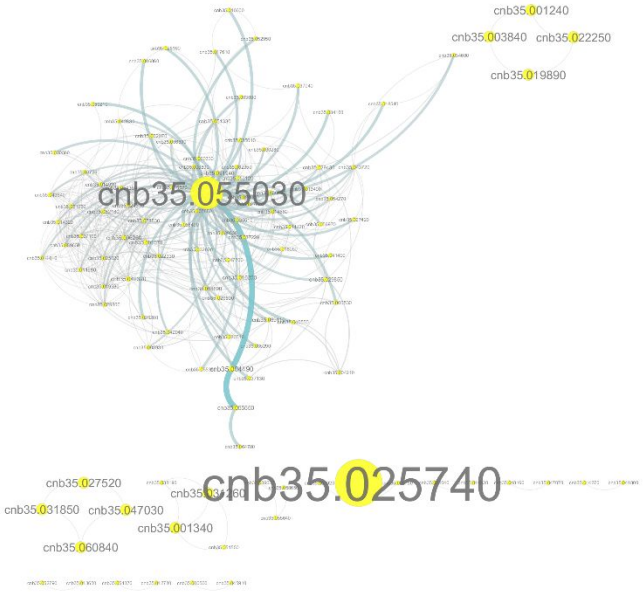

Figura 86: Visualização da rede evidenciando os nós com maior grau dentro dos subgrupos.

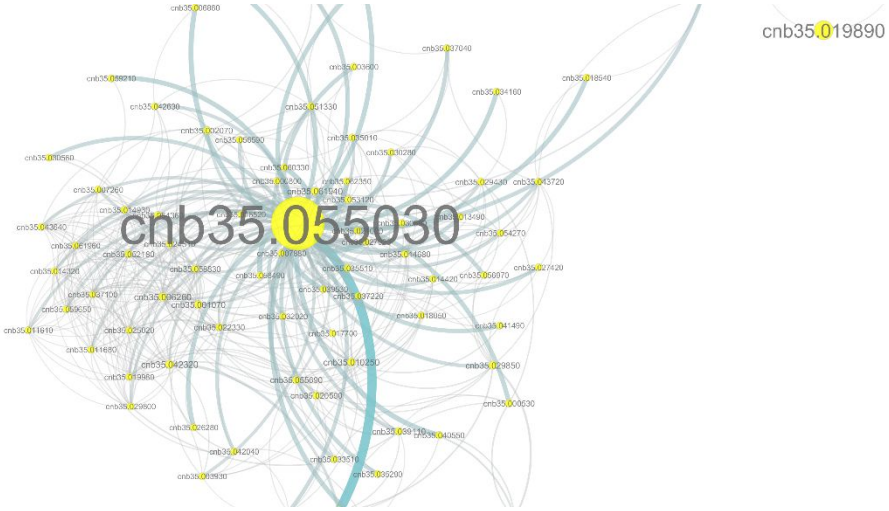

Figura 87: Visualização da rede evidenciando os nós com maior grau dentro da principal rede.

Rede de *Cryptococcus gattii* WM276

Na rede que construímos para *Cryptococcus gattii* WM276, observamos uma maior quantidade de conjuntos de redes ou pares de interações proteína-proteína São 14 conjuntos de interações, porém como nós anteriores somente um possui um grande número de nós. O conjunto gerado possui 118 nós e 476 interações descritas.

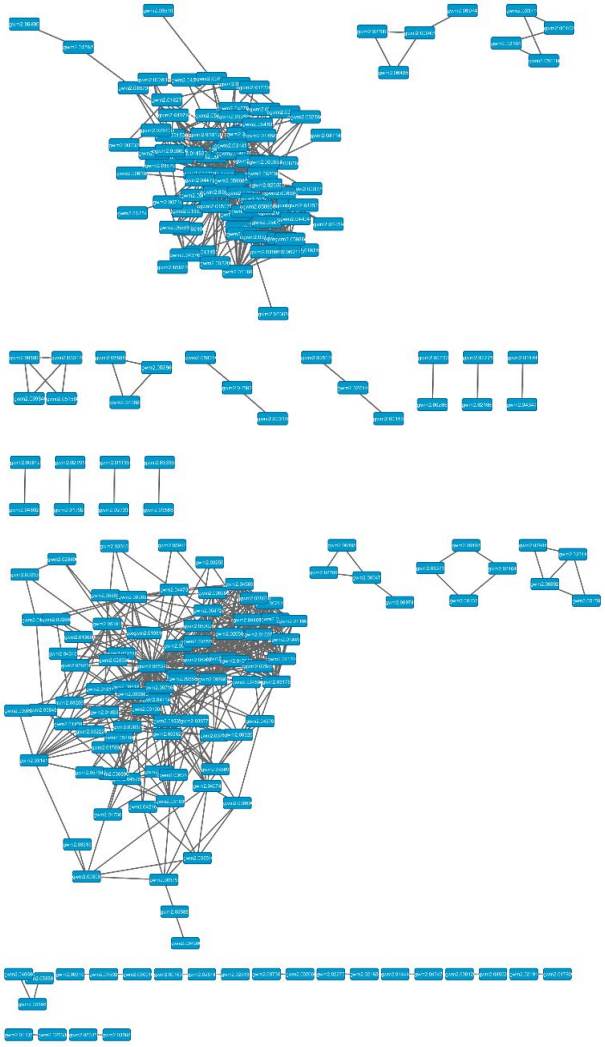

Figura 88: Visualização da rede de interações proteína-proteína de forma geral, onde mostra duas formas de organização para visualziação da rede.

