Universidade de São Paulo
Programa Interunidades de Pós-Graduação em Bioinformática
Faculdade de Ciências Farmacêuticas

**Desenvolvimento de uma Ferramenta Computacional para Análise de
Co-Expressão Gênica e sua Aplicação na Biologia de Sistemas**

Pedro de Sá Tavares Russo

Orientador: Prof. Dr. Helder Takashi Imoto Nakaya

Versão corrigida da dissertação defendida

São Paulo
Data do Depósito Original na CPG:
22/03/2019

Universidade de São Paulo
Programa Interunidades de Pós-Graduação em Bioinformática
Faculdade de Ciências Farmacêuticas

**Desenvolvimento de uma Ferramenta Computacional para Análise de
Co-Expressão Gênica e sua Aplicação na Biologia de Sistemas**

Pedro de Sá Tavares Russo

Orientador: Prof. Dr. Helder Takashi Imoto Nakaya

Dissertação apresentada ao
Programa Interunidades de
Pós-graduação em Bioinformática
da Universidade de São Paulo para
a obtenção do título de Doutor em
Bioinformática

São Paulo
2019

FICHA CATALOGRÁFICA

R969 Russo, Pedro de Sá Tavares
Desenvolvimento de uma ferramenta computacional para análise de co-expressão gênica e sua aplicação na biologia de sistemas / Pedro de Sá Tavares Russo, Helder Takashi Imoto Nakaya, [orient.]. São Paulo : 2019.
106 p.

Tese (Doutorado) - Universidade de São Paulo
Orientador: Prof. Nakaya, Helder Takashi Imoto

Programa Interunidades de Pós-Graduação em Bioinformática
Área de concentração: Bioinformática

1. Bioinformática. 2. Expressão Gênica 3. Transcrição Gênica. 4. Banco de Dados Relacionais. I. Nakaya, Helder Takashi Imoto, orientador. II. Universidade de São Paulo. III. Título.

CDD – 572.8

Elaborada pelo Serviço de Informação e Biblioteca Carlos Benjamin de Lyra do IME-USP, pela bibliotecária Eliana Mara Martins Ramalho CRB-8/ 4819

Agradecimentos

Apesar de ser meu o nome que encabeça a primeira página deste trabalho, aproveito este espaço para agradecer e nomear aqueles que de alguma maneira marcaram o desenvolvimento desta tese, ou de seu autor.

Primeiramente, a meus pais, Ana e Antonio, pelo amor, educação e por me ensinarem que a maior virtude que alguém pode ter é a ambição pelo auto-desenvolvimento, e à minha irmã, Mariana, pela certeza de um eterno ombro para me apoiar.

Aos grandes amigos que eu tive o prazer de conhecer cedo na vida: Daniel, Felipe, Patrick, Raul, Rodrigo e Thiago, que mesmo depois de anos, mudando de opiniões, pontos de vista e estilos de vida, continuamos mantendo viva essa segunda família que cultivamos juntos.

Ao meu orientador Helder, por apesar de sua posição como chefe sempre ter me deixado totalmente à vontade para perguntar, discutir, debater, e conversar. Por sempre prezar pelo nosso ambiente de trabalho e o entrosamento do laboratório acima de tudo. Por se preocupar sempre pelo nosso aprendizado, ainda que custe mais tempo. Pela orientação e liderança no decorrer do projeto. Pelos *scripts* reaproveitados. Por focar naquilo que é importante e não deixar que me perdesse nos milhares de detalhes, e sempre ter a noção de qual seria o próximo passo a ser tomado. Pela amizade, pelas festas e pelo churrasco de fim de ano.

Aos amigos e colegas do CSBL adquiridos durante o doutorado, desde a turma mais antiga, que pacientemente me ajudaram com o começo, sempre mais difícil: Thiago, Matheus, Lucas, Gustavo, Diógenes, Fred, Léo - até o pessoal mais novo, que pude ter o prazer de ver aprender cada vez mais e crescer, que pude ensinar, e que também aprendi muito: Fernando,

Jaque, Mindy, César, Fábio, Melissa, Flávio, Paty, Edu, Diogo, Mari, Nathy, Bruno, Vivi, Tiago, Isa, André, Sara, Alysso, Deney, Bruna, Leandro, Alan, Felipe, Ana.

À FAPESP, pelo apoio financeiro (processo 2017/05762-7), à USP pela infraestrutura oferecida para o desenvolvimento do trabalho, e aos professores das matérias que cursei.

Finalmente, à minha incrível namorada Gabi, que sempre faz de tudo para tentar me acalmar nos meus momentos de ansiedade e que sacrifica finais de semana e horários de lazer para me ajudar quando preciso trabalhar. Por me ajudar a me cobrar menos, aumentar minha auto-estima e me aturar mesmo quando estou de mau humor. Por ser minha companheira de todo momento, minha melhor amiga e minha cúmplice.

Change everything you are

And everything you were

Your number has been called

Fights, battles have begun

Revenge will surely come

Your hard times are ahead

Best, you've got to be the best

You've got to change the world

And you use this chance to be heard

Your time is now

Resumo

A Biologia de Sistemas proporciona um olhar holístico sobre os processos biológicos, integrando os diversos componentes intracelulares através de redes altamente complexas. Em particular, redes de co-expressão tem permitido nos últimos anos uma compreensão cada vez maior dos sistemas biológicos e dos mecanismos moleculares que os regem. Por outro lado, as ferramentas matemáticas e estatísticas já desenvolvidas para a análise destas redes e sistemas são, em geral, densas e pouco familiares para profissionais das áreas biológicas e da saúde. Portanto, a fim de possibilitar uma análise ao mesmo tempo relevante e facilitada, nosso grupo criou a ferramenta CEMiTool, que tem por objetivo identificar módulos de coexpressão de genes de modo automático, de maneira fácil e intuitiva para usuários com pouca ou nenhuma experiência com linguagens de programação. A fim de demonstrar a facilidade de uso da ferramenta, aplicamos o CEMiTool a mais de 1000 estudos de transcriptômica, cujos resultados foram utilizados para a confecção de um banco de dados, permitindo a integração de informações entre estudos. Além disso, para facilitar ainda mais o acesso a este tipo de análises, foi criada uma versão online da ferramenta, denominada webCEMiTool, que permite realizar as análises no navegador. Finalmente, criou-se também a ferramenta annotator, permitindo a definição automática de grupos de amostras de estudos de transcriptômica a partir do agrupamento de cadeias de caracteres presentes em dados de anotação. Todo o código está livremente disponível à comunidade.

Palavras-chave: módulos de co-expressão, redes gênicas, análises modulares, transcriptômica, banco de dados.

Abstract

System biology methods provide a holistic view of biological processes, integrating the several intracellular molecular components via the use of highly complex networks. In particular, co-expression networks have allowed for an increasing understanding of biological systems and the complex molecular mechanisms driving them. On the other hand, previously described tools for the analysis of biological networks are in general relatively difficult to use for life and health scientists given their high mathematical and computational demand. Therefore, in order to provide at the same time a relevant and easy-to-use analysis, we have developed the CEMiTool package, which aims to identify gene coexpression modules in an automatic, easy and intuitive way for users with little to no prior computational expertise. We applied CEMiTool to over 1000 transcriptomics studies and used the results to create a new gene coexpression database, which allows users to integrate information across analyses. Moreover, to further facilitate analyses we developed an online version of the tool named webCEMiTool, which permits users to run coexpression analysis easily via browser. Finally, we also developed annotator, a package for automatically determining experimental groups based on sample annotation string similarity. All code is freely available to the community.

Keywords: coexpression modules, gene networks, modular analysis, transcriptomics, databases.

Lista de abreviaturas e siglas

CEMiTool - Co-expression Modules identification Tool

DICER - Differential Correlation in Expression for meta-module Recovery

WGCNA - Weighted Gene Coexpression Network Analysis

ORA - Over-representation Analysis

GSEA - Gene Set Enrichment Analysis

GEO - Gene Expression Omnibus

VST - Variance Stabilizing Transformation

TOM - Topological Overlap Matrix

SQL - Structured Query Language

HDBSCAN - Hierarchical Density Based Clustering of Applications with Noise

CES - Combined Enrichment Score

NES - Normalized Enrichment Score

DINGO - Differential Network Analysis in Genomics

CoP - Coexpressed Biological Processes

GZMA - Granzyme A

SW - Smith-Waterman

NW - Needleman-Wunsch

ME - Monge-Elkan

LV - Levenshtein

DL - Damerau-Levenshtein

OSA - Optimal String Alignment

LCS - Longest Common Substring

Índice de figuras

Figura 1 - Análises realizadas pelo CEMiTool.	26
Figura 2 - <i>Workflow</i> do CEMiTool.	27
Figura 3 - Arquivo de expressão.	28
Figura 4 - Arquivo de anotação de amostras.	29
Figura 5 - Arquivo de interações.	30
Figura 6 - Arquivo de <i>gene sets</i> .	30
Figura 7 - Redes livres de escala.	34
Figura 8 - Comparação de curvas de β x R^2 .	36
Figura 9 - Pseudo-código referente ao algoritmo de definição automática do valor de beta.	38
Figura 10 - <i>Workflow</i> utilizado para popular o banco de dados CEMiToolDB	42
Figura 11 - Esquema do banco de dados CEMiToolDB.	43
Figura 12 - <i>Workflow</i> do pacote anotator.	46
Figura 13 - Números de <i>downloads</i> do pacote CEMiTool.	48
Figura 14. Seleção automática do parâmetro beta.	51
Figura 15 - Filtragem automática dos genes do CEMiTool.	53
Figura 16 - Aba “Modules” do relatório do CEMiTool.	55
Figura 17 - Aba “Profile Plot” do relatório do CEMiTool.	56
Figura 18 - Aba “Gene Set Enrichment Analysis”.	58
Figura 19 - Aba “Over Representation Analysis” do relatório do CEMiTool.	60
Figura 20 - Aba “Interaction Network” do relatório do CEMiTool.	61
Figura 21 - Aba “Parameters” do relatório do CEMiTool.	62
Figura 22 - Aba “Sample Clustering” do relatório de diagnóstico do CEMiTool.	64

Figura 23 - Aba “Mean Variance” do relatório diagnóstico do CEMiTool.	66
Figura 24 - Aba “Quantile-quantile plot” do relatório diagnóstico do CEMiTool.	67
Figura 25 - Aba “Expression histogram” do relatório diagnóstico do CEMiTool.	68
Figura 26 - Aba “ β x R^2 plot” do relatório diagnóstico do CEMiTool.	69
Figura 27 - Aba “Mean connectivity plot” do relatório diagnóstico do CEMiTool.	70
Figura 28 - <i>Homepage</i> da plataforma webCEMiTool.	71
Figura 29 - Página de <i>upload</i> de arquivos da plataforma do webCEMiTool.	72
Figura 30 - Seção “Summary” da página de resultados do webCEMiTool.	73
Figura 31 - Seção “Modules” da página de resultados do webCEMiTool.	74
Figura 32 - Seção “Gene Set Enrichment Analysis” da plataforma webCEMiTool.	75
Figura 33 - CEMiTool aplicado a estudos de dengue.	81
Figura 34 - CEMiTool aplicado a um estudo de RNAseq de pacientes com psoríase.	83
Figura 35 - Análise modular de leishmaniose.	87
Figura 36 - Fluxograma de análise da ferramenta.	89
Figura 37 - Encontrando genes co-expressos com o gene GZMA.	91
Figura 38 - Encontrando vias metabólicas de genes co-expressos com o gene GZMA.	92
Figura 39 - Encontrando vias metabólicas de módulos contendo uma lista de genes.	93
Figura 40 - Aplicação do pacote annotator no estudo GSE2125.	94
Figura 41. Curvas R^2 x β hipotéticas para WGCNA.	107
Figura 42 - Relação entre o número de amostras e ϕ .	110

Índice de Tabelas

Tabela 1 - Funcionalidades de programas de identificação de redes de Coexpressão.	79
Tabela 2 - Métricas de avaliação dos <i>clusters</i>	95
Tabela 3 - Dependências do pacote CEMiTool	112
Tabela 4 - Dependências do pacote annotator	112

Sumário

1. Introdução	14
1.1. Redes Gênicas	14
1.2. Biologia de Sistemas	15
1.3. Redes de co-expressão	17
1.4. Método WGCNA e análises complementares	18
2. Objetivos	25
2.1 Objetivos primários	25
2.2 Objetivos secundários	25
3. Métodos	25
3.1. Arquivos de entrada e parâmetros	28
3.2. Pacote CEMiTool para R	31
3.2.1. Instalação	31
3.2.2. Seleção dos genes de entrada	31
3.2.3. Criação das redes de coexpressão	33
3.2.3.1. Determinação da matriz de adjacência	33
3.2.3.2. Determinação do parâmetro beta	35
3.2.3.3. Definição de módulos	38
3.2.4. Análises adicionais	39
3.2.5. Criação de relatório automatizado	40
3.2.6. Relatório de diagnóstico	40
3.3. webCEMiTool	41
3.4. Banco de dados CEMiToolDB	41
3.5. Pacote anotator	44
4. Resultados	47
4.1. Estatísticas de <i>downloads</i> do CEMiTool	47
4.2. Comparação empírica dos métodos de seleção do parâmetro beta em 15 estudos de microarray	49
4.3. Determinação dos parâmetros de filtragem	51
4.4. Geração de <i>report</i> automatizado	54
4.5. Relatório de diagnóstico	63
4.6. webCEMiTool	70
4.7. Comparação do CEMiTool com outros métodos de detecção de redes de coexpressão	75
4.8. Aplicação do CEMiTool a dados reais	79

4.8.1. Dados de dengue	79
4.8.2. Dados de psoríase	82
4.8.3. Dados de leishmaniose visceral	84
4.9. Banco de dados CEMiToolDB	88
4.9.1. Exemplo de uso: Granzima A	90
4.10. Pacote annotator	94
5. Discussão	96
6. Conclusão	101
7. Referências bibliográficas	107
<u>Apêndice I</u>	107
1. Métodos complementares	107
1.1. Comparação dos métodos de seleção do parâmetro beta nos pacotes WGCNA e CEMiTool	107
1.2. Filtragem baseada em variância aplicada a expressão gênica	108
1.2.1. Estimação dos parâmetros α e β	109
1.2.2. <i>Combined Enrichment Score</i> (CES)	109
1.3. Efeito do número de amostras no ajuste ao modelo de topologia livre de escala	110
<u>Apêndice II</u>	112
1. Dependências	112
<u>Apêndice III</u>	114
1. Publicações	114
2. Atividades acadêmicas	114
2.1. Disciplinas cursadas	114
2.2. Organização de congressos/reuniões científicas	115
2.3. Participação de congressos/reuniões científicas	115
2.4. Atividades complementares	117

1. Introdução

1.1. Redes gênicas

Processos biológicos são regulados por vias metabólicas, de sinalização e de expressão gênica. Nestas vias, milhares de componentes como genes, RNA, proteínas ou metabólitos, interagem entre si formando redes regulatórias extremamente complexas. Exemplos de interações incluem: proteína-proteína (homodímeros, heterodímeros, complexos inibidores de enzima, complexos anticorpo-proteína, entre outros) (Jones and Thornton, 1996); proteína-DNA (fatores de transcrição TFIIA, TFIIB, TFIID, TFIIE, TFIIIF, entre outros) (Thomas and Chiang, 2006); proteína-RNA (proteínas ligantes a RNA - sendo as mais comuns em humanos aquelas com domínios RG/RGG, DEAD, zf-CCCH – e riboswitches) (Gerstberger *et al.*, 2014); RNA-RNA (formação do complexo RISC no silenciamento pós-transcricional de RNA mensageiro) (Salmanidis *et al.*, 2014); e metabólito-proteína (moduladores e efetores alostéricos) (Lehninger *et al.*, 2012).

As redes de interações biológicas apresentam três propriedades inerentes interessantes. Primeiro, elas tendem a ser robustas, o que significa que são capazes de manter seu estado e funções frente a diferentes perturbações internas e externas (Friboulet and Thomas, 2005). Segundo, elas possuem propriedades emergentes, ou seja, da interação múltipla entre componentes simples de um sistema é possível emergir uma característica complexa (Konopka, 2006). Terceiro, as redes se organizam de forma modular. A modularidade é um dos fatores que mais influencia na robustez dos sistemas biológicos, uma vez que ela provoca um relativo isolamento físico ou

funcional dos subsistemas de maneira que se houver falha em um módulo, os outros não serão afetados de forma catastrófica (Kitano, 2002). Graças ao agrupamento dos processos biológicos em módulos, os efeitos das interações entre os seus diversos componentes (e não necessariamente os componentes em si) tornam-se passíveis de ser estudados.

1.2. Biologia de Sistemas

A Biologia de Sistemas dedica-se a estudar e conhecer as interações complexas que existem nos sistemas biológicos utilizando um olhar holístico, isto é, focando-se em como todo o conjunto de sistemas biológicos de interesse se comportam, em oposição à metodologia reducionista convencional (Fang and Casadevall, 2011). Assim, o âmbito dessa ciência consiste em adquirir informações de cada nível de informação de sistemas biológicos e integrar estes dados de maneira a gerar modelos matemáticos preditivos dos referidos sistemas.

O desenvolvimento da área tem se dado em grande parte devido à crescente capacidade de se realizar uma perturbação dos sistemas biológicos, ao ponto que hoje é possível prever respostas de vias metabólicas às interferências impostas e confirmá-las experimentalmente (Likić *et al.*, 2010). Além disso, o advento de novas tecnologias como sequenciadores de DNA, *microarrays* e proteômica *high-throughput* têm permitido que as propriedades e a dinâmica dos sistemas biológicos sejam avaliadas mais rápida e compreensivamente. De fato, métodos de Biologia de Sistemas já foram utilizados para estudar diversos tipos de interações biológicas complexas,

como prever funções metabólicas exercidas diretamente por conjuntos de genes específicos (Ghazalpour *et al.*, 2006) ou identificar como coordenam processos regulatórios e assim modular sistemas indiretamente (Xue *et al.*, 2013).

Estes dados da chamada área da transcriptômica são armazenados comumente em bancos de dados públicos como o GEO (Gene Expression Omnibus) (Edgar *et al.*, 2002), o qual proporciona diversos tipos de estruturas de dados para proporcionar a catalogação e armazenamento das informações de amostras biológicas e os valores de expressão de genes. Resumidamente, dados sobre conjuntos de experimentos são armazenados em “plataformas” (código GPLxxx), com descrição dos *arrays* ou sequenciadores, em conjunto com informações sobre as sondas utilizadas em comum. Amostras são armazenadas sob o identificador GSMxxx, e elas mantêm informações sobre procedimentos experimentais particulares aplicados a elas, assim como os valores de intensidade de sinal das mesmas. Dados de um estudo particular são armazenados nas chamadas “séries” (GSExxx), as quais contém uma descrição resumida de todas as etapas experimentais realizadas, assim como análises e conclusões. Finalmente, *datasets* (GDSxxx) são séries GSE curadas manualmente pela equipe do GEO, a fim de garantir a qualidade e a comparabilidade dos dados entre diferentes estudos. Amostras dentro de um mesmo GDS são oriundas de uma mesma plataforma e normalizadas da mesma maneira, sendo portanto facilmente comparadas. Em maio de 2019, o GEO contava com 4348 *datasets*, 113013 séries, 19714 plataformas e 3038267 amostras.

1.3. Redes de co-expressão

Estudos de Biologia de Sistemas podem levar em consideração as chamadas redes de coexpressão. Neste caso, as interações entre os genes refletem os casos em que os mesmos possuem perfil de expressão similares entre amostras diferentes (Navlakha *et al.*, 2014). Ou seja, uma rede de coexpressão interligará os genes (os nós da rede) entre si a partir de alguma medida de similaridade, em geral dada pelo valor da correlação de Pearson entre a expressão dos genes em questão (Langfelder and Horvath, 2008). Após o estabelecimento das redes de coexpressão, agrupam-se os genes mais correlacionados em módulos, e estes por sua vez podem ser correlacionados com informações externas, como dados clínicos, *Gene Ontology*, entre outros, a fim de se encontrar módulos biologicamente interessantes.

Análises de redes de coexpressão gênica têm tido sucesso em elucidar o papel de genes e vias de sinalização em inúmeras condições biológicas, desde câncer (Carter *et al.*, 2004) até AIDS (Yanhong Zhou, 2014). Este tipo de análise pode ser realizada através de diversas ferramentas, entre as quais é possível citar Petal (Petereit *et al.*, 2016), que é um pacote para a linguagem de programação R que procura definir uma rede de coexpressão utilizando um corte de similaridade automaticamente estabelecido. Já a ferramenta DICER também é capaz de identificar módulos de coexpressão, diferenciando grupos de genes que possuem correlações diferentes entre diferentes grupos de amostras (Amar *et al.*, 2013). CoP (Co-expressed Biological Processes; anteriormente, CoexProcess) é um banco de dados de co-expressão focado em plantas, criado a partir de 10022 arrays (Ogata *et al.*, 2010).

Outra ferramenta similar é o GeNET , um framework online para a análise de dados de coexpressão, focado especificamente no compartilhamento *online* dos dados (Desai *et al.*, 2017). Finalmente, DiffCoEx, coXpress e DINGO são ferramentas similares também utilizadas para realizar análises de co-expressão, além de co-expressão diferencial, ou seja, encontrar grupos que apresentem diferenças nos valores de co-expressão entre grupos experimentais (Tesson *et al.*, 2010; Watson, 2006; Ha *et al.*, 2015).

1.4. Método WGCNA e análises complementares

Outra metodologia para a análise de redes de coexpressão gênica é o WGCNA (do inglês, *Weighted Gene Coexpression Network Analysis*) (Zhang and Horvath, 2005). Esta análise baseia-se em redes ponderadas de genes, as quais propiciam resultados mais biologicamente significativos do que redes não-ponderadas, uma vez que quantidades de expressão gênica são relativas, isto é, variam em intensidade. Assim, ao dicotomizar as conexões de rede em conectadas ou não-conectadas, perde-se a informação do quão forte estas interações são (Zhang and Horvath, 2005).

O método WGCNA parte da premissa de que redes biológicas tendem a seguir o chamado modelo *scale-free* (do inglês, livre de escala), segundo o qual muitos genes na rede são esparsamente conectados, enquanto alguns poucos outros são altamente conectados, agindo portanto como elementos centrais da rede, conhecidos como pontos *hubs*. Este modelo já é conhecido na literatura há bastante tempo, sendo originário de um braço da matemática chamado ciência da complexidade (Holme, 2019), tendo sido descrito em alguns campos da ciência da computação, desde pelo

menos 1965 quando de Solla Price descreveu que distribuições de grau de redes de citações bibliográficas tendem a seguir distribuições de Pareto (Price, 1965). Décadas depois, o interesse no modelo ressurgiu quando Barabási e colaboradores o descreveram através do mecanismo que batizaram de *preferential attachment* (do inglês, acoplamento preferencial), e o observaram em algumas redes sociais e biológicas (Barabasi and Albert, 1999).

Apesar da popularidade do modelo *scale-free*, tendo sido descrito em diversos campos após sua popularização, muita controvérsia em torno do mesmo surgiu, e em geral considera-se atualmente que o modelo precisa ser no mínimo adaptado para o uso na maioria dos campos. Em particular, Broido e Clauset propuseram o uso de diferentes tipos de distribuição, como a log-normal, Weinbull, e exponencial após observarem recentemente que o modelo não é tão comum quanto se imaginava, tendo sido observado de maneira realmente significativa em apenas 4% das 937 redes analisadas, embora 57% delas tenham sido classificadas como sendo ao menos similares à distribuição (Broido and Clauset, 2018). Apesar disso, observa-se que as redes biológicas analisadas no estudo são em sua maioria redes ecológicas, não necessariamente estendendo-se a todos os tipos de redes da área, por exemplo redes de co-expressão de genes.

De qualquer forma, a partir do modelo *scale-free*, a metodologia WGCNA é capaz de definir módulos, isto é, grupos de genes altamente correlacionados. Ela então estabelece um perfil de expressão representativo dos mesmos, chamado *eigengene*, e este é utilizado para se comparar as relações entre eles. Finalmente, averigua-se

dentro dos módulos quais os principais genes e aqueles que possuem um potencial valor de interesse através de validações experimentais (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

A fim de unificar dentro de um só pacote todas as metodologias e ferramentas mais comuns para este tipo de análise, Langfelder e Horvath desenvolveram o pacote WGCNA para R (Langfelder and Horvath, 2008). Este pacote possui ferramentas para construção de redes; detecção de módulos (conjuntos de genes com alta correlação de coexpressão); determinação do perfil de um módulo através do estabelecimento de um *eigengene*; determinação de pertencimento a um módulo; detecção automática de módulos por blocos; detecção de módulos-consenso; funções para análise de propriedades topológicas, simulação de dados de *microarrays* com estruturas modulares, visualização de dados, interação com outros pacotes de análise, entre muitas outras.

Apesar da funcionalidade proporcionada pelo pacote WGCNA, até mesmo seguir o seu método pode proporcionar algum trabalho, uma vez que o seu uso demanda conhecimento razoável da linguagem de programação R. Embora os autores do pacote tenham tido o cuidado de criar tutoriais aprofundados sobre o uso do WGCNA, ainda assim os mesmos são relativamente densos. Além disso, o pacote exige que o usuário escolha manualmente parâmetros de análise, e não proporciona uma funcionalidade de filtragem dos genes de entrada. Isso dificulta tentativas de automatização do processo, além de dificultar na reprodutibilidade das análises, uma vez que pesquisadores diferentes podem selecionar parâmetros diferentes, obtendo

resultados distintos para um mesmo conjunto de dados. Finalmente, o pacote WGCNA oferece uma quantidade relativamente limitada de análises posteriores que ele permite fazer tendo-se em mãos os módulos de coexpressão.

