UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

LARA MARINELLI DATIVO DOS SANTOS

Análise de agrupamento de dados de expressão gênica e sua aplicação para o entendimento da relação entre progesterona e diabetes gestacional

São Paulo
2022

LARA MARINELLI DATIVO DOS SANTOS

Análise de agrupamento de dados de expressão gênica e sua aplicação para o entendimento da relação entre progesterona e diabetes gestacional

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Orientador: Profa. Dra. Patrícia Rufino Oliveira

São Paulo

2022

Ficha catalográfica elaborada pela Biblioteca da Escola de Artes, Ciências e Humanidades,
com os dados inseridos pelo(a) autor(a)
Brenda Fontes Malheiros de Castro CRB 8-7012; Sandra Tokarevicz CRB 8-4936

Marinelli Dativo dos Santos, Lara
Análise de agrupamento de dados de expressão gênica e sua aplicação para o entendimento da relação entre progesterona e diabetes gestacional / Lara Marinelli Dativo dos Santos; orientadora, Patrícia Rufino Oliveira. -- São Paulo, 2022.
116 p: il.

Dissertacao (Mestrado em Ciencias) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2022.
Versão corrigida

1. Agrupamento. 2. Dados de expressão gênica. 3. Análise de enriquecimento funcional. I. Oliveira, Patrícia Rufino, orient. II. Título.

Dissertação de autoria de Lara Marinelli Dativo dos Santos, sob o título **"Análise de agrupamento de dados de expressão gênica e sua aplicação para o entendimento da relação entre progesterona e diabetes gestacional"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 24 de outubro de 2022 pela comissão julgadora constituída pelos doutores:

Profa. Dra. Patrícia Rufino Oliveira
EACH/USP
Presidente

Prof. Dr. Fabrício Martins Lopes
UTFPR

Prof. Dr. David Corrêa Martins Junior
UFABC

Prof. Dr. Ana Carolina Lorena
ITA

∞

Aos ancestrais.

♡

Agradecimentos

A YHWH.

Aos meus pais, José (em memória) e Rita e ao meu irmão, Afonso.

À minha vó, Anésia.

À minha tia, Cristina.

Ao meu tio, Antônio (em memória).

À minha madrinha Edy (em memória) e ao meu padrinho Gerson.

Aos Narizinho, Princesa, Laica, Toretto, Barbinha I, II, Pita, Pepe, Leila, Giovanino, Penélope, Ratãozinho, Lili, Mumuzinho, Matilda, Bino, Benjamin, Frido, Jubileu, Cacatua, Flora, Cora, Sabrina, Cléo, Antônio, Alice, e aos demais.

Ao Thiago, Felipe, Josiane, Vinícius, Neemias, Horst, André, Rodolfo, Adriano, Danielli, Williane, Walkíria, Raquel, Isabel, Zenilda, pelas dicas.

À minha orientadora, Profa. Dra. Patrícia Rufino Oliveira, por toda a orientação, pelas correções, ensinamentos, e por todas as valiosas contribuições.

À Profa. Dra. Anna Karenina de Azevedo Martins, pelos ensinamentos, pela disponibilização do problema e pela auto-disponibilização a explicar conceitos, enfim, pela coorientação ainda que informal.

Aos Profs. Drs. Esteban Fernandes Tuesta e Clodoaldo Aparecido Moraes Lima, por suas valiosas contribuições durante o exame de qualificação, e aos Profs. Drs. Ana Carolina Lorena, Fabrício Martins Lopes e David Corrêa Martins Junior pelas contribuições durante a defesa.

Aos Profs. Drs. Marcelo Medeiros Eler, presente coordenador do PPgSI, e Karina Valdivia Delgado, ex-coordenadora, por terem/estarem conferindo ao programa uma postura especialmente humana diante da pandemia.

À Superintendência de Tecnologia da Informação da Universidade de São Paulo pela disponibilização dos recursos de HPC (Computação de Alto Desempenho) sem os quais esta pesquisa não poderia avançar.

À Humanidade.

Ao Universo.

E a quem nos lê.

"Porque, agora, vemos por espelho em enigma; mas, então, veremos face a face..."

(1 Coríntios 13:12)

Resumo

SANTOS, Lara Marinelli Dativo dos. **Análise de agrupamento de dados de expressão gênica e sua aplicação para o entendimento da relação entre progesterona e diabetes gestacional** 2022. 115 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2022.

Devido ao crescente uso farmacológico de progestógenos ao longo da gravidez para a prevenção do parto prematuro, a relação entre esses hormônios e o diabetes gestacional requer atenção. A morte de células beta-pancreáticas está associada aos diabetes tipo I e tipo II, mas ainda precisa ser melhor compreendida no contexto do diabetes gestacional. Para investigar este problema, experimentos de *microarray* foram conduzidos com células da linhagem RINm5F submetidas à progesterona em três doses (0,1 µM, 1 µM e 100 µM) e dois tempos (6h e 24h). A análise de agrupamento em dados de expressão gênica é amplamente utilizada para auxiliar no entendimento da função dos genes e tem sido consistentemente aplicada na literatura. No entanto, diante da variedade de técnicas de agrupamento existentes, a escolha daquela mais adequada a determinado problema torna-se um desafio. O estudo desenvolvido por Saelens, Cannoodt e Saeys, em 2018, buscou preencher tal lacuna avaliando, segundo índices de validação externa e com base em módulos conhecidos, vários métodos de agrupamento e propondo uma metodologia para a realização de estudos comparativos que envolvam detecção de módulos em dados de expressão gênica. No entanto, em cenários do mundo real, o pesquisador muitas vezes só tem à sua disposição os índices de validação internos, não dispondo de módulos conhecidos a respeito dos genes, como no caso do presente trabalho. Desta forma, para proceder com a análise de agrupamento dos dados das células beta-pancreáticas submetidas à progesterona, fez-se necessário estender o estudo mencionado para incluir índices de avaliação internos, de modo a selecionar a técnica e os parâmetros adequados ao problema. Ao fim dos experimentos, foram selecionados para a análise de enriquecimento funcional os resultados de agrupamento de acordo com as pontuações baseadas nos índices de validação externos e internos. Do ponto de vista da compreensão do problema, o resultado mais significativo foi aquele obtido por meio de índices internos, revelando que o gene TXNIP é relevante para a compreensão do diabetes gestacional.

Palavras-chaves: Agrupamento. Dados de expressão gênica. *Microarrays*. Análise de enriquecimento funcional. *Benchmark*.

Abstract

SANTOS, Lara Marinelli Dativo dos. **Clustering analysis of gene expression data and its application to understanding the relationship between progesterone and gestational diabetes** 2022. 115 f. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2022.

Due to the increasing pharmacological use of progestins throughout pregnancy for the prevention of preterm birth, the relationship between these hormones and gestational diabetes requires attention. Pancreatic beta cell death is associated with both type I and type II diabetes, but still needs to be better understood in the context of gestational diabetes. To investigate this problem, microarray experiments were conducted with RINm5F cells subjected to progesterone in three doses (0.1 µM, 1 µM and 100 µM) and two times (6h and 24h). Cluster analysis on gene expression data is widely used to help understand the function of genes and has been consistently applied in the literature. However, given the variety of existing clustering techniques, choosing the most appropriate one for a given problem becomes a challenge. The study developed by Saelens, Cannoodt and Saeys, in 2018, sought to fill this gap by evaluating, according to external validation indices and based on known modules, various clustering methods and proposing a methodology for carrying out comparative studies involving the detection of modules in gene expression data. However, in real world scenarios, the researcher often only has at his disposal the internal validation indices, not having known modules about the genes, as in the case of the present work. Thus, in order to proceed with the cluster analysis of data from pancreatic beta cells submitted to progesterone, it was necessary to extend the aforementioned study to include internal evaluation indices, in order to select the technique and parameters appropriate to the problem. At the end of the experiments, grouping results according to scores based on external and internal validation indices were selected for the functional enrichment analysis. From the point of view of understanding the problem, the most significant result was obtained through internal indices, revealing that the TXNIP gene is relevant for understanding gestational diabetes.

Keywords: Clustering. Gene expression data. Microarrays. Functional enrichment analysis. Benchmark.

Lista de figuras

Lista de algoritmos

Lista de quadros

Lista de tabelas

Sumário

1 Introdução

O Consórcio Internacional de Sequenciamento do Genoma Humano anunciou, em 2004, a finalização do sequenciamento do genoma humano, a primeira dentre os vertebrados. Essa finalização foi fruto de um trabalho colaborativo para converter o rascunho divulgado em 2001 em um sequenciamento com alta acurácia e cobertura próximo de completa, como se pode ver nos trabalhos de Consortium *et al.* (2004) e Mcpherson *et al.* (2001). Tal avanço permitiu que os pesquisadores pudessem investigar funções celulares com base em um panorama sólido e completo do genoma, dispensando a imposição de trabalhos exaustivos para confirmar algumas hipóteses, e possibilitou que estudos mais sofisticados fossem realizados, como por exemplo comparações com genomas de outros mamíferos, identificação de elementos funcionais, de controles regulatórios e de interações entre genes. O conhecimento que é construído a partir desses estudos é compartilhado pela comunidade científica. Uma maneira de se organizar e compartilhar o conhecimento são as ontologias.

Segundo Studer, Benjamins e Fensel (1998), uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada, que deve ser interpretável por computador e aceita por um grupo ou comunidade na área do conhecimento por ela modelado. Nesse contexto, a *Gene Ontology* (GO)[1] é uma ontologia em que o conhecimento a respeito dos genes é especificado, formalizado e compartilhado, disponibilizando um vocabulário unificado e estruturado para a descrição dos genes e de seus produtos nos organismos (HENNIG; GROTH; LEHRACH, 2003).

Para a investigação de novas informações a respeito das funções dos genes, existem diversas abordagens. Na abordagem clássica da genética, por exemplo, uma das maneiras para identificar a função de um gene é analisar o que acontece com o organismo na sua ausência, por meio da análise de organismos mutantes com aparência interessante ou incomum, como por exemplo, moscas da fruta com olhos brancos ou asas curvadas. Desta maneira, é feita uma análise partindo-se do fenótipo - aparência ou comportamento do indivíduo - para entender como a modificação no genótipo expressa tais características. Um outra forma de análise para a descoberta de funções de genes é o rastreamento genético, método em que indivíduos são analisados em busca de mutações de interesse (ALBERTS, 2017).

[1] http://www.geneontology.org/

Por outro lado, o sequenciamento do genoma humano possibilitou também o desenvolvimento de novas tecnologias e avanços metodológicos para a análise das informações dos genes (MOCELLIN; ROSSI, 2007), (BERGER; PENG; SINGH, 2013). Dentre essas tecnologias destacam-se os *microarrays* ou *chips* de DNA, que, diferentemente dos métodos clássicos, permitem a análise simultânea da expressão de diversos genes. De acordo com Baldi e Brunak (2001), os *microarrays* possibilitam o estudo de padrões de expressão gênica em determinado tipo de célula, em determinado tempo e sob determinado conjunto de condições. Nesses *arrays*, o RNA é transcrito para criar um DNA complementar (cDNA), que é combinado com uma ampla biblioteca de fragmentos de genes previamente distribuídos em uma placa de vidro ou suporte. Em seguida, técnicas são utilizadas para medir as expressões de dezenas de milhares de genes sob diversas condições experimentais por meio da interpretação da fluorescência dos genes com equipamento específico (AHMED, 2002). Denomina-se expressão gênica o processo em que a informação contida no gene se transforma em um produto, em sua maioria proteínas, a partir do processo de transcrição e tradução do DNA. Nos *microarrays* a expressão é medida relativamente pelo número de cópias que o gene realiza em determinado fenômeno. Como são os produtos do gene que efetivamente constituem o fenótipo dos indivíduos, entende-se que a atividade do gene pode ser medida pela quantidade de cópias que o mesmo expressou em determinado contexto. A utilização desses *arrays* viabiliza a coleta de dados potencialmente capazes de dar *insights* fundamentais sobre processos biológicos que vão desde a significância do gene para um determinado fenômeno (MAJI; SHAH, 2017) até o desenvolvimento de novos medicamentos.

Para compreender um fenômeno do ponto de vista genômico é fundamental entender a função e a contribuição dos genes nesse contexto, uma vez que um mesmo gene pode desempenhar funções diferentes em situações diferentes. Normalmente, os experimentos com *microarrays* possuem uma quantidade considerável de genes, sendo que nem todos estão ligados ao fenômeno em questão. Em análises manuais de dados de *microarrays* o pesquisador levanta a hipótese de que determinado gene participa e de como participa do fenômeno estudado, em função da variação de sua expressão, e realiza novos experimentos para verificar a hipótese. Esse processo geralmente tem altos custos, pois os experimentos com *microarrays* são de difícil execução e as tentativas de validação são, portanto, limitadas. Utilizar o apoio computacional para detectar padrões imperceptíveis a olho nu para o levantamento de hipóteses sobre os genes mais relevantes em determinado processo é

uma alternativa que tem sido amplamente adotada. Além de permitir a detecção de padrões, outra vantagem da abordagem computacional é a de tornar possível que os próximos experimentos *in vitro* sejam realizados de maneira mais assertiva do que quando comparados às análises feitas a olho nu.

Nesse contexto, a análise de agrupamento em dados de expressão gênica é amplamente utilizada para auxiliar no entendimento da função dos genes, regulação gênica, processos e subtipos celulares e tem sido consistentemente aplicada com a finalidade de identificar e analisar diversas patologias, como câncer, malária e tuberculose (DALTON; BALLARIN; BRUN, 2009). Tal análise permite auxiliar na identificação da função de genes a partir de um princípio, conhecido na literatura como *guilt by association* (OLIVER, 2000) (EISEN *et al.*, 1998), que estabelece que genes com funções semelhantes tendem a ser agrupados juntos. O fato de determinados genes terem sido alocados em um mesmo grupo, que no contexto de expressão gênica é chamado de módulo, pode indicar que tais genes sejam relacionados aos mesmos processos celulares, que apresentem corregulação ou que compartilhem um mecanismo em comum. Além disso, os grupos detectados tendem a ser significativamente enriquecidos com categorias funcionais específicas, fato que pode ser explorado para inferir a função de genes em um determinado contexto (JIANG; TANG; ZHANG, 2004), (DALTON; BALLARIN; BRUN, 2009). O agrupamento de dados de expressão gênica pode ser validado de duas principais maneiras: (i) índices internos, que mensuram o quão distintos ou bem separados os grupos identificados são; e (ii) índices externos, baseados na concordância entre os grupos obtidos e um agrupamento de referência (JIANG; TANG; ZHANG, 2004) (OYELADE *et al.*, 2016).

Devido à existência de um grande número de técnicas de agrupamento, a escolha do método mais adequado a uma dada aplicação torna-se um desafio. Em uma tentativa de mitigar esse problema, o trabalho em (SAELENS; CANNOODT; SAEYS, 2018) apresenta uma visão geral das características e desempenho de técnicas de agrupamento aplicadas a dados de expressão gênica e propõe uma estratégia de *benchmark* para a realização de estudos comparativos. No estudo comparativo foram analisados 49 métodos aplicados a nove conjuntos de dados de expressão gênica. Os métodos foram avaliados somente segundo índices externos, de maneira que se atribuiu uma pontuação para cada método de acordo com a concordância entre os módulos identificados por meio do agrupamento e os módulos de referência estabelecidos com base em redes regulatórias. De acordo com o esquema de pontuação desenvolvido pelos autores, o método que obteve o melhor desempenho foi

aquele baseado na análise de componentes independentes (ICA). Vários estudos posteriores, como aqueles feitos em (LAWLOR; CAO; ELLISON, 2021), Chen *et al.* (2020), Tan *et al.* (2020), Sastry *et al.* (2021), e Poudel *et al.* (2020), escolheram a técnica de análise de componentes independentes (ICA) para seus experimentos, tendo como justificatvia os resultados apresentados por Saelens, Cannoodt e Saeys (2018). O problema aqui é que essa justificativa é apoiada apenas em resultados de avaliação externa, que não levam em conta a natureza do problema e dos grupos inerentes. De fato, como afirma Wiwie, Baumbach e Röttger (2015), em um cenário do mundo real (geralmente sem um agrupamento de referência) o pesquisador que executa uma tarefa de agrupamento em um novo conjunto de dados tem à sua disposição apenas os índices de validação internos.

Nesse sentido, o presente trabalho se dispôs a percorrer, de maneira interdisciplinar, a trajetória composta pelas três principais etapas para a obtenção de conhecimento a partir de dados de expressão gênica, por meio da análise de um conjunto de dados de *microarray* de células beta-pancreáticas submetidas à progesterona cedido por pesquisadoras da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP), aprofundando-se na discussão da avaliação das técnicas de agrupamento do estado da arte de forma que se contemplem também os índices de validação internos e que se suporte a seleção da técnica de agrupamento mais adequada ao problema. Por fim, foi realizada a análise de enriquecimento funcional e interpretação de resultados com o apoio da especialista no domínio para utilizar os resultados obtidos na direção de compreensão do problema do diabetes gestacional.

1.1 Contextualização e motivação

Devido ao crescente uso farmacológico de progestógenos ao longo da gravidez para a prevenção do parto prematuro (PERGIALIOTIS *et al.*, 2019), a relação entre esses hormônios e o diabetes gestacional requer atenção. A partir de experimentos *in vitro* com a linhagem celular RINm5F, verificou-se que a progesterona foi capaz de induzir a oxidação e morte das células beta-pancreáticas, que são produtoras de insulina (NUNES *et al.*, 2014).

A morte de células beta-pancreáticas está associada aos diabetes tipo I e tipo II (ROJAS *et al.*, 2018), mas ainda precisa ser melhor compreendida no contexto do diabetes gestacional. Para investigar este problema, experimentos de *microarray* foram conduzidos

estudando-se células da linhagem RINm5F submetidas à progesterona em três doses (0,1 μM, 1 μM e 100 μM) e dois tempos (6h e 24h). Tal conjunto de dados de expressão gênica é objeto da análise deste trabalho, que se propôs a auxiliar na compreensão do fenômeno realizando a identificação e a aplicação das técnicas de agrupamento mais adequadas ao problema e a interpretação baseada em análises de enriquecimento funcional dos resultados.

1.2 Objetivos

1.2.1 Objetivo geral

O objetivo geral deste trabalho consiste em realizar a análise de agrupamento do conjunto de dados de *microarray* de células beta-pancreáticas submetidas à progesterona fazendo uso da técnica mais adequada ao problema e realizar a interpretação dos resultados no sentido de contribuir para a compreensão da doença do diabetes gestacional.