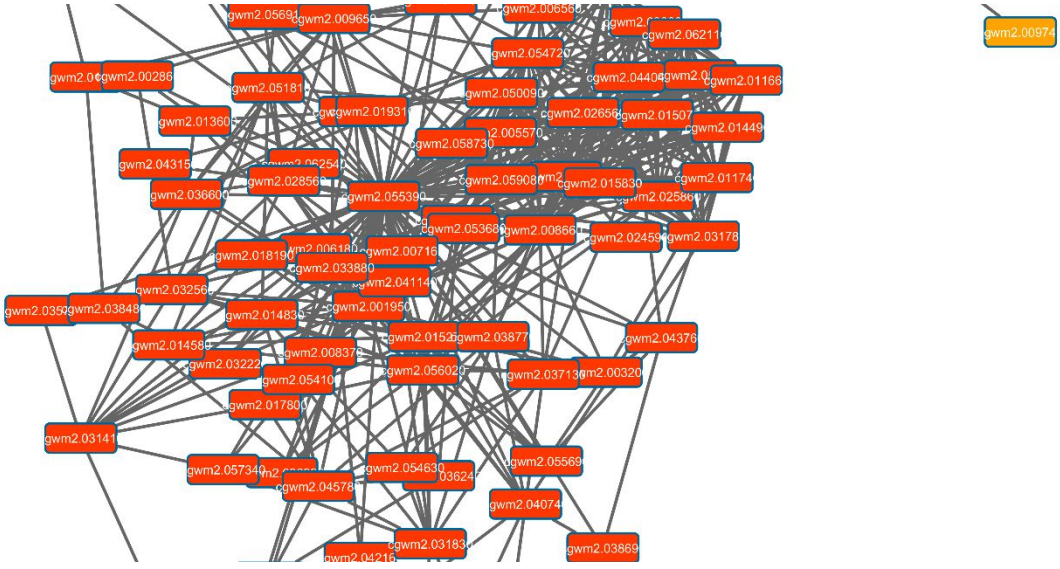

Figura 89: Visualização de parte da rede, evidenciando o nó com maior grau.

Observamos na rede o nó com um maior grau foi o cgwm2.055390, com 75 arestas, cuja a função descrita no subset do NR sem *Cryptococcus* foi ubiquitin [Dacryopinax sp. DJM-731 SS1](gi|402220063|gb|EJU00136.1), no subset do NR somente com *Cryptococcus* foi | ATP-dependent protein binding protein [Cryptococcus gattii WM276] ATP-dependent protein binding protein, putative [Cryptococcus gattii WM276] ubiquitin protein 1 [Cryptococcus gattii R265](gi|321263887|ref|XP_003196661.1).

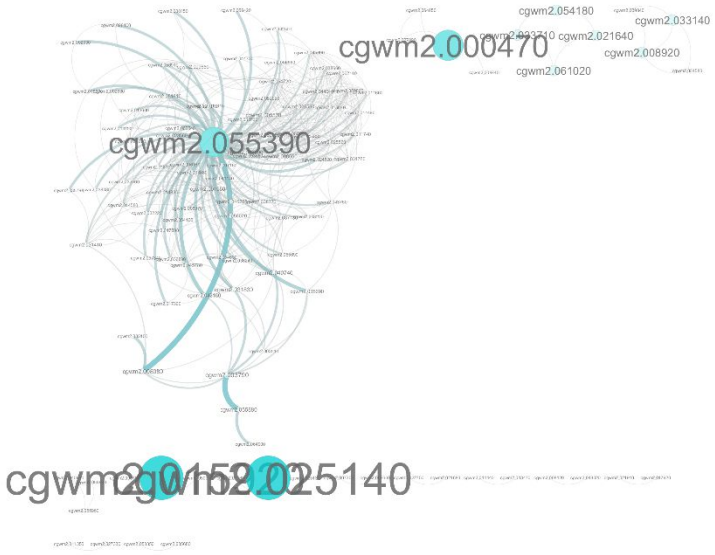

Figura 90: Visualização da rede evidenciando os nós com maior grau dentro dos subgrupos.

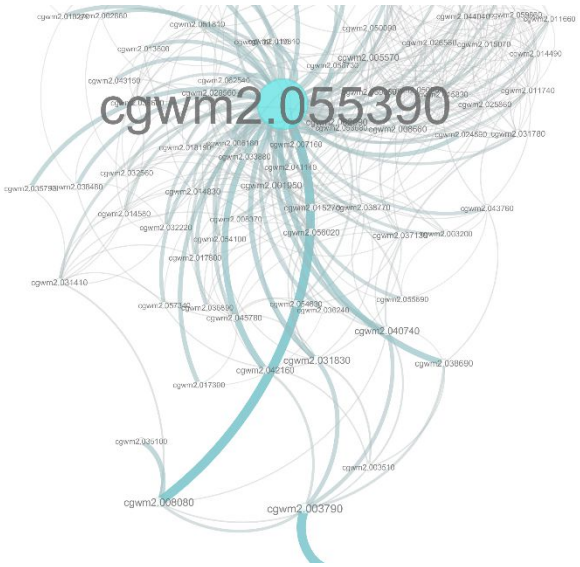

Figura 91: Visualização da rede evidenciando os nós com maior grau dentro da principal rede.

Rede de *Cryptococcus gattii* CBS 7750

Na rede que construímos para *Cryptococcus gattii* CBS 7750, observamos uma maior quantidade de conjuntos de redes ou pares de interações proteína- proteína São 13 conjuntos de interações, porém como nós anteriores somente um possui um grande número de nós. O conjunto gerado possui 113 nós e 473 interações descritas.