Após a divisão de genes em módulos, pode ser interessante, por exemplo, determinar se os genes de um módulo estão enriquecidos por genes de vias conhecidas de sinalização ou metabólicas (*pathways* biológicos). Uma maneira de se realizar isso é utilizando *Over-representation Analysis* (ORA), a qual verifica o enriquecimento utilizando o teste hipergeométrico em vias de bancos de dados conhecidos como o REACTOME (Croft *et al.*, 2011). Além disso, *Gene Set Enrichment Analysis* (GSEA) permite que se associe a atividade de um dado módulo com determinados fenótipos de estudo (por exemplo, grupos de amostra). Finalmente, a integração de dados de coexpressão com informações de interações proteína-proteína pode ser útil para a identificação de reguladores-chave. Nota-se, no entanto, que o uso de todas estas análises requer diversos pacotes e programas, e um domínio razoável de habilidades de bioinformática.

Desta forma, no presente trabalho desenvolvemos o *Co-Expression Modules identification Tool* (CEMiTool), um pacote para a linguagem de programação R que permite que usuários identifiquem e analisem módulos de coexpressão a partir de dados de transcriptoma de maneira totalmente automatizada (Russo *et al.*, 2018). Além disso, desenvolvemos também uma plataforma web para que usuários com nenhuma experiência em programação também possam realizar as análises. O CEMiTool também proporciona um novo método de filtragem de genes, seleção automatizada de

parâmetros de análise, análises de enriquecimento e funcional de módulos, além de integração com dados de interactoma. A ferramenta então disponibiliza os resultados em relatórios HTML, com gráficos de alta definição e tabelas interativas.

Dada a facilidade em realizar análises de coexpressão proporcionada pelo CEMiTool, criamos um banco de dados denominado CEMiToolDB com o intuito de criar um compêndio de informações de coexpressão para milhares de estudos. Esta abordagem possui uma série de benefícios, permitindo a comparação entre estudos de condições similares e reduzindo vieses relacionados a desenhos experimentais específicos. Apesar da existência de uma série de outros bancos de dados baseados em coexpressão (Zhu *et al.*, 2015; van Dam *et al.*, 2015; Obayashi *et al.*, 2019), o CEMiToolDB se destaca justamente por todas as vantagens da ferramenta CEMiTool, isto é, obter informações adicionais sobre vias metabólicas enriquecidas, módulos enriquecidos em determinados fenótipos ou grupos experimentais e interações com redes proteína-proteína.

Por outro lado, mesmo a análise e o pré-processamento de grandes quantidades de estudos ainda é uma tarefa árdua, devido a uma alta heterogeneidade das informações das amostras, mesmo em bancos de dados como o Gene Expression Omnibus (GEO) (Barrett *et al.*, 2013) e ArrayExpress (Kolesnikov *et al.*, 2015). Embora estes recursos empreguem alguma curadoria em relação à anotação das amostras, em geral esta se restringe à padronização dentro de um mesmo estudo ou série de estudos, e para a elaboração de ontologias. Embora hajam esforços no sentido de anotar manualmente as amostras dos bancos de dados, em particular utilizando

recursos de *crowdsourcing* (Hadley *et al.*, 2017; Wang *et al.*, 2016; Shah *et al.*, 2016), estes métodos falham em acompanhar a escala massiva com que os dados de sequenciamento de nova geração vão sendo criados. Alternativas de aprendizado de máquina também tem sido exploradas (Ellis *et al.*, 2018), no entanto, estes métodos exigem conjuntos de treinamento confiáveis prévios a fim de inferir as anotações com precisão alta, e a definição das características a serem usadas para a anotação devem ser inseridas manualmente, impedindo a automatização do processo.

Assim, na presente tese está sendo apresentada a ferramenta de detecção de módulos de coexpressão gênica CEMiTool. Nós comparamos o CEMiTool com outros pacotes de análises de redes de coexpressão e mostramos as diferenças em funcionalidades entre eles. O pacote CEMiTool foi rodado em mais de 1000 estudos de *microarrays* e RNAseq a fim de demonstrar a sua facilidade de uso e automatização de fluxos de trabalho. Utilizamos também o pacote CEMiTool em dados novos de sangue de pacientes infectados por *Leishmania* e descrevemos módulos com genes que não foram descritos anteriormente como sendo associados à doença. Os resultados das análises do CEMiTool em centenas de estudos do banco de dados GEO foram armazenados em um novo banco de dados de coexpressão denominado CEMiToolDB, que necessita da menor interação possível com o usuário, propiciando o acesso às informações de uma maneira fácil e rápida. Estes estudos foram anotados através de uma nova metodologia de anotação automática de amostras através de similaridade de cadeias de caracteres também desenvolvida neste projeto, denominada *annotator*. O pacote CEMiTool para R está disponível como parte do projeto Bioconductor

(10.18129/B9.bioc.CEMiTool), a plataforma web pode ser acessada através do endereço <https://cemitool.sysbio.tools> e uma imagem Docker para o uso da ferramenta está disponível em <https://hub.docker.com/r/csblusp/cemitool>. O pacote anotador pode ser acessado em <https://github.com/pedrostrusso/annotator>. O banco de dados CEMiToolDB está em processo de desenvolvimento do *website* e estará disponível futuramente.

2. Objetivos

2.1. Objetivo primário

Criar uma ferramenta que trace redes de coexpressão gênica provenientes de estudos de transcriptoma de uma forma automática e de fácil utilização.

2.2. Objetivos secundários

- Disponibilizar a ferramenta criada na forma de pacote para a linguagem R;
- Facilitar o uso da ferramenta através da criação de um *website*;
- Disponibilizar um ambiente Docker para uso da ferramenta de maneira a garantir maior reprodutibilidade de resultados;
- Criar uma ferramenta que facilite a anotação automática de amostras (pacote annotator)
- Estabelecer um banco de dados para o armazenamento de resultados.

3. Métodos

O *framework* de análise que criamos é denominado CEMiTool (*Co-expression Modules identification Tool*, do inglês, ferramenta de identificação de módulos de coexpressão). Como implícito no nome, esta ferramenta irá receber os arquivos de entrada e retornar módulos de coexpressão. Um resumo das análises realizadas pode ser visualizado na Figura 1, e o *workflow* da ferramenta está disposto na Figura 2.

Figura 1 - Análises realizadas pelo CEMiTool. (a) O CEMiTool requer um arquivo de expressão de genes para a identificação de módulos e arquivos opcionais para: (b) visualizar o perfil de expressão de genes individuais ao longo de amostras de diferentes grupos, definidos pelo usuário e mostrados com cores diferentes; (c) realizar análise de enriquecimento de módulos, mostrando a atividade dos módulos em cada grupo de amostras, (d) realizar análise funcional dos módulos e (e) criar redes de genes, demonstrando os dez genes mais conectados. (Russo *et al.*, 2018)

Figura 2 - Workflow do CEMiTool. A partir de um arquivo de expressão (cinza), o pacote irá (azul): filtrar os genes de entrada, calcular a matriz de similaridade e correlação, a qual será utilizada para definir uma rede de correlações. Em seguida, é realizado um *clustering* hierárquico, e os módulos de co-expressão são definidos como os galhos do dendrograma resultante. O CEMiTool então retornará o perfil de expressão de cada módulo (rosa). Finalmente, utilizando arquivos opcionais (verde) de anotação de amostras, vias metabólicas, e dados de interações, o pacote é capaz de realizar análises complementares, como GSEA, ORA e redes de interação. Todos os resultados são retornados para o usuário através de um report HTML interativo contendo gráficos de alta qualidade (laranja), com os resultados das análises e uma ferramenta de busca por genes de interesse e detecção de genes *hubs* (rosa).

3.1. Arquivos de entrada

O CEMiTool é capaz de criar módulos de coexpressão gênica utilizando apenas um arquivo com dados já previamente normalizados de expressão gênica (por exemplo, provenientes de *microarrays* e RNAseq), contendo genes em suas linhas, e amostras em suas colunas (Figura 3). O método particular de normalização em geral não possui tanta influência nos resultados das análises, contanto que os valores de correlação entre os genes sejam robustos à transformação utilizada, isto é, se mantenham após sua aplicação.

First column should contain the gene IDs/symbols or probes

Remaining columns are labeled with sample IDs in the header

SamplesName	Bcell_1022	Bcell_2596	Bcell_5852	Bcell_7653	Blast_SU444	Blast_SU501	Blast_SU575	Blast_SU583	Blast_SU654	Blasts_SU048	Blasts_SU070
A1BG	14	9	1	5	13	24	3	3	1	0	23
A1BG-AS1	3	0	1	0	27	6	4	0	0	0	5
A1CF	0	0	0	0	0	1	0	0	0	0	0
A2M	78	192	36	82	66	175	9	74	30	14	97
A2M-AS1	71	76	52	86	49	68	64	55	5	11	98
A2ML1	1	17	3	2	2	1	0	0	0	0	3
A2MP1	0	3	0	6	5	1	1	0	0	0	2
A3GALT2	0	0	0	0	0	0	0	0	0	0	2
A4GALT	0	0	0	0	3	0	0	0	0	0	0
A4GNT	0	1	0	1	0	8	0	0	0	0	0
AAOG	0	0	0	0	0	0	0	0	0	0	0
AAAS	681	317	146	297	684	650	255	316	89	200	1210
AACS	78	93	33	47	130	91	28	33	20	11	118
AACSP1	0	0	0	0	0	0	0	0	0	0	0
AADAC	0	0	0	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0	0	0	0
AADACL2-AS1	0	0	0	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0	0	0	0
AADACL4	0	0	0	0	0	0	0	0	0	0	0
AADACP1	2	0	0	0	0	0	0	0	0	0	0
AADAT	173	113	84	85	82	144	158	69	139	133	227
AAED1	777	303	297	161	518	456	201	119	78	117	238
AAGAB	170	110	93	95	170	162	94	87	135	118	317
AAK1	975	798	625	488	1111	1126	833	481	409	769	1499
AAMDC	76	84	47	48	108	159	31	37	55	59	133
AAMP	47	68	20	32	145	170	39	41	55	14	159
AANAT	0	1	0	0	5	0	0	1	0	0	2

Gene expression data

Figura 3 - Arquivo de expressão. O arquivo de expressão de entrada do CEMiTool deve ser uma tabela separada por tabulação contendo genes nas linhas e amostras nas colunas, com os respectivos valores de expressão nas células da tabela (adaptado de <https://cemitool.sysbio.tools/tutorial>).

Além disso, a ferramenta também pode receber os seguintes arquivos opcionais:

- Arquivo de anotação de amostras, com informações sobre os grupos experimentais em que as amostras se encontram em cada linha, permitindo a análise conjunta destas amostras (Figura 4);
- Arquivo de interação entre proteínas, cujas linhas indicam um par de proteínas - informação utilizada para melhor definir as conexões presentes nos módulos (Figura 5);
- Arquivo contendo dados de módulos biológicos previamente descritos em bancos de dados como REACTOME (arquivo de *gene sets*), com cada linha contendo *pathways* e os genes que pertencem aos mesmos (Figura 6).

SampleName	Class
Bcell_1022	Bcell
Bcell_2596	Bcell
Bcell_5852	Bcell
Bcell_7653	Bcell
Blast_SU444	Blast
Blast_SU501	Blast
Blast_SU575	Blast
Blast_SU583	Blast
Blast_SU654	Blast
Blasts_SU048	Blasts

Figura 4 - Arquivo de anotação de amostras. Tabela separada por tabulação contendo informações sobre o grupo experimental a que cada amostra pertence (adaptado de <https://cemitool.sysbio.tools/tutorial>).

First column contains gene 1

Second column contains gene 2

(IDs/Symbols should match the IDs/Symbols in expression file)

This represents one interaction between A1BG and SNCA genes

Gene1	Gene2
A1BG	ANXA7
A1BG	CDKN1A
A1BG	GRB7
A1BG	LGALS3
A1BG	PRDX4
A1BG	SETD7
A1BG	SMN1
A1BG	SNCA
A1BG	TK1
A2M	ANXA6
A2M	CTSE
A2M	EGLN2

Figura 5 - Arquivo de interações. Tabela separada por tabulação contendo pares de genes que interagem para serem comparados com as interações presentes nos módulos (adaptado de <https://cemitoool.sysbio.tools/tutorial>).

First column contains the gene set name

This column can hold additional information about the gene sets (but it is going to be ignored in the analysis)

TRNA_PROCESSING	http://www.ADAT1	TRNT1	FARS2	METTL1	SARS	AARS	THG1L	SSB	POP4	NSUN2	
REGULATION_OF_BIOLOGICAL_QUALITY	http://www.DLC1	ALS2	SLC9A7	PTGS2	PTGS1	MPV17	SGMS1	AGTR1	AGTR2	APP	FLI1
DNA_METABOLIC_PROCESS	http://www.XRCC5	XRCC4	RAD51C	XRCC3	XRCC2	XRCC6	ISG20	PRIM1	PRIM2	TLK1	TLK2
AMINO_SUGAR_METABOLIC_PROCESS	http://www.UAP1	CHIA	GNPDA1	GNE	CSGALNAC	CHST2	CHST4	CHST5	NAGK	EXTL2	SLC35A3
BIOPOLYMER_CATABOLIC_PROCESS	http://www.BTRC	HNRNPD	USE1	RNASEH1	RNF217	ISG20	CDKN2A	CPA2	FBXO22	ANAPC2	MTOR
RNA_METABOLIC_PROCESS	http://www.HNRNPF	HNRNPD	SYNCRIP	MED24	RORB	MED23	REST	MED21	MED20	ISG20	EPC1
GLUCAN_METABOLIC_PROCESS	http://www.GCK	PYGM	GSK3B	EPM2A	MGAM	GAA	GYS2	GYG2	DYRK2	PYGB	
PROTEIN_POLYUBIQUITINATION	http://www.ERCC8	HUWE1	DZIP3	DDB2	UBE2V1	UBE2V2	AMFR	UBE3C	UBE2D1	TRAF6	STUB1

Genes (IDs/Symbols should match the IDs/Symbols in expression file)

Figura 6 - Arquivo de gene sets. Neste arquivo cada linha corresponde a um dado mecanismo biológico ou via metabólica. A primeira coluna é reservada para o nome da

via em questão, a segunda para metadados (ignorados na análise), enquanto as colunas subsequentes devem conter os genes ou elementos membros da via (adaptado de <https://cemitool.sysbio.tools/tutorial>).

A utilização de cada um dos arquivos de entrada será discutida em maiores detalhes mais adiante.

3.2. Pacote CEMiTool para R

3.2.1. Instalação

A instalação do pacote CEMiTool para R pode ser feita utilizando o pacote BiocManager proporcionado pelo projeto Bioconductor ou baixando o *script* diretamente do site do projeto, utilizando os comandos a seguir:

```
library("BiocManager") # Utilizar o pacote BiocManager  
install("CEMiTool") # Instalar o pacote CEMiTool  
library("CEMiTool") # Carregar o pacote instalado
```

3.2.2. Seleção de genes de entrada

O primeiro passo tomado para as análises pelo pacote CEMiTool consiste em uma etapa de filtragem, na qual serão selecionados os melhores genes para se determinar os módulos de coexpressão. Os autores do pacote WGCNA recomendam uma filtragem baseada na expressão média ou na variância dos genes, ao invés de expressão diferencial, uma vez que a mesma invalidaria a premissa de topologia livre de escala (ver subseção 3.2.3.1 Determinação da matriz de adjacência). No entanto,

um número de genes ideal para a filtragem não é definido, o que pode levar pesquisadores a fazer escolhas arbitrárias para selecionar seus genes, afetando análises posteriores.

Desta maneira, implementamos no CEMiTool um método flexível, porém objetivo, para a filtragem dos genes. O método de filtragem está descrito na seção Métodos Complementares, e a obtenção empírica dos melhores parâmetros de filtragem encontra-se na seção Resultados. Resumidamente, ao se modelar a variância dos genes como uma distribuição *gamma* inversa, conforme sugerido em (Wright and Simon, 2003), é possível selecionar genes baseados em um p-valor (em nossas análises, utilizamos $p = 0,1$ como valor de corte). Para certos tipos de normalizações de dados de RNA-seq, o nosso método permite uma correção da dependência entre a média e a variância (Subramaniam and Hsiao, 2012; Robinson *et al.*, 2010) ao modelarmos os dados de expressão como uma distribuição binomial negativa (Robinson *et al.*, 2010) e então aplicar a transformação estabilizadora de variância apropriada (VST, do inglês, *Variance Stabilizing Transformation*) (Yu, 2009). A fim de remover potenciais fontes de ruído, o nosso pacote também remove por padrão os 25% (valor arbitrariamente definido) dos genes com a menor média de expressão ao longo das amostras antes da filtragem. Esta segunda etapa de filtragem pode apresentar como efeito colateral a remoção de genes baixo expressos potencialmente interessantes, tais como genes reguladores; estes genes podem ser levados em consideração na análise de redes com dados de interação, ilustrada na seção 4.4 “Geração de report automatizado”.

3.2.3. Criação das redes de coexpressão

3.2.3.1. Determinação da matriz de adjacência

Após a filtragem, o CEMiTool define valores de correlação entre os valores de expressão de cada par de genes presente no arquivo de expressão, criando uma matriz de correlação. Em seguida, o CEMiTool analisa se a topologia da rede formada pelos genes analisados é livre de escala. Redes livres de escala são, por definição, aquelas em que a probabilidade de um nó estar conectado com k outros nós ($P(k)$) decai segundo uma distribuição lei de potência $P(k) \sim k^{-\gamma}$ (Zhang and Horvath, 2005; Horvath, 2011), conforme pode ser visualizado na Figura 7a. Consequentemente, uma rede livre de escala possui alguns poucos nós (os chamados *hubs*) com muito mais conexões que o restante (Barabási and Bonabeau, 2003).

Assim, para determinar se a rede de coexpressão segue de fato uma topologia livre de escala (ou seja, a conectividade dos genes segue uma distribuição lei de potência), pode-se observar o coeficiente de determinação da reta R^2 (definido por $-r^2 \cdot \frac{|a|}{a}$, onde r é o coeficiente de correlação, e a é a inclinação da linha de regressão) na curva entre $\log_{10}(p(k))$ e $\log_{10}(k)$. Uma linha reta nesse gráfico (isto é, valores de R^2 próximos de 1) indica topologia livre de escala, como pode ser visto na Figura 7b (Zhang and Horvath, 2005).

Figura 7. Redes livres de escala. (a) O gráfico mostra a probabilidade ($P(k)$) de um determinado nó de uma rede estar conectado a k outros nós desta rede. Verifica-se que esta curva tende a decair segundo uma curva lei de potência $P(k) \sim k^{-\gamma}$, com gamma entre os valores 2 e 3. (b) Transformação logarítmica da curva $P(k) \times k$. Como é esperado que dados biológicos se comportem segundo a distribuição lei de potência exposta em (A), quanto mais próximo de 1 o coeficiente de determinação R^2 da transformação logarítmica dos dados, mais próxima do modelo livre de escala será a rede.

Em seguida, deve-se determinar uma função de adjacência, utilizada para criar uma matriz de adjacência $A = a_{ij}$ da rede, cujas entradas a_{ij} especificam a força da conexão entre os genes i e j (Horvath, 2011). A escolha desta função também determina se a rede é ponderada ou não. Para redes ponderadas (como é do escopo

da análise por WGCNA), utiliza-se a função de adjacência $a_{ij} = s_{ij}^\beta$. O parâmetro β é um inteiro positivo que transforma o valor da correlação entre os genes i e j em uma matriz de adjacência (Zhang and Horvath, 2005). Esta transformação pode ser realizada de forma a manter o sinal das correlações entre os genes i e j ($s_{ij} = (1 + \text{cor}(i,j))/2$) ou removendo-se o sinal ($s_{ij} = |\text{cor}(i,j)|$). Por padrão, o CEMiTool utiliza a versão sem sinal da transformação, porém o usuário tem a liberdade de definir o método que preferir.

3.2.3.2. Determinação do parâmetro beta

A seleção do parâmetro β possui influência direta no quão aderente ao modelo livre de escala a rede será. Desta forma, é uma etapa crucial do processo, e sua seleção deve ser tomada com cuidado. Para determinar o valor de β , os autores do pacote WGCNA recomendam o critério de topologia livre de escala, segundo o qual se testa um conjunto de possíveis valores para β (por padrão, os inteiros de 1 a 20) e os respectivos valores de R^2 para a rede formada. Seleciona-se como β então o menor valor cujo R^2 seja acima de 0,85. Assim, o CEMiTool observa a forma da curva $\beta \times R^2$ (Figura 8). O ideal é que a curva se aproxime de 1 mais rapidamente que β se aproxima de seu valor máximo. Em outras palavras, a curva deve assumir valores próximos de 1 já para valores de β baixos, conforme ilustrado na Figura 8.

Figura 8 - Comparação de curvas de β x R^2 . Dados mais adequados à análise devem ter valores de R^2 mais próximos de 1 para valores pequenos de β .

Embora o pacote WGCNA para R desenvolvido por Langfelder e Horvath possua uma função (*pickSoftThreshold*) para a determinação automática do valor β , nós criamos um outro algoritmo para determinar β que tem encontrado resultados mais consistentes e confiáveis. Além disso, nosso algoritmo permite um valor de corte para R^2 menor ($R^2 > 0.8$) quando comparado com o valor de corte padrão do WGCNA ($R^2 > 0.85$) - o que, por sua vez, permite a seleção de valores menores de β . Esta comparação está detalhada melhor na seção Métodos complementares.

Nosso algoritmo de determinação do parâmetro β se baseia no conceito matemático de sequências de Cauchy, definido como: uma sequência (x_m) de números reais é dita de Cauchy se, e somente se, existe um número natural N tal que

$$|x_m - x_n| < \varepsilon, \forall \varepsilon > 0 \in \mathbb{R}, \forall m, n \geq N.$$

Ou seja, de maneira mais intuitiva, a partir de um valor N , todos os elementos da sequência estarão arbitrariamente próximos (as distâncias entre os valores são menores que um ε definido). Assim, escolhe-se um valor para β de tal forma que ele e os seguintes valores do parâmetro possuam correspondentes R^2 dentro de uma faixa pré-especificada (em nosso algoritmo, definimos uma faixa de largura 0,05). Contudo, segundo os autores do pacote de WGCNA, a fim de garantir ainda uma alta conectividade média na rede, o valor de R^2 correspondente ao β escolhido deverá ser o menor valor de β cujo R^2 ainda seja superior a 0,8 (Langfelder and Horvath, 2008). Feitas estas ressalvas, descrevemos o algoritmo de determinação do parâmetro β abaixo:

Sejam $P = (p_1, p_2, \dots, p_n)$ o vetor que contém os números “candidatos” a β , $R^2(p)$ o valor de R^2 calculado em p , $d(x, y)$ a distância entre os números reais x e y e $K(p)$ a conectividade média da rede gerada usando $\beta = p$. O algoritmo irá determinar um parâmetro $\varepsilon > 0$, e, em seguida, para valores i pertencentes ao conjunto $\{1, 2, \dots, n-2\}$ determinará as distâncias entre os valores de $R^2(p_i)$, $R^2(p_{i+1})$ e $R^2(p_{i+2})$ e os armazenará no vetor D . Se o maior valor do vetor D for menor do que ε , β receberá o valor de p em D que proporcione o maior valor de conectividade média $K(p)$. Caso esta condição nunca seja satisfeita, o programa retorna uma mensagem de erro, constando que não foi possível encontrar um valor satisfatório para o parâmetro β . O usuário é então orientado a avaliar diversos parâmetros de seus dados utilizando a função *diagnostic_report* (ver seção “4.4. - Relatório de diagnóstico”). O pseudocódigo do algoritmo encontra-se disposto na Figura 9.

```

#' Este programa calcula o melhor beta para ser utilizado para transformar a
#' matriz de adjacencia a partir de uma matriz contendo valores de beta x R2.
#'
function get_cemitoool_r2_beta(objeto_cem, eps){
  fit_indices recebe a matriz de valores beta x R2
  se(fit_indices vazia){
    entao levantar erro
  }
  inicializar com coluna contendo valores de beta
  inicializar fit com -1 * coluna do angulo da reta * R2
  inicializar st com vetor (NA, NA) para armazenar resultado final (R2, beta)

  para(cada valor de r2){
    se(o valor de fit correspondente ao r2 >= 0.8){
      d (vetor de diferencas) recebe um vetor de 3 elementos{
        1) absoluto do (valor de fit - o proximo valor de fit)
        2) absoluto do (valor de fit - o segundo proximo valor de fit)
        3) absoluto do (proximo valor de fit - o segundo proximo valor de fit)
      }
      se(o maior valor de d < eps){
        st recebe (valor de beta do r2, valor do r2)
      }
    }
  }
}

```

Figura 9 - Pseudo-código referente ao algoritmo de definição automática do valor de beta. O algoritmo itera em cima de todos os valores de beta, calcula a diferença entre cada conjunto de três valores de R2, e verifica o menor valor de beta acima de 0.8 que mantém os subsequentes valores de beta dentro da faixa epsilon pré-especificada.

3.2.3.3. Definição de módulos

A partir da determinação do valor de β a ser utilizado, o CEMiTool calculará então a matriz de adjacência (A_{ij}) através da função $a_{ij} = s_{ij}^{\beta}$. Esta matriz codifica a força de conexão entre cada par de nós da rede, determinando todas as ligações e conectando os elementos do grafo, constituindo em uma descrição completa do grafo (Zenil *et al.*, 2016). Em redes ponderadas, seus valores variam entre 0 e 1.