1.2.2 Objetivos específicos

Para que se cumprisse o objetivo geral, fizeram-se necessários os seguintes objetivos específicos:

- Reproduzir o procedimento consolidado no estudo de *benchmark* proposto por Saelens, Cannoodt e Saeys (2018) para as principais técnicas utilizadas na literatura, incluindo aquelas mais bem avaliadas no esquema de pontuação proposto, a saber, agrupamento hierárquico aglomerativo, k-médias, agrupamento de deslocamento médio, análise de componentes independentes e bi-agrupamento espectral.
- Estender a discussão do estudo de *benchmark* por meio do cálculo e da discussão da pontuação, segundo índices de validação internos, a saber, índice de Dunn, índice da silhueta média e índice de Davies-Bouldin.
- Reproduzir o procedimento do estudo de *benchmark* incluindo o conjunto de células beta-pancreáticas e a discussão dos índices internos. Para tanto, uma vez que não se tem disponível agrupamento de referência para o conjunto de dados células beta-pancreáticas submetidas à progesterona, se fez necessário realizar o levantamento de módulos (grupos de genes) de referência.

- Realizar a análise de enriquecimento funcional dos melhores resultados obtidos para o conjunto de células beta-pancreáticas submetidas à progesterona segundo índices de validação internos e externos e realizar a validação com a especialista do domínio, a a Profa. Dra. Anna Karenina Azevedo Martins, fornecendo a interpretação e seleção dos resultados relevantes para o problema.

1.3 Organização do trabalho

O presente documento é organizado como segue. No capítulo 2, são descritos os conceitos necessários para a compreensão do estudo, enquanto no capítulo 3 é retratado o estado da arte. O capítulo 4 descreve o procedimento realizado para a avaliação e interpretação dos métodos de agrupamento aplicados a dados de expressão gênica. Na seção 4.2 são discutidos os resultados experimentais obtidos em cada etapa da realização deste trabalho e, por fim, o capítulo 5 conclui este trabalho.

2 Fundamentação teórica

O trabalho de Nagi, Bhattacharyya e Kalita (2011), aborda o processo de análise de dados de expressão gênica em três grandes etapas, a saber, (i) experimentos laboratoriais; (ii) análise computacional; e (iii) interpretação dos resultados. Seguindo esse esquema, primeiramente, os dados são coletados em experimentos laboratoriais, como por exemplo com o uso de *microarrays*. Essa etapa envolve desde a execução do experimento até a obtenção dos valores de expressão gênica. A segunda etapa consiste na análise computacional desses dados, que pode ser realizada de diversas maneiras a depender do propósito e, neste trabalho, especificamente, trata-se da análise de agrupamento. Por fim, é executada a etapa de análise e interpretação dos resultados, com base no conhecimento existente.

Nas seções que seguem, estão explicados detalhes a respeito das técnicas envolvidas no processo de obtenção de conhecimento sobre os genes.

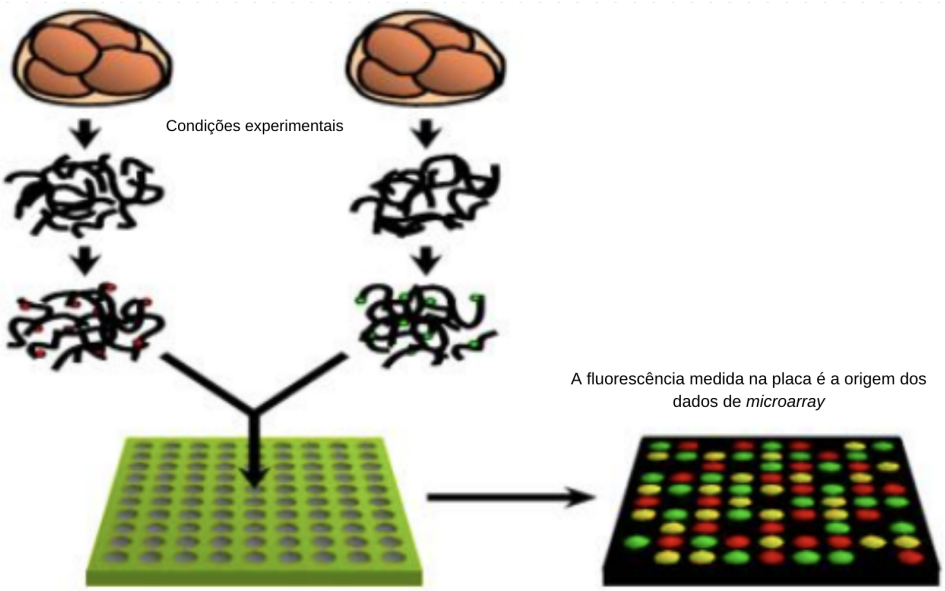
2.1 Microarrays

A tecnologia de *microarrays* transformou a maneira pela qual os experimentos científicos com a finalidade de obter conhecimento sobre os genes eram conduzidos. Na abordagem clássica da genética, por exemplo, uma das maneiras para identificar a função de um gene é analisar o que acontece com o organismo na sua ausência, por meio da análise de organismos mutantes com aparência interessante ou incomum, como por exemplo, moscas da fruta com olhos brancos ou asas curvadas. Desta maneira, é feita uma análise partindo-se do fenótipo - aparência ou comportamento do indivíduo - para entender como a modificação no genótipo expressa tais características. Conforme Tuimala e Laine (2003), a principal vantagem da tecnologia de *microarrays* quando comparada aos métodos tradicionais é que, diferentemente dos experimentos tradicionais, a tecnologia permite a análise simultânea da expressão de uma grande quantidade de genes.

De acordo com Baldi e Brunak (2001), os *microarrays* permitem que sejam estudados padrões de expressão genética em determinado tipo de célula, em determinado tempo e sob determinado conjunto de condições. Nessa tecnologia, o RNA é transcrito para criar um DNA complementar (cDNA), que é combinado com uma ampla biblioteca de fragmentos de genes previamente distribuídos em uma placa de vidro ou suporte, conforme ilustrado

na figura 1. Em seguida, as expressões dos genes sob diversas condições experimentais são medidas por meio da interpretação da fluorescência dos genes com equipamento específico. Os resultados das análises de fluorescência são as medidas fornecidas pela tecnologia de *microarrays* e representam, portanto, a expressão dos genes nas condições experimentais. A utilização desses *arrays* produz grandes quantidades de dados, potencialmente capazes de dar *insights* fundamentais sobre processos biológicos que vão desde a descoberta de função dos genes ao desenvolvimento de novos medicamentos.

Figura 1 – Ilustração do experimento de *microarray*. Nessa tecnologia, o RNA é transcrito para criar um DNA complementar (cDNA), que é combinado com uma ampla biblioteca de fragmentos de genes previamente distribuídos em uma placa de vidro ou suporte. Em seguida, as expressões de centenas de genes sob diversas condições experimentais são medidas por meio da interpretação da fluorescência dos genes com equipamento específico.



Fonte: adaptado de (YUVARAJ; MANJULA, 2018)

Os experimentos de *microarray* possuem um grupo de controle e um ou mais grupos que foram submetidos às condições estudadas. A variação de expressão dos genes (quantificação de proteína que o gene produziu sob as circunstâncias do experimento) é uma medida relativa, calculada em relação ao grupo de controle, e essa mensuração pode ser realizada de formas variadas (TUIMALA; LAINE, 2003). A abordagem mais simples para o cálculo dessa variação consiste em dividir a intensidade da expressão do gene obtida no experimento pela intensidade da expressão do mesmo gene no grupo de controle. Os dados que resultam dos experimentos laboratoriais, após passarem por tal cálculo, são exibidos em tabela cujas linhas contêm os genes e as colunas contêm a medida da expressão

daquele gene em diversos experimentos, conforme ilustrado na figura 2. Nessa figura, o valor m_{ln} representa a expressão do gene l medido no experimento n.

Figura 2 – Matriz de expressão gênica. As linhas representam os genes enquanto as colunas representam os experimentos. O valor m_{ln} representa a variação da expressão do gene l em relação ao grupo controle quando medido no experimento n.

m_{11}	m_{12}	m_{13}	.....	m_{1n}
m_{21}	m_{22}	m_{23}	.....	m_{2n}
...	...	...	...	...
...	...	...	...	...
m_{l1}	m_{l2}	m_{l3}	.....	m_{ln}

Fonte: Adaptado de Nagi, Bhattacharyya e Kalita (2011)

Em relação ao grupo controle, considera-se que um gene foi: (i) neutro, quando manteve a mesma quantidade de expressão em relação ao grupo controle; (ii) superexpresso, ou expresso para cima, quando o gene produziu mais proteína na condição experimental; e (iii) subexpresso, ou expresso para baixo, quando produziu uma quantidade menor de proteína em relação ao grupo controle. Após a coleta das informações do suporte e o cálculo da expressão dos genes realizada por *software* específico que faz parte do equipamento do experimento, os genes são submetidos às etapas de pré-processamento e normalização antes da análise propriamente dita. Essas etapas serão descritas na seção 2.2.

2.2 Técnicas de pré-processamento de dados

Segundo Tuimala e Laine (2003), os métodos de pré-processamento para dados de *microarrays* incluem procedimentos analíticos ou de transformação que precisam ser aplicados aos dados antes que estes sejam analisados, com o intuito de evitar distorções nos resultados. O tratamento de valores faltantes e o cálculo da variação de expressão são algumas das técnicas que podem ser empregadas e serão detalhadas na seção 2.2.1.

2.2.1 Valores faltantes

Na análise de dados de *microarrays*, é possível encontrar algumas observações com valores faltantes. Um valor faltante corresponde a uma observação (intensidade da expressão de determinada proteína) cuja medida quantitativa não está disponível. No contexto dos *microarrays*, as medidas podem estar ausentes por dois principais motivos: (1)

a proteína não se expressou (intensidade igual a zero); ou (2) a intensidade do *background* (base da placa de vidro) é maior que a intensidade da proteína em questão.

Devido à interferência nos testes estatísticos e computacionais, os valores faltantes podem levar a problemas na análise dos dados. Por esse motivo, esses valores precisam ser tratados na fase de pré-processamento. Os valores faltantes podem ser substituídos por valores estimados, em um processo chamado imputação, ou podem ser retirados do conjunto original. Para a remoção dos dados dos valores faltantes pode-se considerar excluir todas as observações do conjunto de dados que tenham ao menos uma variável (coluna) com valor faltante. Por outro lado, a imputação substitui o valor faltante de uma variável pela média de todos os valores que a mesma assume no conjunto (imputação pela média). Como os dados faltantes são substituídos por valores artificiais na imputação, pode haver alteração dos valores de correlação entre as variáveis observadas.

2.3 Técnicas de agrupamento de dados

Cinco métodos de agrupamento foram analisados neste trabalho: k-médias (LLOYD, 1982), agrupamento hierárquico aglomerativo (JAIN; DUBES, 1988), bi-agrupamento espectral (KLUGER *et al.*, 2003), análise de componentes independentes (ICA z-score) (HYVÄRINEN; OJA, 2000), deslocamento médio (mean shift) (COMANICIU; MEER, 2002), e aleatório (SAELENS; CANNOODT; SAEYS, 2018). Essas técnicas estão resumidas na tabela 1 e classificadas de acordo com: (i) sobreposição (se um gene pode ser atribuído a mais de um módulo); (ii) determinístico (se o resultado do agrupamento será sempre o mesmo para todas as execuções); e (iii) parâmetros requeridos pela técnica (além do próprio conjunto de dados).

Tabela 1 – Visão geral dos métodos de agrupamento utilizados.

Nome	Parâmetro(s)	Sobreposição	Determinístico
k-médias	No. de grupos	Não	Não
Aglomerativo	No. de grupos	Não	Sim
Bi-agrupamento espectral	No. de grupos	Não	Não
ICA	No. de grupos, valor de corte	Sim	Não
Deslocamento médio	*Bandwidth*	Sim	Sim
Aleatório	No. de grupos	Não	Não

É importante ressaltar que os resultados de determinada técnica de agrupamento variam conforme os parâmetros de entrada utilizados. Neste trabalho, utilizou-se a estratégia

de busca em grade, que consiste na variação dos parâmetros dentro de determinado intervalo e na posterior seleção dos melhores resultados, de acordo com valores das medidas de avaliação. A metodologia de avaliação das técnicas será descrita mais adiante neste trabalho, na seção 2.7, enquanto a descrição de cada técnica de agrupamento é realizada na seção 2.3.1.

2.3.1 Algoritmo k-médias

O algoritmo k-médias é uma técnica de agrupamento particional que tem por parâmetros de entrada o conjunto de dados e o número de grupos (k) em que esse conjunto será dividido. Em sua formulação, é utilizado o conceito de centroide, que é o centro representativo (protótipo) de cada grupo, inicializado aleatoriamente com algum dos pontos do conjunto de dados. De forma iterativa, o algoritmo aloca cada elemento do conjunto de dados no grupo mais próximo, de acordo com uma medida de distância (ver seção 2.3.7), e os centroides são então recalculados, considerando a nova configuração de grupos. O procedimento é repetido até que se atinja um dado número de iterações ou até que não haja mudança nos centroides após a atualização. Os passos do método estão listados no algoritmo 1.

Algoritmo 1 Algoritmo k-médias.

1: Selecione aleatoriamente k pontos do conjunto de dados como centroides iniciais (protótipos) dos grupos.
2: **repeat**
3: aloque cada ponto do conjunto de dados no grupo com centroide mais próximo.
4: atualize os protótipos dos grupos, calculando o vetor de médias para os objetos em cada grupo.
5: **until** não haver mudança na configuração de grupos.

Fonte: Tan, Steinbach e Kumar (2005)

2.3.2 Agrupamento hierárquico aglomerativo

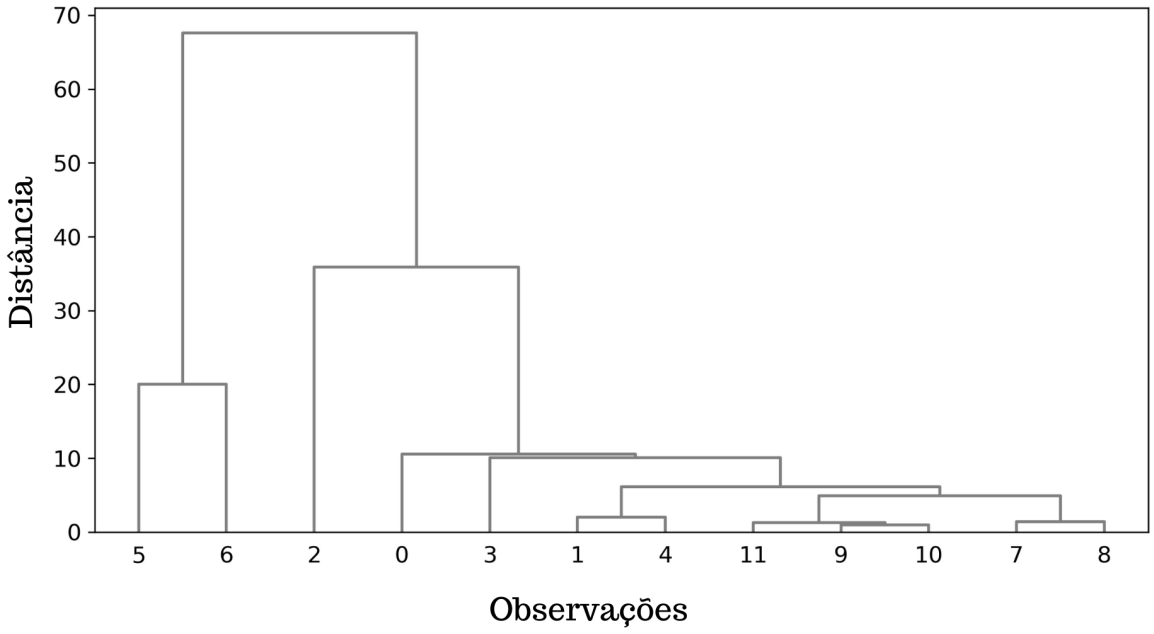
Segundo Tan, Steinbach e Kumar (2005), os métodos hierárquicos, de modo geral, tratam o conjunto de dados como um conjunto de subgrupos, organizados hierarquicamente, de acordo com a similaridade entre os grupos. Os resultados do agrupamento são dispostos na forma de uma árvore, também conhecida como dendrograma. Nesta representação, a raiz da árvore corresponde a um único grupo contendo todos os dados e cada nó representa

um subgrupo e corresponde à união dos seus filhos, exceto os nós folha. Dessa forma, o agrupamento hierárquico resulta em um conjunto de partições aninhadas.

O agrupamento hierárquico pode ser subdividido em divisivo (*top-down*) e aglomerativo (*bottom-up*). O método divisivo consiste em considerar, a princípio, a existência de um grupo único que é subdividido em grupos menores, enquanto o método aglomerativo parte do pressuposto de que cada dado é um grupo e une os grupos em conjuntos maiores até que seja obtido um único grupo com todos os dados.

Um exemplo de dendrograma, resultado de um agrupamento hierárquico, pode ser visto na figura 3.Os valores exibidos no eixo x são os índices das observações conjunto de dados analisado, enquanto os valores do eixo y representam a distância entre os grupos. Para a obtenção dos grupos propriamente ditos, realiza-se um corte em algum nível da árvore, segundo algum critério. Por exemplo, se o corte deste dendrograma fosse feito na altura = 40, o conjunto seria particionado em dois grupos. Sendo assim, um mesmo dendrograma pode gerar agrupamentos diferentes de acordo com a escolha do nível da árvore, realizada pelo usuário. Os agrupamentos hierárquicos, apesar de apresentarem a vantagem de permitir examinar o dendrograma para estimar o número de grupos, possuem a desvantagem de serem impraticáveis quando a quantidade de dados é muito alta, devido ao alto custo computacional do método (JAIN; DUBES, 1988).

Figura 3 – Exemplo de dendrograma.



Neste trabalho, será dado destaque à técnica de agrupamento hierárquico aglomerativo, descrita a seguir.

As técnicas de agrupamento hierárquico aglomerativo são derivações de uma abordagem básica: inicia-se com os grupos compostos por dados individuais e sucessivamente mesclam-se os grupos mais próximos até que se tenha um único grupo (TAN; STEINBACH; KUMAR, 2005). O funcionamento da técnica baseia-se no cálculo da matriz de distâncias, que inicialmente contém as distâncias entre todos os pares de objetos do conjunto de dados. Essa abordagem é formalmente expressa no algoritmo 2.