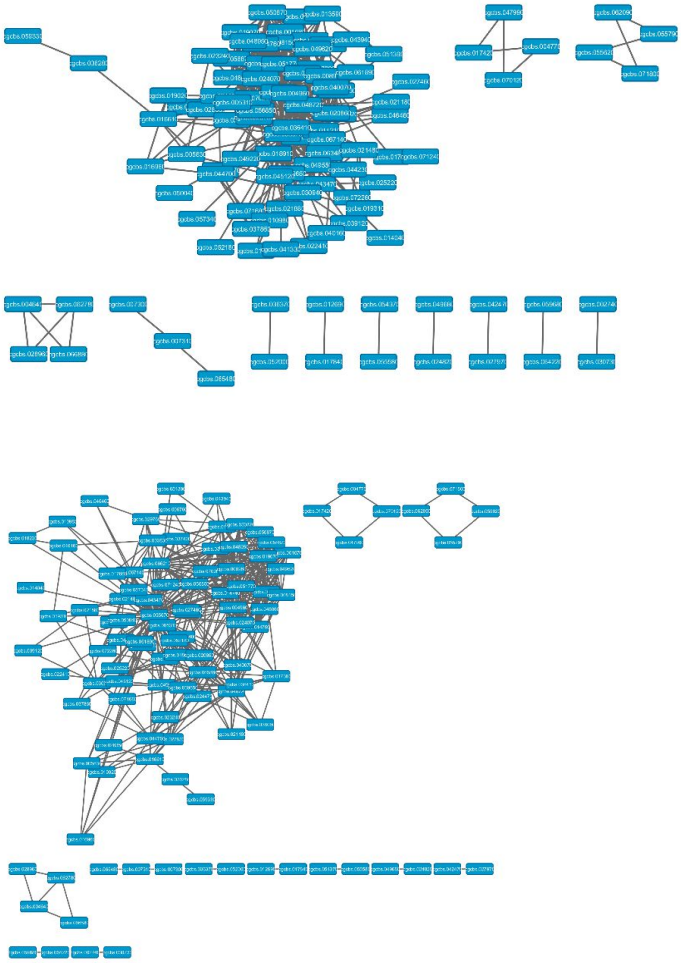

Figura 92: Visualização da rede de interações proteína-proteína de forma geral, onde mostra duas formas de organização para visualziação da rede.

Observamos na rede o nó com um maior grau foi o cgcbs.035670, com 78 arestas, cuja a função descrita no subset do NR sem *Cryptococcus* foi ubiquitin [Dacryopinax sp. DJM-731 SS1](gi|402220063|gb|EJU00136.1|). No subset do NR somente com *Cryptococcus* foi ATP-dependent protein binding protein [Cryptococcus gattii WM276] ATP-

dependent protein binding protein, putative [Cryptococcus gattii WM276] ubiquitin protein 1 [Cryptococcus gattii R265] (gi|321263887|ref|XP_003196661.1|).

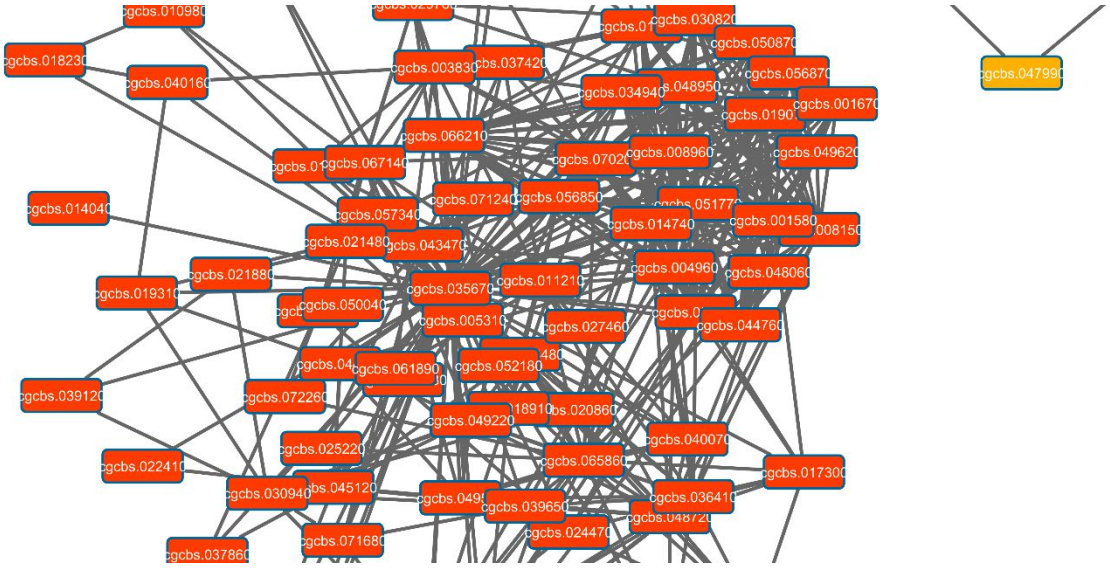

Figura 93: Visualização de parte da rede, evidenciando o nó com maior grau.