Dada a matriz de adjacência A , o CEMiTool inicia a etapa de determinação de módulos da rede. Através da função *TOMsimilarity* do pacote WGCNA para R, o programa CEMiTool transforma a matriz de adjacência em uma Matriz de Sobreposição Topológica (TOM, do inglês, *Topological Overlap Matrix*). A matriz TOM ($\Omega = [\omega_{ij}]$) é uma medida de similaridade, e reflete a interconectividade relativa entre dois nós, podendo ser utilizada para assim definir dissimilaridades definidas como $1 - \omega_{ij}$. Um valor de $\omega_{ij} = 1$ implica que os nós i e j estão conectados aos mesmos nós, enquanto que um valor $= 0$ indica que i e j não possuem conexões comuns entre si (Ravasz *et al.*, 2002).

Para separar os genes em módulos, o programa utiliza clusterização hierárquica de conexão com base em médias em conjunto com a dissimilaridade baseada em TOM, utilizando a função *hclust* com a medida $1 - \omega_{ij}$ como critério de distância. Os galhos resultantes deste dendrograma são então os módulos, isto é, conjuntos de genes altamente coexpressos (Zhang and Horvath, 2005), separados utilizando a função *cutreeDynamic* (Langfelder *et al.*, 2008).

3.2.4. Análises adicionais

O CEMiTool possui algumas funcionalidades adicionais, a depender dos arquivos providos pelo usuário. Entre elas estão a análise GSEA, a qual é utilizada para identificar se genes de um módulo podem ser associados coletivamente com condições biológicas distintas (Mishra *et al.*, 2014), descritas no arquivo opcional de anotação das amostras.

Além disso, caso o usuário inclua um arquivo de *pathways* de genes, o CEMiTool realizará uma análise de enriquecimento funcional, isto é, uma análise estatística da fração de genes encontrada em um dado *pathway*, dado o conjunto inicial de genes analisados (Khatri *et al.*, 2012). Finalmente, caso o usuário proporcione um arquivo de interações (por exemplo, interações proteína-proteína), o CEMiTool é capaz de levar estas informações em consideração ao executar a análise, inserindo-as no desenho das redes dos módulos.

3.2.5. Criação de relatório automatizado

Terminadas as etapas de análise do CEMiTool, o usuário pode utilizar a função *generate_report* do pacote para criar automaticamente um relatório interativo em formato HTML, ou um relatório estático em formato PDF contendo todos os resultados das análises descritas acima.

3.2.6. Relatório de diagnóstico

Durante a etapa de análise, é possível que o CEMiTool não encontre um valor razoável para o parâmetro β . Nestes casos, pode ser utilizada a função *diagnostic_report*, a qual proporciona diversos gráficos para que se tente aferir se existe algum problema subjacente aos dados.

3.3. webCEMiTool

A fim de tornar as análises do CEMiTool acessíveis para todos os tipos de usuário, nós decidimos implementar uma versão *online* da ferramenta (Cardozo *et al.*, 2019). Esta plataforma permite aos usuários executar as mesmas análises disponibilizadas pelo pacote em R do CEMiTool, com a vantagem de não ser necessário possuir conhecimentos em programação. Por outro lado, a fim de facilitar ao máximo o uso da ferramenta *online*, o webCEMiTool não permite aos usuários a mesma flexibilidade que a versão para R, tendo alguns parâmetros engessados, como a seleção do sinal da rede, um valor de beta a ser utilizado ao invés daquele automaticamente encontrado, valores de corte de p-valores, entre outros.

3.4. Banco de dados CEMiToolDB

Para a criação do banco de dados CEMiToolDB, foram baixados todos os 5080 estudos pertencentes à plataforma GPL570 do banco de dados GEO, já normalizados. A expressão de cada gene foi considerada como sendo equivalente à da sua sonda de maior expressão média. Foram mantidos apenas os estudos que tivessem entre 18 e 300 amostras e que fossem pertencentes a apenas uma série. Os estudos foram então anotados automaticamente utilizando a ferramenta annotator (seção 3.5, “Pacote annotator”). Finalmente, foram mantidos somente os estudos nos quais foi possível obter um valor aceitável para o parâmetro beta na análise pelo CEMiTool, resultando em 1001 estudos finais. Este *workflow* está disposto na Figura 10.

Figura 10 - Workflow utilizado para popular o banco de dados CEMiToolDB. Os dados foram baixados, filtrados baseado no número de amostras, processados para transformar de sondas de *microarray* para símbolos de genes, anotados automaticamente com a ferramenta *annotator*, analisados segundo o CEMiTool, e os outputs foram processados e então incluídos no banco de dados.

O banco de dados CEMiToolDB foi desenvolvido utilizando o sistema de gerenciamento de banco de dados MySQL. Para a criação do banco foi utilizada uma interface com a linguagem de programação R através do pacote RMySQL (Ooms *et al.*, 2018) a fim de automaticamente popular o banco de dados à medida que as análises com o CEMiTool foram sendo efetuadas. Na Figura 11 está disposto o esquema do banco de dados.

Foram estabelecidas 9 tabelas para o banco de dados, *study_info*, *cemitool_run*, *sample_annot*, *keywords*, *modules*, *ora*, *module_genes*, *interactions* e *enrichment*. As tabelas *study_info*, *keywords* e *sample_annot* armazenam informações a respeito dos dados de cada estudo, coletados através do pacote GEOquery para R (Davis and

Meltzer, 2007). A tabela *cemitoool_run* é utilizada para agregar informações a respeito de cada uso da ferramenta CEMiTool, e informações dos módulos obtidos são armazenados na tabela *modules*. Finalmente, as tabelas *ora*, *module_genes*, *enrichment* e *interactions* contêm informações a respeito dos resultados das análises *downstream* do CEMiTool.

Figura 11 - Esquema do banco de dados CEMiToolDB. Cada tabela armazenará informações a respeito das análises do CEMiTool, sendo cada uma particular para um determinado módulo, específico para cada rodada do CEMiTool, as quais por sua vez são específicas para um determinado estudo. Imagem criada com a ferramenta dbDesigner (www.dbdesigner.net).

3.5. Pacote anotador

Durante o desenvolvimento do banco de dados CEMiTool, verificou-se que um grande entrave para a automatização do processo é o fato de os estudos de transcriptômica em geral não possuírem uma anotação consistente e significativa das amostras. Esta informação é utilizada pelo CEMiTool em análises comparativas entre grupos experimentais, como o enriquecimento dos módulos entre classes. Assim, a fim de evitar a necessidade de anotar manualmente centenas de estudos, buscamos criar uma ferramenta, o pacote anotador, baseada em similaridade de cadeias de caracteres para automaticamente agrupar amostras dos mesmos grupos experimentais baseado apenas nas informações presentes sobre as mesmas no banco de dados GEO.

Previamente, outros autores (Wang *et al.*, 2010) buscaram utilizar o algoritmo de Needleman-Wunsch para avaliar a similaridade entre os nomes de linhagens celulares, mas o escopo de utilização se limitou a corrigir inconsistências de anotação, não uma tentativa de identificar grupos de amostras. Outras abordagens de realizar a anotação das amostras de estudos do GEO foram baseadas em *crowdsourcing* (Wang *et al.*, 2016; Hadley *et al.*, 2017) (Shah *et al.*, 2016). No entanto, o sucesso destas investidas dependem do engajamento de participantes, aperfeiçoamento de algoritmos de aprendizado de máquina e conjuntos de treinamento criados de maneira supervisionada para serem efetivos (Ellis *et al.*, 2018).

O *workflow* do anotador encontra-se disposto na Figura 12. Dado um arquivo de anotação de amostras, o pacote anotador primeiramente extrai as colunas

Sample_title, *Sample_source_name* e *Sample_characteristics_ch1*, uma vez que verificou-se empiricamente que as informações de grupo nos estudos tendem a se encontrar em pelo menos uma destas colunas. O pacote então verifica se em alguma das colunas existem elementos invariantes, isto é, se todos os elementos da coluna são iguais, e se forem, a coluna é excluída.

As informações destas três colunas são então concatenadas em uma quarta coluna, e mede-se a similaridade entre as cadeias de caracteres obtidas. Em seguida é realizada uma análise de agrupamento utilizando uma versão modificada da implementação do algoritmo HDBSCAN (Campello *et al.*, 2013) presente no pacote *dbscan* para R, utilizando 2 como o parâmetro de número mínimo de elementos por *cluster*, observando que, em geral, estudos científicos buscam realizar experimentos no mínimo em duplicata. Onze métricas de distância foram testadas com o agrupamento: Jaro (Jaro, 1989), Smith-Waterman (Smith and Waterman, 1981), Needleman-Wunsch (Needleman and Wunsch, 1970), Monge-Elkan (Monge and Elkan, 1997), Levenshtein (Levenshtein, 1966), Damerau-Levenshtein (Damerau, 1964), Jaccard (Jaccard, 1901), *Optimal String Alignment* (distância Damerau-Levenshtein restringida), *Longest Common Substring*, distância q-gram e distância de cossenos.

Para as distâncias de Jaro, Levenshtein, Damerau-Levenshtein, Jaccard, *Optimal String Alignment*, *Longest Common Substring*, q-gram e de cossenos, foi utilizada a implementação destes algoritmos no pacote *stringdist* em R (van der Loo, 2014). Para o algoritmo de Monge-Elkan foi implementado um script customizado na linguagem R. Finalmente, para os algoritmos de Smith-Waterman e

Needleman-Wunsch, foram utilizadas respectivamente as implementações presentes nos pacotes *textreuse* (Mullen, 2015) e *NameNeedle* (Wang *et al.*, 2010). Uma vez que estes algoritmos tradicionalmente não representam uma distância no sentido matemático formal, foi implementada uma transformação (Jelinek, 2007) a fim de garantir que as mesmas mantenham as propriedades esperadas de uma métrica.

Finalmente, o pacote automaticamente define um nome para o grupo de amostras utilizando as palavras repetidas em todos os elementos do grupo e excluindo pontuação, caracteres especiais e números. Nós rodamos o pacote em 100 estudos da plataforma GPL570 do banco de dados GEO anotados manualmente a fim de avaliar a capacidade do pacote em corretamente atribuir as amostras a grupos biologicamente relevantes.

Figura 12 - Workflow do pacote anotator. (1) Dado um arquivo de anotação de amostras, o pacote extrai as colunas “Sample_title”, “Sample_source_name_ch1” e “Sample_characteristics_ch1”. Colunas com elementos não-informativos são removidas, e as restantes, concatenadas. (2) As *strings* obtidas são agrupadas utilizando o algoritmo HDBSCAN, o qual define um número para cada *cluster*. (3) É definido um rótulo para o grupo utilizando as palavras recorrentes em todos os elementos do grupo.

4. Resultados

4.1. Estatísticas de *downloads* do CEMiTool e boas práticas de engenharia de *software*

A Figura 13 ilustra o número de vezes que foi feito o download do pacote CEMiTool desde o seu lançamento no projeto Bioconductor em agosto de 2017. Até o momento de confecção da presente tese (março de 2019), no total, foram 4997 *downloads*, sendo 2382 feitos por IPs distintos. Percebe-se que o CEMiTool tem encontrado um número crescente de usuários, e está em fase de disseminação e aquisição de novos usuários.

Figura 13 - Números de *downloads* do pacote CEMiTool. Em rosa, está disposto o número de *downloads* da ferramenta feito por IPs distintos, e em azul, o número de *downloads* total. É possível perceber o crescimento progressivo do número de *downloads*.

Durante o desenvolvimento do CEMiTool, buscou-se implementar o pacote seguindo uma série de diretrizes de boas práticas engenharia de *software*. Na presente data (maio de 2019), o CEMiTool:

- Implementa 106 testes unitários, que testam o código a cada nova modificação;
- Implementa integração contínua (CI, do inglês, *continuous integration*) através da plataforma Travis-CI, o qual executa os testes unitários acima citados;
- Apresenta documentação extensa, contando com 66 itens documentados, entre funções, datasets de exemplo e objetos;

- Possui um manual introdutório detalhado para auxiliar novos usuários (<https://github.com/csbl-usp/CEMiTool/blob/master/vignettes/CEMiTool.Rmd>)
- Utiliza objetos S4, conforme diretrizes estabelecidas pelo projeto Bioconductor (<https://www.bioconductor.org/developers/how-to/efficient-code/#essential-s4-interface>)

4.2. Comparação empírica dos métodos de seleção do parâmetro β em 15 estudos de *microarray*

A fim de comparar os métodos de seleção do parâmetro β , rodamos o CEMiTool e o WGCNA em 15 estudos públicos de *microarray* (Figura 14). Utilizando os mesmos genes de entrada, utilizamos 3 métodos para a seleção de β : a função *pickSoftThreshold* do pacote WGCNA com valores de $R^2 > 0.8$ e > 0.85 , assim como a função *cemitool* do CEMiTool utilizando $R^2 > 0.8$. Com a exceção do estudo GSE53441, o valor de β retornado pelo CEMiTool foi consistentemente menor ou igual àquele retornado pelo WGCNA.

Ressalta-se que a escolha do parâmetro β afeta não somente a topologia da rede, mas também seu conteúdo de informação: quanto maior o valor de β , menor será a sua conectividade média, uma vez que as intensidades das conexões na matriz de adjacência são limitadas por $[0,1]$ (Horvath, 2011). Consequentemente, um balanço entre a conectividade média da rede e a adequação ao modelo livre de escala deve ser considerado. Desta maneira, no contexto do presente trabalho, valores mais baixos de

β serão considerados de maior interesse que valores mais altos, contanto que seus valores de R^2 sejam similares.

As diferenças entre o WGCNA e o CEMiTool na escolha do parâmetro β podem ser, em sua maior parte, explicadas pelo menor valor de corte para R^2 implementado no CEMiTool (0.8 no CEMiTool contra 0.85 no WGCNA). Esta escolha foi feita baseando-se na opção original dos autores do WGCNA (Zhang and Horvath, 2005). Além disso, o CEMiTool utiliza um algoritmo estrigente, baseado em sequências de Cauchy, a fim de selecionar o menor valor de β que estabiliza a sequência (isto é, mantém os valores de R^2 dentro de uma faixa pré-definida de valores) e que ainda se encontre acima do valor de corte. Quando o mesmo valor de corte (0.8) é aplicado, o CEMiTool geralmente encontra o mesmo valor para o parâmetro β que a função *pickSoftThreshold* do WGCNA. No entanto, existem vários casos em que o WGCNA retorna um valor inadequado de $\beta = 1$. Maiores detalhes na comparação dos dois pacotes podem ser encontrados na seção Métodos Complementares.

Figura 14. Seleção automática do parâmetro beta. (a) Parâmetros beta escolhidos pelo WGCNA (barras vermelhas e laranjas) ou pelo CEMiTool (barras pretas) em 15 estudos de *microarray* utilizando os mesmos genes de entrada. (b) Curva $\beta \times R^2$ para 3 estudos representativos. Valores de β escolhidos pelo WGCNA (linhas vermelhas) e pelo CEMiTool (linha preta) (Russo *et al.*, 2018).

4.3. Determinação dos parâmetros de filtragem

A fim de estabelecer os melhores parâmetros para o filtro do CEMiTool, nós executamos as análises do pacote em 300 estudos de *microarray* obtidos do banco de dados *Gene Expression Omnibus* (GEO, <https://www.ncbi.nlm.nih.gov/gds>) utilizando

diferentes valores de corte para o p-valor e avaliamos o significado biológico dos módulos resultantes. Para tal, calculamos o escore de enriquecimento combinado (*Combined Enrichment Score*, CES) dos módulos utilizando os dados dos *pathways* do banco de dados Reactome. Resumidamente, o CES permite que se condense os resultados de enriquecimento globais em um único número; quanto menor este número, mais enriquecidos serão estes módulos. À medida que o p-valor de corte para o filtro aumenta de 0.05 para 0.3, o CES atinge um mínimo global em $p \approx 0.1$, sugerindo que o ruído introduzido por genes não-correlacionados sobrepuja o ganho em informação (Figura 15). Assim, o p-valor do filtro foi definido como 0.1 por padrão.

Figura 15 - Filtragem automática dos genes do CEMiTool. O efeito do p-valor de corte no enriquecimento dos módulos foi analisado rodando o CEMiTool em 300 estudos. (a) O *Combined Enrichment Score* (CES) dos módulos resultantes para cada estudo e para cada p-valor de filtro foi calculado utilizando *pathways* do Reactome. (b) O número de genes escolhido para cada p-valor do filtro. Em ambos os painéis, a linha preta representa o CES médio de todos os 300 estudos, enquanto a região verde representa o intervalo de confiança de 95% da média (Russo *et al.*, 2018).

4.4. Geração de report automatizado

Após executar as análises, o pacote CEMiTool proporciona aos usuários a opção de gerar um relatório em formato HTML ou PDF. Em particular, o formato HTML permite aos usuários visualizar os resultados de forma interativa, com tabelas dinâmicas e os resultados dispostos em forma de abas. Este relatório também está disponível para usuários da versão *online* do CEMiTool. Conforme mostrado na Figura 16, cada uma das etapas de análise encontra-se disposta em uma das abas do relatório HTML, sendo que a primeira aba, “Modules”, contém informações sobre o número de genes presente em cada módulo e quais os genes mais conectados neles (*hubs*).

Figura 16 - Aba “Modules” do relatório do CEMiTool. Nesta aba encontram-se dispostas informações sobre cada módulo de coexpressão, seu número de genes e os cinco genes mais conectados (*hubs*).

Na Figura 17 é possível visualizar a aba “Profile Plot” do relatório do CEMiTool, na qual é possível ver a variação dos valores de expressão dos genes de cada módulo, assim como uma linha preta indicando alguma medida de centralidade da expressão, podendo ser a média ou a mediana. Além disso, caso o usuário tenha proporcionado um arquivo de anotação de amostras com informações sobre grupos experimentais, os mesmos serão indicados pela coloração do fundo do gráfico.

Figura 17 - Aba “Profile Plot” do relatório do CEMiTool. Nesta aba constam gráficos com os perfis de expressão de cada gene pertencente aos módulos (em verde) e uma

linha com uma medida de centralidade da expressão geral do módulo (em preto, a média da expressão).

Os resultados da análise de GSEA podem ser encontrados na aba “Gene Set Enrichment Analysis”, conforme disposto na Figura 18. Nesta figura, enriquecimento significativo de um determinado módulo para algum grupo experimental está disposto como uma circunferência cujo diâmetro e cor refletem a magnitude do valor do enriquecimento. Este valor é representado pelo escore de enriquecimento normalizado (NES, do inglês, *Normalized Enrichment Score*), que representa o valor normalizado do quanto um dado conjunto de genes está sobrerepresentado na parte superior ou inferior de uma lista ranqueada de genes, casos nos quais o conjunto será considerado enriquecido. Por outro lado, o conjunto de genes não será considerado enriquecido se seus elementos tenderem a se encontrar distribuídos aleatoriamente pela lista de genes. Em casos de módulos que não se encontram enriquecidos para nenhum grupo experimental, o pacote os remove da figura. Os valores de NES e o p-valor de significância ajustado (Benjamini-Hochberg) da análise também são disponibilizados pelo CEMiTool, encontrando-se na parte inferior da aba.

Report

Figura 18 - Aba “Gene Set Enrichment Analysis”. Enriquecimento das listas de genes dos módulos encontrados pelo CEMiTool. O diâmetro e as cores das circunferências refletem os valores de NES do GSEA. O p-valor ajustado

(Benjamini-Hochberg) da análise e do NES de cada módulo em cada grupo experimental encontram-se dispostos na parte inferior da aba.

Na aba “Over Representation Analysis”, encontra-se o resultado da análise funcional dos módulos (Figura 19). Nesta etapa é utilizado o teste hipergeométrico a fim de se determinar o quanto os genes de cada módulo estão enriquecido em uma dada lista de *pathways*. O valor em log10 ajustado (Benjamini-Hochberg) do p-valor dos 10 *pathways* mais enriquecidos pela análise são dispostos em um gráfico de barras. Além disso, uma linha tracejada cinza atravessa o eixo x do gráfico a fim de ilustrar um valor sugerido de corte de significância (p-valor 0,05). Na parte inferior da aba encontram-se outras informações da análise, como o número de genes do módulo em cada *pathway* (Count), a fração dos genes do *pathway* no número total de genes enriquecidos no módulo (GeneRatio), a fração dos genes do *pathway* no número total de genes do arquivo de *pathways* (BgRatio) e o p-valor ajustado da análise (p.adjust).

Report

Modules Profile Plot Gene Set Enrichment Analysis Over Representation Analysis Interaction Network Parameters

M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12

Show 5 entries

Search:

ID	Count	GeneRatio	BgRatio	p.adjust
Cell Cycle, Mitotic	103	103/280	445/7644	0
Mitotic Prometaphase	41	41/280	108/7644	0
Resolution of Sister Chromatid Cohesion	36	36/280	100/7644	0
M Phase	52	52/280	266/7644	0
Mitotic G1-G1/S phases	35	35/280	135/7644	0

Showing 1 to 5 of 50 entries

Previous

1

2

3

4

5

...

10

Next

Figura 19 - Aba “Over Representation Analysis” do relatório do CEMiTool. O gráfico de barras indica o log₁₀ do p-valor da análise de enriquecimento dos genes do módulo para cada *pathway*. A linha tracejada cinza indica o corte de significância (p-valor 0,05). A tabela na parte inferior ilustra outros resultados da análise (ver texto).

A aba “Interaction Network” ilustra a integração das redes de coexpressão com dados de interactoma (Figura 20). Nesta figura, o grau de conexão dos genes, isto é, o número de genes com os quais interagem, está indicado pelo diâmetro dos pontos. A figura também ilustra os 10 genes mais conectados do módulo (*hubs*), colorindo-os diferencialmente a depender da origem da informação do gene.

Figura 20 - Aba “Interaction Network” do relatório do CEMiTool. Rede de interações dos módulos encontrados pelo CEMiTool. O tamanho do nó reflete o seu grau, enquanto que os genes destacados são aqueles com maior número de conexões. As cores das caixas demonstram a origem da informação do gene: azul, se a interação veio da análise de coexpressão; vermelho se a interação veio puramente do arquivo de interações inserido; verde se a interação consta em ambos.

Finalmente, a última aba, “Parameters” está reservada para disponibilizar ao usuário alguns dos parâmetros de análise, como o valor escolhido para o parâmetro β , o p-valor de corte escolhido para o método de filtragem, entre outros. Um exemplo da aba “Parameters” está disposto na Figura 21.

CEMiTool Code ▾

Report

Modules Profile Plot Gene Set Enrichment Analysis Over Representation Analysis Interaction Network **Parameters**

Show entries Search:

Parameter	Description	Value
beta	Value of beta chosen.	6
cor_method	Correlation method.	pearson
diss_thresh	Dissimilarity threshold to be used as cutoff on hierarchical clustering.	0.8
filter_pval	p-value used on filtering.	0.1
merge_similar	Should similar modules be merged?	TRUE
min_ngen	Minimum number of genes per module.	30
n_genes	Number of genes after filtering.	2129
n_mods	Number of modules returned by CEMiTool.	13
phi	Area under curve / Total area in the Beta vs R squared graph.	0.708
r2	Determination coefficient. Reflects the "scale-freeness" of the resulting network.	0.823

Showing 1 to 10 of 10 entries Previous 1 Next

Figura 21 - Aba “Parameters” do relatório do CEMiTool. Diversos parâmetros da análise encontram-se dispostos na aba, tais como o valor selecionado para o parâmetro β e o p-valor de filtro escolhido.

4.5. Relatório de diagnóstico

Existem casos nos quais o CEMiTool pode não conseguir selecionar um valor adequado para o parâmetro β , o que impede a definição de módulos e, portanto, o restante das análises. Isto pode ser devido a diversos fatores, como a simples não-aderência ao modelo livre de escala (Broido and Clauset, 2018), o que impede a determinação de valores satisfatórios para R^2 , e por sua vez, β . Esta não-aderência ao modelo, por sua vez, pode ser causada por fatores que promovam altas correlações entre grupos de genes, como efeitos de lote, alterações muito bruscas entre grupos de amostras (por exemplo em estudos temporais), ou heterogeneidade biológica muito elevada (por exemplo, grupos de amostras de tecidos diferentes) (Langfelder, 2017).

A fim de auxiliar o usuário nestes casos, foi implementada a função *diagnostic_report* no pacote CEMiTool, a qual proporciona ferramentas para visualmente se aferir potenciais problemas nos dados e eventualmente corrigi-los. A estrutura do relatório diagnóstico é semelhante ao relatório de resultados do CEMiTool, sendo disposto na forma de abas. Na Figura 22 é possível ver a primeira aba, “Sample clustering”, a qual realizará um agrupamento hierárquico das amostras baseado naquelas que possuírem maiores correlações entre os valores de expressão de cada gene. Esta também é uma forma de se visualizar potenciais amostras *outliers*, e ponderar se as mesmas devem ou não ser removidas. Além disso, caso o usuário tenha inserido um arquivo de anotação das amostras, o CEMiTool colorirá a área embaixo de cada ramo do dendrograma com cores representando o grupo

experimental de cada amostra. Assim, é possível visualizar se existe um grupo de amostras com maior grau de agregação (mais correlacionadas), o que invalidaria a hipótese de topologia livre de escala (Langfelder and Horvath, 2014).

CEMiTool diagnostics

[Code ▾](#)

Diagnostics

[Sample clustering](#)[Mean Variance](#)[Quantile-quantile plot](#)[Expression histogram](#)[Beta x R2 plot](#)[Mean connectivity plot](#)

Figura 22 - Aba “Sample Clustering” do relatório de diagnóstico do CEMiTool.