Algoritmo 2 Agrupamento hierárquico aglomerativo.

1: Compute a matriz de distâncias entre os elementos (grupos iniciais) do conjunto de dados.
2: **repeat**
3: mescle os dois grupos mais próximos (com menor distância).
4: recalcule a matriz de distâncias.
5: **until** haver apenas um grupo.

Fonte: Tan, Steinbach e Kumar (2005)

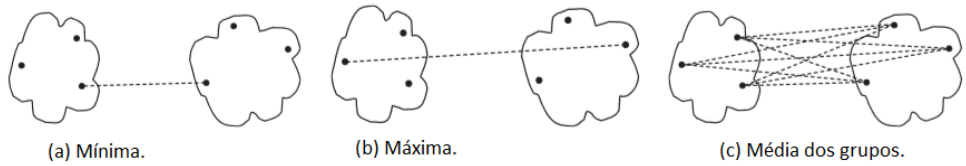
A operação chave do algoritmo 2 é o cálculo da distância entre os grupos, utilizado na etapa de mesclagem dos grupos. Tal cálculo pode ser feito de diversas maneiras e costuma ser o principal fator que diferencia os algoritmos de agrupamento hierárquico aglomerativo existentes.

De uma visão baseada em grafo dos grupos, decorrem os critérios de distância *mínimo*, *máximo* e *média dos grupos*. Para as medidas de distância entre dois grupos, o critério *mínimo* utiliza como representante da distância entre dois grupos a menor distância entre dois pontos que estejam em grupos diferentes. Em termos de uma visão baseada em grafo, *mínimo* é a menor linha reta que une dois pontos que estejam em grupos distintos.

Alternativamente, *máximo* utiliza, para representar a medida da distância entre dois grupos, o maior valor de distância entre dois pontos em grupos distintos, ou, em termos de uma visão baseada em grafo, a maior linha reta que une dois pontos que estejam em grupos distintos. Já a abordagem referente à *média dos grupos* toma como distância entre dois grupos a média das distâncias entre todos os pares de pontos em conjuntos distintos.

Essas medidas são ilustradas na figura 4.

Figura 4 – Critérios para o cálculo da distância entre grupos.



Fonte: adaptado de Tan, Steinbach e Kumar (2005)

2.3.3 Análise de componentes independentes (ICA)

A análise de componentes independentes é um método de decomposição que visa separar fontes de sinais tão estatisticamente independentes quanto possível de um determinado modelo de mistura das fontes originais. A motivação clássica do problema é conhecida por *cocktail-party problem*, descrita conforme segue.

Motivação da técnica

Imagine que você está em uma sala em que duas pessoas estão falando simultaneamente e nessa sala existem dois microfones gravando o som do ambiente, posicionados em locais diferentes. Os microfones fornecem duas gravações em determinado período de tempo, que podemos denotar por $\mathbf{grav}_1(t)$ e $\mathbf{grav}_2(t)$, em que $\mathbf{grav}_1$ e $\mathbf{grav}_2$ são amplitudes e t é o instante de tempo. Cada uma das gravações é uma soma ponderada dos sinais de fala emitidos pelas duas pessoas, que podem ser denotados por $\mathbf{s}_1(t)$ e $\mathbf{s}_2(t)$. Isso pode ser expresso na forma da equações lineares:

$$\mathbf{grav}_1(t) = \mathbf{a}_{11}s_1 + \mathbf{a}_{12}\mathbf{s}_2 \tag{1}$$

e

$$\mathbf{grav}_2(t) = \mathbf{a}_{21}s_1 + \mathbf{a}_{22}\mathbf{s}_2. \tag{2}$$

Em que $\mathbf{a}_{11}$, $\mathbf{a}_{12}$, $\mathbf{a}_{21}$ e $\mathbf{a}_{22}$ são parâmetros que dependem das distâncias dos microfones dos alto-falantes. A técnica de análise de componentes independentes visa estimar, a partir dos sinais gravados $\mathbf{x}_1(t)$ e $\mathbf{x}_2(t)$, os dois sinais de fala originais $\mathbf{s}_1(t)$ e $\mathbf{s}_2(t)$. Para tanto, o objetivo é encontrar uma mistura de sinais independentes em dados decompondo-os em duas matrizes: uma matriz de origem e uma matriz de mistura.

Análise de componentes independentes (ICA) com pós-processamento utilizando escores z (z-scores)

A análise de componentes independentes pode ser utilizada para agrupamento de dados de expressão gênica, com o objetivo de extrair os componentes correspondentes aos módulos de coexpressão decompondo a matriz de expressão em um produto de matrizes menores (ROTIVAL *et al.*, 2011). Desta forma, a expressão de cada gene pode ser escrita como uma função linear dos componentes, e as influências de cada componente possuem dependências estatísticas mínimas.

Neste trabalho, a detecção de grupos de genes com base nessa técnica, neste estudo chamada de ICA z-scores, é composta por duas etapas principais: (i) a decomposição da matriz de expressão em um conjunto de componentes independentes usando o algoritmo *FastICA*, uma implementação da biblioteca *scikit-learn* (PEDREGOSA *et al.*, 2011) baseada em (HYVÄRINEN; OJA, 2000); e (ii) a detecção de módulos potencialmente sobrepostos dentro de cada sinal utilizando o pós-processamento baseado em escores z (SAELENS; CANNOODT; SAEYS, 2018).

Nesse pós-processamento, os pesos de cada sinal de origem são primeiro padronizados para escores z (número de desvios padrão em relação à média) de um gene é atribuído a um módulo se o seu escore $z$absoluto é maior do que um valor de corte, chamado de *stdcutoff*. Portanto esta técnica também requer como parâmetro de entrada, além do número de grupos que deseja-se obter, o valor de corte que será utilizado no pós-processamento da matriz de origem (*stdcutoff*).

2.3.4 Agrupamento de deslocamento médio (*mean shift*)

O método do agrupamento de deslocamento médio (*mean shift*) foi proposto inicialmente por Fukunaga e Hostetler (1975) e posteriormente estendido por Comaniciu e Meer (2002) e, diferentemente das demais técnicas avaliadas neste estudo, não requer como parâmetro de entrada o número de grupos. Para isso, a técnica, baseada em centroides, conta com um parâmetro denominado largura de banda (*bandwidth*), que determina o raio da região em que a densidade local será estimada.

A noção por trás desse método é tratar os pontos do espaço d-dimensional de características como uma densidade de probabilidade empírica na qual regiões densas

correspondem a máximos locais ou modas das distribuições subjacentes. Neste trabalho, a implementação utilizada foi aquela do pacote *scikit-learn* (PEDREGOSA *et al.*, 2011) e pode ser resumida em duas principais etapas: (i) primeiramente há um cálculo da estimativa de densidade do *kernel* usando, neste caso, um *kernel* gaussiano; e (ii) em seguida, os pontos são iterativamente deslocados para regiões próximas de maior densidade até a convergência. Neste trabalho, os conjuntos de dados foram padronizados antes de aplicar este algoritmo.

2.3.5 Bi-agrupamento espectral

No contexto de dados de expressão gênica, as técnicas de bi-agrupamento visam resolver o problema não resolvido pelas técnicas tradicionais de detectar padrões de genes coexpressos em "recortes"do conjunto de dados, compostos por subconjuntos de experimentos e subconjuntos de condições experimentais.

Esse "recorte"de condições experimentais e de genes é o que compõe um bi-grupo. No bi-agrupamento espectral, dentro de um bi-grupo, a expressão gênica é relativamente constante. O objetivo do bi-agrupamento espectral é reordenar os genes e amostras da matriz de expressão de forma que seja revelada uma estrutura "quadriculada"de bi-grupos, problema que pode ser resolvido usando uma decomposição de valor singular (SVD) (KLUGER *et al.*, 2003). Nesta técnica, a matriz de expressão gência é decomposta em vetores que são posteriormente agrupados para identificar os bi-grupos aplicando-se o algoritmo k-médias. Neste trabalho, a implementação utilizada foi aquela do pacote *scikit-learn* (PEDREGOSA *et al.*, 2011) e os conjuntos de dados foram padronizados antes da aplicação do algoritmo.

2.3.6 Agrupamento aleatório

O agrupamento aleatório é utilizado como referência (*baseline*) para comparação entre diferentes métodos de agrupamento. Este agrupamento é feito sorteando-se um grupo, no intervalo de 1 a k, para cada atribuir a cada um dos n elementos do conjunto de dados.

2.3.7 Medidas de distância

As medidas de distância mais comumente consideradas pelos métodos de agrupamento para são a distância euclidiana e distância de Manhattan, que serão brevemente descritas a seguir. As fórmulas estão exemplificadas para distância entre pontos, mas podem ser consideradas utilizando vetores n-dimensionais.

- **Distância euclidiana**

A distância euclidiana $de(\mathbf{p}, \mathbf{q})$ entre dois pontos $\mathbf{p} = (p_1, p_2)$ e $\mathbf{q} = (q_1, q_2)$ é dada pela distância geométrica entre estes, ou seja, pela reta que os liga no espaço cartesiano de acordo com a seguinte equação:

$$de(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - p_2)^2 + (p_2 - q_2)^2}. \tag{3}$$

- **Distância de Manhattan**

A distância de Manhattan $dm(\mathbf{p}, \mathbf{q})$ entre dois pontos $\mathbf{p} = (p_1, p_2)$ e $\mathbf{q} = (q_1, q_2)$ em um espaço euclidiano com um sistema cartesiano de coordenadas bidimensional pode ser definida como a soma dos comprimentos da projeção da linha que une os pontos com os eixos das coordenadas e é dada pela seguinte equação (DEZA; DEZA, 2009):

$$dm(\mathbf{p}, \mathbf{q}) = \mid p_1 - q_1 \mid + \mid p_2 - q_2 \mid . \tag{4}$$

2.4 Avaliação dos resultados de agrupamento

O agrupamento, conforme dito anteriormente, pode ser avaliado utilizando informações externas, o que são os chamados índices externos, ou por meio de medidas denominadas índices internos, que não se utilizam de tais informações.

2.4.1 Índices de validação internos

De um modo geral, a avaliação da qualidade do agrupamento segundo índices de validação internos é fundamentada em dois principais conceitos: coesão, que se refere

ao quão distantes entre si estão os itens de um mesmo grupo (distância intra-grupo), e separação, que visa quantificar o quão distantes os grupos encontrados uns dos outros (distância entre grupos) (TAN; STEINBACH; KUMAR, 2005). As distâncias intra-grupos e entre grupos podem ser calculadas usando um critério de distância determinado, como os listados na seção 2.3.7. Segundo os índices de validação internos, o agrupamento melhor avaliado para determinado conjunto de dados é aquele que minimiza a distância intra-grupos e maximiza a distância entre grupos e as medidas de validação internas baseiam-se nesse princípio. As equações de coesão e separação são descritas conforme as equações:

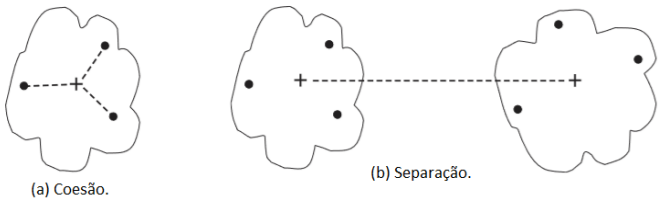
$$\text{Coesão}(\mathbf{c}_i) = \sum_{\mathbf{x}\epsilon\mathbf{C}_i} \text{distância}(\mathbf{x}, \mathbf{c}_i) \tag{5}$$

e

$$\text{Separação}(\mathbf{c}_i, \mathbf{c}_j) = \text{distância}(\mathbf{c}_i, \mathbf{c}_j), \tag{6}$$

em que $\mathbf{x}$ representa todos os pontos do conjunto de dados agrupado e $\mathbf{c}_i$ representa o protótipo do grupo $\mathbf{C}_i$ e i representa um dos k grupos resultantes do agrupamento. A distância calculada irá depender do critério estabelecido. A figura 5 ilustra os conceitos de coesão e separação em um agrupamento baseado em protótipos, considerando que a distância entre grupos é dada pela distância entre os protótipos dos grupos e a distância intra-grupo, entre os elementos pertencentes ao grupo e o protótipos correspondente, sendo o protótipo dos grupos representado pelo sinal de "+".

Figura 5 – Ilustração dos conceitos de coesão e separação.



Fonte: adaptado de (TAN; STEINBACH; KUMAR, 2005)

Destas noções derivam as medidas de validação internas de grupo. Neste trabalho, três medidas de validação internas foram empregadas, cuja interpretação é sumarizada no quadro 1. Essas medidas serão descritas nas seções que seguem.

Quadro 1 – Resumo dos índices de validação internos

Índice	Melhor se...	Intervalo
Dunn (DUNN, 1974)	Alto	[-∞,+∞]
Davies-Bouldin (DAVIES; BOULDIN, 1979a)	Baixo	[-∞,+∞]
Silhueta média(ROUSSEEUW, 1987)	Alto	[-1,+1]

Índice de Dunn

O índice de Dunn (D) (DUNN, 1974) avalia o desempenho do agrupamento relacionando o diâmetro máximo do grupo à mínima distância entre grupos, conforme a equação 7:

$$D = \frac{\min\limits_{c_i \neq c_j \epsilon \mathbf{C}} \left(\min\limits_{x_i \epsilon c_i, x_j \epsilon c_j} d(x_i, x_j) \right)}{\max\limits_{c_k \epsilon \mathbf{C}} \left(\max\limits_{x_i \epsilon c_k, x_j \epsilon c_k} d(x_i, x_j) \right)}. \tag{7}$$

É desejável que em um agrupamento os grupos estejam o mais bem separados e o mais coesos possível. Assim, a distância mínima entre grupos representa uma medida da separação desse agrupamento em seu pior cenário. Por outro lado, o diâmetro máximo de um grupo permite que seja possível ter uma ideia da coesão desses grupos, em seu pior caso.

Esse índice, por ter no numerador a distância mínima entre grupos, que é melhor que seja a maior o possível e no denominador o diâmetro do grupo, que é melhor que seja o menor possível, apresenta valores mais altos quanto mais coesos e separados forem os grupos.

É importante destacar que esta medida está sujeita a interferência de *outliers*, uma vez que é baseada em distâncias mínimas e máximas.

Índice de Davies-Bouldin

O índice de Davies Bouldin (DB) (DAVIES; BOULDIN, 1979b) é definido com base na distância média entre objetos e os seus respectivos e pode ser calculado conforme:

$$DB = \frac{1}{n} \cdot \sum_{\mathbf{c_i} \epsilon \mathbf{C}} \max_{c_i \neq c_j \epsilon \mathbf{C}} \left(\frac{\sigma_i + \sigma_j}{d(\bar{c}_k, \bar{c}_j)} \right) \tag{8}$$

em que

$$\sigma_i = \sqrt{\frac{1}{|c_i|} \sum_{\mathbf{x_i} \epsilon \mathbf{c_i}} |x_i - c_i|}, \tag{9}$$

e

$$d(\bar{c}_i, \bar{c}_j) = ||\bar{c}_i - \bar{c}_j||. \tag{10}$$

Ressalta-se que como esta medida tem em seu numerador a noção implícita de coesão e no denominador a noção de separação, tão melhor avaliados serão os valores obtidos quanto menores.

Índice da silhueta média

O índice da silhueta média (Sil) (ROUSSEEUW, 1987) é uma das medidas mais populares de validação interna de agrupamentos e relaciona distâncias de elementos de diferentes grupos às distâncias dos elementos do mesmo grupo. Esta é uma medida menos sensível aos *outliers* que o índice de Davies Bouldin. Sejam $a(x_i)$ a distância média do objeto $\mathbf{x}_i$ aos outros elementos do seu grupo; $b(\mathbf{x}_i)$ a distância média do objeto $\mathbf{x}_i$ ao grupo mais próximo $o(\mathbf{x}_i)$; e $c(\mathbf{x}_i)$ o grupo ao qual pertence o objeto $\mathbf{x}_i$, o valor da silhueta média é definido como:

$$Sil = \frac{1}{n} \cdot \sum_i sv_i \tag{11}$$

em que

$$sv_i = \frac{1}{n} \cdot \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}}, \tag{12}$$

$$a(\mathbf{x}_i) = \frac{1}{|c(\mathbf{x}_i)|} \sum_{\mathbf{x}_j \epsilon c(\mathbf{x}_i)} d(\mathbf{x}_i, \mathbf{x}_j), \tag{13}$$

e

$$b(\mathbf{x}_i) = \frac{1}{|o(\mathbf{x}_i)|} \sum_{\mathbf{x}_j \epsilon o(\mathbf{x}_i)} d(\mathbf{x}_i, \mathbf{x}_j). \tag{14}$$

Deve-se notar que o numerador do índice da silhueta de um ponto (sv_i) é melhor avaliado tanto maior forem os valores e que, se o valor do índice for negativo é um indício de que o ponto está mal posicionado, pois encontra-se mais próximo a um grupo distinto do que aquele ao qual ele foi atribuído.

Além dos índices de validação internos, há os índices calculados com base em informações externas ao agrupamento e que serão descritos na seção 2.4.2.

2.4.2 Índices de validação externos

Os índices de avaliação externos assumem um padrão de agrupamento de comparação, que pode ser chamado de *gold standard* (padrão ouro), como referência (WIWIE; BAUMBACH; RÖTTGER, 2015), (SAELENS; CANNOODT; SAEYS, 2018). A avaliação é feita de maneira análoga àquela realizada no aprendizado supervisionado e é baseada essencialmente em quatro quantidades: verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). Para um dado agrupamento C e um dado agrupamento de referência K, há duas principais estratégias para a comparação entre o agrupamento obtido e o agrupamento de referência, a de emparelhamento e a de mapeamento, definidas a seguir:

- **Emparelhamento (*pairwise strategy*):** Nesta abordagem são comparadas as relações entre todos os pares de objetos no agrupamento de referência com todos os objetos no agrupamento sendo avaliado. As quantidades de TP, TN, FP e FN são definidas para um par de dois elementos a e b, conforme segue:
  TP: se ambos estão no mesmo grupo, tanto em C como em K;
  TN: se ambos estão em grupos diferentes, tanto em C como em K;
  FP: se ambos estão no mesmo grupo em C, mas em grupos diferentes em K; e
  FN, se os elementos forem encontrados em grupos diferentes em C, mas no mesmo grupo em K.
- **Mapeamento:** Essa abordagem requer um mapeamento entre os grupos de C e de K. Para cada grupo $k_i \in K$, seleciona-se o grupo c_j com o maior número de elementos em comum: $k_i \cap c_j \rightarrow \max$. Um elemento a é contabilizado como:
  TP: se $a \epsilon k_i \cap a \epsilon c_j$;
  TN: não há definição significativa;
  FP: se $a \notin k_i \cap a \epsilon c_j$; e
  FN, se $a \epsilon k_i \cap a \notin c_j$.