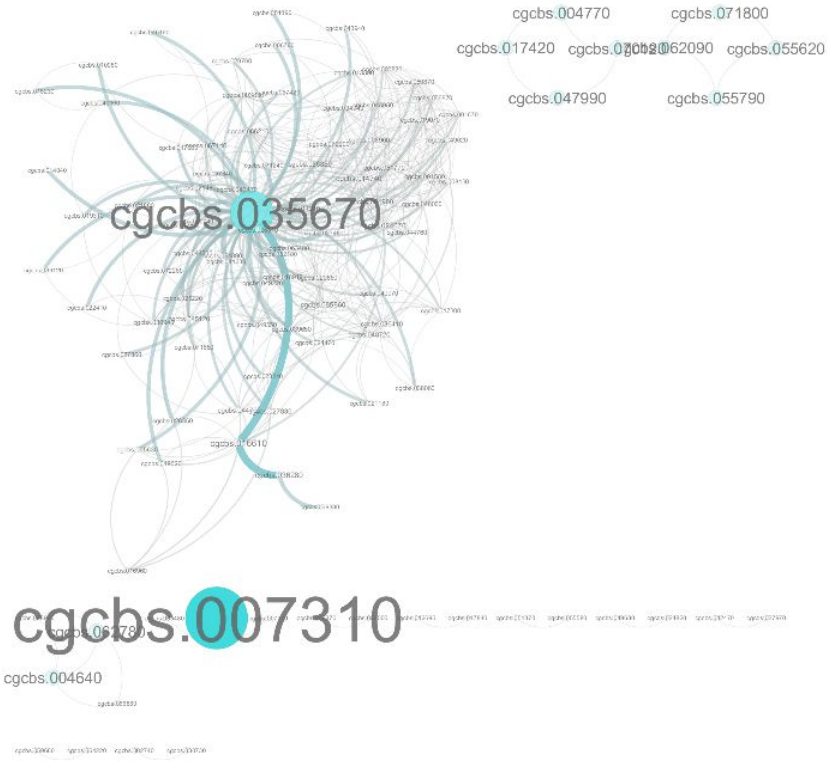

Figura 94: Visualização da rede evidenciando os nós com maior grau dentro dos subgrupos.

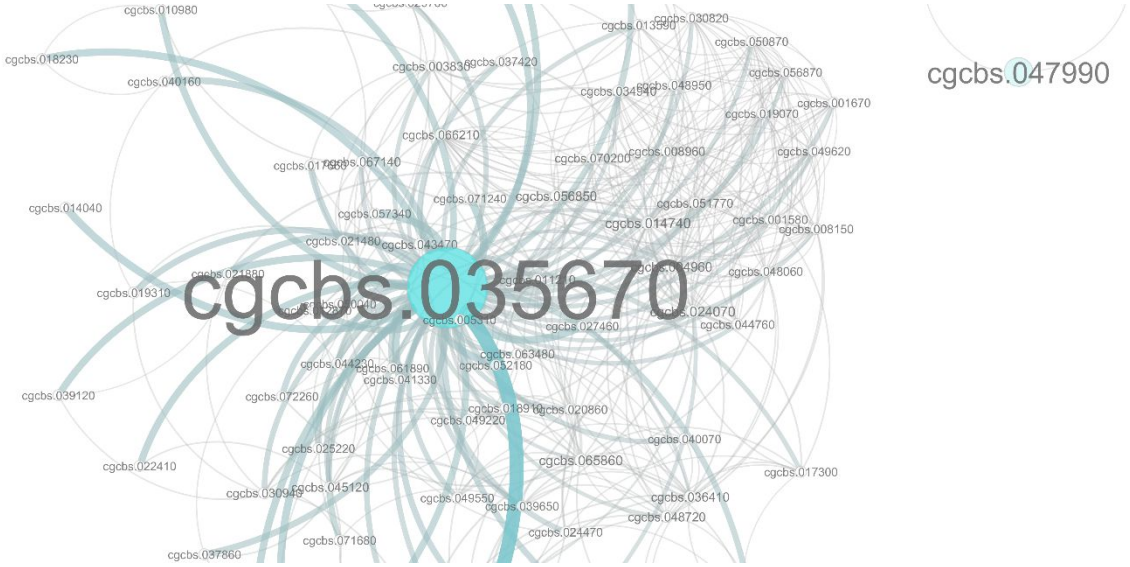

Figura 95: Visualização da rede evidenciando os nós com maior grau dentro da principal rede.

Rede de *Cryptococcus gattii* R265

Na rede que construímos para *Cryptococcus gattii* R265observamos uma maior quantidade de conjuntos de redes ou pares de interações proteína- proteína São 12 conjuntos de interações, porém como nós anteriores somente um possui um grande número de nós. O conjunto gerado possui 114 nós e 480 interações descritas.

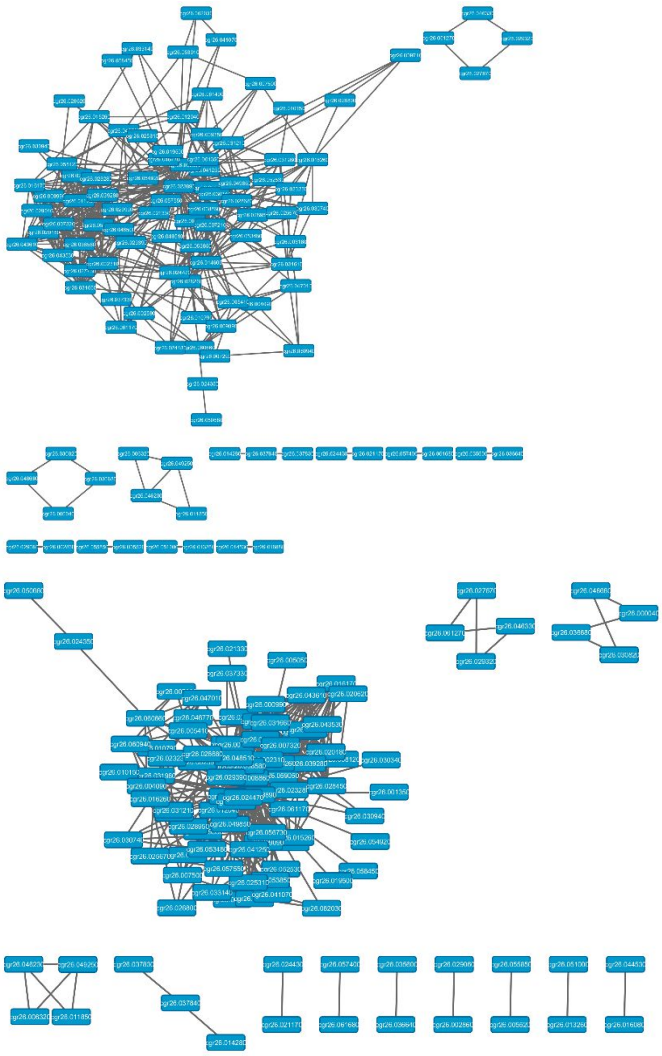

Figura 96: Visualização da rede de interações proteína-proteína de forma geral, onde mostra duas formas de organização para visualziação da rede.