Dendrograma resultante da análise de agrupamento hierárquico aplicado às amostras do conjunto de dados inserido. As amostras são baseadas na correlação entre os valores de expressão de cada gene. Dado um arquivo de anotação das amostras, o

CEMiTool colorirá a área embaixo de cada amostra baseado no grupo experimental a que cada amostra pertence.

A segunda aba do relatório de diagnóstico, “Mean Variance” possui um diagrama de pontos no qual a expressão média e a variância dos genes são plotadas em escala log10 (Figura 23), e uma regressão linear é realizada a fim de verificar a relação entre as grandezas. Em particular, para dados de RNAseq, caso seja encontrada dependência entre a média e a variância (indicada pelo coeficiente de determinação R^2 , também incluído no gráfico junto à equação da reta), recomenda-se utilizar o argumento `apply_vst` da função *cemitoool* com o valor “TRUE”, a fim de aplicar uma transformação estabilizadora da variância (VST, do inglês, *Variance Stabilizing Transformation* (Yu, 2009)) e remover esta dependência (Subramaniam and Hsiao, 2012; Robinson *et al.*, 2010).

CEMiTool diagnostics

[Code ▾](#)

Diagnostics

[Sample clustering](#)[Mean Variance](#)[Quantile-quantile plot](#)[Expression histogram](#)[Beta x R2 plot](#)[Mean connectivity plot](#)

Figura 23 - Aba “Mean Variance” do relatório diagnóstico do CEMiTool. Diagrama de pontos entre a média e a variância da expressão em log10 dos genes de entrada do CEMiTool. Recomenda-se aplicar VST quando houver uma dependência linear elevada entre as duas grandezas.

Nas Figuras 24 e 25 estão dispostos, respectivamente, um gráfico Q-Q e um histograma da expressão dos genes. Estes gráficos podem ser utilizados para se visualmente ter uma idéia de se os dados possivelmente seguem uma distribuição teórica como a normal. Esta informação pode então ser utilizada para a seleção

apropriada do coeficiente de correlação (Pearson ou Spearman) para a análise do CEMiTool, que pode ser escolhido com o parâmetro `cor_method` da função `cemitoool`, sendo que o coeficiente de Pearson é o mais recomendado caso os dados sigam uma distribuição normal devido ao maior poder estatístico; caso contrário, a correlação de Spearman pode ser a melhor opção (Petereit *et al.*, 2016).

CEMiTool diagnostics

[Code ▾](#)

Diagnostics

[Sample clustering](#)[Mean Variance](#)[Quantile-quantile plot](#)[Expression histogram](#)[Beta x R2 plot](#)[Mean connectivity plot](#)

Figura 24 - Aba “Quantile-quantile plot” do relatório diagnóstico do CEMiTool.

Valores de expressão dos dados de entrada são comparados com dados de uma distribuição normal. Caso os pontos se aproximem de uma linha reta, a distribuição pode ser considerada aproximadamente normal.

Diagnostics

[Sample clustering](#)[Mean Variance](#)[Quantile-quantile plot](#)[Expression histogram](#)[Beta x R2 plot](#)[Mean connectivity plot](#)

Figura 25 - Aba “Expression histogram” do relatório diagnóstico do CEMiTool.

Histograma ilustrando a distribuição dos valores de expressão inseridos no CEMiTool a fim de visualmente se aferir se a mesma se assemelha a uma distribuição normal.

Finalmente, o CEMiTool também proporciona aos usuários o gráfico $R^2 \times \beta$, utilizado na etapa de seleção do parâmetro β e um gráfico ilustrando a conectividade média da rede em função do parâmetro β (Figuras 26 e 27, respectivamente). Verificando-se o gráfico $R^2 \times \beta$, é possível ter uma idéia do quão bem os dados se encaixam em uma topologia livre de escala. Já o gráfico de conectividade média da

rede ilustra o balanço que deve ser feito entre valores selecionados de β e a conectividade resultante da rede.

CEMiTool diagnostics

[Code ▾](#)

Diagnostics

[Sample clustering](#)[Mean Variance](#)[Quantile-quantile plot](#)[Expression histogram](#)[Beta x R2 plot](#)[Mean connectivity plot](#)

Figura 26 - Aba “ β x R2 plot” do relatório diagnóstico do CEMiTool. Gráfico ilustrando a relação entre o valor de β escolhido e a adequação da rede ao modelo de topologia livre de escala. O valor de β selecionado pelo CEMiTool é mostrado em vermelho.

Diagnostics

[Sample clustering](#)[Mean Variance](#)[Quantile-quantile plot](#)[Expression histogram](#)[Beta x R2 plot](#)[Mean connectivity plot](#)

Figura 27 - Aba “Mean connectivity plot” do relatório diagnóstico do CEMiTool.

O gráfico ilustra claramente o compromisso entre a conectividade média da rede e valores do parâmetro β . Enquanto valores altos de β tendem a representar maior aderência ao modelo de topologia livre de escala, os mesmos tendem a gerar redes esparsas (conectividade média baixa).

4.6. webCEMiTool

A plataforma online webCEMiTool foi criada a fim de permitir que usuários com absolutamente nenhuma experiência com programação possam utilizar as análises

proporcionadas pelo pacote CEMiTool. A página pode ser acessada através da URL <http://cemitool.sysbio.tools/>. A *homepage* da plataforma está disposta na Figura 28.

Figura 28 - *Homepage* da plataforma webCEMiTool. A partir dessa página o usuário pode iniciar as análises, visualizar o tutorial para uso da ferramenta ou ver informações sobre a ferramenta.

A Figura 29 mostra a página para realizar o *upload* de arquivos. Nela, usuários podem inserir um arquivo de expressão segundo o modelo disposto, assim como os arquivos opcionais de anotação de amostras, interações entre genes e de *pathways*. Diversos parâmetros podem ser configurados opcionalmente ao clicar na aba

“Parameters”, como por exemplo o nome a ser dado à análise, p-valor de corte do filtro, o uso de VST, entre outros. Um e-mail também pode ser inserido para que o usuário receba uma notificação quando a análise terminar.

Upload your files

Expression file
.txt/.csv - example

Drag-and-drop or click here to upload your Expression file

genes	sample1	sample2	sample3
IL2	10.883	9.723	10.513
BRCA1	7.432	6.231	7.502
CXCR4	11.432	12.894	9.500
ALB	6.013	8.875	6.970
EPGF	7.543	6.123	10.513
ZNF	13.854	14.700	10.002
... (more)	...	...	...

Choose File | No file chosen

Optional files

Sample phenotypes (.txt/.csv - example)

Choose File

No file chosen

Gene interactions (.txt/.csv - example)

Choose File

No file chosen

Gene sets (.gmt - example)

Choose File

No file chosen

Parameters + Click here to [change default parameters](#)

Email

Email for sending results (optional)

Run CEMiTool

Figura 29 - Página de *upload* de arquivos da plataforma webCEMiTool. Nesta página podem ser inseridos o arquivo de expressão de genes, assim como os arquivos opcionais. Diversos parâmetros podem ser customizados, e o usuário também pode inserir um e-mail para receber uma notificação quando a análise terminar.

Terminadas as análises, o webCEMiTool resume os principais resultados na página mostrada nas Figuras 30 a 32. Nesta página constam informações sobre o número de genes selecionados após realizar a filtragem do webCEMiTool, o número de módulos encontrados, um gráfico de barras com o número de genes encontrados por

módulo e um link para *download* do relatório de resultados descrito na seção 4.3. “Geração de report automatizado” (Figura 30). Além disso, como é possível ver na Figura 31, a página também mostra informações sobre a mediana da atividade dos genes de cada módulo, o número de genes em cada módulo, os principais genes *hubs*, assim como um *link* para realizar o enriquecimento funcional dos genes do módulo na plataforma Enrichr (Kuleshov *et al.*, 2016; Chen *et al.*, 2013). Nota-se que, diferentemente da ferramenta CEMiTool para R, o webCEMiTool permite realizar o enriquecimento dos módulos comparando com todo o banco de dados do Enrichr, e não apenas com um arquivo gmt de entrada.

Finalmente, a página ainda mostra o resultado da análise GSEA em cima dos módulos, ilustrando em um *heatmap* o valor NES do enriquecimento de cada módulo em cada grupo experimental. A análise é rodada utilizando os genes do módulo como um *gene set* e utilizando a mediana dos valores de z-score de cada grupo experimental como *rank* (Figura 32).

Figura 30 - Seção “Summary” da página de resultados do webCEMiTool. Nesta seção estão dispostas informações sobre o número de genes filtrados, o número de

módulos obtidos e o número de genes em cada módulo, assim como um link para fazer o *download* do relatório completo da análise.

Figura 31 - Seção “Modules” da página de resultados do webCEMiTool. A seção detalha informações sobre cada módulo encontrado, em particular a mediana da atividade dos genes pertencendo a cada módulo, o número de genes presentes neles, os principais *hubs* e um link para se realizar o enriquecimento dos genes do módulo na plataforma Enrichr.

Figura 32 - Seção “Gene Set Enrichment Analysis” da plataforma webCEMiTool. A análise mostra a atividade dos módulos encontrados nos grupos experimentais inseridos pelo usuário.

4.7. Comparação do CEMiTool com outros métodos de detecção de redes de coexpressão

Nós comparamos as funcionalidades e os métodos de detecção de redes de coexpressão do CEMiTool com dois outros pacotes para R: WGCNA e Petal (Petereit *et al.*, 2016) em dois estudos públicos previamente normalizados de *microarrays* de pacientes com Dengue (GSE18090 e GSE43777). O CEMiTool foi rodado utilizando seus parâmetros padrão e todos os arquivos opcionais. Após o filtro, as análises foram executadas em cima de 2129 genes para o estudo GSE18090 e 1765 genes para o

estudo GSE43777. Para fins de comparação foi assumido que um maior enriquecimento em *pathways* relevantes à doença analisada é um bom parâmetro de qualidade de uma análise de redes de coexpressão.

Para o estudo GSE18090, o CEMiTool selecionou um parâmetro β igual a 6 e identificou 12 módulos de coexpressão, dos quais 9 possuíam pelo menos um *pathway* significativamente enriquecido na análise ORA. Notavelmente, os módulos M4 e M6 estavam enriquecidos significativamente com *pathways* de interferon, sinalização de citocinas e mecanismos antivirais, conforme esperado de uma doença infecciosa como a Dengue. Além disso, o módulo M2 foi enriquecido significativamente para cascatas de *toll-like receptors* (TLR), capazes de induzir a liberação de citocinas pró-inflamatórias e citocinas em infecção por Dengue. Estes resultados refletem o que foi encontrado nos 7 módulos significativamente enriquecidos (de um total de 11) do estudo GSE43777 ($\beta = 5$). Rodando o CEMiTool com todos os possíveis arquivos opcionais levou em torno de 3 minutos para cada estudo (Tabela 1).

A fim de se comparar o WGCNA ao CEMiTool, rodou-se o WGCNA nos estudos de dengue utilizando os 4000 genes mais variantes de cada conjunto de dados. Este número foi escolhido com base no número de genes selecionados como exemplo no tutorial da ferramenta, uma vez que o método não possui uma sugestão para a quantidade de genes de entrada. A análise identificou 18 módulos para o estudo GSE18090 com um parâmetro β igual a 9. É interessante notar, no entanto, que mais da metade deles (10 módulos) não apresentaram nenhum *pathway* significativamente enriquecido após a análise ORA ($p\text{-valor} < 0,01$). Em contraste com

os resultados do CEMiTool, o WGCNA não reportou vias relacionadas a cascatas de TLR. Já para o estudo GSE43777 ($\beta = 6$), o WGCNA retornou 10 módulos significativamente enriquecidos, de um total de 16. Estes resultados sugerem que, apesar de rodar em um número menor de genes, o CEMiTool é capaz de filtrar com sucesso genes irrelevantes e construir módulos com os genes mais importantes para a dada condição de interesse. O nosso *script* customizado para rodar a análise WGCNA foi capaz de executar a análise em tempo similar ao do CEMiTool (em torno de 3 minutos; Tabela 1). No entanto, isso não leva em consideração o tempo necessário para se inserir manualmente todas as etapas necessárias para executar as análises por WGCNA, selecionar os parâmetros a serem definidos pelo usuário e a curva de aprendizado necessária para se compreender todo o procedimento.

A fim de levar em consideração a diferença no número de genes de entrada, executou-se novamente a análise com o WGCNA utilizando os mesmos genes retornados pelo método de filtro do CEMiTool. Para o estudo GSE18090, o WGCNA identificou 16 módulos, com um valor de β igual a 7. Destes, 10 módulos possuíam pelo menos um *pathway* significativamente enriquecido na análise ORA. Conforme esperado, os módulos se tornaram mais similares àqueles encontrados pelo CEMiTool, com a inclusão de um módulo relacionado a atividade de TLR (M2), assim como diferentes módulos para interferon tipos gamma (M4) e alfa/beta (M5). Já para o estudo GSE43777, o WGCNA (com valor de β igual a 6) foi capaz de identificar 6 módulos significativamente enriquecidos de um total de 12, o que faz com que ele tenha 2 módulos não-significativamente enriquecidos a mais que o CEMiTool. Estas diferenças

sutis são provavelmente decorrentes da diferença na escolha do valor de β , e demonstram a importância de uma escolha criteriosa do parâmetro, assim como a capacidade do CEMiTool de obter resultados comparáveis ao WGCNA de maneira mais facilitada e conveniente.

Utilizando os mesmos genes de entrada que a primeira análise com WGCNA (4000 genes mais variantes), fizemos a mesma análise utilizando desta vez a ferramenta Petal. O Petal é um pacote para R que busca definir uma rede de coexpressão gênica utilizando um valor de corte automaticamente definido para indicar expressão similar entre genes. No entanto, após 20 minutos para o estudo GSE18090 e 40 minutos para o estudo GSE43777, o programa foi incapaz de selecionar qualquer valor de corte. Isto ocorreu novamente quando os datasets resultantes do filtro do CEMiTool, embora o tempo de execução tenha diminuído (9 minutos para o estudo GSE18090 e 4 minutos para o estudo GSE43777). Outros problemas da ferramenta Petal incluem um *output* confuso na linha de comando, nenhum resultado gráfico da análise ou análises complementares, geração de diversos arquivos intermediários sem a opção de redirecionamento dos mesmos, ausência de tutorial ou manual de instruções, esquemas de nomeação inconsistentes, entre outros.

Outros pacotes, como CoXpress, DINGO e DiffCoEx não foram considerados na comparação, uma vez que eles não são capazes de analisar mais de 2 grupos de amostras (Tabela 1).

Tabela 1 - Funcionalidades de programas de identificação de redes de coexpressão. *Over Representation Analysis* dos programas CoP e GeNET foi considerada “limited” devido ao fato de estes programas permitirem apenas o uso de *gene sets* específicos (GO, Pfam ou KEGG). O tempo de execução de 2 estudos utilizando o mesmo computador e parâmetros padrão está ilustrado para as ferramentas CEMiTool, WGCNA e Petal.

Features	CEMiTool	WGCNA	Petal	CoP	GeNET	DiffCoEx	CoXpress	DICER	DINGO
Automatic Gene Filtering	yes	no	no	no	no	no	no	no	no
Over representation analysis is	yes	yes	no	limited	limited	no	no	no	no
Gene set enrichment analysis is	yes	no	no	no	yes	no	no	no	no
Integration with interactome	yes	no	no	no	no	no	no	no	no
Report in HTML	yes	no	no	yes	no	no	no	no	no
Native plots	yes	yes	no	no	yes	no	no	no	no
Search for a gene or gene list	yes	no	yes	yes	yes	no	no	no	no
Merging modules	yes	yes	no	no	no	yes	no	yes	no
Allows 2+ sample groups	yes	yes	yes	yes	yes	no	no	no	no
R package	yes	yes	yes	no	no	no	yes	no	yes
Year of last update	2017	2017	2017	2010	Unknown	Unknown	2013	Unknown	Unknown
Runtime for study GSE18090	2min12s	3min10s	17min18s	-	-	-	-	-	-
Runtime for study GSE43777	3min03s	4min33s	40min10s	-	-	-	-	-	-

4.8. Aplicação do CEMiTool a dados reais

4.8.1. Dados de dengue

Aproveitando dados de pacientes com dengue para a comparação apresentada na seção anterior (4.6 “Comparação do CEMiTool com outros métodos de detecção de redes de coexpressão”), apresentamos aqui a utilidade do CEMiTool para a geração potencial de novos *insights* biológicos.

Os dados dos estudos GSE18090 e GSE43777 foram anotados utilizando os dados fornecidos pelos autores originais dos estudos: controle (pacientes

não-infectados), DF (pacientes diagnosticados com febre) e DHF (pacientes diagnosticados com febre hemorrágica). Dados de interação de proteínas foram obtidos do banco de dados GeneMania (<http://genemania.org/>) e os dados de vias foram obtidos do Reactome (<http://www.reactome.org/>). Os resultados estão dispostos na Figura 33. A análise de ORA mostra que o pacote identificou módulos (M4 no estudo GSE18090 e M3 no estudo GSE43777) relacionados à imunidade antiviral, como sinalização de interferon e o mecanismo antiviral ISG15. Além disso, a análise GSEA mostra que a atividade destes módulos é maior em DF ou DHF comparado ao controle. O grafo da rede de coexpressão do estudo GSE43777 também destaca genes importantes como *hubs*, entre eles o gene CCL2 (*chemokine C-C motif ligand 2*), associado com dengue severa e síndrome do choque da dengue (Lee *et al.*, 2006). No entanto, este gene não foi destacado como gene-chave na publicação original do estudo GSE43777 (Sun *et al.*, 2013).

Figura 33 - CEMiTool aplicado a estudos de dengue. (a) Dados de expressão de dois estudos de *microarray* de pacientes infectados ou não com dengue foram baixados do banco de dados GEO. CEMiTool foi rodado nestes estudos utilizando um arquivo de anotação de amostras, um arquivo de genes de *pathways* e um arquivo de interações proteína-proteína. Resultados selecionados das 4 análises do CEMiTool estão dispostos nos painéis (b) a (e). (b) GSEA mostrando a atividade de cada módulo em cada classe de amostras. (c) Gráficos de perfil de expressão dos módulos M4 (GSE18090) e M3 (GSE43777). (d) Análise ORA dos módulos M4 (GSE18090) e M3 (GSE43777). (e) Redes gênicas dos módulos M4 (GSE18090) e M3 (GSE43777) (Russo *et al.*, 2018).

4.8.2. Dados de psoríase

O estudo GSE54456 (Tsoi *et al.*, 2015) mediu 174 transcriptomas de amostras de pacientes com psoríase e amostras controle. Entre os 8 módulos identificados pelo CEMiTool (Figura 24), o módulo M1 estava enriquecido para vias do sistema imune, incluindo sinalização de interferon alfa, sabidamente relacionado à doença (Kim *et al.*, 2013; Afshar *et al.*, 2013). Um gene *hub* de destaque para o módulo M1 foi o S100A7A. Embora este gene não tenha sido mencionado na publicação original, outros já demonstraram que a sua expressão é aumentada em pacientes de psoríase com lesão de pele (Keermann *et al.*, 2015). A análise do CEMiTool também revelou um módulo relacionado à organização da matriz extracelular e formação de colágeno (Figura 34),

sugerindo que a expressão de genes responsáveis por manter a estrutura da pele pode ser alterada pela doença.

Figura 34 - CEMiTool aplicado a um estudo de RNAseq de pacientes com psoríase. Dados de RNAseq (normalização RPKM) de pacientes de psoríase com lesão e controle foram baixados do banco de dados GEO (GSE54456). (a) Report

interativo do CEMiTool mostrando os resultados das análises principais com os arquivos opcionais de anotação de amostra, *pathways* e interações proteína-proteína. Na aba principal, os genes *hubs* para cada módulo podem ser vistos. (b) Vias significativamente enriquecidas para o módulo M2. Processos metabólicos relacionados à psoríase, como “organização da matriz extracelular”, estão enriquecidos no módulo M2 (Russo *et al.*, 2018).

4.8.3. Dados de leishmaniose visceral

Finalmente, utilizamos o CEMiTool para investigar o transcriptoma de sangue de pacientes com leishmaniose visceral (VL), um grande problema de saúde no Brasil e no mundo, tendo 3038 novos casos sido notificados em território nacional somente no ano de 2012 (Marcondes and Rossi, 2014). Para tanto, foram realizados 17 experimentos de RNAseq utilizando o sangue de 6 pacientes infectados com *Leishmania infantum*, antes e depois do tratamento, assim como 5 indivíduos saudáveis. O CEMiTool gerou 14 módulos, contendo 1700 genes no total (Figura 35). Destes, os módulos M7 e M10 se referem a respostas imunes mediadas por interferon (IFN γ e IFN I, respectivamente), que são bem descritos como estando envolvidos em modelos experimentais da doença (Varma and Naseem, 2010) (Figura 35b). No entanto, embora respostas de IFN γ (M7) já tenham sido descritas em estudos clínicos, pouco se sabe sobre o papel de IFN tipo I (M10) em pacientes. De fato, genes de IFN tipo I são geralmente evocados em infecções virais, e não parasitoses como a leishmaniose. Assim, experimentos futuros devem ser executados a fim de aferir como os genes de

IFN I podem desencadear a resposta de células do sistema imune inato e adaptativo durante infecção por VL.

A análise GSEA revelou novas descobertas sobre as perturbações moleculares causadas pela infecção (Figura 35a). Por exemplo, a atividade transcricional do módulo M4, cujos genes estão associados com categorias como “degranulação de plaquetas” e “hemostase”, está significativamente elevada em pacientes VL antes de receberem tratamento. De fato, VL está associada com diversas manifestações hematológicas, incluindo anemia, leucopenia e coagulação intravascular disseminada (Varma and Naseem, 2010). Além disso, o padrão de atividade do módulo M9, associado ao ciclo celular, sugere que uma proliferação intensa de células expressando CCR5 (como macrófagos, células dendríticas e células T de memória) pode ocorrer durante a infecção por VL.

O CEMiTool também foi capaz de ilustrar a dinâmica de genes envolvidos com imunidade mediada por células B durante o tratamento de VL, conforme ilustrado pelo módulo M3. A integração com dados de interações proteína-proteína revela que os genes CD79A e CD79B são *hubs* no módulo M3 (Figura 35c). Ambas estas proteínas formam um dímero associado com o receptor de antígeno de célula B (BCR, do inglês *B Cell Receptor*), e são críticos para a imunização por linfócitos B. Este resultado, assim como a coexpressão com outros genes relacionados ao desenvolvimento de célula B, e o fato de que este módulo possui atividade aumentada em amostras de pacientes em tratamento de VL demonstram que o CEMiTool é capaz de proporcionar novos *insights* sobre a resposta do hospedeiro ao tratamento e à doença. Tomadas em

conjunto, nossas descobertas podem ajudar a definir quais genes estão desencadeando estas manifestações hematológicas, e então sugerir novos tratamentos eficazes contra a VL.

Figura 35 - Análise modular de leishmaniose. (a) Análise GSEA ilustrando a atividade dos módulos em cada classe de amostras. “Healthy” = pacientes

não-infectados; “SickBeforeTreat” = pacientes infectados com *Leishmania* antes do tratamento; “SickAfterTreat” = pacientes infectados com *Leishmania* após tratamento. (b) Análise ORA do módulo M7. O gráfico de barras mostra o $-\log_{10}$ do p-valor ajustado (Benjamini-Hochberg) do enriquecimento de genes dos módulos e os *gene sets* do banco de dados Reactome. A linha vertical tracejada indica um p-valor ajustado de 0,01. (c) Redes gênicas dos módulos M3 e M9. Os dez genes com maior conectividade (hubs) estão identificados e coloridos baseado em sua “origem”: caso originalmente presentes no módulo do CEMiTool, estão em azul; se inseridos pelo arquivo de interações, estão em vermelho. O tamanho do nó é proporcional ao seu grau (Russo *et al.*, 2018).

4.9. Banco de dados CEMiToolDB

Para a criação do banco de dados CEMiToolDB, foram baixados todos os 5080 estudos da plataforma GPL570 do banco de dados GEO. Em seguida, foi feita uma seleção a fim de manter apenas os estudos relacionados com uma única série, e foi especificada uma faixa de número de amostras limite entre 18 e 300 amostras. Após estas restrições, o banco de dados foi populado com os resultados do pacote CEMiTool aplicado aos 1001 estudos remanescentes.

Na interface com o CEMiToolDB, o usuário pode prover à ferramenta um determinado gene de interesse e o banco de dados retornará uma lista de genes coexpressos e uma lista de módulos de coexpressão contendo estes genes e os respectivos estudos em que foram determinados, assim como metadados sobre os

estudos. Alternativamente, o usuário poderá fornecer diretamente uma lista de genes, e a ferramenta irá retornar em seguida a lista de módulos envolvidos com os genes presentes na lista. Estas etapas estão ilustradas na Figura 36.

Figura 36 - Fluxograma de análise da ferramenta. Dado um determinado gene de interesse ou uma lista de genes, a ferramenta retornará uma lista dos genes que possuem interações com o gene de interesse no banco de dados. A partir desta lista, são retornados os módulos encontrados pela ferramenta CEMiTool que possuem enriquecimento para as interações listadas.

4.9.1. Exemplo de uso: Granzima A

A fim de demonstrar a utilidade de se definir o banco de dados CEMiToolDB, utilizamos como exemplo de caso o gene GZMA, codificador da Granzima A. Este gene está sabidamente envolvido com processos inflamatórios, inibindo o sistema imune durante infecção por *Streptococcus pneumoniae* (van den Boogaard *et al.*, 2016), sendo um biomarcador eficiente de pacientes com maior chance de responder mais favoravelmente a tratamento em casos de colite ulcerativa (Tew *et al.*, 2016), induzindo citocinas pro-inflamatórias (Metkar *et al.*, 2008), entre outros.