Dadas essas definições de TP, TN, FP e FN e a estratégia de contagem desses valores (mapeamento ou emparelhamento), as medidas são definidas conforme segue.

- **Precisão**

A precisão é uma medida que representa a proporção de identificações positivas que estão corretas e é baseada na estratégia de mapeamento, conforme a equação:

$$Precisão = \frac{TP}{TP + FP}. \tag{15}$$

Note que na ausência de falsos positivos, a precisão terá valor igual a um.

• **Revocação**

Essa medida é baseada na estratégia de mapeamento e representa a proporção dos verdadeiros positivos que foi identificada corretamente pelo método de agrupamento. Portanto, um agrupamento que não teve identificações de falsos negativos terá a revocação igual a um. A Revocação pode ser definida definida conforme a equação:

$$Revocação = \frac{TP}{TP + FN}. \tag{16}$$

• **Escore F (F_β)**

O escore F é definido como a média harmônica entre as medidas de precisão e revocação, anteriormente definidas, e β é um fator de controle das influências dessas métricas. Essa medida é baseada na estratégia de mapeamento e é definida conforme segue:

$$F_\beta = \frac{(1 + \beta^2) \cdot Precisão \cdot Revocação}{\beta^2 \cdot Precisão + Revocação} \tag{17}$$

$$= \frac{(1 + \beta^2) \cdot TP}{(1 + \beta^2) \cdot TP + \beta^2 \cdot FN + FP}. \tag{18}$$

• **Taxa de descoberta falsa (FDR)**

Esse valor relaciona o número de falsos positivos ao número total de elementos preditos como positivos. É uma estimativa da probabilidade de um elemento ser predito como positivo caso o elemento seja negativo. Essa medida é baseada na abordagem de emparelhamento e é definida conforme segue:

$$FDR = \frac{FP}{FP + TP}. \tag{19}$$

• **Taxa de falsos positivos (FPR)**

A taxa de falsos positivos relaciona o número de falsos positivos ao número total de elementos negativos. Estima, portanto, a probabilidade de um elemento ser predito como positivo se o elemento é negativo. Esta medida é baseada na abordagem de emparelhamento.

$$FPR = \frac{FP}{FP + TN}. \tag{20}$$

• **Índice de Jaccard (J)**

O índice de Jaccard negligencia os TNs e relaciona os TPs ao número de pares que pertencem ou à mesma classe ou ao mesmo grupo. Esta medida é baseada na abordagem de emparelhamento.

$$J = \frac{TP}{TP + FP + FN}. \tag{21}$$

• **Índice de Rand (R)**

O Índice de Rand é a taxa entre os pares de objetos agrupados corretamente e o número de pares possíveis. Essa medida é baseada na abordagem de emparelhamento e é definida conforme segue:

$$R = \frac{TP + TN}{TP + FP + FN + TN}. \tag{22}$$

• **Sensibilidade** (S_n)

A sensibilidade (*recall*) relaciona o número de verdadeiros positivos ao número total de elementos positivos. Esta medida é baseada na abordagem de emparelhamento e é definida conforme a equação:

$$S_n = \frac{TP}{TP + FN}. \tag{23}$$

• **Especificidade** (S_p)

A especificidade relaciona o número de verdadeiros negativos ao número total de elementos negativos. Essa medida é baseada na abordagem de emparelhamento e é definida conforme a equação:

$$S_p = \frac{TN}{TN + FP}. \tag{24}$$

• **Recuperação**

A medida de recuperação visa representar quão bem os módulos conhecidos do padrão ouro foram recuperados em relação aos módulos observados no resultado do agrupamento. Na definição da medida dada pela equação a seguir, são empregadas as convenções de que M representa um conjunto de módulos conhecidos e M' um conjunto de módulos observados, e J é o índice de Jaccard:

$$\text{Recuperação} = \frac{1}{|M|} \sum_{m \in M} \max_{m' \in M'} \mathrm{J}\left(m', m\right) \tag{25}$$

.

• **Relevância**

A medida de relevância visa representar quão bem são os módulos observados no resultado do agrupamento já são conhecidos no padrão ouro e é definida como segue:

$$\text{Relevância} = \frac{1}{|M'|} \sum_{m' \in M'} \max_{m \in M} \mathrm{J}\left(m', m\right), \tag{26}$$

em que M representa um conjunto de módulos conhecidos e M' um conjunto de módulos observados e J é o índice de Jaccard.

• **F1rr**

Esta medida é composta pela média harmônica entre a recuperação e a relevância, anteriormente definidas e é dada por:

$$F1rr = \frac{2}{\frac{1}{Recuperação} + \frac{1}{Relevância}}. \tag{27}$$

• **F1rprr**

Esta medida é composta pela média harmônica entre revocação, precisão, recuperação e relevância, anteriormente definidas, e é dada por:

$$F1rprr = \frac{4}{\frac{1}{Revocação} + \frac{1}{Precisão} + \frac{1}{Recuperação} + \frac{1}{Relevância}}. \tag{28}$$

- **F1rp**

Já a medida $F1rp$ é composta pela média harmônica entre revocação e precisão, conforme a equação:

$$F1rp = \frac{2}{\frac{1}{Revocação} + \frac{1}{Precisão}}. \tag{29}$$

Uma vez calculados os índices de validação, internos ou externos, esses resultados precisam ser apresentados. Na seção 2.5 serão discutidas as técnicas para visualização dos resultados.

2.5 Técnicas de visualização dos resultados

Seja para a apresentação dos resultados dos índices de validação internos ou externos, seja para fornecer uma visualização da própria matriz de expressão gênica, um dos recursos visuais mais utilizados é o mapa de calor. Esse mapa fornece uma representação visual que favorece a percepção de padrões nos valores exibidos e auxilia na interpretação dos resultados. Outras maneiras de apresentação dos resultados empregadas neste trabalho são os já consolidados recursos visuais de tabelas, gráficos de barras e diagramas de caixas. Na seção 2.5.1 será detalhado o uso dos mapas de calor.

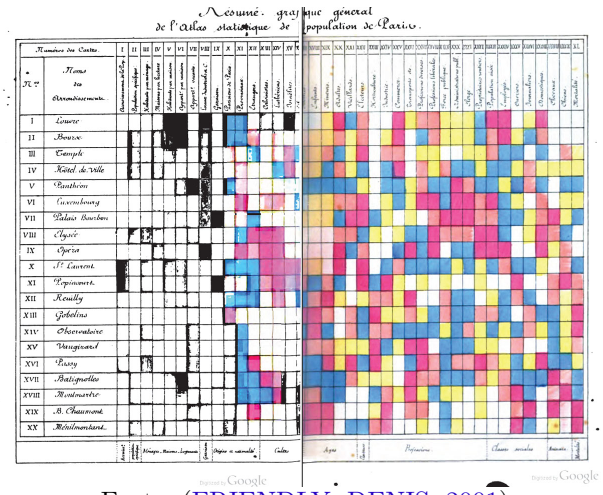
2.5.1 Mapas de calor

Os mapas de calor, também conhecidos como *heat maps*, são, em princípio, uma maneira de utilizar cores para representar números em uma tabela. Gehlenborg e Wong (2012), em um artigo sobre métodos, mencionam o exemplo de 1873, conforme ilustra a figura 6, também disponível de maneira ampliada no anexo B, feito pelo economista francês Toussaint Loua (LOUA, 1873) (note-se que, por se tratar de um documento antigo, parte dele não parece ter sido perfeitamente preservada e está em preto e branco). Nesse tipo de visualização, centenas de linhas e de colunas podem ser exibidas em uma única página e, uma vez que os mapas de calor se baseiam fundamentalmente na codificação de cores e na reordenação significativa das linhas e de colunas, se qualquer um desses dois componentes estiver comprometido, a utilidade da visualização fica comprometida. Há algumas outras observações que se deve ter em mente ao criar ou visualizar mapas de calor, como por exemplo, o fato de que as cores podem deixar de ser confiáveis se estiverem

representando valores discretos. Além disso, as cores podem passar impressões visuais diferentes, a depender da sua proximidade. A escala de cores escolhida para representar o problema deve ser adequada aos do intervalo numérico no qual os dados estão contidos.

Quando a técnica de visualização é combinada ao agrupamento, isso pode melhorar a habilidade de percepção dos resultados no mapa de calor. Conforme será descrito mais adiante, a proposta de Eisen *et al.* (1998), uma das mais empregadas para a finalidade de visualização de expressão gênica, trabalha justamente nesse aspecto.

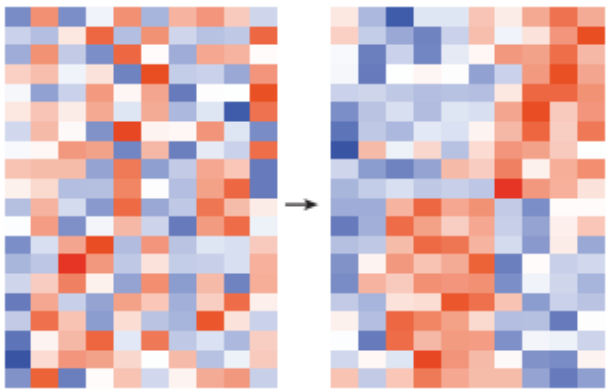
Figura 6 – Exemplo de tabela colorida, datada de 1873, semelhante aos mapas de calor



Fonte: (FRIENDLY; DENIS, 2001)

Na figura 7 é possível perceber o efeito da ordenação das linhas e das colunas exibidas nos mapas de calor de maneira que facilite a percepção de padrões nos dados.

Figura 7 – Mapas de calor. À esquerda, mapa de calor com disposição aleatória, à direita, o mesmo mapa de calor reorganizado.



Fonte: (GEHLENBORG; WONG, 2012)

No contexto de agrupamento de dados de expressão gênica, os mapas de calor são amplamente utilizados para auxiliar na interpretação dos grupos resultantes. A percepção

de padrões de expressão gênica aliada ao enriquecimento funcional dos genes é capaz de gerar *insights* sobre os problemas investigados. Os conceitos utilizados na interpretação dos resultados do agrupamento serão descritos na seção 2.6.

2.6 Interpretação dos resultados do agrupamento

Os grupos resultantes da aplicação das técnicas de agrupamento de dados de expressão gênica só terão a sua contribuição revelada quando interpretados. A interpretação dos resultados do agrupamento geralmente é apoiada por recursos visuais, como os mapas de calor, e por análises de enriquecimento funcional. As análises de enriquecimento funcional, por sua vez, são compostas por anotações funcionais dos genes consultadas em diversas fontes de dados, como a ontologia de genes. Na seção 2.6.1, serão descritas as anotações funcionais de genes, a Ontologia de Genes e como funciona o processo de análise de enriquecimento funcional utilizando a ferramenta *Enrichr*[1].

2.6.1 Anotações funcionais dos genes

Segundo Lewis, Ashburner e Reese (2000), a anotação funcional é o resultado do processo de interpretar dados de sequência bruta em informação biológica. As anotações descrevem o genoma e transformam sequências genômicas em informações biológicas, integrando análises computacionais e dados biológicos auxiliares. As anotações funcionais fornecem uma perspectiva ampla e uma visão geral de todo o genoma, mas são superficiais e descrevem de forma incompleta genes individuais. A visão aprofundada das funções dos genes dentro de cada contexto é obtida em pesquisas específicas.

Conforme descrito em (REED *et al.*, 2006), existem três níveis de anotação funcional. A anotação unidimensional detalha a posição dos genes dentro do genoma e descreve a função celular dos produtos genéticos. A anotação bidimensional envolve os componentes da anotação unidimensional com a adição de interações químicas e físicas. Por fim, a reconstrução de redes é uma descrição da interação dos genes. As anotações funcionais costumam ser disponibilizadas ao público na forma de bancos de dados *online*, como é o caso da Ontologia de Genes (*Gene Ontology*), que será descrita na seção 2.6.2.

[1] https://maayanlab.cloud/Enrichr/

2.6.2 Ontologia de Genes

Segundo Studer, Benjamins e Fensel (1998), uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada, que deve ser interpretável por computador e aceita por um grupo ou comunidade na área do conhecimento modelado pela ontologia. Nesse contexto, a *Gene Ontology* (GO)[2] é uma ontologia voltada a informações de genes e possui, entre outros, dados sobre as anotações funcionais dos genes.

Ao realizar o agrupamentos de dados de *microarrays*, espera-se que genes com funções similares sejam alocados no mesmo grupo (MAJI; SHAH, 2017) (ABU-JAMOUS *et al.*, 2013), e isso pode ser verificado comparando-se o resultado do agrupamento com o conhecimento disponível na ontologia. Por indução, genes com funções desconhecidas possuem chances de possuir papel similar aos dos genes (com funções conhecidas) que se encontram no mesmo grupo. Dessa forma, por meio da realização de consultas à ontologia GO, pode ser possível inferir funções desconhecidas de genes.

As anotações funcionais dos genes obtidas, por exemplo, na Ontologia de Genes, são comumente empregadas para validação de resultados da detecção de grupos na literatura.

2.6.3 Análise de enriquecimento funcional

A análise de enriquecimento funcional de genes é um método computacional para inferir conhecimento sobre um conjunto de genes de entrada comparando-o com conjuntos de genes anotados que representam conhecimento biológico prévio, como por exempo, aqueles descritos na Ontologia de Genes. Essa análise pode ser utilizada para identificar se um conjunto de genes de entrada se sobrepõe significativamente a conjuntos de genes anotados, calculando ainda um valor estatístico que resume a probabilidade de esta sobreposição não ocorrer ao acaso (HUANG; SHERMAN; LEMPICKI, 2009).

A partir da análise de enriquecimento funcional é possível inferir funções de genes pertencentes a um grupo com base no princípio conhecido na literatura como *guilt by association* (OLIVER, 2000) (EISEN *et al.*, 1998), que estabelece que genes com funções semelhantes tendem a ser agrupados juntos. Desta forma, se um grupo de genes foi funcionalmente enriquecido com determinado termo do banco de dados de genes, ou da

[2] http://www.geneontology.org/

ontologia, muito provavelmente os outros genes pertencentes a este grupo estão relacionados com esse termo.

Desta forma, portanto, o especialista do domínio consegue, a partir do conhecimento prévio sobre o problema (SOLLERO; GRYNBERG, 2020), da interpretação e da seleção das contribuições relevantes a partir dos resultados da análise de enriquecimento funcional, juntamente com a análise das representações visuais dos grupos, obter *insights* para a compreensão do fenômeno estudado. Vale ressaltar que, devido ao fato de a análise de enriquecimento funcional ser baseada no conhecimento existente sobre os genes, o resultado desta análise tende a variar ao longo do tempo conforme a evolução das anotações funcionais (TOMCZAK *et al.*, 2018).

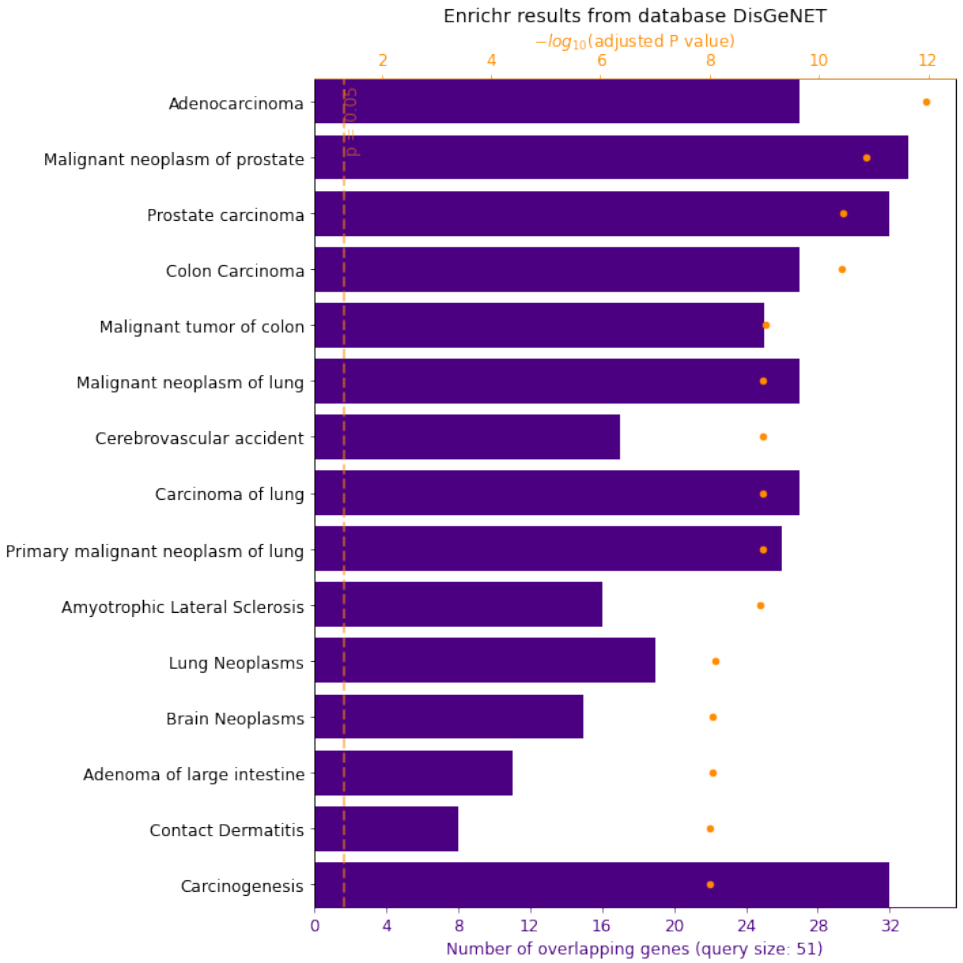
Dentre as ferramentas disponíveis para se realizar a análise, destaca-se o *Enrichr* (KULESHOV *et al.*, 2016). As bibliotecas (fontes de dados) de conjuntos de genes da ferramenta incluem genes expressos diferencialmente após perturbações de drogas, genes, doenças e patógenos. O *Enrichr* implementa quatro pontuações para relatar os resultados de enriquecimento: valor p, valor q, escore z (z-score) e pontuação combinada. O valor p é calculado usando um método estatístico padrão usado pela maioria das ferramentas de análise de enriquecimento funcional: o teste exato de Fisher ou o teste hipergeométrico. Este é um teste de proporção binomial que assume uma distribuição binomial e independência para probabilidade de qualquer gene pertencente a qualquer conjunto. Já o valor q é um valor p ajustado usando o método de Benjamini-Hochberg para correção para teste de múltiplas hipóteses. O escore de classificação, ou escore z, é calculado usando uma modificação do teste exato de Fisher, no qual é calculado um escore z para desvio de uma classificação esperada. Por fim, a pontuação combinada é uma combinação do valor p e da pontuação z calculada pela multiplicação das duas pontuações da seguinte forma:

$$c = ln(p) * z. \tag{30}$$

Uma vez que o teste exato de *Fisher* produz valores de p mais baixos para conjuntos mais longos, mesmo quando os conjuntos de entrada são aleatórios, a ferramenta aplica uma correção estatística para evitar que o enriquecimento retorne um valor-p significativo ($p <$ 0,05), sem que haja efeito real. A ferramenta utiliza o procedimento de Benjamini-Hochberg para corrigir múltiplas hipóteses, de maneira que os resultados retornados não recebam destaque se não atenderem ao limite mínimo de $p = 0.05$ antes da correção estatística.