Observamos na rede o nó com um maior grau foi o cgr26.023890, com 79 arestas, cuja a função descrita no subset do NR sem *Cryptococcus* foi ubiquitin [Dacryopinax sp. DJM-731 SS1](gi|402220063|gb|EJU00136.1). No subset NR somente com *Cryptococcus ATP-dependent protein bindingprotein [Cryptococcus gattii WM276],ATP-dependent protein binding protein, putative [Cryptococcusgattii WM276], ubiquitin protein 1 [Cryptococcus gattii R265] (gi|321263887|ref|XP_003196661.1|).*

Figura 97: Visualização de parte da rede, evidenciando o nó com maior grau.

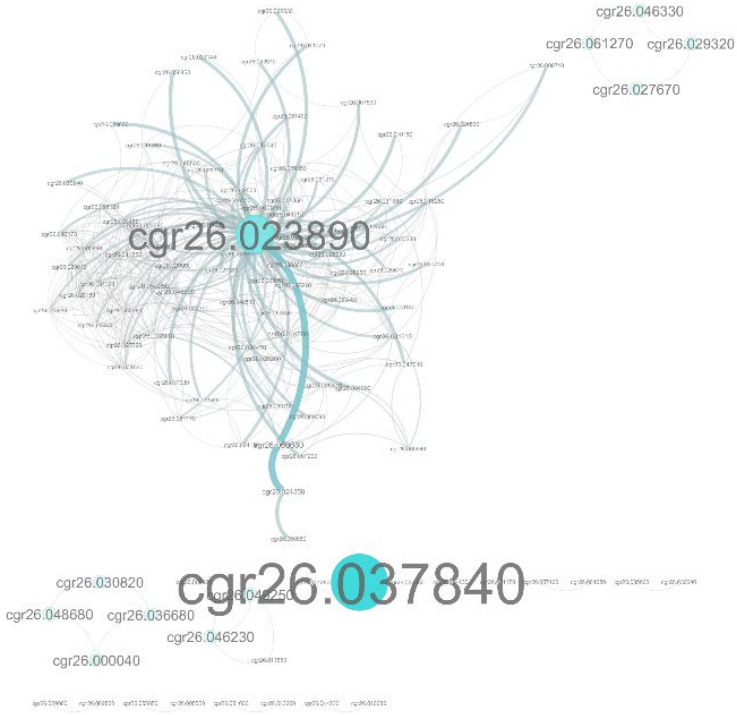

Figura 98: Visualização da rede evidenciando os nós com maior grau dentro dos subgrupos.

Figura 99: Visualização da rede evidenciando os nós com maior grau dentro da principal rede, indicando a parte superior e a parte inferior.

Rede de *Cryptococcus pinus* CBS 10737

Na rede que construímos para *Cryptococcus pinus* CBS 10737observamos uma maior quantidade de conjuntos de redes ou pares de interações proteína- proteína São 12 conjuntos de interações, porém como nós anteriores somente um possui um grande número de nós. O conjunto gerado possui 79 nós e 425 interações descritas.

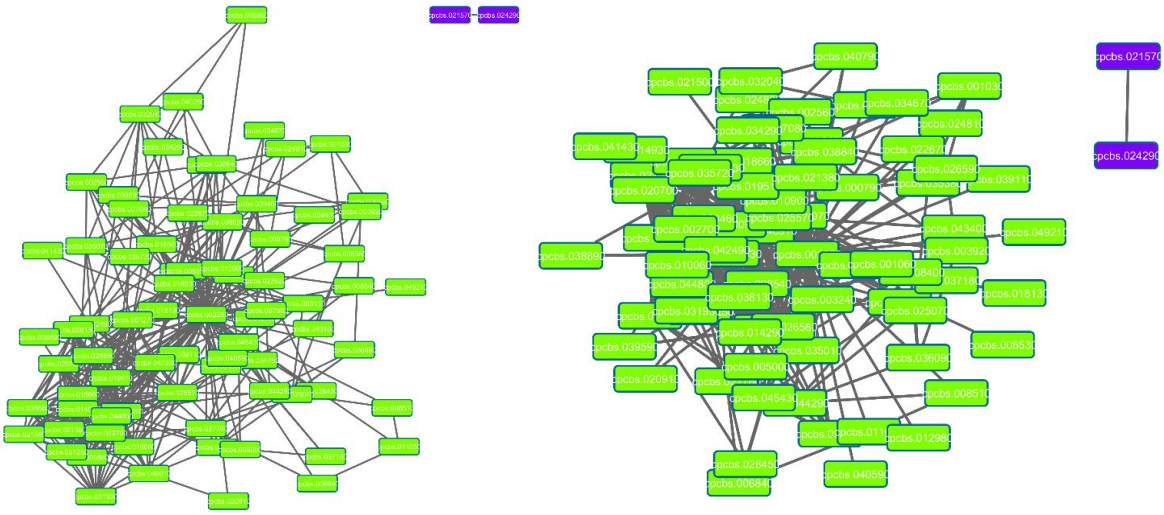

Figura 100: Visualização da rede de interações proteína-proteína de forma geral, onde mostra duas formas de organização para visualziação da rede.