Utilizando o CEMiToolDB, podemos conferir o número de vezes que um determinado gene aparece co-expresso (isto é, encontra-se no mesmo módulo) com o gene GZMA, utilizando a *query* demonstrada na Figura 37. Como esperado, outros genes codificadores de granzimas encontram-se entre os genes mais co-expressos, tais como GZMK, GZMB e GZMH. Além disso, outros genes relacionados à função da granzima A no sistema imune também encontram-se nos resultados, particularmente genes centrais na função efetora de linfócitos T tais como IL7R, TRBC1, KLRB1 e SELL (Perreault, 2011).

```
mysql> SELECT a.gene_symbol a, b.gene_symbol b, COUNT(*) cnt
-> FROM module_genes a JOIN module_genes b ON b.moduleID = a.moduleID
-> AND b.gene_symbol > a.gene_symbol AND a.gene_symbol="GZMA"
-> GROUP BY a.gene_symbol, b.gene_symbol
-> ORDER BY cnt DESC limit 20;
```

a	b	cnt
GZMA	GZMK	267
GZMA	GZMB	249
GZMA	IL7R	208
GZMA	GZMH	205
GZMA	TRBC1	198
GZMA	NKG7	197
GZMA	KLRB1	195
GZMA	PRF1	190
GZMA	TARP	185
GZMA	LCK	184
GZMA	SLAMF7	177
GZMA	TCRDV2	172
GZMA	PTPRC	170
GZMA	ITK	169
GZMA	SELL	166
GZMA	KLRC4-KLRK1	165
GZMA	HLA-DQA1	163
GZMA	IGHG1	159
GZMA	XCL1	156
GZMA	IGLC1	156

20 rows in set (0.37 sec)

Figura 37 - Encontrando genes co-expressos com o gene GZMA. A *query* exposta na parte superior da figura permite que o usuário consulte os genes que mais aparecem nos mesmos módulos juntamente com o gene de interesse GZMA.

A natureza tabular dos sistemas gerenciadores de bancos de dados relacionais como o utilizado pelo CEMiToolDB permite que se combine tabelas diferentes a fim de obter resultados mais informativos do que aqueles presentes nas tabelas inicialmente projetadas. Na Figura 38, está disposta a tabela da Figura 37, porém com informações das vias metabólicas com enriquecimento mais significativo para os genes em questão. Nota-se que o resultado tende a demonstrar vias metabólicas mais gerais; isto é

esperado, dado que vias mais gerais abrangem um número maior de genes, e portanto tendem a ser mais enriquecidas.

```
mysql> SELECT mg.moduleID, mg.a, mg.b, mg.cnt, ora.pathway_name, min(ora.pvalue) AS min_p
-> FROM (SELECT a.moduleID, a.gene_symbol a, b.gene_symbol b, COUNT(*) cnt
-> FROM module_genes a JOIN module_genes b ON b.moduleID = a.moduleID AND b.gene_symbol > a.gene_symbol AND a.gene_symbol="GZMA"
-> GROUP BY a.gene_symbol, b.gene_symbol ORDER BY cnt DESC) AS mg
-> INNER JOIN ora ON mg.moduleID = ora.moduleID
-> GROUP BY mg.a, mg.b
-> ORDER BY cnt DESC
-> LIMIT 30;
```

moduleID	a	b	cnt	pathway_name	min_p
5	GZMA	GZMK	267	Hemostasis	0.000114953852715831
5	GZMA	GZMB	249	Hemostasis	0.000114953852715831
13	GZMA	IL7R	208	TCR signaling	0.000293120648435059
5	GZMA	GZMH	205	Hemostasis	0.000114953852715831
58	GZMA	TRBC1	198	Phosphorylation of CD3 and TCR zeta chains	0.000108763973132245
5	GZMA	NKG7	197	Hemostasis	0.000114953852715831
5	GZMA	KLRB1	195	Hemostasis	0.000114953852715831
5	GZMA	PRF1	190	Hemostasis	0.000114953852715831
5	GZMA	TARP	185	Hemostasis	0.000114953852715831
58	GZMA	LCK	184	Phosphorylation of CD3 and TCR zeta chains	0.000108763973132245
13	GZMA	SLAMF7	177	TCR signaling	0.000293120648435059
5	GZMA	TCRDV2	172	Hemostasis	0.000114953852715831
5	GZMA	PTPRC	170	Hemostasis	0.000114953852715831
22	GZMA	ITK	169	Chemokine receptors bind chemokines	0.000857302791132183
5	GZMA	SELL	166	Hemostasis	0.000114953852715831
5	GZMA	KLRC4-KLRK1	165	Hemostasis	0.000114953852715831
13	GZMA	HLA-DQA1	163	TCR signaling	0.000293120648435059
13	GZMA	IGHG1	159	TCR signaling	0.000293120648435059
22	GZMA	XCL1	156	Chemokine receptors bind chemokines	0.000857302791132183
13	GZMA	IGLC1	156	TCR signaling	0.000293120648435059
13	GZMA	IGK	154	TCR signaling	0.000293120648435059
13	GZMA	IGJ	153	TCR signaling	0.000293120648435059
58	GZMA	TRAC	150	Phosphorylation of CD3 and TCR zeta chains	0.000108763973132245
5	GZMA	KLRK1	149	Hemostasis	0.000114953852715831
13	GZMA	LYZ	148	TCR signaling	0.000293120648435059
13	GZMA	IGKC	146	TCR signaling	0.000293120648435059
48	GZMA	MNDA	145	Phosphorylation of CD3 and TCR zeta chains	0.000243709478549033
22	GZMA	IDO1	145	Chemokine receptors bind chemokines	0.000857302791132183
13	GZMA	SAMSN1	144	TCR signaling	0.000293120648435059
5	GZMA	IL2RB	144	Hemostasis	0.000114953852715831

30 rows in set (8.74 sec)

Figura 38 - Encontrando vias metabólicas de genes co-expressos com o gene GZMA. A *query* exposta na parte superior da figura permite que o usuário consulte os genes que mais aparecem nos mesmos módulos juntamente com o gene de interesse GZMA, assim como a via metabólica com enriquecimento mais significativo na análise ORA do CEMiTool.

Caso seja dada uma lista de genes para o CEMiToolDB, uma *query* similar pode ser efetuada, retornando os módulos com o maior número de genes presentes na lista

e a via metabólica mais significativamente enriquecida nele, conforme ilustrado na Figura 39. Neste caso, foram procurados módulos contendo os genes GZMA, SET, GZMK, NME1, TREX1 e GZMB.

```
mysql> SELECT t1.moduleID, t1.count, t1.genes, ora.pathway_name, ora.pvalue
-> FROM (SELECT moduleID, COUNT(1) AS count, GROUP_CONCAT(gene_symbol ORDER BY gene_symbol) AS genes
-> FROM module_genes
-> WHERE gene_symbol IN ("GZMA", "SET", "GZMK", "NME1", "TREX1", "GZMB")
-> GROUP BY moduleID) AS t1
-> INNER JOIN ora
-> ON t1.moduleID = ora.moduleID
-> GROUP BY moduleID
-> ORDER BY count DESC
-> LIMIT 20;
```

moduleID	count	genes	pathway_name	pvalue
720	4	GZMA, GZMB, GZMK, NME1	Chemokine receptors bind chemokines	2.66977126368858e-09
6230	4	GZMA, GZMB, GZMK, NME1	Cell Cycle, Mitotic	1.09780812502845e-13
6627	4	GZMA, GZMB, GZMK, NME1	Mitotic G1-G1/S phases	1.46042558126177e-11
5	3	GZMA, GZMB, GZMK	Hemostasis	3.22461417191578e-05
13	3	GZMA, GZMB, GZMK	TCR signaling	4.98410502638568e-13
22	3	GZMA, GZMB, GZMK	Chemokine receptors bind chemokines	1.55436927001472e-06
58	3	GZMA, GZMB, GZMK	Phosphorylation of CD3 and TCR zeta chains	8.58822743797915e-18
78	3	GZMA, GZMB, GZMK	Interferon alpha/beta signaling	2.64616697370017e-10
99	3	GZMA, GZMB, GZMK	Hemostasis	1.08986498801727e-09
102	3	GZMA, GZMB, GZMK	Cell surface interactions at the vascular wall	6.10000035306752e-10
138	3	GZMA, GZMB, GZMK	Chemokine receptors bind chemokines	7.73169098939379e-13
150	3	GZMA, GZMB, GZMK	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	2.18269446166015e-08
197	3	GZMA, GZMB, GZMK	Hemostasis	3.22461417191578e-05
205	3	GZMA, GZMB, GZMK	TCR signaling	4.98410502638568e-13
214	3	GZMA, GZMB, GZMK	Chemokine receptors bind chemokines	1.55436927001472e-06
250	3	GZMA, GZMB, GZMK	Phosphorylation of CD3 and TCR zeta chains	8.58822743797915e-18
270	3	GZMA, GZMB, GZMK	Interferon alpha/beta signaling	2.64616697370017e-10
291	3	GZMA, GZMB, GZMK	Hemostasis	1.08986498801727e-09
294	3	GZMA, GZMB, GZMK	Cell surface interactions at the vascular wall	6.10000035306752e-10
330	3	GZMA, GZMB, GZMK	Chemokine receptors bind chemokines	7.73169098939379e-13

20 rows in set (4.53 sec)

Figura 39 - Encontrando vias metabólicas de módulos contendo uma lista de genes. A *query* exposta na parte superior da figura permite que o usuário consulte as vias metabólicas mais significativamente enriquecidas para módulos contendo o maior número de genes presentes em uma dada lista, de acordo com a análise ORA do CEMiTool.

4.10. Pacote anotator

Um exemplo dos resultados da anotação automática realizada com o pacote anotator pode ser encontrado na Figura 40. Neste exemplo, criado com os dados do estudo GSE2125 do banco de dados GEO, as informações das colunas “Sample_title”, “Sample_source_name_ch1” e “Sample_characteristics_ch1” são concatenadas na coluna “mega_col”, onde caracteres especiais e pontuação são removidos e letras maiúsculas são convertidas em minúsculas. O algoritmo HDBSCAN então utiliza a métrica de distância escolhida (por padrão, a distância de Levenshtein, ver abaixo) para realizar o agrupamento das *strings* da coluna mega_col, e então definir os *clusters* das amostras, dispostos na coluna “lv_cluster”. As palavras comuns a todos os elementos de cada *cluster* são depois utilizadas para definir o seu nome.

Sample_title	Sample_source_name_ch1	Sample_characteristics_ch1	mega_col	lv_cluster	lv_label1
18.Woodruff_8sa1.CEL	isolated alveolar macrophage	status: Asthmatic	18 woodruff 8sa1 cel status asthmatic	1	woodruff cel status asthmatic
19.Woodruff_13saA1.CEL	isolated alveolar macrophage	status: Asthmatic	19 woodruff 13saa1 cel status asthma	1	woodruff cel status asthmatic
20.Woodruff_18saA1.CEL	isolated alveolar macrophage	status: Asthmatic	20 woodruff 18saa1 cel status asthma	1	woodruff cel status asthmatic
13.Woodruff_32_C.CEL	isolated alveolar macrophage	status: Smoker	13 woodruff 32 c cel status smoker	2	woodruff cel status smoker
15.Woodruff_51_C.CEL	isolated alveolar macrophage	status: Smoker	15 woodruff 51 c cel status smoker	2	woodruff cel status smoker
16.Woodruff_56_HS.CEL	isolated alveolar macrophage	status: Smoker	16 woodruff 56 hs cel status smoker	2	woodruff cel status smoker
11.Woodruff_20_N.CEL	isolated alveolar macrophage	status: Nonsmoker	11 woodruff 20 n cel status nonsmoke	3	woodruff n cel status nonsmoker
12.Woodruff_22_N.CEL	isolated alveolar macrophage	status: Nonsmoker	12 woodruff 22 n cel status nonsmoke	3	woodruff n cel status nonsmoker
14.Woodruff_33_N.CEL	isolated alveolar macrophage	status: Nonsmoker	14 woodruff 33 n cel status nonsmoke	3	woodruff n cel status nonsmoker
03.Koth_29_HS.CEL	isolated alveolar macrophage	status: Smoker	03 koth 29 hs cel status smoker	4	koth cel status smoker
05.Koth_26_C.CEL	isolated alveolar macrophage	status: Smoker	05 koth 26 c cel status smoker	4	koth cel status smoker
09.Koth_59_HS.CEL	isolated alveolar macrophage	status: Smoker	09 koth 59 hs cel status smoker	4	koth cel status smoker
01.Koth_24_N.CEL	isolated alveolar macrophage	status: Nonsmoker	01 koth 24 n cel status nonsmoker	5	koth n cel status nonsmoker
02.Koth_25_N.CEL	isolated alveolar macrophage	status: Nonsmoker	02 koth 25 n cel status nonsmoker	5	koth n cel status nonsmoker
04.Koth_61_N.CEL	isolated alveolar macrophage	status: Nonsmoker	04 koth 61 n cel status nonsmoker	5	koth n cel status nonsmoker

Figura 40 - Aplicação do pacote anotator no estudo GSE2125. As informações fenotípicas ou de experimento são concatenadas em uma única coluna, a qual é então utilizada pelo algoritmo HDBSCAN para fazer uma análise de agrupamento de *strings* e

então definir um *cluster* para cada amostra. O nome do *cluster* é depois definido como as palavras comuns a todos os seus elementos.

A fim de definir a métrica de similaridade que melhor extrai as informações dos grupos experimentais no agrupamento pelo HDBSCAN, 100 estudos de transcriptômica foram anotados manualmente. Em seguida, calculou-se a taxa de falsos-positivos, a taxa de falsos negativos e o índice de Rand (Rand, 1971) a fim de aferir o quanto que as partições encontradas pelo agrupamento com cada métrica de distância concordou com os grupos manualmente anotados. Os resultados da análise estão dispostos na Tabela 2. Escolheu-se como métrica padrão a distância de Levenshtein para o pacote anotador dado que apresenta uma combinação relativamente balanceada de baixos FPR e FNR, e um elevado índice de Rand.

Tabela 2 - Métricas de avaliação dos *clusters* - Cada cluster retornado pelo algoritmo HDBSCAN automaticamente anotado foi avaliado com respeito à anotação realizada manualmente, e então calculou-se a taxa de falsos-positivos, falsos-negativos e o índice de Rand para cada métrica de distância utilizada. SW: Smith-Waterman; NW: Needleman-Wunsch; ME: Monge-Elkan; LV: Levenshtein; DL: Damerau-Levenshtein; OSA: *Optimal String Alignment*; LCS: *Longest Common Substring*.

Métrica	Média FPR	Média FNR	Média Rand
Jaro	0.084	0.215	0.881
JW	0.074	0.222	0.887
SW	0.132	0.135	0.864
NW	0.113	0.145	0.88
ME	0.22	0.564	0.699
LV	0.094	0.176	0.88
DL	0.096	0.175	0.879
OSA	0.096	0.175	0.879
LCS	0.096	0.168	0.89
Qgram	0.093	0.186	0.885
Cosine	0.079	0.193	0.891
Jaccard	0.3	0.316	0.681

5. Discussão

Neste trabalho, foi desenvolvido um conjunto de ferramentas a fim de facilitar o trabalho de pesquisadores interessados em realizar análises de co-expressão de genes. O pacote CEMiTool foi criado com o intuito de automatizar o máximo possível o *workflow* desenvolvido pelo pacote WGCNA para R, de maneira a tornar o processo o mais fácil e acessível possível para usuários com pouca ou nenhuma experiência em programação. Embora o WGCNA disponha de tutoriais bem escritos e o pacote seja muito bem documentado (a versão mais atual de seu manual possui mais de 400 páginas), o uso do pacote pode ser intimidador para novos usuários pouco familiares com a análise, e difícil de saber por onde começar.

O CEMiTool retorna informações sobre módulos de genes, os quais ele define através de valores de correlações entre pares de genes. Já neste passo, ele se

diferencia do pacote WGCNA, automatizando a etapa de definição do parâmetro beta, a qual demonstramos neste trabalho que é feita de uma maneira mais reproduzível e coerente pelo CEMiTool. Esta diferença encontrada nos valores de beta se dá uma vez que o algoritmo padrão utilizado pelo WGCNA utiliza redes sem sinal, e quando o valor de R^2 se encontra abaixo de 0, o mesmo é transformado em positivo. No entanto, não é deixado claro para o usuário que esta alteração é feita, e além disso, faz com que o *scale-free topology criterion* (Zhang and Horvath, 2005) proposto pelos autores deixe de ser aplicável em casos como os mostrados. Nota-se também que apesar de haver uma certa discordância na literatura quanto à real aplicabilidade do modelo livre de escala como um todo em redes biológicas (Broido and Clauset, 2018), optamos por seguir o modelo de forma a poder proporcionar uma ferramenta automatizada que seguisse a metodologia WGCNA.

Tendo definido um valor para o parâmetro beta, os módulos então definidos por um agrupamento hierárquico são submetidos a diversas análises complementares opcionais, como GSEA, ORA e a inclusão de interações, tudo utilizando apenas um comando e arquivos opcionais.

Apesar de toda a facilidade oferecida pela ferramenta, existem ainda muitos desafios a serem cumpridos pelo CEMiTool. Embora muito de nosso esforço tenha se direcionado para deixar as funções do pacote o mais utilizáveis possível para os casos gerais, existe muito que pode ser melhorado quanto à definição de parâmetros padrão, como as melhores maneiras de se filtrar os genes de entrada, os melhores modelos de rede para as interações entre genes (redes com ou sem sinal), métodos de correlação,

entre outros. Além disso, nós propositadamente limitamos o escopo do CEMiTool à definição de módulos e análises de enriquecimento dos mesmos, embora hajam várias outras análises que possam ser feitas, como a correlação da atividade dos módulos a fatores clínicos, que o pacote WGCNA proporciona, ou a definição de módulos sobrepostos, atualmente em desenvolvimento em nosso laboratório.

Nós introduzimos ainda na presente tese a ferramenta *online* webCEMiTool, uma versão para navegador das análises do CEMiTool. Esta ferramenta permite que as análises de coexpressão geradas pelo CEMiTool sejam acessíveis a usuários de qualquer nível de experiência em programação, sendo necessário apenas que possuam um arquivo de expressão de genes. Além disso, por ser hospedado num servidor próprio, as análises realizadas são totalmente independentes de fatores dependentes do usuário, como por exemplo o versionamento dos pacotes-dependência que ele possua instalado em seu computador.

Embora existam ferramentas *online* similares ao webCEMiTool, em geral, estas não proporcionam resultados comparáveis (Cardozo *et al.*, 2019). Um exemplo é o GeNET (Desai *et al.*, 2017), ferramenta *online* criada para facilitar análises de co-expressão, de enriquecimento e redes gene a gene. No entanto, apesar de suas funcionalidades serem bastante similares às do webCEMiTool, a ferramenta é limitada em seu escopo, disponibilizando suas análises apenas para três organismos, *R. capsulatus*, *M. tuberculosis* e *O. sativa*. De maneira similar ao GeNET, o CoExpNetViz (Tzfadia *et al.*, 2015) proporciona ao usuário a possibilidade de visualizar e construir redes gênicas, no entanto, também é limitado com respeito aos arquivos de entrada,

tendo sido desenhado especificamente para analisar dados de transcriptômica de plantas, enquanto o webCEMiTool proporciona análises independentemente do organismo. Além disso, embora o CoExpNetViz seja tratado como uma ferramenta *online*, os usuários devem baixar seus dados da plataforma, e então os colocarem manualmente na ferramenta Cytoscape para só assim visualizá-los. O webCEMiTool proporciona os resultados da análise já no próprio navegador, de maneira muito mais conveniente. Ainda assim, o webCEMiTool continua uma ferramenta relativamente nova, e no momento estamos trabalhando para melhorar a experiência do usuário e estabelecer a atualização automática do *site* a cada nova versão do CEMiTool.

Dada a facilidade de uso do pacote CEMiTool, percebeu-se a utilidade de criar um repositório com os resultados análises realizadas em cima de milhares de estudos. Assim, desenvolvemos o banco de dados CEMiToolDB, o qual proporciona através da interface com a altamente difundida linguagem SQL o acesso a estes resultados e portanto a uma quantidade enorme de possibilidades de questionamentos e análises. No presente trabalho foi mostrado apenas alguns exemplos de pesquisas que o CEMiToolDB poderia proporcionar, de forma rápida, segura e confiável.

Embora o CEMiToolDB não esteja sozinho como banco de dados de coexpressão, até a extensão de nosso conhecimento ele é o único que proporciona todas as análises que o CEMiTool retorna, em particular o enriquecimento funcional de módulos, e a comparação de enriquecimento entre grupos experimentais. Apesar disso, o banco ainda se encontra em fase de desenvolvimento, sendo necessário incluir os resultados das análises de outras plataformas e de dados de RNAseq, assim como

a criação de uma interface *front-end* amigável para o acesso através de um navegador *web*.

Na presente tese, foi descrito também o pacote anotator, uma iniciativa que surgiu durante o desenvolvimento do CEMiToolDB quando se percebeu o esquema inconsistente de anotação de amostras presente no banco de dados GEO. Este pacote utiliza critérios de similaridade de *strings* compostas pelas informações presentes nos dados de anotação de amostras do GEO para inferir grupos de amostras. Conforme mostrado neste trabalho, esta abordagem simples é capaz de estabelecer os grupos amostrais com confiança relativamente alta.

No entanto, o pacote anotator, como as outras ferramentas, ainda é relativamente nova, e portanto possui ampla oportunidade para melhoria. Diferentes algoritmos de agrupamento, por exemplo, podem encontrar os grupos experimentais de forma ainda melhor. Outra possibilidade é o uso de algoritmos de aprendizado de máquina para melhor definir as colunas a serem concatenadas para a criação das *strings*.

Uma funcionalidade interessante da ferramenta decorrente do uso de algoritmos de agrupamento que não fora inicialmente contemplada é a sua relativa robustez a erros ortográficos nas *strings* clusterizadas. Esta propriedade decorre do fato de que em geral estes erros ocorrerão em poucas posições nas *strings*, e portanto a tendência é que os erros inseridos não serão suficientes para definirem grupos separados.

6. Conclusão

A análise de redes de genes coexpressos possui um alto potencial de auxiliar na elucidação de novos mecanismos moleculares envolvidos em diferentes doenças e condições. Além disso, ela permite a compreensão das diferentes interações entre os componentes presentes nas análises e como cada um pode estar envolvido no contexto biológico de interesse.

A fim de estabelecer um *framework* de análise para redes de coexpressão gênica que fosse robusto e confiável, mas ao mesmo tempo intuitivo e fácil de analisar foi criado o CEMiTool. Este pacote para a linguagem de programação R permite que usuários com pouca experiência em programação sejam capazes de analisar seus próprios dados de transcriptômica. Em seguida, a fim de maximizar a facilidade de uso e a acessibilidade, foi criada também a plataforma *online* do CEMiTool.

Através de diversas comparações, foi mostrado como o CEMiTool é capaz de produzir resultados tão interessantes quanto ou mais que outras ferramentas já estabelecidas, em particular o WGCNA, no qual ele é baseado. Além disso, o CEMiTool proporciona muito mais funcionalidades que a maior parte das ferramentas de análises de redes de coexpressão. Também foi mostrado como o CEMiTool pode ser utilizado para encontrar informações biológicas de interesse mesmo em dados públicos já previamente publicados e discutidos.

A fim de permitir um processo totalmente automatizado de redes de coexpressão, foi implementado um novo pacote para a anotação automatizada de dados de amostras de transcriptômica denominado *annotator*. Este pacote utiliza as

informações já presentes no banco de dados GEO para agrupar amostras e automaticamente definir grupos de amostras.

Tendo em mãos os resultados das análises em cima das redes de coexpressão gênica, e as anotações automáticas executadas em cima das amostras de centenas de estudos, foi desenvolvido um banco de dados no qual foram armazenados os resultados de várias iterações do CEMiTool. Esta ferramenta permite a inferência de redes de coexpressão de genes em diferentes condições e permite que as mesmas sejam disponibilizadas para que a comunidade possa fazer buscas e incrementar os resultados de suas próprias análises.

7. Referências bibliográficas

- Afshar,M. *et al.* (2013) Induction and exacerbation of psoriasis with Interferon-alpha therapy for hepatitis C: a review and analysis of 36 cases. *J. Eur. Acad. Dermatol. Venereol.*, 27, 771–778.
- Amar,D. *et al.* (2013) Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput. Biol.*, 9, e1002955.
- Barabasi,A.L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, 286, 509–512.
- Barabási,A.-L. and Bonabeau,E. (2003) Scale-free networks. *Sci. Am.*, 288, 50–59.
- Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.*, 41, D991-5.
- van den Boogaard,F.E. *et al.* (2016) Granzyme A impairs host defense during *Streptococcus pneumoniae* pneumonia. *Am. J. Physiol. Lung Cell. Mol. Physiol.*, 311, L507-16.
- Broido,A.D. and Clauset,A. (2018) Scale-free networks are rare.
- Campello,R.J.G.B. *et al.* (2013) Density-Based Clustering Based on Hierarchical Density Estimates. In, Pei,J. *et al.* (eds), *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 160–172.
- Cardozo,L.E. *et al.* (2019) webCEMiTool: Co-expression modular analysis made easy.

Frontiers in Genetics.