A ferramenta, que também conta com uma interface *web* publicamente disponível [3], recebe como entrada um grupo de genes e uma opção dentre os conjunto de dados disponíveis na ferramenta que será utilizado como fonte de consulta para o enriquecimento funcional e retorna os termos relacionados a esse conjunto de genes, conforme a figura 8. Note que na figura retornada pelo enriquecimento funcional há uma linha linha tracejada para o valor-$p = 0.05$. Os pontos em laranja sinalizam o valor de $-\log_{10} p$ para cada termo retornado. Quanto mais altos esses valores, mais significativos os resultados. À esquerda da figura há o termo relacionado ao conjunto de genes, e na parte inferior do gráfico, o número de genes do conjunto de entrada que são relacionados a este termo.

Figura 8 – Exemplo de retorno de enriquecimento funcional de genes.



Neste trabalho a interpretação dos resultados das análises de enriquecimento funcional foi realizada com o apoio da pesquisadora especialista no domínio do problema, a Profa. Dra. Anna Karenina Azevedo Martins. Porém, em uma etapa anterior à análise de enriquecimento funcional, foi necessária seleção das técnicas de agrupamento e dos

[3] https://maayanlab.cloud/Enrichr/

parâmetros ideais. Saelens, Cannoodt e Saeys (2018) propuseram uma metodologia para avaliação de técnicas de agrupamento aplicadas a dados de expressão gênica, que será descrita na seção 2.7.

2.7 Metodologia para avaliação de técnicas de agrupamento aplicados à dados de expressão gênica

Saelens, Cannoodt e Saeys (2018) propuseram uma metodologia para avaliar as técnicas de agrupamento aplicadas a dados de expressão gênica. Nesse contexto, um grupo de genes é denominado módulo. É central para a compreensão de tal estudo o conceito de agrupamento de referência, ou *gold standard*. Um agrupamento de referência contém uma configuração desejada de módulos de genes, ou uma saída esperada de um agrupamento. Com base na avaliação da concordância entre os resultados obtidos a partir da aplicação das técnicas de agrupamento e os módulos de referência, que é dado pelos índices de validação externos, os autores elaboraram uma pontuação para cada método.

Para a avaliação dos métodos, os autores utilizaram conjuntos de dados de expressão gênica publicamente disponíveis e bem conhecidos na literatura. Para cada um desses conjuntos, os autores construíram os agrupamentos de referência com base em informações externas ao agrupamento, encontradas em redes regulatórias, e estabeleceram três diferentes critérios para a construção dos módulos: (i) mínimo, critério pelo qual genes que possuem ao menos um regulador em comum são considerados pertencentes ao mesmo módulo; (ii) estrito, critério pelo qual apenas genes que possuem exatamente os mesmos reguladores em comum são considerados um módulo; e (iii) subgrafos interconectados, critério segundo o qual os módulos do agrupamento de referência são construídos com base em uma análise de grafos da rede regulatória.

Uma vez estabelecidos os módulos de referência, os autores aplicaram uma busca em grade, de forma que foram executados os experimentos com as possibilidades de combinação de parâmetros e de métodos para cada um dos conjunto de dados avaliados.

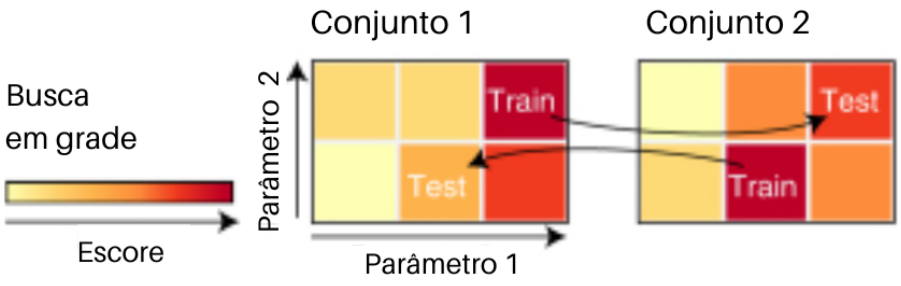
Foram, então, calculados índices de validação externos, anteriormente descritos, para cada método, experimento e agrupamento de referência. Para evitar o sobreajuste do modelo, os autores atribuiram para cada método e conjunto de dados, uma pontuação de treinamento e de teste, com base nos índices de validação externos. O procedimento

utilizado para o cálculo da pontuação de treinamento e de teste será descrito na seção 2.7.1.

2.7.1 Cálculo da pontuação (escore) de treinamento e de teste

O cálculo da pontuação de treinamento e de teste pode ser compreendido com base na figura 9. Nesse mapa de calor, a posição de cada elemento quadriculado representa determinada configuração de parâmetros utilizada na busca em grade, e é equivalente entre as figuras, ou seja, um elemento com a mesma posição no mapa de calor para ambos os conjuntos representa um experimento feito com os mesmos parâmetros, mas em conjunto de dados distintos. A cor com que cada célula está preenchida simboliza o valor observado segundo um índice de validação, nesse caso, externo ($f1rprr$), para o respectivo experimento. Note-se que, em ambos os mapas de calor, "conjunto 1"e "conjunto 2", a observação mais alta obtida, representada pela cor mais intensa na escala de cores, é nomeada "treinamento". Por outro lado, um valor é considerado como "teste"é aquele observado para o experimento com os mesmos parâmetros da pontuação de treinamento do conjunto distinto, o que no mapa de calor é representado por uma posição equivalente.

Figura 9 – Esquema da da pontuação de treinamento e de teste.



Fonte: adaptado de (SAELENS; CANNOODT; SAEYS, 2018)

Esse procedimento foi feito para evitar sobreajuste do modelo e por tal motivo, no momento de discutir a pontuação dos métodos, a discussão é baseada nas pontuações de teste.

Desta forma, o cálculo da pontuação de treinamento e de teste feito por Saelens, Cannoodt e Saeys (2018) pode ser dividido em duas etapas. Primeiramente, são calculados os valores brutos, ou provisórios, de treinamento e de teste para todas as combinações de conjuntos de dados, experimentos e parâmetros de referência, representadas aqui na figura 9. Em um segundo momento, os valores brutos observados para treinamento e teste são reduzidos cada um a um único valor, chamado de pontuação. Para tanto, as pontuações são calculadas por meio da média dos valores provisórios mencionados.

Fato importante é que a pontuação, seja de treinamento ou de teste, é calculada em determinado contexto. Por exemplo, os métodos podem ser pontuados (i) de maneira geral, sem fazer distinção entre os conjuntos de dados, e nesse caso, a média é calculada considerando todos os valores brutos de treinamento ou de teste obtidos; ou (ii) por conjunto de dados, cenário em que a média dos valores brutos obtidos será calculada separadamente para cada conjunto de dados.

Por fim, Saelens, Cannoodt e Saeys (2018) utilizaram-se de mapas de calor para sumarizar os resultados das pontuações de treinamento e de teste para cada combinação de método e de conjunto de dados e de gráficos de barras quando visaram representar a pontuação dos métodos sem fazer distinção por conjunto de dados.

Neste trabalho, se fez necessária a reprodução dos experimentos descritos em (SAELENS; CANNOODT; SAEYS, 2018) por dois principais motivos: (i) o referido estudo atribui a pontuação aos métodos apenas com base em índices de validação externos; e (ii) os conjuntos de dados utilizados pelo estudo possuem dimensões maiores do que as do problema analisado neste estudo, que tem dimensões de 89 genes por seis condições experimentais.

Sendo assim, neste trabalho, (i) os experimentos de Saelens, Cannoodt e Saeys (2018) foram reproduzidos, (ii) as discussões foram estendidas aos índices de validação internos, (iii) uma nova execução dos de experimentos foi feita e adaptada para incluir o conjunto de dados de células beta-pancreáticas. Esses passos foram utilizados para apoiar a seleção da técnica e dos parâmetros mais adequados ao problema de agrupamento de dados de células beta-pancreáticas submetidas à progesterona e serão descritos no capítulo 4. No capítulo 3 será apresentada uma revisão bibliográfica sobre os trabalhos relacionados.

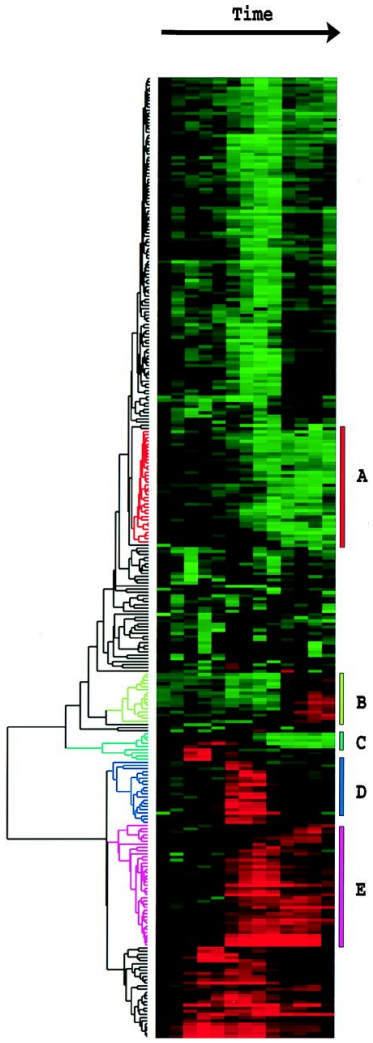
3 Revisão bibliográfica

Um dos primeiros trabalhos que trata do assunto de agrupamento de dados de expressão gênica é aquele realizado por Eisen *et al.* (1998), que aplicou a técnica de agrupamento hierárquico. A maneira de se visualizar os resultados do agrupamento é ainda uma das mais aplicadas até os dias de hoje. Tal visualização combina o dendrograma resultante do agrupamento hierárquico a um mapa de calor e permite a visualização dos padrões de expressão gênica e pode ser visualizada na figura 10, disponível em tamanho maior no anexo A. Este trabalho também constatou, a partir da verificação das anotações funcionais obtidas na literatura, que genes com funções semelhantes tendem a ser agrupados juntos, fato que norteia as análises de agrupamento de dados de expressão gênica até os dias de hoje. Os autores mencionam que um passo natural na análise dos genes dos *microarrays* é avaliar os genes com mudanças significativas de expressão. Essa técnica simples pode ser extremamente eficiente para indicar potenciais marcadores de tumores ou indicadores de drogas, por exemplo. Porém, considera-se que tais análises não se utilizam do potencial completo dos dados de *microarrays*, sendo necessária uma abordagem holística para um entendimento integrado do processo estudado.

Em (VOEHRINGER *et al.*, 2000) é avaliada a sensibilidade e a resistência de células B de linfoma de ratos, antes e depois de sofrerem irradiação, a partir de dados de 11000 genes. Para o agrupamento dos dados, os autores utilizaram a mesma técnica aplicada em (EISEN *et al.*, 1998), anteriormente descrita neste trabalho. As anotações funcionais dos genes selecionados foram levantadas e a partir dessas análises, os resultados apontaram achados sobre células sensíveis à apoptose (morte celular) e indicaram proteínas que estão envolvidas na interrupção da função mitocondrial normal, além de identificarem células resistentes à apoptose. Além disso, os autores apresentaram a contribuição de expandir o conhecimento relativo à suscetibilidade à apoptose, definindo os perfis de expressão gênica de células relacionados, ou não, à morte celular.

Vanichayobon, Siriphan e Wiphada (2007) introduziram uma nova técnica para predição de câncer a partir de dados de *microarray*. O trabalho apresenta duas contribuições principais: a seleção dos genes significativos por meio de metodologia estatística (redução da dimensionalidade) e o passo de seleção de genes por meio do agrupamento com mapas auto-organizáveis. Após isso, foi proposto um processo de criação de regras do tipo *se-então*,

Figura 10 – Representação visual que combina o dendrograma à imagem na qual os genes são representados por linhas e as colunas representam os experimentos. As cores são em função da expressão; tons de preto para neutro, tons crescentes de vermelho para valores positivos e tons de verde para os negativos.



Fonte: (EISEN *et al.*, 1998)

as quais foram subsequentemente avaliadas para compor um modelo para predição do câncer, sendo que os autores realizaram experimentos com diversos conjuntos de dados.

Gazi e Kayis (2012) realizam um estudo comparativo entre técnicas de agrupamento em dados de *microarrays*. Nesse trabalho, os autores consideraram os métodos de agrupamento hierárquico aglomerativo, k-médias, PAMSAM (*Partitioning Around Medoids with Sammon Mapping*) e HIPAM (*Hierarchical PAM*), sendo os três últimos considerados métodos particionais. A partir dos resultados das análises de *microarray*, os autores concluíram que, se o objetivo do estudo é dividir os genes em grupos disjuntos, de modo a destacar suas propriedades funcionais, o método PAMSAM pode ser sugerido como uma alternativa ao algoritmo k-médias. Mas, se o relacionamento funcional entre os

grupos também for necessário, o método HIPAM pode ser aplicado, pois fornece resultados globais mais estáveis com relação às demais abordagens.

O trabalho de Gonçalves *et al.* (2014) buscou descobrir como os genes interagem quando submetidos ao estresse, explorando técnicas de análise de dados de expressão gênica. O objetivo do trabalho consistiu em validar trabalhos anteriores sobre os genes envolvidos em caminhos de estresse e também em expandir esse conhecimento acrescentando informações de genes. Foram utilizados vários tipos de dados, incluindo dados de *microarrays*. Futuramente, os autores pretendem recriar redes regulatórias plausíveis com base nesses resultados, utilizando um modelo lógico probabilístico. Os dados de entrada utilizados foram dados de *microarrays* e dados de sequência de RNA (*RNA - Seq*). Para os dados de *microarray*, a análise foi feita utilizando o pacote *Limma (Linear Models for Microarray Data)*[1]. Os autores analisaram três situações de estresse e interceptaram os resultados para encontrar os genes mais representativos. Os dados foram validados tanto pela sobreposição entre os resultados dos conjuntos quanto comparando-os com estudos anteriores.

Segundo Ochieng, Tarigan e Didik (2016), a expressão gênica é um fenômeno biológico contínuo que pode ser representado por funções contínuas (curvas). Nesse caso, os genes com comportamento modelado pela mesma função frequentemente compartilham formas funcionais similares. No entanto, padrões como números, formas e identidades dos genes que compartilham formas funcionais semelhantes permanecem desconhecidos. Para identificar essas formas funcionais, o artigo apresenta um modelo de agrupamento para identificação de padrões de expressão gênica no tempo. O método utiliza uma abordagem *S-spline* para modelar as curvas funcionais e uma abordagem de log-verossimilhança com penalização para ajustar o modelo. Um algoritmo EM (*Expectation–Maximization*) foi também desenvolvido para minimizar o erro e o custo computacional durante a estimativa da curva média. Além disso, os autores utilizam uma técnica de validação cruzada para selecionar parâmetros de suavização e medir a incerteza de agrupamento, aplicando um critério de informação bayesiano. A importância do método é ilustrada por sua aplicação aos conjuntos de dados do ciclo de vida de uma espécie de mosca, denominada *D. melanogaster*. Os resultados da simulação indicaram que a técnica estima com precisão a curva de expressão média para formas funcionais verdadeiras, prevendo a curva média com 95% de

[1] Limma é um pacote de *software* para a linguagem *R* que propicia uma solução integrada para a análise de experimentos de expressão gênica e é descrito em (SMYTH, 2005)

confiança para cada grupo. Com base na descrição das ontologias dos genes disponíveis na literatura, a curva média estimada em cada grupo reflete anotações funcionais verdadeiras de genes com padrões de expressão gênica biologicamente significativos.

Em (MAJI; SHAH, 2017) é apresentado um novo algoritmo de seleção de genes denominado SiFS (*Significance and Functional Similarity*), cuja finalidade é identificar genes associados a doenças. Para isso, tal algoritmo elimina um subconjunto de genes dos dados de *microarrays*, maximizando tanto a significância quanto a similaridade funcional do subconjunto de genes restantes, fazendo com que os genes selecionados estejam todos associados a uma mesma função. O algoritmo utiliza como entrada a rede de interação das proteínas e os perfis de expressão dos genes, sendo a similaridade funcional dada por uma nova medida baseada nas redes de interação de proteínas (PPIN - *Protein Protein Interaction Network*). O algoritmo proposto teve o seu desempenho comparado com outras técnicas para diversos conjuntos de dados de *microarray*. Os genes indicados pelas diferentes técnicas foram avaliados quanto à sua ontologia disponível na literatura e quanto à sobreposição entre listas de genes indicadas e as listas conhecidas para determinadas funções. Desta forma, os resultados foram avaliados pela proximidade com os dados já existentes.

Devido à existência de um grande número de técnicas de agrupamento, a escolha do método mais adequado a uma dada aplicação torna-se um desafio. Nesse sentido, o trabalho em (SAELENS; CANNOODT; SAEYS, 2018) apresenta uma visão geral das características e desempenho de técnicas de agrupamento aplicadas a dados de expressão gênica e propõe uma estratégia de *benchmark* para a realização de estudos comparativos. No estudo, os métodos foram pontuados apenas segundo índices externos. De acordo com o esquema de pontuação desenvolvido pelos autores, o método que obteve o melhor desempenho foi aquele baseado em análise de componentes independentes (ICA). Estudos posteriores, como os realizados em (LAWLOR; CAO; ELLISON, 2021), (CHEN *et al.*, 2020), (TAN *et al.*, 2020), (SASTRY *et al.*, 2021) e (POUDEL *et al.*, 2020) escolheram a técnica de análise de componentes independentes (ICA) para seus experimentos, tendo como justificativa os referidos resultados do estudo de *benchmark*.