Observamos na rede o nó com um maior grau foi o cpcbs.002250, com 75 arestas, cuja a função descrita no subset do NR sem *Cryptococcus* foi hypothetical protein V496_10504, partial [Pseudogymnoascus pannorum VKM F-4515 (FW-2607)] (gi|682358681|gb|KFY47659.1|). No subset somente com *Cryptococcus ATP-dependent protein binding protein [Cryptococcus neoformans var. neoformans JEC21], hypothetical protein CNBK0920 [Cryptococcus neoformans var.neoformans B-3501A], hypothetical protein CNBK0920 [Cryptococcus neoformans var.neoformans B-3501A], ATP-dependent protein binding protein, putative [Cryptococcus neoformans var. neoformans JEC21], polyubiquitin [Drechslerella stenobrocha 248] (gi|58260904|ref|XP_567862.1|).*

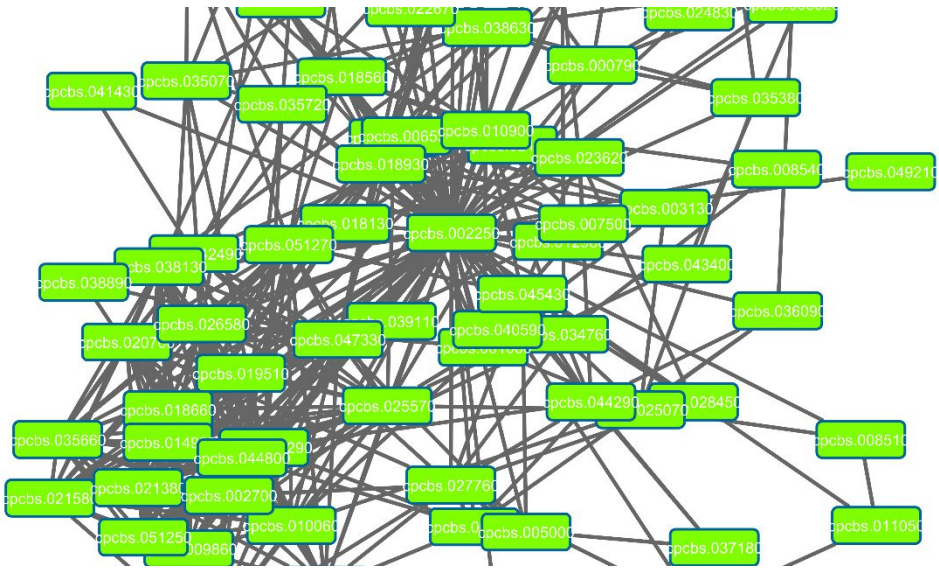

Figura 101: Visualização de parte da rede, evidenciando o nó com maior grau.

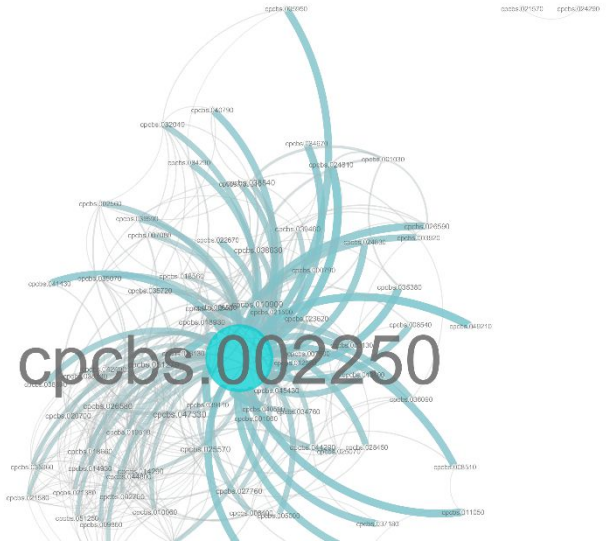

Figura 102: Visualização da rede evidenciando os nós com maior grau dentro dos subgrupos.

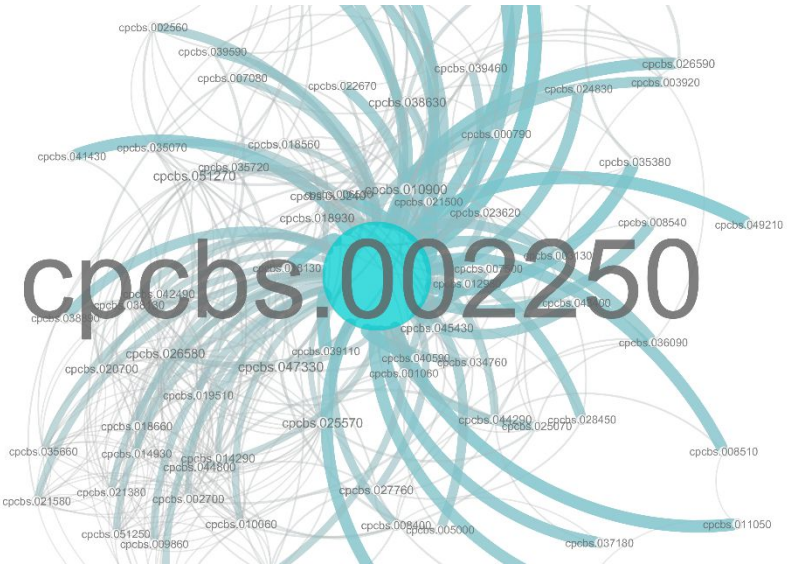

Figura 103: Visualização da rede evidenciando os nós com maior grau dentro da principal rede.

Rede de *Cryptococcus heveanensis* BCC8398

Na rede que construímos para *Cryptococcus heveanensis* BCC8398 observamos uma maior quantidade de conjuntos de redes ou pares de interações proteína-

proteína São 12 conjuntos de interações, porém como nós anteriores somente um possui um grande número de nós. O conjunto gerado possui 81 nós e 393 interações descritas.