- Carter, S.L. *et al.* (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20, 2242–2250.
- Chen, E.Y. *et al.* (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14, 128.
- Croft, D. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39, D691–7.
- van Dam, S. *et al.* (2015) GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Res.*, 43, D1124–32.
- Damerau, F.J. (1964) A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7, 171–176.
- Davis, S. and Meltzer, P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, 23, 1846–1847.
- Desai, A.P. *et al.* (2017) GeNET: a web application to explore and share Gene Co-expression Network Analysis data. *PeerJ*, 5, e3678.
- Edgar, R. *et al.* (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30, 207–210.
- Ellis, S.E. *et al.* (2018) Improving the value of public RNA-seq expression data by phenotype prediction. *Nucleic Acids Res.*, 46, e54.
- Fang, F.C. and Casadevall, A. (2011) Reductionistic and holistic science. *Infect. Immun.*, 79, 1401–1404.
- Fisher, R.A. (1925) Statistical methods for research workers Oliver and Boyd, Edinburgh, Oliver and Boyd, 1970.
- Friboulet, A. and Thomas, D. (2005) Systems Biology-an interdisciplinary approach. *Biosens. Bioelectron.*, 20, 2404–2407.
- Gerstberger, S. *et al.* (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, 15, 829–845.
- Ghazalpour, A. *et al.* (2006) Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.*, 2, e130.
- Ha, M.J. *et al.* (2015) DINGO: differential network analysis in genomics. *Bioinformatics*, 31, 3413–3420.
- Hadley, D. *et al.* (2017) Precision annotation of digital samples in NCBI's gene expression omnibus. *Sci. Data*, 4, 170125.
- Holme, P. (2019) Rare and everywhere: Perspectives on scale-free networks. *Nat. Commun.*, 10, 1016.
- Horvath, S. (2011) Weighted network analysis: Application in genomics and systems biology Hardcover; 2011-04-01. Springer, New York, NY.
- Jaccard, P. (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Societe Vaudoise des Sciences Naturelle*, 37.

- Jaro,M.A. (1989) Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *J. Am. Stat. Assoc.*, 84, 414–420.
- Jelinek,J. (2007) Metrics in Similarity Search.
- Jones,S. and Thornton,J.M. (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci USA*, 93, 13–20.
- Keermann,M. *et al.* (2015) Transcriptional landscape of psoriasis identifies the involvement of IL36 and IL36RN. *BMC Genomics*, 16, 322.
- Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, 8, e1002375.
- Kim,G.-W. *et al.* (2013) Extensive psoriasis induced by pegylated interferon alfa-2a and ribavirin in the treatment of chronic hepatitis C. *Ann. Dermatol.*, 25, 479–482.
- Kitano,H. (2002) Systems biology: a brief overview. *Science*, 295, 1662–1664.
- Kolesnikov,N. *et al.* (2015) ArrayExpress update--simplifying data submissions. *Nucleic Acids Res.*, 43, D1113-6.
- Konopka,A.K. (2006) Systems biology: Principles, methods, and concepts CRC Press/Taylor & Francis, Boca Raton.
- Kuleshov,M.V. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, 44, W90-7.
- Langfelder,P. *et al.* (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24, 719–720.
- Langfelder,P. (2017) WGCNA package FAQ.
- Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 559.
- Langfelder,P. and Horvath,S. (2014) WGCNA package: Frequently Asked Questions.
- Lee,Y.-R. *et al.* (2006) MCP-1, a highly expressed chemokine in dengue haemorrhagic fever/dengue shock syndrome patients, may cause permeability change, possibly through reduced tight junctions of vascular endothelium cells. *J. Gen. Virol.*, 87, 3623–3630.
- Lehninger,A.L. *et al.* (2012) Lehninger principles of biochemistry 6th ed. W.H. Freeman, New York.
- Levenshtein,V.I. (1966) Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics - Doklady*, 10.
- Likić,V.A. *et al.* (2010) Systems biology: the next frontier for bioinformatics. *Adv. Bioinformatics*, 268925.
- van der Loo,M. (2014) The stringdist Package for Approximate String Matching. *R J*, 6.
- Marcondes,M. and Rossi,C.N. (2014) Leishmaniose visceral no Brasil. *Braz. J. Vet. Res. Anim. Sci.*, 50, 341.
- Metkar,S.S. *et al.* (2008) Human and mouse granzyme A induce a proinflammatory cytokine response. *Immunity*, 29, 720–733.

- Mishra,P. *et al.* (2014) Gene set analysis: limitations in popular existing methods and proposed improvements. *Bioinformatics*, 30, 2747–2756.
- Monge,A. and Elkan,C. (1997) An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records.
- Mullen,L. (2015) textreuse: Detect Text Reuse and Document Similarity.
- Navlakha,S. *et al.* (2014) Topological properties of robust biological and computational networks. *J. R. Soc. Interface*, 11, 20140283.
- Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48, 443–453.
- Obayashi,T. *et al.* (2019) COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res.*, 47, D55–D62.
- Ogata,Y. *et al.* (2010) CoP: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics*, 26, 1267–1268.
- Ooms,J. *et al.* (2018) RMySQL: Database Interface and “MySQL” Driver for R.
- Perreault,C. (2011) Mechanisms and Implications of Immunodominance in CD8+ T-Cell Responses. In, Medin,J. and Fowler,D. (eds), *Experimental and applied immunotherapy*. Humana Press, Totowa, NJ, pp. 195–206.
- Petereit,J. *et al.* (2016) petal: Co-expression network modelling in R. *BMC Syst. Biol.*, 10 Suppl 2, 51.
- Price,D.J. (1965) Networks of scientific papers. *Science*, 149, 510–515.
- Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, 66, 846–850.
- Ravasz,E. *et al.* (2002) Hierarchical organization of modularity in metabolic networks. *Science*, 297, 1551–1555.
- Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26, 139–140.
- Ross,S.M. (2010) A First Course in Probability 9 ed. Pearson Custom Publishing, Boston, MA; Pearson Custom Publishing.
- Russo,P.S.T. *et al.* (2018) CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics*, 19, 56.
- Salmanidis,M. *et al.* (2014) Direct transcriptional regulation by nuclear microRNAs. *Int. J. Biochem. Cell Biol.*, 54, 304–311.
- Shah,N. *et al.* (2016) A crowdsourcing approach for reusing and meta-analyzing gene expression data. *Nat. Biotechnol.*, 34, 803–806.
- Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, 147, 195–197.
- Subramaniam,S. and Hsiao,G. (2012) Gene-expression measurement:

- variance-modeling considerations for robust data analysis. *Nat. Immunol.*, 13, 199–203.
- Sun,P. *et al.* (2013) Sequential waves of gene expression in patients with clinically defined dengue illnesses reveal subtle disease phases and predict disease severity. *PLoS Negl. Trop. Dis.*, 7, e2298.
- Tesson,B.M. *et al.* (2010) DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*, 11, 497.
- Tew,G.W. *et al.* (2016) Association between response to etrolizumab and expression of integrin α e and granzyme A in colon biopsies of patients with ulcerative colitis. *Gastroenterology*, 150, 477–87.e9.
- Thomas,M.C. and Chiang,C.-M. (2006) The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.*, 41, 105–178.
- Tsoi,L.C. *et al.* (2015) Analysis of long non-coding RNAs highlights tissue-specific expression patterns and epigenetic profiles in normal and psoriatic skin. *Genome Biol.*, 16, 24.
- Tzfadia,O. *et al.* (2015) CoExpNetViz: Comparative Co-Expression Networks Construction and Visualization Tool. *Front. Plant Sci.*, 6, 1194.
- Varma,N. and Naseem,S. (2010) Hematologic changes in visceral leishmaniasis/kala azar. *Indian J. Hematol. Blood Transfus.*, 26, 78–82.
- Wang,J. *et al.* (2010) Blasted cell line names. *Cancer Inform.*, 9, 251–255.
- Wang,Z. *et al.* (2016) Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun.*, 7, 12846.
- Watson,M. (2006) CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics*, 7, 509.
- Wright,G.W. and Simon,R.M. (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, 19, 2448–2455.
- Xue,Z. *et al.* (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, 500, 593–597.
- Yanhong Zhou,C.M. (2014) Measuring Inequalities in Gene Co-expression Networks of HIV-1 Infection Using the Lorenz Curve and Gini Coefficient. *J. Data Mining Genomics Proteomics*, 05.
- Yu,G. (2009) Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. *Stat. Probab. Lett.*, 79, 1621–1629.
- Zenil,H. *et al.* (2016) Methods of information theory and algorithmic complexity for network biology. *Semin. Cell Dev. Biol.*, 51, 32–43.
- Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, 4, Article17.

Zhu,Q. *et al.* (2015) Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nat. Methods*, 12, 211–4, 3 p following 214.

Apêndice I

1. Métodos complementares

1.1. Comparação dos métodos de seleção do parâmetro β nos pacotes

WGCNA e CEMiTool

Definimos $R^2(\beta)$ para se referir ao valor de R^2 obtido para um dado β . Tendo em mãos a curva $R^2 \times \beta$, escolhe-se um valor de corte τ . O método WGCNA consiste em selecionar o menor β de maneira que $R^2(\beta) > \tau$. Suponhamos que este valor seja β_1 e que ele corresponda a $R^2(\beta_1) = \tau + \delta$ para algum $\delta > 0$ (isto é, o R^2 equivalente encontra-se apenas ligeiramente acima de τ). Tomemos agora o próximo valor, β_2 . Se $R^2(\beta_2) \approx R^2(\beta_1)$, então a escolha do WGCNA foi boa, uma vez que há pouco a se ganhar ao se aumentar mais o valor de corte escolhido, porém muito a se perder quanto à conectividade da rede, conforme mostra a Figura 41. Quando τ está próximo de 1, isto ocorre na maior parte dos casos, mas ao custo de valores de β desnecessariamente altos quando R^2 se aproxima de 1 de forma devagar (ou, pior, não seleciona um valor de corte).

Figura 41. Curvas R^2 x β hipotéticas para WGCNA. (a) Boa escolha. (b) Escolha ruim.

Por outro lado, escolhendo um valor baixo de τ leva a uma perda de adequação ao modelo livre de escala (indicado por R^2) se R^2 for significativamente maior que $R^2(\beta_1)$ – por exemplo, se $R^2(\beta_2) = R^2(\beta_1) + \Delta$, para $\Delta > \delta$. O algoritmo do CEMiTool evita estes casos ao avaliar se $|R^2(\beta_1) - R^2(\beta_2)| < \epsilon$ para um valor pré-definido de ϵ – caso contrário, o CEMiTool rejeitará β_1 como valor de corte e seguirá até o próximo, parando quando a curva aparentar estabilizar. Isso significa que ele testa todos os potenciais ganhos de adequação ao modelo livre de escala antes de escolher um valor, o que permite a escolha de um valor menor de τ sem prejudicar o valor de R^2 .

1.2. Filtragem baseada em variância aplicada a expressão gênica

Seja g_{ij} o valor de expressão do gene i na amostra j . Assumimos, conforme proposto em (Wright and Simon, 2003), que $g_{ij} \sim N(\mu_i, \sigma_i^2)$ e que $\sigma_i^2 \sim \text{IG}(\alpha, \beta)$, onde IG denota a distribuição gamma inversa. Assim, um p-valor é associado a cada gene correspondendo à hipótese nula $H_0 : g_{ij} \sim N(\mu_i, \sigma_i^2)$, calculado como $p_i = \text{Pr}[\sigma^2 \geq s_i^2]$, onde s_i^2 é a variância estimada para o i -ésimo gene.

Como $\sigma_i^2 \sim \text{IG}(\alpha, \beta)$, isso significa que p_i pode ser escrito como

$$p_i = \int_{s_i^2}^{\infty} f(x|\alpha, \beta) dx,$$

onde $f(x|\alpha, \beta)$ é a função densidade de probabilidade de uma distribuição gamma inversa com parâmetros α e β . Então re-escrevemos p_i utilizando as funções gamma:

$$p_i = 1 - \frac{\Gamma(\alpha, \frac{\beta}{s_i^2})}{\Gamma(\alpha)}.$$

Aqui, $\Gamma(\cdot, \cdot)$ é a função gamma superior incompleta e $\Gamma(\cdot)$ é a função gamma euleriana. Para calcular estes valores, basta estimar α e β .

1.2.1. Estimação dos parâmetros α e β

Nós utilizamos um estimador de método dos momentos (MM) para os parâmetros α e β , fazendo uso do fato que $X \sim IG(\alpha, \beta)$ implica que $E[X] = \frac{\beta}{\alpha-1}$ e $Var[X] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$. Assim, para uma amostra $(S_1^2, \dots, S_n^2)$ da variância estimada de cada gene, o MM nos dá:

$$\{\bar{S}^2 = \frac{\hat{\beta}}{\hat{\alpha}-1} Var[S^2] = \frac{\hat{\beta}^2}{(\hat{\alpha}-1)^2(\hat{\alpha}-2)}\}$$

onde $\bar{S}^2$ representa a média amostral de S^2 e $Var[S^2]$ é sua variância estimada. Então:

$$\{\hat{\alpha} = \frac{\bar{S}^2}{Var[S^2]} \hat{\beta} = \bar{S}^2(\hat{\alpha}-1) + 2\}$$

1.2.2. Combined Enrichment Score (CES)

A fim de averiguar o impacto do filtro do CEMiTool nos aspectos biológicos da análise dos módulos, os mesmos foram sujeitos à análise ORA implementada no CEMiTool, e os p-valores resultantes foram combinados através do procedimento abaixo, denominado *Combined Enrichment Score* (CES):

Lema: Seja p_{jk} o p-valor da análise ORA do j-ésimo módulo relativo ao k-ésimo *pathway* interrogado ($j = 1, 2, \dots, M$ e $k = 1, 2, \dots, L$). Então o valor

$$y = - \sum_{j=1}^M \ln(1 - \min_{1 \leq k \leq L} p_{jk})$$

possui uma distribuição gamma com parâmetros M e L .

Prova: De acordo com (Fisher, 1925), para cada módulo j os p-valores $p_{j1}, p_{j2}, \dots, p_{jL}$ seguem uma distribuição uniforme no intervalo unitário. Se definirmos agora $x_j = \min_{1 \leq k \leq L} p_{jk}$, um teorema conhecido em estatísticas de ordem (Ross, 2010) nos diz que $x_j \sim \text{Beta}(1, L)$ para cada j , e portanto $-\ln(1 - x_j) \sim \text{Exp}(L)$. Finalmente, a soma de M distribuições exponenciais i.i.d. com parâmetro L é uma distribuição gamma com parâmetros M e L .

Seguindo este lema, o CES é definido como $CES = \frac{L}{\sqrt{M}}(y - \frac{M}{L})$. Esta grandeza possui média zero e variância unitária, permitindo-nos comparar estudos com diferentes números de módulos; uma vez que a transformação envolvida em calcular y é monotonicamente crescente, quanto menor o seu valor, mais significativo é o enriquecimento dos módulos.

1.3. Efeito do número de amostras no ajuste ao modelo de topologia livre de escala

A fim de avaliar o número mínimo ótimo de amostras para o CEMiTool, foi criado um parâmetro de controle de qualidade, ϕ , para a curva $\beta \times R^2$. Definimos ϕ como a razão da área sob a curva e a área do retângulo definido por $\beta \times 1$, que é o maior valor possível para R^2 . Valores altos de ϕ significam que a topologia da rede converge rapidamente para uma distribuição de grau livre de escala. Para estimar o número mínimo de amostras que retornam o maior valor de ϕ possível antes de estabilizar, foi realizada uma análise de *bootstrapping* em cima das amostras de 3 estudos de *microarray*, selecionando-se inicialmente 5 amostras aleatoriamente e então

incrementando em 5 o número de amostras a cada etapa. Para cada etapa, o CEMiTool foi rodado 10 vezes utilizando os parâmetros padrão. Conforme ilustrado pela Figura 29, o parâmetro ϕ tende a se estabilizar em torno de 20 amostras, indicando que a curva $\beta \times R^2$ - e por conseguinte, a topologia da rede - não deve variar tanto a partir desse número de amostras.

Figura 42 - Relação entre o número de amostras e ϕ . Diferentes quantias de amostras aleatórias foram selecionadas de 3 estudos de *microarray* para análises do CEMiTool. A linha preta representa a média para 10 conjuntos de amostragem. A região verde representa os intervalos de confiança de 95% da média. A linha cinza tracejada representa o número mínimo de amostras requeridas para se definir uma rede cuja topologia não varia significativamente com um número crescente de amostras, conforme indicado pelo parâmetro ϕ [\(Russo et al. 2018\)](#).

Apêndice II

1. Dependências

Tabela 3 - Dependências do pacote CEMiTool. Tabela gerada no dia 20/05/2019.

Pacotes	Versão	Data instalado
clusterProfiler	3.10.1	20-12-18
data.table	1.11.8	30-09-18
dplyr	0.7.8	10-11-18
DT	0.5	05-11-18
ff	2.2.14	15-05-18
ffbase	0.12.7	06-11-18
fgsea	1.8.0	31-10-18
GeneOverlap	1.18.0	30-10-18
ggdendro	0.1.20	27-04-16
ggplot2	3.1.0	25-10-18
ggpmisc	0.3.0	16-07-18
ggrepel	0.8.0	09-05-18
ggthemes	4.0.1	24-08-18
gRbase	1.8.3.4	07-08-18
gridExtra	2.3	09-09-17
gtable	0.2.0	26-02-16
htmltools	0.3.6	28-04-17
igraph	1.2.2	27-07-18
intergraph	2.0.2	05-12-16
knitr	1.21	10-12-18
limma	3.38.3	02-12-18
matrixStats	0.54.0	23-07-18
network	1.13.0.1	02-04-18
plyr	1.8.4	08-06-16
pracma	2.2.2	02-12-18
RColorBrewer	1.1.2	07-12-14
rmarkdown	1.11	08-12-18
scales	1.0.0	09-08-18
sna	2.4	08-08-16
stringr	1.3.1	10-05-18
tidyr	0.8.2	28-10-18
WGCNA	1.66	24-10-18

Tabela 4 - Dependências do pacote anotator. Tabela gerada no dia 20/05/2019.

Pacote	Versão	Data instalado
data.table	1.11.8	30-09-18
dbscan	1.1.3	13-11-18
dplyr	0.7.8	10-11-18
FastKNN	0.0.1	12-02-15
fossil	0.3.7	30-03-12
NameNeedle	1.2.4	11-07-17
plyr	1.8.4	08-06-16
PTXQC	0.92.5	15-03-19
qualV	0.3.3	12-06-17
rentrez	1.2.1	05-03-18
rvest	0.3.2	17-06-16
stringdist	0.9.5.1	08-06-18
textreuse	0.1.4	28-11-16
xml2	1.2.0	24-01-18

Apêndice III

1. Publicações

- Cardozo, L.E., **Russo, P.S.T.**, Gomes-Correia, B., Araujo-Pereira, M., Sepúlveda-Hermosilla, G., Maracaja-Coutinho, V., and Nakaya, H.I. (2019). webCEMiTool: Co-expression modular analysis made easy. *Frontiers in Genetics*.
- Lever, M., Hirata, T.D.C., **Russo, P.S.T.**, Nakaya, H.I. (2018) *Systems Immunology*. In, da Silva, F. A. B., Carels, N., Silva Junior, F. P. (eds), *Theoretical and Applied Aspects of Systems Biology*. Springer Press, Berlin, pp. 159-173.
- **Russo, P.S.T.**, Ferreira, G.R., Cardozo, L.E., Bürger, M.C., Arias-Carrasco, R., Maruyama, S.R., Hirata, T.D.C., Lima, D.S., Passos, F.M., Fukutani, K.F., et al. (2018). CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics* 19, 56.
- Kasturi, S.P., Kozlowski, P.A., Nakaya, H.I., Burger, M.C., **Russo, P.S.T.**, Pham, M., Kovalenkov, Y., Silveira, E.L.V., Havenar-Daughton, C., Burton, S.L., et al. (2017). Adjuvanting a Simian Immunodeficiency Virus Vaccine with Toll-Like Receptor Ligands Encapsulated in Nanoparticles Induces Persistent Antibody Responses and Enhanced Protection in TRIM5 α Restrictive Macaques. *J. Virol.* 91.
- Serafim, K.R., **Russo, P.S.T.**, Fernandes, C.E.M., Gianlorenço, A.C.L., and Mattioli, R. (2016). Intra-amygdala microinjections of chlorpheniramine impair memory formation or memory retrieval in anxiety- and fear-mediated models. *Brain Res. Bull.* 125, 127–133.

2. Atividades acadêmicas

2.1. Disciplinas cursadas

O aluno completou os 56 créditos de disciplinas exigidos pelo Programa

Interunidades de Pós-Graduação em Bioinformática, sendo elas:

- IBI5013 - Banco de Dados para Bioinformática (conceito: “A”)
- IBI5035 - Biologia Molecular Computacional (conceito: “A”)

- IBI5037 - Algoritmos em Bioinformática (conceito: “A”)
- IBI5011 - Introdução à Computação para Bioinformática (conceito: “A”)
- PCS5031 - Introdução à Ciência dos Dados (conceito: “A”)
- IBI5073 - Biologia de Sistemas para Ciências da Vida (conceito: “A”)
- FBA5905 - Planejamento Experimental e Análise Multivariada (conceito: “A”)
- PCS5024 - Aprendizado Estatístico (conceito: “A”)

2.2. Organização de congressos/reuniões científicas

O aluno participou da organização nos seguintes congressos e reuniões científicas, nos quais apresentou resultados obtidos no presente projeto e assuntos pertinentes na forma de apresentações orais:

- CBAB – Bioinformatics Applied to Vaccinology (2015)
- IME-USP - VIII Curso de Verão Bioinformática (2016)
- IME-USP - IX Curso de Verão Bioinformática (2017)

2.3. Participação de congressos/reuniões científicas

O aluno participou dos seguintes congressos e reuniões científicas:

- 2018
 - 2018 International Conference on Systems Biology of Human Diseases (pôster) - UCLA - Los Angeles, EUA.
 - 2018 SPCell Workshop (pôster) – São Paulo School for Advanced Science in Cell Biology, 2018, São Paulo - SP.

- 2017
 - X-Meeting 2017 - 13th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (pôster)
 - Prêmio de menção honrosa para o pôster “*CEMiTool: Co-Expression Modules identification Tool*”
- 2016
 - Brazilian Summer School on Machine Learning 2016 (ouvinte)
 - Hospital Innovation Show 2016
 - Vencedor do *hackathon* “Hackathon Hack4Health”
 - Workshop Beyond Zika - A Tripartite Initiative - USP/Institut Pasteur/FioCruz (ouvinte)
 - X-Meeting 2016 - 12th International Conference of the Brazilian Association of Bioinformatics and Computational Biology (participação oral e pôster)
- 2015
 - University of Surrey - UGPN Doctoral Seminar – Guildford, UK (participação oral)
 - AB3C + BSB – X-Meeting + BSB 2015 (pôster)
 - CEFAP-ICB – 2º Minicurso: Transcriptômica com RNAseq (ouvinte)
 - FCF-USP – II International Symposium on Pathophysiology and Toxicology (ouvinte)

- FCF-USP – VII Simpósio de Pós-Graduação em Análises Clínicas
(ouvinte)
- CBAB - Use of Bioinformatics for the Study of Vaccines (participação oral)

2.4. Atividades complementares

O aluno exerceu as seguintes atividades complementares:

- Atual bolsista FAPESP, modalidade Doutorado Direto, processo número 2017/05762-7
- Ex-bolsista FAPESP, modalidade Mestrado, processo número 2015/20897-0
- Representante Discente suplente do Programa Interunidades de Pós-Graduação em Bioinformática (vigência 2016)
- Participou da organização dos Cursos de Verão do Programa Interunidades de Pós-Graduação em Bioinformática dos anos de 2016 e 2017.
- Ex-bolsista CAPES do Programa Interunidades de Pós-Graduação em Bioinformática

SOFTWARE

Open Access

CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses

Pedro S. T. Russo^{1†}, Gustavo R. Ferreira^{1†}, Lucas E. Cardozo¹, Matheus C. Bürger¹, Raul Arias-Carrasco², Sandra R. Maruyama³, Thiago D. C. Hirata¹, Diógenes S. Lima¹, Fernando M. Passos¹, Kiyoshi F. Fukutani³, Melissa Lever¹, João S. Silva³, Vinicius Maracaja-Coutinho² and Helder I. Nakaya^{1*}

Abstract

Background: The analysis of modular gene co-expression networks is a well-established method commonly used for discovering the systems-level functionality of genes. In addition, these studies provide a basis for the discovery of clinically relevant molecular pathways underlying different diseases and conditions.

Results: In this paper, we present a fast and easy-to-use Bioconductor package named CEMiTool that unifies the discovery and the analysis of co-expression modules. Using the same real datasets, we demonstrate that CEMiTool outperforms existing tools, and provides unique results in a user-friendly html report with high quality graphs. Among its features, our tool evaluates whether modules contain genes that are over-represented by specific pathways or that are altered in a specific sample group, as well as it integrates transcriptomic data with interactome information, identifying the potential hubs on each network. We successfully applied CEMiTool to over 1000 transcriptome datasets, and to a new RNA-seq dataset of patients infected with *Leishmania*, revealing novel insights of the disease's physiopathology.

Conclusion: The CEMiTool R package provides users with an easy-to-use method to automatically implement gene co-expression network analyses, obtain key information about the discovered gene modules using additional downstream analyses and retrieve publication-ready results via a high-quality interactive report.

Keywords: Co-expression modules, Gene networks, Modular analysis, Leishmaniasis, Transcriptomics

Background

Cellular processes are controlled by a host of interacting molecules whose activity and levels are frequently co-regulated or co-expressed. Detecting the groups (i.e. modules) of co-expressed genes in a myriad of biological conditions has generated important insights in brain evolution [1], coronary artery disease [2], and macrophage activation [3], among many other biological conditions.