Tratando-se de índices de validação internos para agrupamento de dados de expressão gênica, Bolshakova e Azuaje (2003) realizaram uma análise sobre os índices de validação internos, e enumeram os índices da silhueta média, Dunn e Davies-Bouldin como estratégias robustas para a seleção de partições ótimas nos agrupamentos em dados

de expressão gênica. Esses mesmos índices também são avaliados em (YANG; WAN; GAO, 2006), (WIWIE; BAUMBACH; RÖTTGER, 2015), (FRATELLO *et al.*, 2022) e (BHANDARI *et al.*, 2022).

Tratando-se de seleção de métodos de agrupamento para dados de expressão gênica, o trabalho mais completo encontrado nesta revisão bibliográfica foi aquele realizado em (SAELENS; CANNOODT; SAEYS, 2018). No entanto, o trabalho não aborda os índices de validação internos, o que, em um cenário do mundo real, em que o pesquisador não tem agrupamentos de referência à sua disposição, torna-se essencial (WIWIE; BAUMBACH; RÖTTGER, 2015). Neste trabalho, a discussão sobre o estudo de *benchmark* do estado da arte feito em (SAELENS; CANNOODT; SAEYS, 2018) é estendida aos índices de validação internos mais utilizados na literatura para agrupamento de dados de expressão gênica, a saber, (i) índice de Dunn; (ii) índice de Davies-Bouldin; e (iii) índice da silhueta média.

4 Estudo da relação entre progesterona e diabetes gestacional

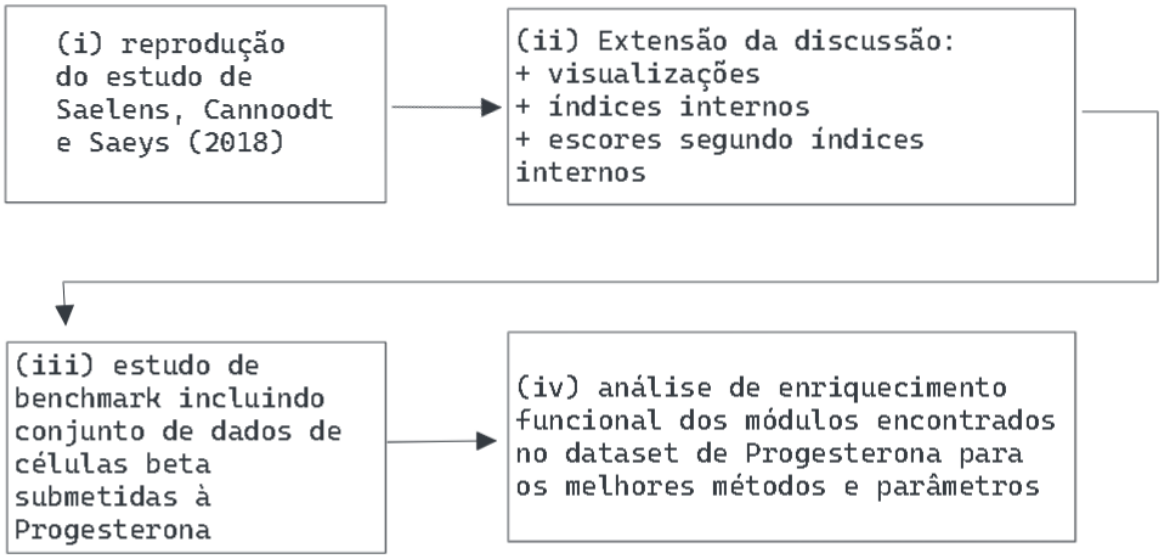
4.1 Descrição do problema

Devido ao crescente uso farmacológico de progestógenos ao longo da gravidez para a prevenção do parto prematuro (PERGIALIOTIS *et al.*, 2019), a relação entre esses hormônios e o diabetes gestacional requer atenção. Apesar de já ser conhecido que a morte de células beta-pancreáticas está associada aos diabetes tipo I e tipo II (ROJAS *et al.*, 2018), essa relação ainda precisa ser melhor compreendida no contexto do diabetes gestacional. No trabalho desenvolvido em (NUNES *et al.*, 2014), experimentos *in vitro* com a linhagem celular RINm5F revelaram que a progesterona foi capaz de induzir a oxidação e morte das células beta-pancreáticas, que são produtoras de insulina. Para investigar mais a fundo este problema, experimentos de *microarray* foram conduzidos, por esses mesmos autores, com células da linhagem RINm5F submetidas à progesterona em três doses (0,1 μM, 1 μM e 100 μM) e dois tempos (6h e 24h). O objetivo principal da presente pesquisa consiste, portanto, em analisar esses dados no intuito de obter *insights* acerca de como os genes envolvidos no balanço *redox* (genes relacionados ao estresse oxidativo) dessas células poderiam estar envolvidos na morte celular induzida por progesterona.

A análise de agrupamento em dados de expressão gênica é amplamente utilizada para auxiliar no entendimento da função dos genes, regulação gênica, processos e subtipos celulares e tem sido consistentemente aplicada com a finalidade de identificar e analisar diversas patologias, como câncer, malária e tuberculose (DALTON; BALLARIN; BRUN, 2009). No entanto, diante da variedade de técnicas de agrupamento existentes, a escolha daquela mais adequada ao problema de análise de dados de expressão gênica torna-se um desafio. O estudo desenvolvido por Saelens, Cannoodt e Saeys (2018) buscou preencher tal lacuna, avaliando, segundo índices de validação externa e com base em módulos conhecidos, vários métodos de agrupamento e propôs uma metodologia para a realização de estudos comparativos envolvendo detecção de módulos em dados de expressão gênica. No entanto, em cenários do mundo real, o pesquisador muitas vezes só tem à sua disposição os índices de validação internos, não dispondo de módulos conhecidos a respeito dos genes (WIWIE; BAUMBACH; RÖTTGER, 2015), como é o caso do problema abordado no presente trabalho. Além disso, os conjuntos de dados utilizados no estudo citado possuem dimensões bem maiores daquelas do problema aqui investigado, que possui dados de 89 genes em seis

medições distintas. Desta forma, para proceder com a análise de agrupamento dos dados das células beta-pancreáticas submetidas à progesterona, fez-se necessário: (i) reproduzir o estudo de *benchmark* proposto em (SAELENS; CANNOODT; SAEYS, 2018); (ii) calcular os índices de validação internos para os resultados analisados no estudo de *benchmark* e estender a discussão do estudo de *benchmark* aos índices de validação internos, propondo-se que a pontuação do desempenho dos métodos seja feita de acordo também com estes índices; (iii) refazer o procedimento de avaliação segundo índices externos e segundo índices internos adicionando aos conjuntos de dados avaliados o conjunto, analisado e publicado pela primeira vez neste estudo, de células beta-pancreáticas submetidas à progesterona e (iv) a partir dos resultados obtidos, selecionar a técnica adequada ao problema de agrupamento no conjunto de dados de células beta-pancreáticas e interpretar os resultados por meio da técnica do enriquecimento funcional de genes e da validação com um especialista do domínio, visto que trata-se de um trabalho interdisciplinar. Estas etapas estão esquematizadas na figura 11 e serão descritas nas seções que seguem.

Figura 11 – Etapas da metodologia do trabalho



4.1.1 Reprodução do estudo de *benchmark* proposto em (SAELENS; CANNOODT; SAEYS, 2018)

O estudo em (SAELENS; CANNOODT; SAEYS, 2018) avalia vários métodos de agrupamento, que, no contexto de expressão gênica, também são conhecidos como métodos

de detecção de módulos, aplicados a dados de expressão gênica. A pontuação para cada método avaliado foi calculada com base em módulos conhecidos, também chamados de *gold standards*, que são agrupamentos de referência, ou seja, são grupos de genes que foram construídos com informações externas ao agrupamento. Os detalhes do procedimento executado são descritos adiante.

- **Métodos avaliados**

Neste trabalho, os experimentos do estudo de *benchmark* foram reproduzidos para os seguintes métodos: (i) aleatório, para ser usado como referência, (ii) agrupamento hierárquico aglomerativo, selecionado por ter apresentado, no estudo de referência, o melhor desempenho entre os métodos de agrupamento clássicos; (iii) Análise de componentes independentes, na implementação do ICA z-scores, por ter apresentado o melhor desempenho no estudo de *benchmark*; o (iv) bi-agrupamento espectral, que foi o método de bi-agrupamento melhor avaliado, e os métodos (v) agrupamento de deslocamento médio (*mean shift*) e (vi) k-médias, por serem métodos muito utilizados na literatura. Para cada conjunto de dados e método, os parâmetros dos experimentos foram variados em combinações segundo a tabela 2, em um procedimento de busca em grade.

Tabela 2 – Métodos e parâmetros combinados na busca em grade

Método(s)	Parâmetros
k-médias, hierárquico aglomerativo, aleatório	k ∈ {25, 50, 75, ..., 275, 300 }
Deslocamento médio	*bandwidth* ∈ {'auto', 2.5, 5.0, 7.5, ..., 67.5, 70.0 }
ICA z-score:	k ∈ {50, 100, 150, ..., 550, 600 } *stdcutoff* ∈ {0.5, 1.0, 1.5, 2.0, 2.5, ..., 6.5, 7.0 }
Bi-agrupamento espectral	n ∈ {10, 20, 50, 100, 200, 300, 400, 500 } ngenes ∈ {10, 20, 50, 100, 200, 300, 400, 500 }

- **Conjuntos de dados utilizados**

Saelens, Cannoodt e Saeys (2018) utilizaram os conjuntos de dados (i) *E. Coli Colombos* e (ii) *E. Coli Precise 2* (MEYSMAN *et al.*, 2014); (iii) *E. Coli Dream 5*, que originalmente foi obtido pelos autores a partir do site do desafio de inferência da rede *DREAM5*[1] (MARBACH *et al.*, 2012); (iv) *Yeast GPL 2529*, que foi resultado da agregação

[1] synapse.org/#!Synapse:syn2787209/wiki/70349

de um compêndio de expressão integrando dados de 127 experimentos (filtrados em amostras de *S. cerevisae*), usando a plataforma *GPL2529* da *Gene Expression Omnibus*[2]; (v) *Yeast Dream 5*, conjunto obtido a partir do *site DREAM5*; (vi) *Synth. Ecoli RegulonDB*, conjunto de dados sintéticos obtidos a partir da rede *E. coli RegulonDB* usando o simulador *GeneNetWeaver* (SCHAFFTER; MARBACH; FLOREANO, 2011), simulador que modela a regulação gênica usando um modelo termodinâmico detalhado e simula esse modelo usando equações diferenciais ordinárias. Diferentes condições experimentais foram simuladas usando a configuração *"Multifactorial Perturbations"*, em que as taxas de transcrição para um subconjunto de genes são perturbadas aleatoriamente; (vii) *Synth Yeast Macisaac*, conjunto de dados sintéticos obtido por meio do uso do simulador GeneNetWeaver a partir da rede *MacIsaac* de levedura; (viii) *Human TCGA*, obtidos pelos autores com base em um estudo de 12 tipos de câncer [3]; (ix) *Human GTEX*, conjunto que contém perfis de expressão de diferentes órgãos de centenas de doadores, foi originalmente baixado do site *GTEX* [4]; e (x) *Human SEEK GPL5175*, que foi composto por meio de uma agregação de conjuntos de dados públicos usando a plataforma de *microarray GPL5175*[5] e foi recuperado do estudo feito em (CONSORTIUM *et al.*, 2015). Na tabela 3 são apresentadas as dimensões (quantidade de genes e de experimentos) de cada um dos conjuntos de dados utilizados.

Tabela 3 – Quantidades de genes e de experimentos por conjunto de dados

Conjunto de dados	**nº de genes**	**nº de condições experimentais**
E. coli (COLOMBOS)	2093	2470
E. coli (DREAM5)	2442	805
E. coli (PRECISE2)	4211	815
Human (GTEX)	5177	8555
Human (SEEK)	4437	2308
Human (TCGA)	5888	3602
Synthetic (E. coli)	1509	1509
Synthetic (Yeast)	1790	1790
Yeast (DREAM5)	3292	536
Yeast (GPL2529)	3178	3025

Para o cálculo da avaliação segundo índices externos foram utilizados agrupamentos de referência. A obtenção dos agrupamentos de referência para cada um dos conjuntos de dados será descrita adiante.

[2] ncbi.nlm.nih.gov/geo
[3] synapse.org/#!Synapse:syn1715755
[4] gtexportal.org
[5] seek.princeton.edu

- **Agrupamentos de referência**

Para os conjuntos de dados de *benchmark* foram utilizados os *gold standards* fornecidos pelos autores, que foram obtidos segundo as definições de módulos: (i) *minimal*, que definem como módulos dados que tenham ao menos um elemento em comum; (ii) *strict*, segundo a qual os módulos têm exatamente os mesmos reguladores em comum; e (iii) *interconnected subgraphs*, baseados na construção de grafos a partir das redes regulatórias. Essas definições foram aplicadas para todos os conjuntos de dados, exceto aos de organismo humanos, para os quais Saelens, Cannoodt e Saeys (2018) elaboraram grupos de referência baseados em circuitos regulatórios. Os conjuntos de dados, bem como os agrupamentos de referência, foram disponibilizados por Saelens, Cannoodt e Saeys (2018).

- **Avaliação dos métodos**

Para o cálculo dos escores, foi utilizado o mesmo esquema de pontuação com base em índices de validação externa do estudo de *benchmark* (ver o capítulo 2). Para a visualização dos resultados dos escores dos métodos reproduziu-se a figura do mapa de calor em (SAELENS; CANNOODT; SAEYS, 2018) com as pontuações obtidas por método e por conjunto de dados. Esses resultados serão discutidos na seção 4.2.

4.1.2 Extensão da discussão do estudo de *benchmark* proposto em (SAELENS; CANNOODT; SAEYS, 2018)

Para estender a discussão do estudo de *benchmark* aos índices internos, foram calculados os índices de Dunn, Davies-Bouldin e índice da silhueta média para cada experimento. Em seguida, foram calculadas as pontuações segundo cada índice, de maneira análoga ao que foi feito para a avaliação externa ($F1rprr$). A diferença aqui é que foi utilizado, no lugar do $F1rprr$, cada índice interno mencionado, resultando em três escores (Davies-Bouldin, silhueta média, e Dunn). Para a visualização dos resultados dos escores segundo índices internos dos métodos, foram criados mapas de calor com as pontuações obtidas por método e por conjunto de dados segundo cada índice.

4.1.3 Experimentos com o conjunto de dados de células beta-pancreáticas submetidas à progesterona

Nesta etapa, os experimentos foram feitos para todos os conjuntos de dados de maneira a se ter uma pontuação para o conjunto de dados de células beta-pancreáticas submetidas à progesterona. Como a estratégia de treinamento e de teste envolve o uso de diversos conjuntos de dados de expressão gênica, foi necessário reproduzir todos os experimentos da etapa anterior com os parâmetros que contemplassem a comparação com o conjunto de dados da progesterona. Para isso, foram utilizados os mesmos conjuntos de dados da etapa anterior, exceto aqueles de organismos humanos, pelo fato de utilizar padrões de referência diferentes dos demais conjuntos.

O conjunto dados de células beta-pancreáticas submetidas à progesterona, contendo informações de 89 genes, foi obtido por meio de experimentos com *microarrays* utilizando o *kit RT2 Profiler PCR Arrays*, do fabricante *Qiagen*[6]. Foram realizados quatro experimentos, sem réplicas, sendo que um deles foi considerado como grupo de controle, sem a submissão dos genes à progesterona, a fim de medir as variações de expressão dos genes em relação às condições normais das células. As células foram submetidas às concentrações de 0.1 μM, 1 μM e 100 μM e as expressões foram coletadas com 6h e 24h a partir do início do experimento, resultando em seis medições. Os seis experimentos foram numerados para fins de visualização. Sendo assim, os experimentos *0.1 μM 6h,0.1 μM 24h,1 μM 6h,1 μM 24h,100 μM 6h,100 μM 6h* são referenciados por I, II, III, IV, V e V, respectivamente. O conjunto de dados completo está disponível no apêndice H deste trabalho.

Além dos módulos já descritos utilizados por Saelens, Cannoodt e Saeys (2018), foi necessário obter módulos de referência para o conjunto de dados de *microarray* de células beta-pancreáticas submetidas à progesterona. Visando obter os melhores agrupamentos de referência possíveis para conjunto de dados original deste trabalho, diante da ausência de redes regulatórias, como aquelas aplicadas em (SAELENS; CANNOODT; SAEYS, 2018), os módulos de referência foram obtidos por meio de duas formas: (i) a partir da aplicação da técnica conhecida como análise de rede de coexpressão de genes ponderada, por meio do algoritmo WGCNA (LANGFELDER; HORVATH, 2008) (utilizou-se o tamanho mínimo de módulo como um gene); e (ii) por meio da construção de módulos a partir de uma busca na ferramenta *Cytoscape* (SHANNON *et al.*, 2003), que é uma plataforma de *software*

[6] https://www.qiagen.com/us/

destinada à visualização de redes de interação molecular e vias biológicas e à integração dessas redes com anotações e perfis de expressão gênica. Nesse caso, foi selecionada a base de dados CTD (*Comparative Toxicogenomics Database*) (DAVIS *et al.*, 2021), por ter retornado maior correspondência na busca por meio da ferramenta. O CTD é um banco de dados robusto e disponível publicamente que visa avançar no entendimento sobre como as exposições ambientais afetam a saúde humana, fornecendo informações selecionadas manualmente sobre interações química-gene/proteína, relações química-doença e gene-doença. Os dados relacionados aos genes do presente estudo foram obtidos do CTD e analisados segundo os critérios de corregulação mínima (definiu-se como módulos os genes que tivessem ao menos um elemento da rede regulatória em comum) e estrita (definiu-se como módulos os genes que tivessem exatamente os mesmos elementos da rede regulatória em comum).

Para a análise dos genes submetidos à progesterona, também foram executados os experimentos para os seguintes métodos: (i) aleatório, (ii) agrupamento hierárquico aglomerativo, (iii) Análise de componentes independentes, na implementação do ICA z-score; (iv) bi-agrupamento espectral; e (v) agrupamento de deslocamento médio (*mean shift*) e (vi) k-médias. A variação dos parâmetros no procedimento da busca em grade precisou ser adaptada para contemplar as dimensões do conjunto de dados das células beta-pancreáticas submetidas à progesterona, conforme a tabela 4. Por fim, foram calculados os escores com base em índices externos e internos para esses novos experimentos.