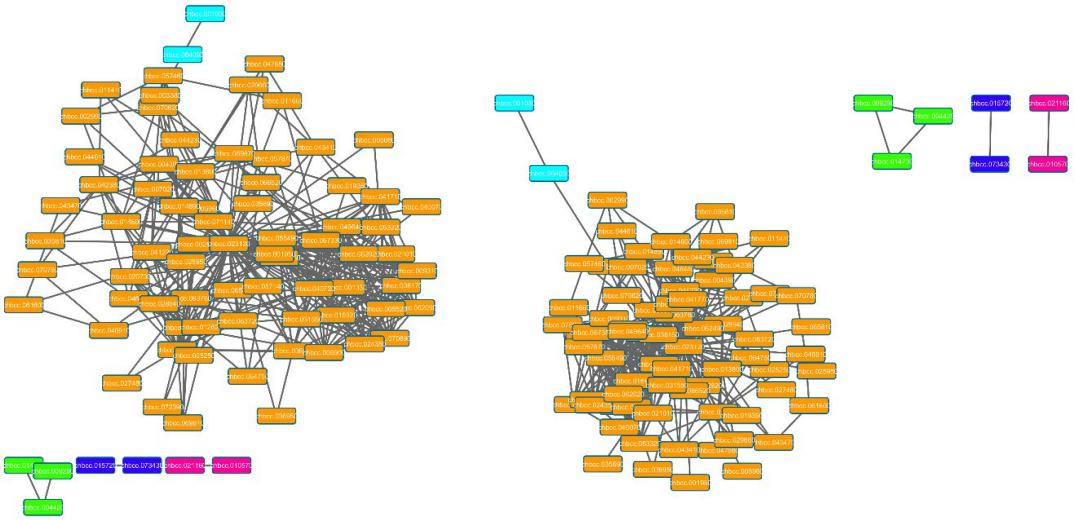

Figura 104: Visualização da rede de interações proteína-proteína de forma geral, onde mostra duas formas de organização para visualziação da rede.

Observamos na rede o nó com um maior grau foi o chbcc.023120, com 70 arestas, cuja a função descrita no subset do NR sem *Cryptococcus* foi hypothetical protein V492_04946 [Pseudogymnoascus pannorum VKM F-4246] (gi|682296803|gb|KFY10578.1). No subset somente com *Cryptococcus | ATP-dependent protein binding protein [Cryptococcus gattii WM276] ATP-dependent protein binding protein, putative [Cryptococcusgattii WM276], ubiquitin protein 1 [Cryptococcus gattii R265] (gi|321263887|ref|XP_003196661.1).*

Figura 105: Visualização de parte da rede, evidenciando o nó com maior grau.

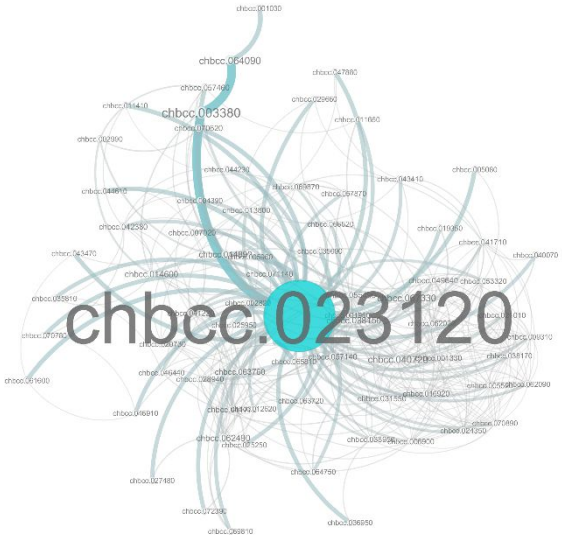

Figura 106: Visualização da rede evidenciando os nós com maior grau dentro dos subgrupos.

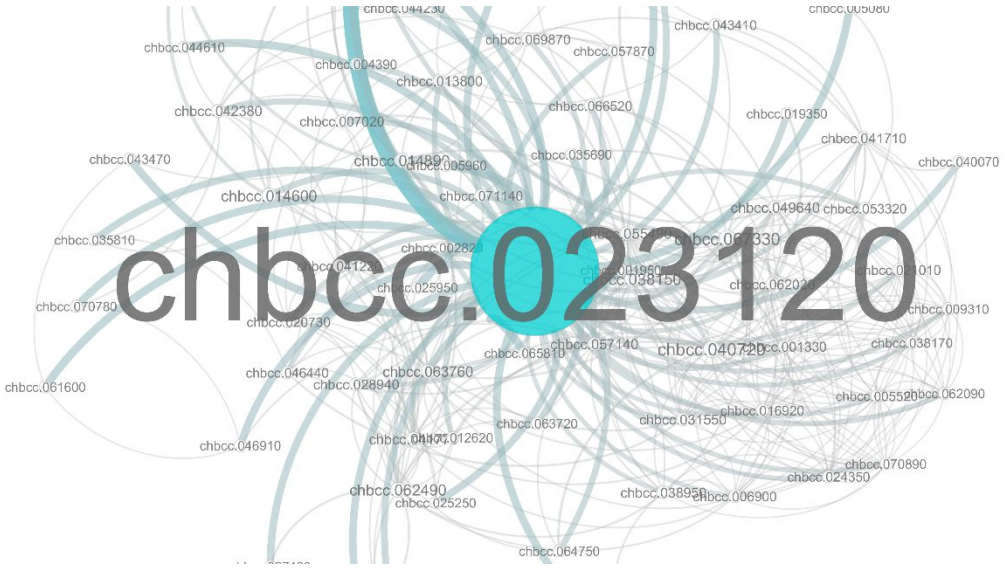

Figura 107: Visualização da rede evidenciando os nós com maior grau dentro da principal rede.

Rede de *Cryptococcus heveanensis* CBS569

Na rede que construímos para *Cryptococcus heveanensis* CBS569 observamos uma maior quantidade de conjuntos de redes ou pares de interações proteína- proteína São 12 conjuntos de interações, porém como nós anteriores somente um possui um grande número de nós. O conjunto gerado possui 84 nós e 415 interações descritas.