Following evidence that genes interact with each other in a scale-free fashion [4], Zhang and Horvath developed an R package named WGCNA (Weighted

Gene-Coexpression Network Analysis) that identifies co-expressed gene modules [5]. Although tutorials and examples are available for using the package, following its workflow verbatim is time-consuming and tiresome. Moreover, users are often required to manually select parameters and to filter the input genes prior running WGCNA. This hinders workflow automation and can impact reproducibility since different researchers may utilize different parameters, obtaining distinct results for the same data set. More importantly, WGCNA is limited in terms of the functional analyses available for the package users.

After identifying co-expressed gene modules, researchers are often interested in performing functional and integrative analyses. Over-representation analysis (ORA) can be used to reveal if a set of co-expressed genes is enriched for

* Correspondence: hnakaya@usp.br

[†]Equal contributors

¹Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, SP 05508-900, Brazil
Full list of author information is available at the end of the article

genes belonging to known pathways or functions. In addition, gene set enrichment analysis (GSEA) [6] can associate the activity of a module with the study phenotypes (i.e. sample group). Finally, integrating co-expression information with protein-protein interaction data can be useful to identify main regulators or hubs. Such analyses, however, require the combination of several packages and programs, and considerable bioinformatics skills.

To address these issues, we developed the Co-Expression Modules identification Tool (CEMiTool), an R package that allows users to easily identify and analyze co-expression modules in a fully automated manner. CEMiTool provides users with a novel unsupervised gene filtering method, automated parameter selection for identifying modules, enrichment and module functional analyses, as well as integration with interactome data. Our tool then reports everything in HTML web pages with high-quality plots and interactive tables.

Using the same real datasets, we compared the features of CEMiTool against existing tools, and showed that our tool outperforms them in several aspects. We also applied CEMiTool to over 1000 microarrays and RNA-seq datasets, demonstrating its power in automating the generation of co-expression gene modules and subsequent analyses. Finally, to gain a better insight of the pathophysiology of Leishmania infection, we ran CEMiTool on a novel RNA-seq dataset, which was generated from the blood of infected patients. Our analyses revealed that several modules contained genes not previously associated with Leishmaniasis. The R package is freely available in Bioconductor (DOI: <https://doi.org/10.18129/B9.bioc.CEMiTool>), and as a Docker image file as well (<https://hub.docker.com/r/csblusp/cemitool>).

Implementation

CEMiTool is an easy-to-use package, automating within a single R function (*cemitool*) the entire module discovery process - including gene filtering and functional analyses (Fig. 1). The process begins with a gene expression file containing the genes as rows and the samples as columns. This file is the only required input for CEMiTool's analyses. An unsupervised filtering method based on the inverse gamma distribution (Additional file 1: Text) will then select the genes used in the analyses. Next, a soft-thresholding power β [5] is chosen using our modified algorithm (Additional file 1: Text), and this value is used to determine a similarity criterion between pairs of genes. The genes are then separated into modules using the Dynamic Tree Cut package [5, 7]. If an optional file containing gene interactions (e.g. protein-protein interaction data) is provided, the package will return network graphs composed of interacting genes within the same module. Additionally, if the user provides a sample annotation file, CEMiTool can

perform gene set enrichment analysis (GSEA), allowing users to visualize which modules are induced or repressed in the different phenotypes. Finally, given an optional file containing gene sets, CEMiTool will perform an over representation analysis (ORA) based on the hypergeometric test to determine the most significant module functions.

Over representation analysis of modules

To determine the biological functions possibly related to each module, CEMiTool is able to take a user-provided gene pathway list and perform an over representation analysis (ORA) via the clusterProfiler R package [8]. CEMiTool will then report the adjusted *p*-value negative logarithm for the top gene sets enriched on each co-expression module based on the hypergeometric test. This analysis is also available in the WGCNA package via the *userListEnrichment* function, however its output is in tabular form, while CEMiTool returns both a table and a bar graph of the most significantly enriched pathways for each module.

Association of module activity to sample phenotypes

If the user submits a sample annotation file describing the phenotypes (i.e. disease, healthy, treated, etc) of samples, CEMiTool performs a gene set enrichment analysis using the fgsea (Fast Gene Set Enrichment Analysis) R package [9]. In this analysis, genes from co-expression modules will be treated as gene sets and the z-score normalized expression of the samples within each phenotype will be treated as rankings on the analysis. The results will assess if the activity of a module is altered across different phenotypes.

Adding gene interactions to modules

Users can also provide a gene interaction file to visualize the interactions between the genes in each co-expression module. This allows users to customize their module graphs according to different interaction databases. The top ten network hubs (genes with the highest connectivities) are highlighted in the graph. The resulting network is provided as a graph (one per module) in the HTML report.

We compared the features provided by CEMiTool with existing tools for co-expression module identification and analysis, namely WGCNA, Petal [10], CoP [11], GeNET [12], DiffCoEx [13], CoXpress [14], DICER [15] and DINGO [16], as shown in Table 1. However, none of the tools evaluated have all the features provided by CEMiTool.

Results and discussion

Co-expressed gene module selection and benchmark

We utilized two publicly available microarray studies of Dengue infection (GSE18090 and GSE43777) to compare CEMiTool to two R packages: WGCNA and Petal [10].

Fig. 1 Overview of CEMiTool. **a** CEMiTool requires a gene expression file to identify the modules and optional files to: **(b)** visualize the expression profile of individual genes across samples from different groups, which are defined by the user and shown as different colors; **(c)** perform Gene Set Enrichment Analyses, showing the module activity on each group of samples; **(d)** run over representation analysis to define module functions; and **(e)** create gene networks, displaying the top ten most connected genes (hubs)

CEMiTool was run using its default parameters and all optional files. After filtering, the analyses were performed on 2129 genes for study GSE18090, and 1765 genes for study GSE43777. Our assumption is that greater gene set enrichment in pathways relevant to the diseases are good proxies for the quality of a co-expression network analysis. For study GSE18090, CEMiTool selected a soft-threshold value of 6 and identified 12 different co-expression modules, out of which 9 had at least one significantly enriched pathway in the Over Representation Analysis. Notably, modules M4 and M6 were significantly enriched with interferon and cytokine signaling pathways, along with antiviral mechanisms, as expected from an infectious disease such as dengue. Furthermore, module M2 was

significantly enriched for toll-like receptor cascades, which have been shown to lead to and induce the release of proinflammatory cytokines and chemokines in Dengue infections. These findings mirror what was found in the 7 significantly enriched (of a total of 11) co-expression modules observed for study GSE43777 (beta = 5). Running CEMiTool analyses with all possible optional files for both studies in an average computer took around 3 min (Table 1).

In order to compare WGCNA to CEMiTool, WGCNA was run on the Dengue studies using the top 4000 most variant genes of each dataset. Since WGCNA does not specify the optimal number of input genes, we utilized the same number of genes suggested in their tutorial.

Table 1 Features Provided by Programs that Identify Co-Expression Modules. Over representation analysis of CoP, and GeNET programs is considered “limited” because they only allow the usage of specific gene sets (GO, Pfam or KEGG). The runtime of 2 studies using the same computer and default settings are shown for CEMiTool, WGCNA, and Petal

Features	CEMiTool	WGCNA	Petal	CoP	GeNET	DiffCoEx	CoXpress	DICER	DINGO
Automatic Gene Filtering	yes	no	no	no	no	no	no	no	no
Over representation analysis	yes	yes	no	limited	limited	no	no	no	no
Gene set enrichment analysis	yes	no	no	no	yes	no	no	no	no
Integration with interactome	yes	no	no	no	no	no	no	no	no
Report in HTML	yes	no	no	yes	no	no	no	no	no
Native plots	yes	yes	no	no	yes	no	no	no	no
Search for a gene or gene list	yes	no	yes	yes	yes	no	no	no	no
Merging modules	yes	yes	no	no	no	yes	no	yes	no
Allows 2+ sample groups	yes	yes	yes	yes	yes	no	no	no	no
R package	yes	yes	yes	no	no	no	yes	no	yes
Year of last update	2017	2017	2017	2010	Unknown	Unknown	2013	Unknown	Unknown
Runtime for study GSE18090	2min12s	3min10s	17min18s	–	–	–	–	–	–
Runtime for study GSE43777	3min03s	4min33s	40min10s	–	–	–	–	–	–

The analysis identified 18 modules for study GSE18090 using a soft-threshold of 9. Interestingly, however, over half of them (10) presented no significantly enriched pathways after Over Representation Analysis (p -value < 0.01). In contrast to the CEMiTool results, WGCNA did not report pathways related to toll-like receptor cascades. As for study GSE43777 (beta = 6), WGCNA returned 10 significantly enriched modules out of a total of 16. These results suggest that, despite running on a smaller number of genes, CEMiTool is able to successfully filter irrelevant genes and construct modules using the most important genes. Our custom WGCNA script was able to run the analysis in a similar time as CEMiTool (around 3 min, Table 1). However, this did not take into account the considerable time required to manually insert all steps needed to perform WGCNA analyses, select the user-specified parameters, and the steep learning curve necessary in order to understand the whole procedure.

To account for the difference in the number of input genes, we also ran WGCNA using the filtered datasets returned by CEMiTool's filter. For study GSE18090, WGCNA identified 16 modules, with a soft-threshold of 7. Out of these, 10 modules had at least one significantly enriched pathway in the Over Representation Analysis. As expected, results became more similar to CEMiTool's, with the inclusion of a module related to Toll-like receptor activity (M2), and different modules for interferon types gamma (M4) and alpha/beta (M5). As for study GSE43777, WGCNA (beta value of 6) was able to identify 6 significantly enriched modules out of a total of 12, giving it 2 more non-significantly enriched modules than CEMiTool. These subtle differences are likely to be

derived from the difference in the selected beta values and showcase CEMiTool's ability to produce results comparable to established tools such as WGCNA with greater ease and convenience.

We ran Petal using the same input genes utilized by WGCNA analysis (4000 most variant genes). Petal is a software which attempts to define a co-expression network using an automatically defined threshold to indicate similar expression between genes [10]. However, after 20 min for study GSE18090 and 40 min for study GSE43777, the program was unable to select any threshold for either study. This happened again when the filtered datasets from CEMiTool were attempted, albeit with lower runtimes (9 min for study GSE18090 and 4 min for study GSE43777). We encountered several other problems, such as confusing command line output; no output plots or complementary analyses; massive cluttering of user's workspace with no option to redirect the several output files; lack of user tutorial or vignette; and inconsistent naming schemes, resulting in an unpleasant user experience.

Other packages, such as CoXpress, DINGO and DiffCoEx were not considered for benchmarking since they analyze more than 2 groups of samples (Table 1). Given these results, we chose to focus the remainder of our benchmarking on the differences between CEMiTool and WGCNA.

The WGCNA method [5] receives an input “m x n” gene expression matrix, containing n samples under specific conditions and m genes, where each element in the matrix gives the expression of one gene in a particular sample. The correlation between each pair of genes is then transformed into an m x m adjacency matrix through an adjacency function. The adjacency matrix

may be signed or unsigned. In the former, correlations in the $[-1, 1]$ interval are scaled into the $[0, 1]$ interval, while in the latter, negative correlations are made positive. During the process, these values are then raised to a power of β , called the soft-threshold, which effectively adjusts how smoothly the connection strengths transition from their lowest to their highest values. The selection of β directly impacts on how adherent to the scale-free model the network will be. In general, the WGCNA authors recommend to use the “scale-free topology criterion” [5], in which the chosen β value is the one that leads the network’s topology to be, at least approximately, scale-free. Adherence to a scale-free topology is measured by a linear regression fit (R^2) that quantifies the extent to which the degree distribution of the genes in the network follows a power law. Thus, for WGCNA, the chosen β value is the lowest one with which an $R^2 > 0.85$ (or $R^2 > 0.8$ in the original paper [5]).

However, the selection of the best soft-threshold is relatively arbitrary and can differ from study to study. By looking at a plot showing R^2 values for each β ranging from 1 to 20, WGCNA users are required to

manually define the value of β by considering the trade-off between R^2 and connectivity - a higher β may make the network more scale-free, but also lowers the mean connectivity.

Despite the WGCNA authors have demonstrated that networks are relatively robust to the selection of the soft-thresholding parameter [5], a more rigorous framework for the selection of beta is still lacking, being usually defined visually by the user, hindering reproducibility and workflow automation. Although WGCNA provides a function named *pickSoftThreshold* that can automatically select the β value, we have created an alternative algorithm, which is based on the concept of Cauchy sequences [17], that improves the automatic selection of the β value, allowing for more reliable and consistent results (See Methods).

Briefly, our method investigates if all possible pairs of β values (in a certain range) possess a difference between their R^2 values within a pre-defined range ϵ , and selects the first beta value in this sequence to satisfy this property. Moreover, our algorithm allows for a lower threshold for R^2 ($R^2 > 0.8$) when compared to WGCNA default

Fig. 2 Automatic Selection of Beta parameter. **a** Beta parameters selected by WGCNA (red and brown bars) or CEMiTool (black bars) for 15 microarray studies using the same input genes. **b** $\beta \times R^2$ curve for 3 representative studies. Beta values selected by WGCNA (red lines) and by CEMiTool (black line) are shown

threshold ($R^2 > 0.85$) - which, in turn, allows for lower values of β . Once the β value is defined, the remaining steps for creating the modules follow the standard WGCNA procedure.

To benchmark the selection of β , we compared the method implemented in WGCNA (*pickSoftThreshold* function) with our algorithm (Additional file 1: Text) on 15 publicly available microarray studies. Using the same genes as input, we utilized three different methods for module identification: WGCNA's *pickSoftThreshold* function with R^2 values > 0.8 and > 0.85 (WGCNA's default), as well as CEMiTool's *cemito* function with $R^2 > 0.8$. Figure 2a shows the value of β for each implementation. With the exception of study GSE53441, the value of β returned by CEMiTool was always equal to or lower than the one returned by WGCNA.

It is worth mentioning that the soft-thresholding impacts not only the network's topology, but also its information content: the higher the β value, the lower its mean connectivity - since connection strengths in the adjacency matrix are bounded by $[0,1]$ [18]. Consequently, a trade-off between the network's connectivity and its adherence to a scale-free topology must be considered. Therefore, in the context of this work we consider lower β values to be of more interest than higher values, as long as their R^2 values are similar.

The difference between WGCNA and CEMiTool in selecting the β parameter can be largely explained by the lower R^2 threshold implemented in our tool (0.8 in CEMiTool versus 0.85 in WGCNA). We picked this lower R^2 threshold observing the WGCNA authors' original recommendation [5]. Also, CEMiTool utilizes a stringent algorithm, based on Cauchy sequences, to select the lowest β parameter that stabilizes the sequence (i.e. keeps its R^2 values within a pre-defined range), while

Fig. 3 Automating filtering of genes to CEMiTool. The effect of the filter p -value threshold on the enrichment of modules was tested by running CEMiTool on 300 studies. **a** The combined enrichment score (CES, see Additional file 1: Text) of the resultant modules for each study and for each filter p -value was calculated using Reactome Pathways as gene sets. The black line represents the mean CES of all 300 studies, while the green shaded area represents the 95% confidence interval from the mean. **b** The number of genes selected for CEMiTool for each filter p -value. The black line represents the mean number of all 300 studies, and the green shaded area represents the 95% confidence interval from the mean

Fig. 4 Relationship between the number of samples and ϕ . Different numbers of random samples from 3 studies (GSE43777, GSE18123 and GSE34205) were picked for CEMiTool analyses. The bold line represents the mean for 10 "sampling" sets. The shaded green area represents the 95% confidence intervals of the mean. The vertical grey line points to the minimum number of samples required to provide a network whose topology does not vary much with increasing sample number, as indicated by the ϕ parameter

keeping the R^2 above the threshold. When the same R^2 threshold (0.8) is applied, CEMiTool usually returns the same β parameter value as WGCNA's *pickSoftThreshold* function. In several cases, however, WGCNA returned an inappropriate β value of 1 (Fig. 2).

Input gene selection

Prior to identifying co-expression modules, it is recommended to filter input genes by either mean expression or variance, rather than by differential expression since this

would invalidate the scale-free topology assumption [19]. Nevertheless, the number of genes to be chosen is left undetermined, leading to arbitrary choices that might affect downstream analyses. We thus opted for a flexible, yet objective method of gene selection (Additional file 1: Text). Briefly, by modeling the variance of genes as an inverse gamma distribution, as suggested in [20], we can select genes based on a p -value (in our analyses, we set $p = 0.1$ as cutoff). For certain types of RNA-seq data normalizations, our method allows for a correction of

Fig. 5 CEMiTool applied to an RNA-seq study of patients with psoriasis. RNA-seq expression data (RPKM normalization) of lesional psoriatic and normal skin samples were download from the GEO database (accession number GSE54456). **a** CEMiTool interactive report showing the results of the main analyses using the optional annotation, pathways and protein-protein interaction files. On the main page, the most connected network hubs can be seen for each module. **b** Significantly enriched pathways for module M2. Metabolic processes such as ‘Extracellular matrix organization’, related to psoriasis, are enriched for module M2

the mean-variance dependency [21, 22] by modeling the expression data as a negative binomial distribution [22], and then performing the adequate Variance Stabilizing Transformation (VST) [23] (Additional file 1: Text). To remove potential noise, our package also removes by default the 25% genes with lowest mean expression across all samples prior to filtering.

In order to determine the most suitable default filtering parameters, we applied CEMiTool to 300 microarray studies obtained from the GEO (Gene Expression Omnibus [24, 25]) database using differing filter p -value thresholds, and assessed the biological significance of the resulting modules (Additional file 2: Table S1). This was determined by calculating the Combined Enrichment Score (CES) of the output modules with respect to the Reactome pathways (Fig. 3). Briefly, the CES allows us to condense the overall enrichment results into a single number - the lower this number is, the more enriched the modules are (Additional file 1: Text). As the filtering p -value increases from 0.05 to 0.3, the CES reaches a global minimum at $p \approx 0.1$, suggesting that the noise introduced by non-correlated genes outweighs the gain in information (Fig. 3). The filtering p -value is therefore set to 0.1 as a default, but is also easily adjustable by the user via the `filter_pval` argument to the `cemitool` function to allow the analysis to be more or less stringent, as needed.

Influence of the number of samples on the scale-free topology model fit

To assess the minimum optimal number of samples for analyses, we devised a quality control parameter for the $\beta \times R^2$ curve, ϕ . We define ϕ as the ratio of the area under the curve relative to the area of the rectangle made

by $\beta \times 1$, which is the highest possible value for R^2 . Higher values of ϕ mean that the topology of the network converges sharply to a scale-free degree distribution. To estimate the minimum number of samples that returns the highest ϕ value before reaching a plateau, we bootstrapped the number of samples for 3 microarray studies, selecting at first 5 random samples, and then incrementing the sample number by 5 at each step. CEMiTool was run 10 times at each step using default parameters. As shown in Fig. 4, the parameter ϕ tends to stabilize at around 20 samples (which is in accordance to previous findings [26]), indicating that the $\beta \times R^2$ curve, and thus network topology, should not vary so much in behavior starting at that sample number.

Application to RNA-seq datasets

We also ran CEMiTool on 8 RNAseq studies, 4 of which had been previously normalized by log2 CPM (GSE69015, GSE77926, GSE92754, GSE94855), 2 normalized by RPKM (GSE44183 and GSE54456), 1 by FPKM (GSE77564) and 1 only adjusted for fragment and length biases (GSE65540).

The study GSE54456 [27] has measured 174 transcriptomes of lesional psoriatic and normal skin samples. Among the 8 modules identified by CEMiTool (Fig. 5), the module M1 was enriched for immune system pathways, including interferon alpha signaling, which is known to be related to the disease [28, 29]. One notable hub gene for module M1 was S100A7A. Although this gene was not mentioned in the original publication [27], others have shown that the expression of S100A7A is upregulated in lesioned-skin psoriasis patients [30]. CEMiTool analyses also revealed a

module related to extracellular matrix organization and collagen formation (Fig. 5), suggesting that the expression of genes responsible for maintaining the structure of the skin may be coordinately altered by the disease.

CEMiTool applied to over 1000 publicly available microarray studies

To demonstrate that CEMiTool can be easily automated, we ran the package on 1094 microarray studies obtained from the GEO database. For each study,

Fig. 7 CEMiTool applied to study Dengue infection. **a** Expression data from two microarray studies of patients infected or not with Dengue virus were downloaded from GEO. CEMiTool was independently run on those studies using an annotation file, pathway gene set list and protein-protein interaction file. Selected results of the 4 CEMiTool analyses are displayed in panels **(b)** to **(e)**. **b** Gene Set Enrichment Analyses showing the module activity on each class of samples. **c** Profile plots of modules M4 (GSE18090) and M3 (GSE43777). The expression levels of individual genes from each module are shown as colored lines. The black line represents the mean expression of all genes inside the module. Samples are shown in the x-axis and colored by classes. **d** Over Representation Analysis of modules M4 (GSE18090) and M3 (GSE43777). Bar graphs shows the $-\log_{10}$ Adjusted P-value of the enrichment between genes in modules and gene sets from Reactome Pathway database. The vertical dashed grey line indicates an adjusted P-value of 0.01. **e** Gene networks of modules M4 (GSE18090) and M3 (GSE43777). The top ten most connected genes (hubs) are labeled and colored based on their "origin": if originally present in the CEMiTool module, they are colored blue; if inserted from the interactions file, they are colored red. The size of the node is proportional to its degree

we downloaded the authors' normalized data and ran the *cemito* function using the default parameters. Figure 6 shows the distribution of β values, and the number of modules and filtered genes selected for the analyses.

Almost 12,000 gene modules were identified by CEMiTool, containing in total over 2 million genes. The studies span hundreds of different biological conditions, including cancer, drug treatments, infectious diseases, and

inflammatory and neurological pathologies. The list of all studies can be found in Additional file 3: Table S2.

Applying CEMiTool to study dengue

To gain novel insights about immunity to infectious diseases, we ran the package on two publicly available microarray studies containing the blood transcriptome of patients infected or not with the Dengue virus (GEO accession numbers GSE18090 and GSE43777). We then

Fig. 8 Modular analysis of Leishmaniasis. **a** Gene Set Enrichment Analyses showing the module activity on each class of samples. “Healthy” = uninfected subjects; “SickBeforeTreat” = Leishmania-infected patients before treatment; “SickAfterTreat” = Leishmania-infected patients after treatment. **b** Over Representation Analysis of modules M7. Bar graphs shows the $-\log_{10}$ Adjusted P-value of the enrichment between genes in modules and gene sets from Reactome Pathway database. The vertical dashed grey line indicates an adjusted P-value of 0.01. **c** Gene networks of modules M3 and M9. The top ten most connected genes (hubs) are labeled and colored based on their “origin”: if originally present in the CEMiTool module, they are colored blue; if inserted from the interactions file, they are colored red. The size of the node is proportional to its degree

annotated samples using the phenotypes provided by the original authors: control (non-infected patients), DF (patients diagnosed as Dengue fever) and DHF (patients diagnosed as dengue hemorrhagic fever). Protein-protein interaction data from GeneMania [31] and gene sets from Reactome Pathways Database [32] were also used in the CEMiTool analyses. The results obtained by such analyses are partially displayed in Fig. 7. The over representation analysis shows that the package identified modules (M4 in GSE18090 and M3 in GSE43777) related to anti-viral immunity, such as interferon signaling and the ISG15 antiviral mechanism. Moreover, Gene Set Enrichment Analyses show that the activity of these modules is higher in DF or DHF compared to Control samples (Fig. 7). Also, the module network graph of study GSE43777 highlights important genes as network hubs, such as CCL2, (coding for chemokine C-C motif ligand 2), which is known to be associated with severe dengue and dengue shock syndrome [33]. However, CCL2 was not highlighted as a key gene for Dengue infection in the original article associated with the study GSE43777 [34].

Modular analyses of visceral leishmaniasis

Finally, we used CEMiTool to investigate the blood transcriptome of patients infected with visceral leishmaniasis (VL), a major public health problem in Brazil and worldwide. For this, we performed 17 RNA-seq experiments using the whole-blood obtained from 6 patients infected with *Leishmania infantum*, before and after treatment, as well as 5 uninfected healthy individuals (Additional file 1: Text). CEMiTool has generated 14 modules containing 1700 genes (Fig. 8 and Additional file 4: html report). Of those, modules M7 and M10 refer to interferon-mediated immune responses (IFN γ and type I IFN, respectively), which are well known to be involved in experimental models of leishmaniasis [35]. However, although

IFN γ response (M7) has been reported in clinical studies as well, little is known about the role of type I IFN (M10) in VL patients. In fact, type I IFN genes are typically elicited in viral infections, and not by protozoan parasites such as *Leishmania infantum*. Further experiments must be conducted to assess how the type I IFN genes may drive the functions of innate and adaptive immune cells during VL infection. Also, CEMiTool was able to unravel the dynamics of genes involved in B cell-mediated immunity during VL treatment, as shown by module M3 (Fig. 8). Integration with protein-protein interaction data revealed CD79A and CD79B as potential hubs in module M3 (Fig. 8). Both proteins form a dimer associated with the B-cell antigen receptor (BCR), and are critical for B cell immunity. The finding that CD79A and CD79B genes, as well as other members of the modules related to B cell development, are co-expressed and that the module activity is increased on VL treatment demonstrate that CEMiTool can provide new insights about the host response to treatment and to the disease.

Gene set enrichment analysis revealed novel insights about the molecular disturbances caused by the infection (Fig. 8). For instance, the transcriptional activity of module M4, which is enriched by genes associated with “platelet degranulation” and “hemostasis”, is significantly high in VL patients before receiving treatment (Additional file 4: html report). Indeed, VL is associated with several haematological manifestations, including anaemia, leucopenia, and disseminated intravascular coagulation [36]. In addition, the pattern of activity of module M9, which is associated with cell cycle (Additional file 4: html report and Fig. 8), suggests that an intense proliferation of cells expressing CCR5 (such as macrophages, dendritic cells and memory T cells) may be occurring during VL infection. Taken together, our findings may define which genes are driving these haematological manifestations, and thus suggesting effective drug treatments to VL.