Tabela 4 – Métodos e parâmetros combinados na busca em grade (incluindo os dados de Progesterona)

Método(s)	Parâmetros
k-médias, hierárquico aglomerativo, aleatório	k ∈ {2, 3, 4, 5, 6, 7, 8 }
Deslocamento médio	*bandwidth* ∈ {'auto', 2.5, 5.0, 7.5, ..., 67.5, 70.0 }
ICA z-score:	k ∈ {2, 3, 4, 5, 6, 7, 8 } *stdcutoff* ∈ {0.5, 1.0, 1.5, 2.0, 2.5, ..., 6.5, 7.0 }
Bi-agrupamento espectral	n ∈ {2, 3, 4, 5, 6, 7 } ngenes ∈ {2, 3, 4, 5, 6, 7 }

4.1.4 Análise de enriquecimento funcional dos grupos detectados no conjunto de células beta-pancreáticas submetidas à progesterona

A análise de enriquecimento funcional é etapa fundamental na obtenção de conhecimento a partir dos resultados do agrupamento, por tratar-se de uma maneira organizada de acrescentar aos módulos obtidos o conhecimento existente sobre os genes. Para isso, foram utilizadas as ferramentas *Enrichr* (KULESHOV *et al.*, 2016) e *gget* (LUEBBERT; PACHTER, 2022). As bases consultadas foram *DisGeNET* (PIÑERO *et al.*, 2020), *KEGG 2021 Human* (KANEHISA *et al.*, 2016), *GO Biological Process 2021*, *GO Cellular Component 2021*, *GO Molecular Function 2021* (CONSORTIUM, 2015) (CONSORTIUM, 2021), *Jensen TISSUES* (PALASCA *et al.*, 2018), *Jensen COMPARTMENTS* (BINDER *et al.*, 2014) e *Jensen DISEASES* (PLETSCHER-FRANKILD *et al.*, 2015). Conforme recomendado por Sollero e Grynberg (2020), foram utilizados, para a discussão do enriquecimento funcional de genes, somente aqueles resultados que acrescentaram informações do ponto de vista da especialista no domínio do problema, a Profa. Dra. Anna Karenina Azevedo Martins.

Vale ainda ressaltar que, para a execução dos experimentos, foi necessária a utilização dos recursos de computação de alto desempenho[7] da Universidade de São Paulo. As linguagens de programação utilizadas nos experimentos foram *Python* e R.

4.2 Discussão de resultados

Nesta seção, serão discutidos os resultados obtidos durante cada etapa da realização dos experimentos, que partiu da reprodução do estudo comparativo feito por Saelens, Cannoodt e Saeys (2018), seguida da extensão do estudo para incluir índices de validação internos e da proposta da pontuação do desempenho dos métodos com base nesses índices. Uma vez que, no estudo comparativo de referência, não foram utilizados conjuntos de dados com dimensões próximas daquelas do conjunto de dados de células beta-pancreáticas submetidas à progesterona, o estudo foi refeito considerando-se este conjunto. Por fim, as técnicas e parâmetros mais adequados à análise de agrupamento dos dados de células beta-pancreáticas submetidas à progesterona puderam, então, ser selecionados e interpretados

[7] https://www.usp.br/hpc/index.php

de acordo com a análise de enriquecimento funcional dos genes, aliada à validação da especialista do domínio, a Profa. Dra. Anna Karenina Azevedo Martins.

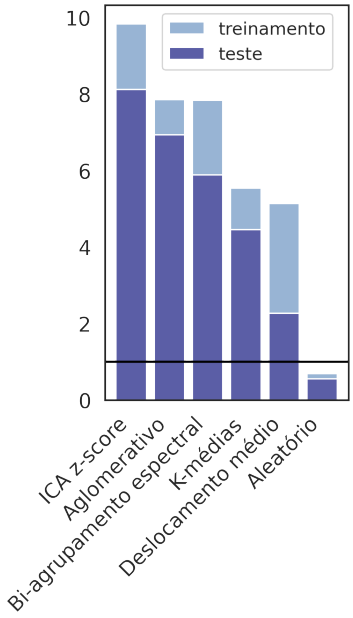
4.2.1 Discussão de resultados da reprodução do estudo de *benchmark* proposto em (SAELENS; CANNOODT; SAEYS, 2018)

Os experimentos realizados nesta etapa resultaram em 26.789 execuções, considerando conjuntos de dados, métodos, parâmetros e agrupamento de referência distintos, conforme detalhes apresentados no apêndice A. Após a execução dos experimentos, foram calculadas as pontuações de treinamento e de teste, com base em índices de validação externa, considerando a medida $f1rprr$ (ver seção 2), para cada método e conjunto de dados. A pontuação geral dos métodos, sem fazer distinção por conjunto de dados, pode ser vista na representação de um gráfico de barras na figura 12 ou na forma de representação tabular, na tabela 5.

Tabela 5 – Escores $f1rprr$ de treinamento e de teste para os métodos de agrupamento aplicados na etapa de reprodução do estudo de *benchmark*

Método	Treinamento	Teste
Aleatório	0.687661	0.549581
Deslocamento médio	5.134528	2.264274
k-médias	5.539047	4.452492
Bi-agrupamento espectral	7.839478	5.887325
Aglomerativo	7.854651	6.931690
ICA z-score	9.831074	8.116071

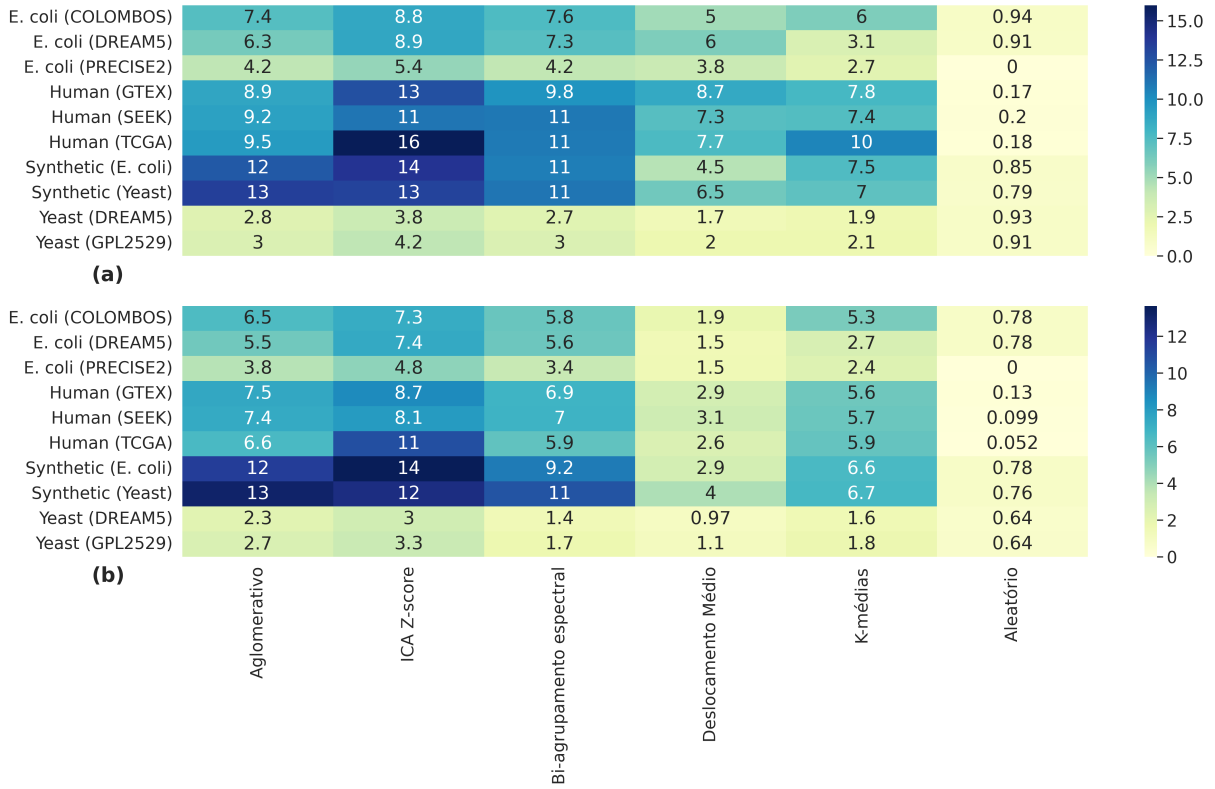
Figura 12 – Gráfico de barras contendo os escores $f1rprr$ de treinamento e de teste para os métodos de agrupamento aplicados na etapa de reprodução do estudo de *benchmark*



Os resultados dispostos na figura 12 estão ordenados de forma decrescente de acordo com a pontuação de teste pelo fato de que tal pontuação visa evitar o sobreajuste dos modelos. Dentre os métodos avaliados, é possível perceber que o agrupamento de deslocamento médio foi aquele que teve a maior diferença entre a pontuação de treinamento e de teste, mostrando uma maior sensibilidade ao sobreajuste dos modelos. Tais resultados estão de acordo com aqueles obtidos por Saelens, Cannoodt e Saeys (2018), conforme o esperado, e segundo essa avaliação, o método que obteve o melhor desempenho é o ICA z-scores, seguido pelos métodos de agrupamento hierárquico aglomerativo, bi-agrupamento espectral, k-médias e agrupamento de deslocamento médio (*mean shift*).

A distribuição das medidas dos escores de treinamento e de teste utilizados no cálculo da pontuação podem ser conferidos, respectivamente, nos apêndices B e C. Os resultados da pontuação de treinamento e de teste para cada método e conjunto de dados podem ser observados no mapa de calor apresentado na figura 13.

Figura 13 – Escores $f1rprr$ **(a)** de treinamento e **(b)** de teste por método e por conjuntos de dados



Com base nos resultados representados na figura 13, foram feitas as seguintes considerações sobre o desempenho dos métodos avaliados: (i) para todos os conjuntos de dados, conforme o esperado, o agrupamento aleatório foi aquele com o pior desempenho; e (ii) ao comparar o desempenho dos métodos entre os conjuntos de dados, no geral todos os métodos tiveram um menor desempenho quando avaliados com base nos conjuntos de dados *E. coli (PRECISE2)*, *Yeast (GPL2529)* e *Yeast (DREAM5)*. Apesar disso, a pontuação geral dos métodos quando comparados entre si manteve a ordem vista na figura 12.

A etapa de reprodução dos experimentos do estudo de *benchmark* e o fato de ter-se conseguido atingir resultados condizentes com o do estudo de Saelens, Cannoodt e Saeys (2018) foi fundamental para que se pudesse expandir a discussão aos índices de validação internos, que serão discutidos adiante.

4.2.2 Extensão da discussão do estudo de *benchmark* proposto em (SAELENS; CANNOODT; SAEYS, 2018)

O estudo comparativo em (SAELENS; CANNOODT; SAEYS, 2018) considera, para atribuir a pontuação aos métodos de agrupamento, somente índices de validação externos. No entanto, em cenários do mundo real, muitas vezes o pesquisador só tem à sua disposição os índices de validação internos (WIWIE; BAUMBACH; RÖTTGER, 2015). Portanto, nesta etapa calculou-se também para todos os experimentos decorrentes da reprodução do estudo de *benchmark*, os índices de validação internos Dunn, Davies-Bouldin e silhueta média. Em seguida, atribuiu-se uma pontuação para os métodos considerando-se cada um dos índices, seguindo o mesmo procedimento feito para a pontuação com índices externos. As análises segundo cada pontuação baseada em índices de validação internos serão descritas a seguir.

Os resultados para o índice de Dunn (que é melhor quanto maior for o valor do índice) estão dispostos na figura 14 e na tabela 6. Contrariando o que foi obtido segundo os índices de validação externos, para a pontuação do índice de Dunn, o método de análise de componentes independentes (ICA) foi o método com o pior desempenho, inclusive em comparação com o agrupamento aleatório. Sob essa óptica, o método de agrupamento melhor avaliado é o k-médias. Observando-se a figura 14 é possível perceber que, de maneira similar ao o que ocorreu na pontuação segundo os índices externos para estes experimentos, o resultado mais sensível ao sobreajuste dos modelos foi o agrupamento de deslocamento médio.

O mapa de calor para os resultados de pontuação segundo o índice de Dunn, por método e por conjunto de dados, pode ser observado na figura 15.

Tabela 6 – Escores de treinamento e de teste por método, segundo o índice de Dunn

	Escores treinamento	**Escores teste**
Aglomerativo	0.119082	0.105372
Aleatório	0.073948	0.072018
Bi-agrupamento espectral	0.112746	0.094540
Deslocamento médio	0.401690	0.100050
ICA z-score	0.085502	0.062835
k-médias	0.196465	0.166811

Figura 14 – Gráfico de barras representando os escores de treinamento e de teste por método, segundo o índice de Dunn

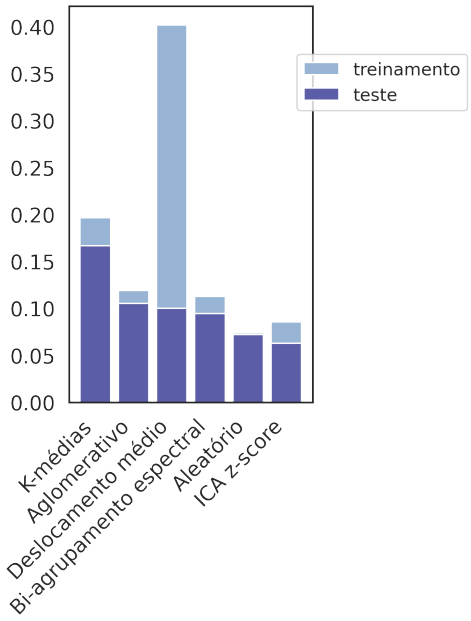
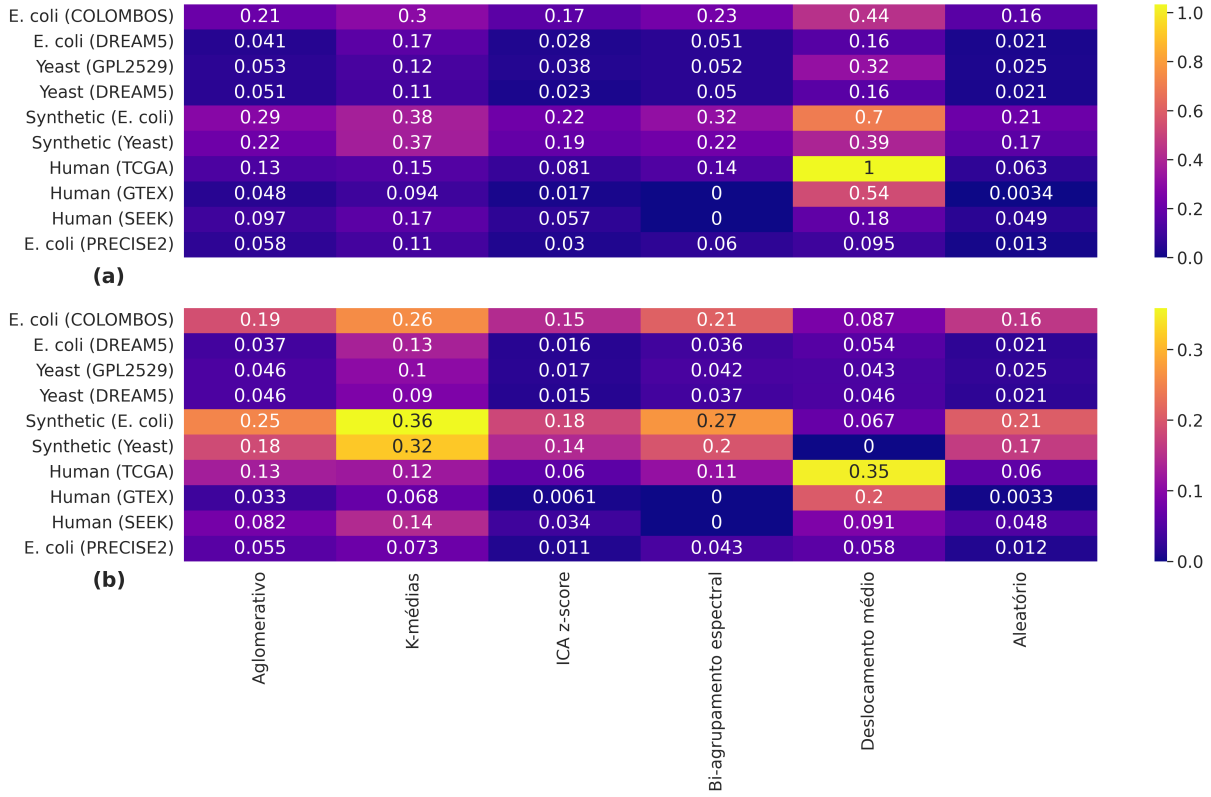


Figura 15 – Mapas de calor contendo os escores (a) de treinamento e (b) de teste por método e por conjunto de dados segundo o índice de Dunn

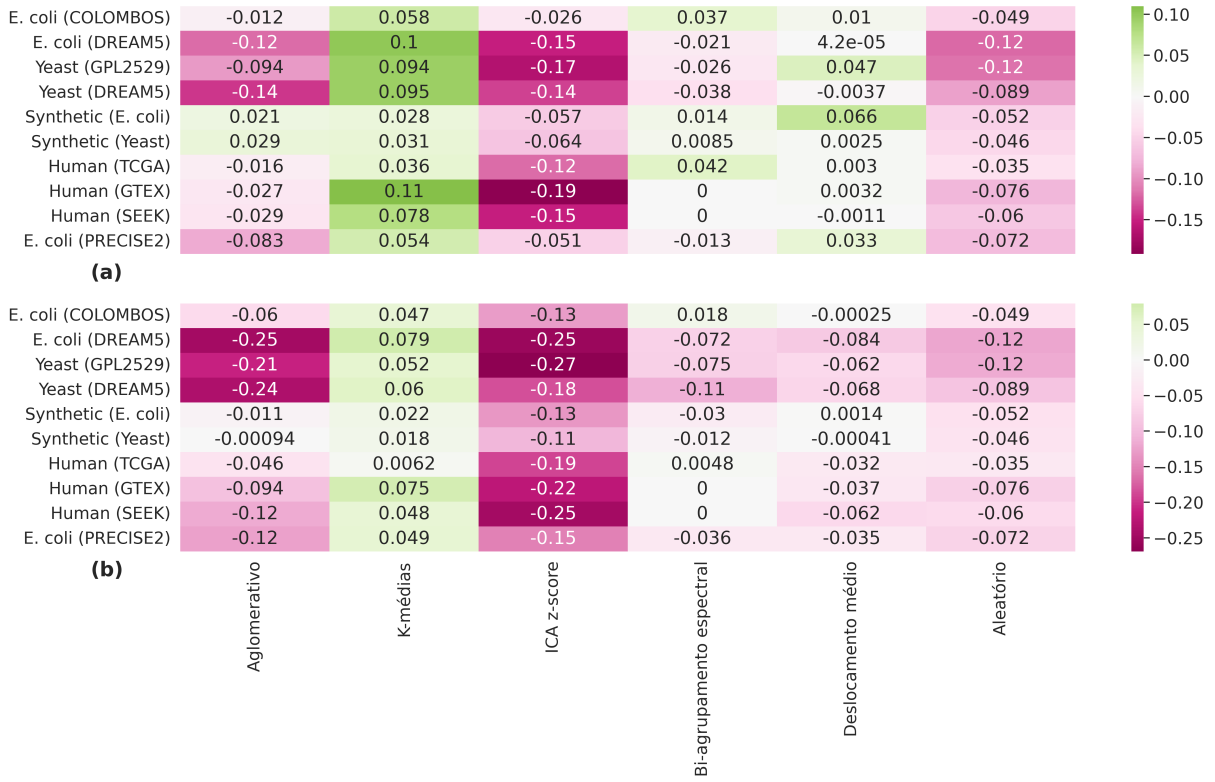


Conforme é possível observar na figura 15 por meio do auxílio do recurso visual das cores, o método que foi predominantemente avaliado com o melhor desempenho nos

dados de teste foi o k-médias, seguido pelo agrupamento hierárquico aglomerativo. O agrupamento de deslocamento médio obteve o melhor desempenho para dois dos conjuntos de dados, *Human (GTEX)* e *Human (TCGA)*.