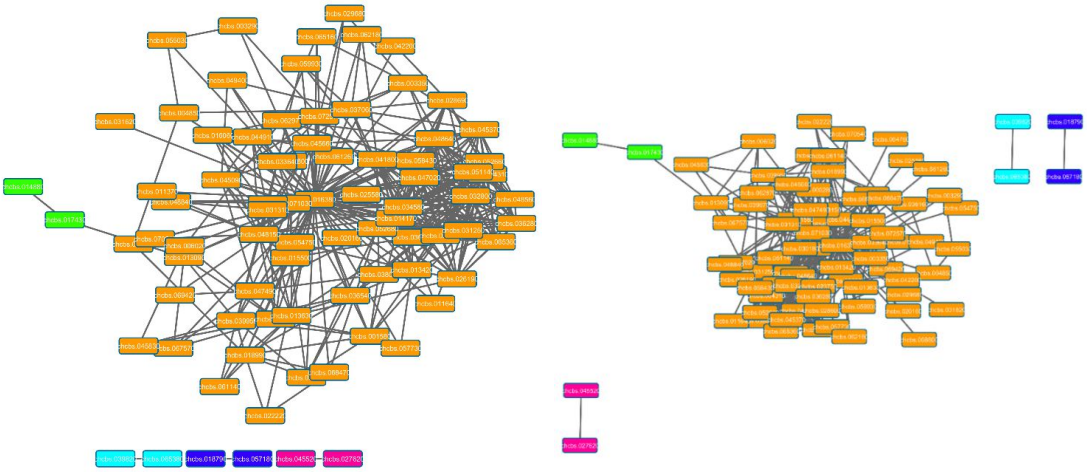

Figura 108: Visualização da rede de interações proteína-proteína de forma geral, onde mostra duas formas de organização para visualziação da rede.

Observamos na rede o nó com um maior grau foi o chcbs.016380, com 72 arestas, cuja a função descrita no subset do NR sem *Cryptococcus* foi hypothetical protein

V492_04946 [Pseudogymnoascus pannorum VKM F-4246](gi|682296803|gb|KFY10578.1). No subset NR somente com *Cryptococcus | ATP-dependent protein binding protein [Cryptococcus gattii WM276] ATP-dependent protein binding protein, putative [Cryptococcus gattii WM276] ubiquitin protein 1 [Cryptococcus gattii R265] (gi|321263887|ref|XP_003196661.1).*

Figura 109: Visualização de parte da rede, evidenciando o nó com maior grau.

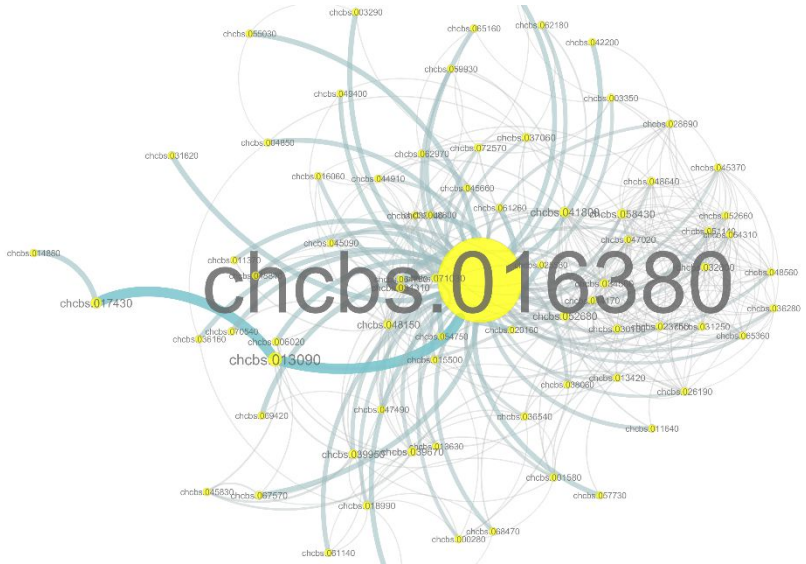

Figura 110: Visualização da rede evidenciando os nós com maior grau dentro dos subgrupos.

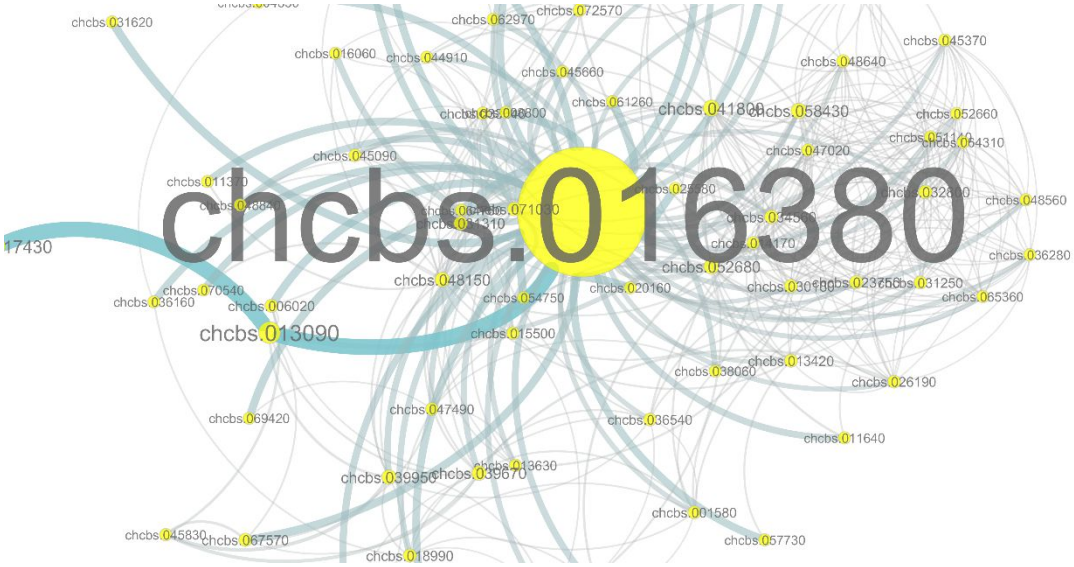

Figura 111: Visualização da rede evidenciando os nós com maior grau dentro da principal rede.