Fig. 9 Hypothetical $R^2 \times \beta$ curves for two hypothetical studies. **a** A good choice for β parameter since the corresponding R^2 value is above a specified threshold (indicated by the blue line), and subsequent β values brings no additional significant increase in adherence to the scale-free model **(b)** A poor choice for β parameter since even though it's the smallest possible beta value above the given threshold, it presents significantly less adherence to the scale-free model than subsequent values of beta

Conclusions

Given the inherently modular profile of biological systems, gene co-expression networks have been extensively used in order to better understand how specific groups of genes are able to orchestrate the several different metabolic pathways present in organisms, as well as identify how they change in response to different conditions and diseases. CEMiTool can identify biologically relevant gene co-expression modules in an automated and easy-to-use way, as well as to perform a comprehensive set of analyses to better understand the biological functions present in the underlying system.

Methods

Soft-threshold selection and gene module discovery

Although WGCNA provides a function named *pickSoft-Threshold* that can automatically select the soft-threshold β value, we have created an alternative algorithm, which is based on the concept of Cauchy sequences [17], that improves the automatic selection of this parameter, allowing for more reliable and consistent results. Moreover, our algorithm allows a lower threshold for R^2 ($R^2 > 0.8$) when compared to WGCNA's default threshold ($R^2 > 0.85$). This, in turn, allows for lower values of β . Once a β value is chosen, subsequent steps for creating modules follow standard WGCNA procedure.

Assuming the reader is familiar with the language of Langfelder and Horvath [37], we define the notation $R^2(\beta)$ to denote the value of R^2 obtained for a given β . With the $R^2 \times \beta$ curve in hand, we should pick a threshold τ . WGCNA's method consists of taking the smallest β such that $R^2(\beta) > \tau$. Suppose that this value is β_1 , and it corresponds to $R^2(\beta_1) = \tau + \delta$ for some $\delta > 0$ (i.e., the equivalent R^2 value is only slightly above τ). Take now the next value, β_2 . If $R^2(\beta_2) \approx R^2(\beta_1)$, then WGCNA's choice was a good one, as there is little to be gained by raising the soft-threshold any further and much to be lost in terms of network connectivity – this case is shown in Fig. 9. This becomes the majority of cases when τ is close to one, but this comes at the price of unnecessarily high values of β when $R^2 \rightarrow 1$ in a slow fashion (or worse, it fails to select a soft-threshold).

On the other hand, selecting a lower value of τ leads to a loss of “scale-freeness” (indicated by R^2) if $R^2(\beta_2)$ is significantly larger than $R^2(\beta_1)$ – say, $R^2(\beta_2) = R^2(\beta_1) + \Delta$, for $\Delta > \delta$, as shown in Fig. 9. CEMiTool's algorithm avoids such cases by interrogating whether $|R^2(\beta_1) - R^2(\beta_2)| < \epsilon$ for a pre-defined value of ϵ – if not, then CEMiTool rejects β_1 as a soft-threshold and moves on, stopping when the curve appears to stabilize. This means that it exploits all possible significant gains of “scale-freeness” before settling on a value, which allows us to use a lower value of τ without settling for poor values of R^2 .

Availability

The CEMiTool package is available at Bioconductor (DOI: <https://doi.org/10.18129/B9.bioc.CEMiTool>) and can be downloaded using the command `biocLite` (“CEMiTool”) (package `BiocInstaller` v. $\geq 1.28.0$). A Docker image with an environment specifically tailored for CEMiTool analyses is also available at DockerHub (<https://hub.docker.com/r/csblusp/cemitool/>). RNA-seq data of Leishmania-infected patients have been deposited in the ArrayExpress database at EMBL-EBI under accession number E-MTAB-6137 (<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6137/>).

Additional files

Additional file 1: Text. Detailed description of methods. (DOCX 122 kb)

Additional file 2: Table S1. List of 300 microarray studies utilized in Fig. 3. (XLSX 99 kb)

Additional file 3: Table S2. List of 1000 microarray studies utilized in Fig. 6. (XLSX 37 kb)

Additional file 4: html report. CEMiTool output html file using the RNA-seq data of Leishmania-infected patients. (ZIP 14736 kb)

Abbreviations

BCR: B-cell antigen receptor; CCL2: Chemokine C-C motif ligand 2; CCR5: C-C chemokine receptor type 5; CD79: Cluster of differentiation 79; CEMiTool: Co-expression modules identification tool; CES: Combined enrichment score; CoP: Co-expressed biological processes; CPM: Counts per million; DF: Dengue fever; DHF: Dengue haemorrhagic fever; diceR: Diverse cluster ensemble in R; DINGO: Differential network analysis in genomics; fgsea: Fast gene set enrichment analysis; FPKM: Fragments per kilobase million; GEO: Gene expression omnibus; GSEA: Gene set enrichment analysis; IFN: Interferon; ORA: Over-representation analysis; RPKM: Reads per kilobase million; VL: Visceral leishmaniasis; VST: Variance stabilizing transformation; WGCNA: Weighted gene co-expression network analysis

Acknowledgements

We are grateful to members of the CSBL and LIB labs who have tested CEMiTool and the financial support of the funding agencies.

Funding

FAPESP grants No 2012/19278-6, 2013/08216-2, 2015/20897-0, 2017/05762-7, 2014/19323-7, 2014/24162-2, 2015/25825-8;
CNPq grants 443719/2014-4;
CAPES grants No. 1496888 and No. 1572872
FONDECYT-CONICYT grant 11161020
PAI-CONICYT grant PAI79170021

Availability of data and materials

Project name: CEMiTool
Project home page: <https://github.com/csbl-usp/CEMiTool>
Operating system(s): Platform independent
Programming language: R
License: GPL-3

Authors' contributions

Conceptualization: HIN. Data curation: PSTR, GRF, FMP, ML, RAC, VMC. Formal analysis: PSTR, GRF. Funding acquisition: HIN. Methodology: HIN, PSTR, GRF. Project Administration: HIN. Resources: HIN, JSS, VMC. Software: PSTR, GRF, LEC, MCB, DSL, TDCH, RAC. Supervision: HIN. Validation: PSTR, GRF, TDCH, DSL, RAC, SRM, KFF, JSS. Writing – original draft: HIN, PSTR, GRF. Writing – review & editing: PSTR, GRF, LEC, MCB, TDCH, DSL, FMP, RAC, ML, VMC, HIN. All authors read and approved the final manuscript.

Ethics approval and consent to participate

All procedures involving humans were approved by the Ethics Committee of the University Hospital of the Federal University of Sergipe and Comissão Nacional de Ética em Pesquisa, Brazil (CONEP).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, SP 05508-900, Brazil. ²Advanced Center for Chronic Diseases (ACCDiS), Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Santiago, Chile. ³Department of Biochemistry, Immunology, and Cell Biology, University of São Paulo, Ribeirão Preto, São Paulo, Brazil.

Received: 20 September 2017 Accepted: 7 February 2018

Published online: 20 February 2018

References

- Oldham MC, Horvath S, Geschwind DH. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci U S A*. 2006;103:17973–8. <https://doi.org/10.1073/pnas.0605938103>.
- Liu J, Jing L, Tu X. Weighted gene co-expression network analysis identifies specific modules and hub genes related to coronary artery disease. *BMC Cardiovasc Disord*. 2016;16:54. <https://doi.org/10.1186/s12872-016-0217-3>.
- Xue J, Schmidt SV, Sander J, Draffehn A, Krebs W, Quester I, et al. Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity*. 2014;40:274–88. <https://doi.org/10.1016/j.immuni.2014.01.006>.
- Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet*. 2004;5:101–13. <https://doi.org/10.1038/nrg1272>.
- Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005;4:Article17. <https://doi.org/10.2202/1544-6115.1128>.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102:15545–50. <https://doi.org/10.1073/pnas.0506580102>.
- Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. *Bioinformatics*. 2008;24:719–20. <https://doi.org/10.1093/bioinformatics/btm563>.
- Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16:284–7. <https://doi.org/10.1089/omi.2011.0118>.
- Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv*. 2016. <https://doi.org/10.1101/060012>.
- Petereit J, Smith S, Harris FC, Schlauch KA. Petal: co-expression network modelling in R. *BMC Syst Biol*. 2016;10(Suppl 2):51. <https://doi.org/10.1186/s12918-016-0298-8>.
- Ogata Y, Suzuki H, Sakurai N, Shibata D. CoP: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics*. 2010;26:1267–8. <https://doi.org/10.1093/bioinformatics/btq121>.
- Desai AP, Rameghin M, Meruvia-Pastor O, Peña-Castillo L. GeNET: a web application to explore and share gene co-expression network analysis data. *Peer J*. 2017;5:e3678. <https://doi.org/10.7717/peerj.3678>.
- Tesson BM, Breitling R, Jansen RC. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*. 2010;11:497. <https://doi.org/10.1186/1471-2105-11-497>.
- Watson M. CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics*. 2006;7:509. <https://doi.org/10.1186/1471-2105-7-509>.
- Chiu DS, Talhouk A. diceR: an R package for class discovery using an ensemble driven approach. *BMC Bioinformatics*. 2018;19:11. <https://doi.org/10.1186/s12859-017-1996-y>.
- Ha MJ, Baladandayuthapani V, Do K-A. DINGO: differential network analysis in genomics. *Bioinformatics*. 2015;31:3413–20. <https://doi.org/10.1093/bioinformatics/btv406>.
- Apostol TM. *Calculus*. 1st ed. New York: Wiley; 1991. p. 1967–9.
- Horvath S. *Weighted network analysis: application in genomics and systems biology*. Hardcover; 2011-04-01. New York: Springer; 2011.
- Langfelder P, Horvath S. WGCNA package: Frequently Asked Questions. 2014. <https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/faq.html>. Accessed 11 Aug 2017.
- Wright GW, Simon RM. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*. 2003;19:2448–55.
- Subramaniam S, Hsiao G. Gene-expression measurement: variance-modeling considerations for robust data analysis. *Nat Immunol*. 2012;13:199–203. <https://doi.org/10.1038/ni.2244>.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139–40. <https://doi.org/10.1093/bioinformatics/btp616>.
- Yu G. Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. *Stat Probab Lett*. 2009;79:1621–9. <https://doi.org/10.1016/j.spl.2009.04.010>.
- Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30:207–10. <https://doi.org/10.1093/nar/30.1.207>.
- Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2013;41(Database issue):D991–5. <https://doi.org/10.1093/nar/gks1193>.
- Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*. 2015;31:2123–30. <https://doi.org/10.1093/bioinformatics/btv118>.
- Tsoi LC, Iyer MK, Stuart PE, Swindell WR, Gudjonsson JE, Tejasvi T, et al. Analysis of long non-coding RNAs highlights tissue-specific expression patterns and epigenetic profiles in normal and psoriatic skin. *Genome Biol*. 2015;16:24. <https://doi.org/10.1186/s13059-014-0570-4>.
- Kim G-W, Jwa S-W, Song M, Kim H-S, Kim B-S, Kim M-B, et al. Extensive psoriasis induced by pegylated interferon alfa-2a and ribavirin in the treatment of chronic hepatitis C. *Ann Dermatol*. 2013;25:479–82. <https://doi.org/10.5021/ad.2013.25.4.479>.
- Afshar M, Martinez AD, Gallo RL, Hata TR. Induction and exacerbation of psoriasis with interferon-alpha therapy for hepatitis C: a review and analysis of 36 cases. *J Eur Acad Dermatol Venereol*. 2013;27:771–8. <https://doi.org/10.1111/j.1468-3083.2012.04582.x>.
- Keermann M, Köks S, Reimann E, Prans E, Abram K, Kingo K. Transcriptional landscape of psoriasis identifies the involvement of IL36 and IL36RN. *BMC Genomics*. 2015;16:322. <https://doi.org/10.1186/s12864-015-1508-2>.
- Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*. 2010;38(Web Server issue):W214–20. <https://doi.org/10.1093/nar/gkq537>.
- Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. 2011;39(Database issue):D691–7. <https://doi.org/10.1093/nar/gkq1018>.
- Lee Y-R, Liu M-T, Lei H-Y, Liu C-C, Wu J-M, Tung Y-C, et al. MCP-1, a highly expressed chemokine in dengue haemorrhagic fever/dengue shock syndrome patients, may cause permeability change, possibly through reduced tight junctions of vascular endothelium cells. *J Gen Virol*. 2006;87(Pt 12):3623–30. <https://doi.org/10.1099/vir.0.82093-0>.
- Sun P, García J, Comach G, Vahey MT, Wang Z, Forshey BM, et al. Sequential waves of gene expression in patients with clinically defined dengue illnesses reveal subtle disease phases and predict disease severity. *PLoS Negl Trop Dis*. 2013;7:e2298. <https://doi.org/10.1371/journal.pntd.0002298>.
- Silva-Barrios S, Stäger S. Protozoan parasites and type I ifns. *Front Immunol*. 2017;8:14. <https://doi.org/10.3389/fimmu.2017.00014>.
- Varma N, Naseem S. Hematologic changes in visceral leishmaniasis/kala azar. *Indian J Hematol Blood Transfus*. 2010;26:78–82. <https://doi.org/10.1007/s12288-010-0027-1>.
- Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559. <https://doi.org/10.1186/1471-2105-9-559>.

webCEMiTool: Co-expression Modular Analysis Made Easy

Lucas E. Cardozo¹, Pedro S. T. Russo¹, Bruno Gomes-Correia², Mariana Araujo-Pereira¹, Gonzalo Sepúlveda-Hermosilla³, Vinicius Maracaja-Coutinho² and Helder I. Nakaya^{1*}

¹Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil, ²Advanced Center for Chronic Diseases-ACCDiS, Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Santiago, Chile, ³Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile

Co-expression analysis has been widely used to elucidate the functional architecture of genes under different biological processes. Such analysis, however, requires substantial knowledge about programming languages and/or bioinformatics skills. We present webCEMiTool,¹ a unique online tool that performs comprehensive modular analyses in a fully automated manner. The webCEMiTool not only identifies co-expression gene modules but also performs several functional analyses on them. In addition, webCEMiTool integrates transcriptomic data with interactome information (i.e., protein-protein interactions) and identifies potential hubs on each network. The tool generates user-friendly html reports that allow users to search for specific genes in each module, as well as check if a module contains genes overrepresented in specific pathways or altered in a specific sample phenotype. We used webCEMiTool to perform a modular analysis of single-cell RNA-seq data of human cells infected with either Zika virus or dengue virus.

Keywords: co-expression analysis, systems biology, transcriptomics, web tool, data integration

INTRODUCTION

Cellular processes are driven by multiple interacting molecules whose activity level must be dynamically regulated (Kitano, 2002). As a result, genes belonging to the same signaling and metabolic pathway or sharing similar functions will tend to be co-expressed across conditions (Wang et al., 2016). Co-expression gene module analysis creates networks comprising sets of genes (i.e., modules) whose expression is highly correlated. Such analysis was applied to reveal functional modules related to infectious (Janova et al., 2015), inflammatory (Beins et al., 2016), and neurological (Voineagu et al., 2011) diseases, as well as several types of cancer (Sharma et al., 2017).

Weighted gene co-expression network analysis (WGCNA) is a widely used method to identify co-expressed gene modules (Zhang and Horvath, 2005). In order to run WGCNA, however, users are required to be familiar to programming environments, as well as to manually select parameters. These features prevent researchers with insufficient knowledge of R to identify gene modules from transcriptome data sets.

¹<https://cemitool.sysbio.tools/>

OPEN ACCESS

Edited by:

Akira Funahashi,
Keio University,
Japan

Reviewed by:

Takahiro G. Yamada,
Keio University, Japan
Marco Vanoni,
Università degli Studi di Milano
Bicocca, Italy

*Correspondence:

Helder I. Nakaya
hnakaya@usp.br

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 29 August 2018

Accepted: 12 February 2019

Published: 06 March 2019

Citation:

Cardozo LE, Russo PST,
Gomes-Correia B, Araujo-Pereira M,
Sepúlveda-Hermosilla G,
Maracaja-Coutinho V and Nakaya HI
(2019) webCEMiTool: Co-expression
Modular Analysis Made Easy.
Front. Genet. 10:146.
doi: 10.3389/fgene.2019.00146

Based on our Bioconductor R package named CEMiTool (Russo et al., 2018), we developed a user-friendly web-based application that allows scientists with no background in bioinformatics to perform comprehensive co-expression network analysis.

MATERIALS AND METHODS

The web interface of webCEMiTool was developed to allow users to quickly generate comprehensive analyses without the need of installing any specific program or internet browser. The only requirement for running the modular analysis is a data set containing the expression levels of all genes in samples under different biological conditions (herein defined as “classes”).

There is no defined range number of samples but our previous study suggests a minimum of 15 samples per data set (Russo et al., 2018). Although it was primarily designed for transcriptome data (i.e., RNA-seq or microarrays), it can also be potentially used for identifying modules of proteins, cytokines, and even metabolites. webCEMiTool will then automatically select the input genes and identify the co-expression modules. Each module contains a set of genes whose expression follows a similar pattern.

We implemented, within webCEMiTool, a feature that assesses the activity of gene modules on each class of samples. For this, the users only have to provide a sample annotation tab-delimited text file that informs the class of each sample. A “profile plot” showing the median level of individual genes within the module is then displayed in the “Results” section of the tool (Figure 1A).

FIGURE 2 | webCEMiTool applied to single-cell RNA-seq data. **(A)** Profile plot of co-expressed gene modules. We selected one representative module for each time point post-dengue virus infection (left) or post-Zika virus infection (right). The black line represents the median expression activity of genes from the modules across each sample. The colors represent the different amount of virus RNA within the cell. **(B)** Overrepresentation analysis of selected modules at 24 h post-virus infection. The bar graphs were adapted from the Enrichr webtool linked to webCEMiTool. The bars are proportional to the $-\log_{10}$ adjusted p -value (Benjamini-Hochberg) of the enriched pathways in a module.

To enable functional analysis, the users can also check if the gene modules are associated with specific signaling or metabolic pathways (**Figure 1B**). These pathways can easily be extracted from databases, such as KEGG, Reactome, and MySigDB. Finally, users can integrate the results with interactome data (i.e., protein-protein interactions, transcription factors and their transcribed genes, or even miRNAs and their target genes). This feature enables users to identify critical regulators of modules (**Figure 1C**), providing valuable insights for experimental validation or potential targets for drugs. Additional details on how to obtain the optional files can be found in the “Tutorial” page of the website.²

To demonstrate that our method is robust, we performed an unprecedented large-scale modular analysis with over 1,000 publicly available RNA-seq and microarray data sets and new RNA-seq data of patients infected with *Leishmania* using the CEMiTool R package version (Russo et al., 2018). Although webCEMiTool and the package have distinct visualization features and are based

on different platforms, the core co-expression functionality is essentially the same. The online tool we are describing here is built to enable easy access to gene modular analyses for non-programming researchers, while the R library version is geared towards users with greater knowledge of the R programming language. Additionally, the results dashboard is composed of interactive charts that facilitate interpretation. Moreover, taking advantage of the rising ecosystem of bioinformatics web services, our tool establishes an interface with the Enrichr platform (Chen et al., 2013), enabling a richer experience for our users.

RESULTS

We demonstrated that webCEMiTool can be applied to analyze expression data at the single cell level. Publicly available viscRNA-Seq data (virus-including single cell RNA-Seq) were obtained from NCBI GEO database (accession number GSE110496) and used as input for the analysis. The data refer

²<https://cemitool.sysbio.tools/tutorial>

to the transcriptome of individual human hepatoma (Huh7) cells, which were infected with either dengue virus (DENV) or Zika virus (ZIKV), using multiplicity of infection (MOI) 0, 1, or 10 (Zanini et al., 2018). Cells collected on four different time points (4, 12, 24 and 48 h after infection) were then sorted for single cell transcriptomic analysis with an adapted Smart-seq2 protocol (Zanini et al., 2018). The DENV data set comprises 933 infected cells (MOI = 1 or 10) and 303 controls (MOI = 0), while the ZIKV data set is composed of 488 infected cells (MOI = 1) and 403 controls. Before submitting the analysis to the webCEMiTool platform, both data sets were log10 transformed and genes that were not expressed in more than 80% of the samples were removed. The data sets were then split by virus and by time point and used as input ("Expression file" field) to webCEMiTool. In addition to the gene expression data, we also provided to webCEMiTool the sample phenotypes (i.e., viral loads) and Reactome gene sets.

Our webCEMiTool analyses generated an average of six modules per time point in DENV infection and more than eight modules per time point in ZIKV infection. We have selected one module per time point as a representative of our findings (Figure 2A). It is clear that at 24 and 48 h post-infection, the expression activity of representative modules increases according to the viral load (Figure 2A). We next performed the pathway enrichment analysis of the representative modules at 24 h post-infection using the webCEMiTool link for Enrichr (Figure 2B). These findings not only corroborate what was described in the original publication (Zanini et al., 2018) but also provide new insights about the physiopathology of dengue and Zika virus infections.

DISCUSSION

Although few similar web-based applications were developed to perform co-expression gene analysis (Tzfadia et al., 2016; Desai et al., 2017), these tools do not provide comparable results to webCEMiTool. One such application is GeNET (Desai et al., 2017). This webtool was designed to facilitate gene co-expression analyses and provides enrichment analysis and gene-to-gene networks. However, it only performs these analyses

for three organisms (*R. capsulatus*, *M. tuberculosis*, and *O. sativa*). Another example is CoExpNetViz (Tzfadia et al., 2016), a webtool designed for the visualization and construction of gene networks. Similar to GeNET, CoExpNetViz is somewhat limited with respect to the organisms as it is stated to be primarily designed for plant transcriptomes. The webCEMiTool aims to provide co-expression analyses for any organism. Moreover, although CoExpNetViz is presented as a web-based application, its results are returned to users as a compressed folder containing a README.txt file with instructions on how to visualize their results on the Cytoscape app. The users have then to manually insert into Cytoscape the several different output files provided by the tool. These additional steps can also make the process error-prone and possibly daunting to users unfamiliar with Cytoscape. The webCEMiTool offers much more convenient browser-displayed results.

We also showed that webCEMiTool is able to analyze single-cell RNA-seq data faster and efficiently. Our results returned relevant information about the biological processes involved with dengue and Zika virus infection. All this analysis were performed in an automated and practical manner, with no need for the user to have deep understanding on the internal processing of gene co-expression data analysis.

AUTHOR CONTRIBUTIONS

LC, PR, BG-C, and MA-P performed the analyses. LC, GS-H, and VM-C developed the webtool. HN conceived the tool and supervised the work. All authors help in the writing of the paper.

FUNDING

This work was supported by grants from FAPESP (2012/19278-6, 2013/08216-2, 2017/05762-7, 2018/10748-6); CNPq (313662/2017-7); FONDECYT-CONICYT (11161020); and PAI-CONICYT (PAI79170021). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

REFERENCES

- Beins, E., Ulas, T., Ternes, S., Neumann, H., Schultze, J., and Zimmer, A. (2016). Characterization of inflammatory markers and transcriptome profiles of differentially activated embryonic stem cell-derived microglia. *Glia* 64, 1007–1020. doi: 10.1002/glia.22979
- Chen, E., Tan, C., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., et al. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf.* 14:128. doi: 10.1186/1471-2105-14-128
- Desai, A., Razeghin, M., Meruvia-Pastor, O., and Peña-Castillo, L. (2017). GeNET: a web application to explore and share Gene Co-expression network analysis data. *PeerJ*. 5:e3678. doi: 10.7717/peerj.3678
- Janova, H., Böttcher, C., Holtman, I., Regen, T., van Rossum, D., Götz, A., et al. (2015). CD14 is a key organizer of microglial responses to CNS infection and injury. *Glia* 64, 635–649. doi: 10.1002/glia.22955
- Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664. doi: 10.1126/science.1069492
- Russo, P., Ferreira, G., Cardozo, L., Bürger, M., Arias-Carrasco, R., Maruyama, S., et al. (2018). CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinf.* 19:56. doi: 10.1186/s12859-018-2053-1
- Sharma, A., Cinti, C., and Capobianco, E. (2017). Multitype network-guided target controllability in phenotypically characterized osteosarcoma: role of tumor microenvironment. *Front. Immunol.* 8:928. doi: 10.3389/fimmu.2017.00918
- Tzfadia, O., Diels, T., De Meyer, S., Vandepoele, K., Aharoni, A., and Van de Peer, Y. (2016). CoExpNetViz: comparative co-expression networks construction and visualization tool. *Front. Plant Sci.* 6:1194. doi: 10.3389/fpls.2015.01194
- Voineagu, I., Wang, X., Johnston, P., Lowe, J., Tian, Y., Horvath, S., et al. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474, 380–384. doi: 10.1038/nature10110
- Wang, J., Xia, S., Arand, B., Zhu, H., Machiraju, R., Huang, K., et al. (2016). Single-cell co-expression analysis reveals distinct functional modules, co-regulation mechanisms and clinical outcomes. *PLoS Comput. Biol.* 12:e1004892. doi: 10.1371/journal.pcbi.1004892
- Zanini, F., Szu-Yuan, P., Bekerman, E., Einav, S., and Quake, S. R. (2018). Single-cell transcriptional dynamics of flavivirus infection. *Elife* 7:e32942. doi: 10.7554/eLife.32942

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:17. doi: 10.2202/1544-6115.1128

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Cardozo, Russo, Gomes-Correia, Araujo-Pereira, Sepúlveda-Hermosilla, Maracaja-Coutinho and Nakaya. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.