O mapa de calor para o índice da silhueta média pode ser visto na figura 16. Vale lembrar que, para esse índice, um valor é considerado melhor quanto mais próximo estiver de 1. Já os valores negativos são considerados indesejáveis, uma vez que indicam uma má alocação dos pontos nos grupos. Aqui, foram utilizadas cores em tons de rosa para valores negativos (indesejáveis) e em tons de verde para valores positivos (desejáveis).

Figura 16 – Mapas de calor contendo os escores (a) de treinamento e (b) de teste por método e por conjunto de dados segundo o índice da silhueta média



Na figura 16 é possível notar que o método que teve valores positivos do índice da silhueta média para todos os conjuntos de dados foi o k-médias. Aqui, novamente, o método ICA, que havia sido melhor avaliado segundo a pontuação de índices externos, foi o pior avaliado, tendo os maiores valores negativos mais extremos em relação aos demais métodos. Destaca-se que, segundo o índice da silhueta, o método k-médias foi o único método que obteve valores desejáveis.

Por fim, foi calculada a pontuação segundo o índice de Davies-Bouldin (quanto menor for o valor para esse índice, melhor o resultado). O gráfico de barras foi ordenado segundo a ordem crescente de pontuação, de modo que os métodos melhor avaliados permanecem nas primeiras posições. Segundo esta pontuação, conforme exibido na figura 17 e na tabela 7, o agrupamento de deslocamento médio foi o método avaliado com o melhor desempenho, seguido do k-médias. Aqui, o método ICA z-scores foi o penúltimo método pior avaliado, sendo melhor apenas que o agrupamento aleatório. O detalhamento da pontuação, por conjunto de dados e por método, pode ser visto no mapa de calor da figura 18, em que quanto mais próximos de verde estiverem os valores, melhores estes são; ao passo que tons crescentes de laranja representam os valores menos favoráveis de pontuação.

Figura 17 – Escores de treinamento e de teste por método, segundo o índice de Davies-Bouldin.

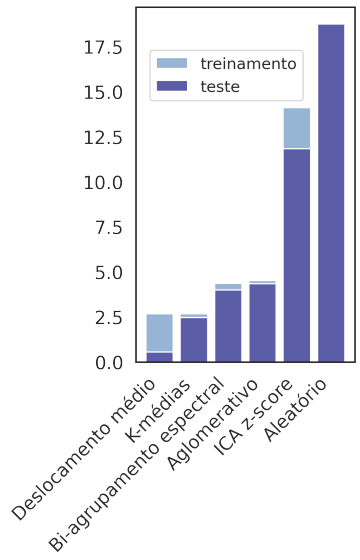
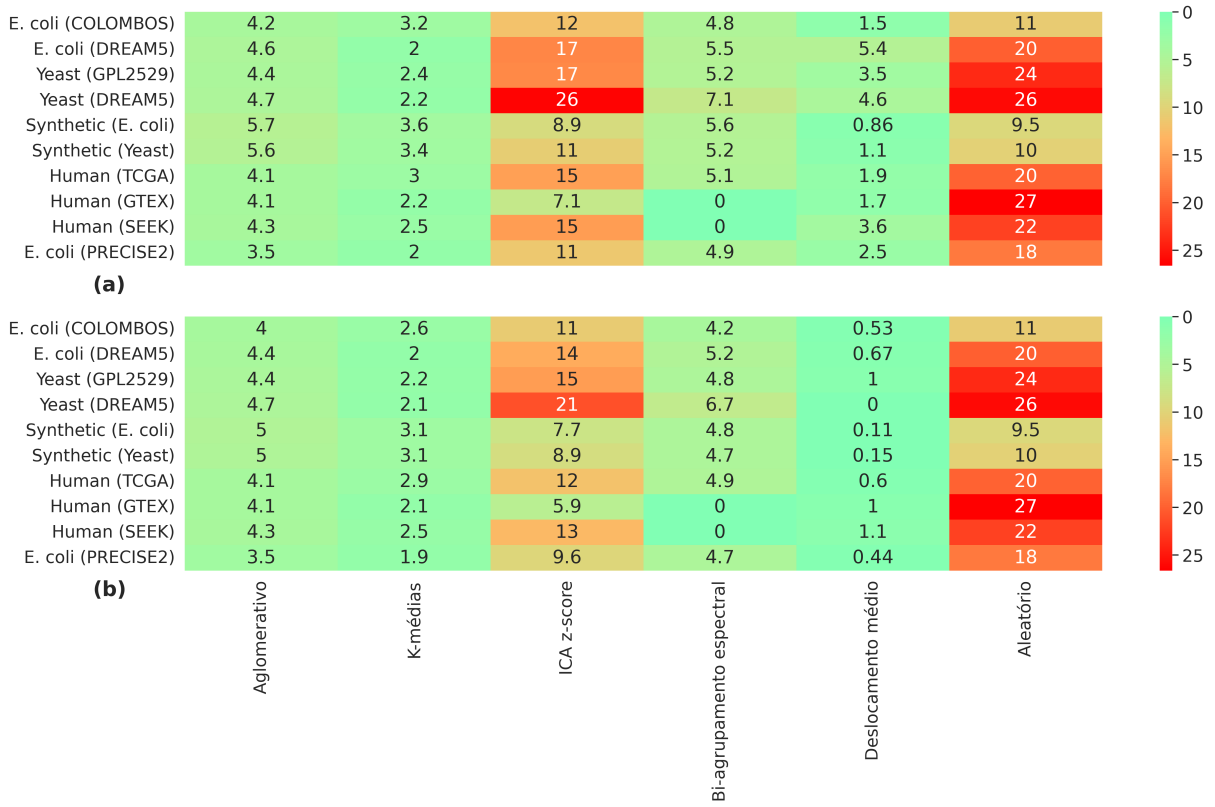


Tabela 7 – Escores de treinamento e teste por método, segundo o índice de Davies-Bouldin

	Escores treinamento	**Escores teste**
Aglomerativo	4.520315	4.338256
k-médias	2.667215	2.454307
ICA z-score	14.107479	11.822192
Bi-agrupamento espectral	4.349029	3.987412
Deslocamento médio	2.672960	0.557490
Aleatório	18.744053	18.744053

Figura 18 – Mapas de calor contendo os escores (a) de treinamento e (b) de teste por método e por conjunto de dados segundo o índice de Davies-Bouldin



É possível observar, na figura 18, que o método ICA z-scores teve o segundo pior desempenho nas pontuações de treinamento quando avaliado em relação a todos os conjuntos de dados. Segundo este índice, o método melhor avaliado foi o agrupamento do deslocamento médio, seguido pelo k-médias.

Uma vez observado que os resultados obtidos segundo a pontuação baseada em índices internos não coincide com aqueles obtidos por meio da pontuação baseada em índices externos, constatou-se a necessidade de verificar se os índices internos e externos possuem correlação. Para tanto, de forma semelhante ao que é proposto em (WIWIE; BAUMBACH; RÖTTGER, 2015), foi calculada a correlação de Pearson entre as medidas de avaliação internas e externas, considerando todos os experimentos realizados. Essa análise é apresentada na figura 19. Com base na análise da correlação de Pearson entre os resultados obtidos para os mesmos agrupamentos quando avaliados segundo índices internos e segundo índices externos, foi possível perceber que esses resultados não são fortemente correlacionados, ou seja, não necessariamente se a pontuação de um método melhora quando analisado segundo índices internos, o mesmo acontece para os índices externos. No

entanto, apesar de não serem fortes correlações, essas medidas se correlacionam. O fato de as correlações entre as medidas de avaliação externa e o índice de Davies-Bouldin serem negativas se deve à interpretação do índice, que é melhor avaliado tanto menor forem os valores obtidos.

Figura 19 – Matriz de correlação de Pearson entre índices internos e externos de todos os experimentos realizados conforme o estudo de *benchmark*

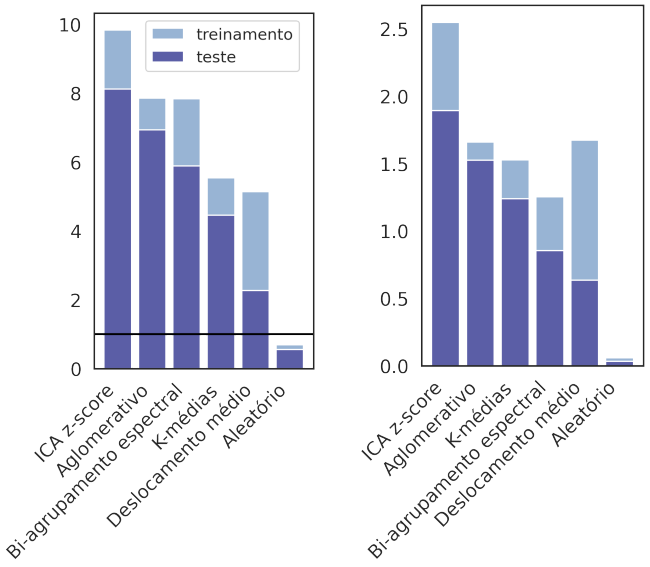
Para proceder com a análise do conjunto de células beta-pancreáticas submetidas à progesterona, integrando-o ao estudo realizado anteriormente, foi necessário: (i) refazer os experimentos, de modo a adequar a análise às dimensões do novo conjunto, especialmente no que diz respeito à seleção de parâmetros pela busca em grade; e (ii) recalcular todas as pontuações, para índices internos e externos, para refletir a adição desse conjunto na avaliação dos métodos de agrupamento. Esses resultados são descritos nas seções que seguem.

4.2.3 Discussão de resultados do estudo de *benchmark* incluindo o conjunto de dados de células beta-pancreáticas submetidas à progesterona

Aqui, refez-se todo o procedimento realizado nas etapas anteriores, considerando também o conjunto de dados de células beta-pancreáticas submetidas à progesterona, com o objetivo de selecionar a melhor técnica e parâmetros para este conjunto. Esse estudo resultou em 15.913 experimentos, envolvendo combinações distintas de conjunto de dados, métodos, parâmetros e módulos de referência, cujos detalhes podem ser vistos no apêndice D. Os métodos foram pontuados novamente segundo os índices de validação externos e internos, gerando os resultados que serão discutidos a seguir.

No que se refere à avaliação segundo índices externos, a pontuação geral dos métodos, sem fazer distinção por conjunto de dados, pode ser vista na representação de um gráfico de barras na figura 20b. Note-se que esses resultados estão de acordo com aqueles obtidos por Saelens, Cannoodt e Saeys (2018) e também com a pontuação obtida no presente trabalho durante a reprodução do estudo de *benchmark*, que pode ser visualizada na figura 20b,

exceto pelo fato de que agora os métodos bi-agrupamento espectral e k-médias tiveram as suas posições invertidas. Antes, o método que obteve o melhor desempenho foi o ICA z-scores, seguido pelos métodos de agrupamento hierárquico aglomerativo, bi-agrupamento espectral, k-médias e agrupamento de deslocamento médio (*mean shift*). Nesta nova etapa, a ordem dos métodos é ICA z-scores, agrupamento hierárquico aglomerativo, k-médias, bi-agrupamento espectral e agrupamento de deslocamento médio (*mean shift*).

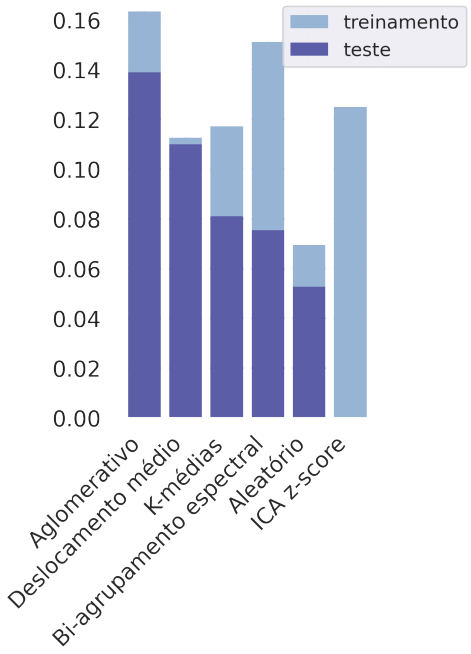


(a) (b)

Figura 20 – Comparação dos resultados do escore $f1rprr$ a partir da (a) reprodução do *benchmark* (etapa I deste trabalho) e (b) reprodução do *benchmark* incluindo o conjunto de células beta-pancreáticas (etapa III deste trabalho).

Observou-se, no entanto que a pontuação para os métodos é diferente quando analisada individualmente para o conjunto de dados de células beta-pancreáticas submetidas à progesterona, conforme a figura 21. Para este conjunto de dados, o método melhor avaliado é o agrupamento hierárquico aglomerativo, seguido pelo agrupamento por deslocamento médio, k-médias, bi-agrupamento espectral, agrupamento aleatório e ICA z-scores.

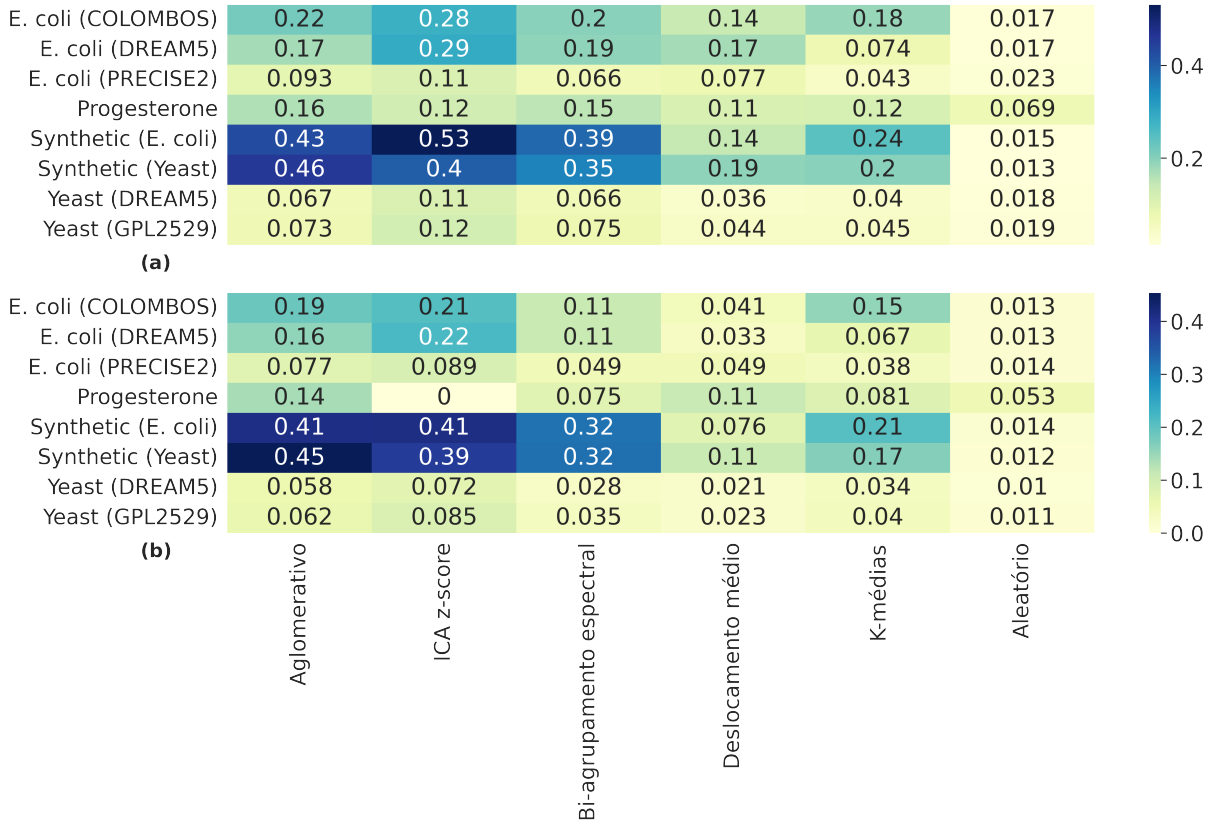
Figura 21 – Escores de treinamento e de teste segundo validação externa ($f1rprr$) somente para o conjunto de dados da progesterona.



Segundo a avaliação baseada em índices de validação externos, o método de agrupamento indicado para o problema de agrupamento de células beta-pancreáticas submetidas à progesterona é o agrupamento hierárquico aglomerativo. O pior método, contrariando a tendência para todos os conjuntos de dados quando avaliados segundo índices de validação externas, é o ICA z-scores.

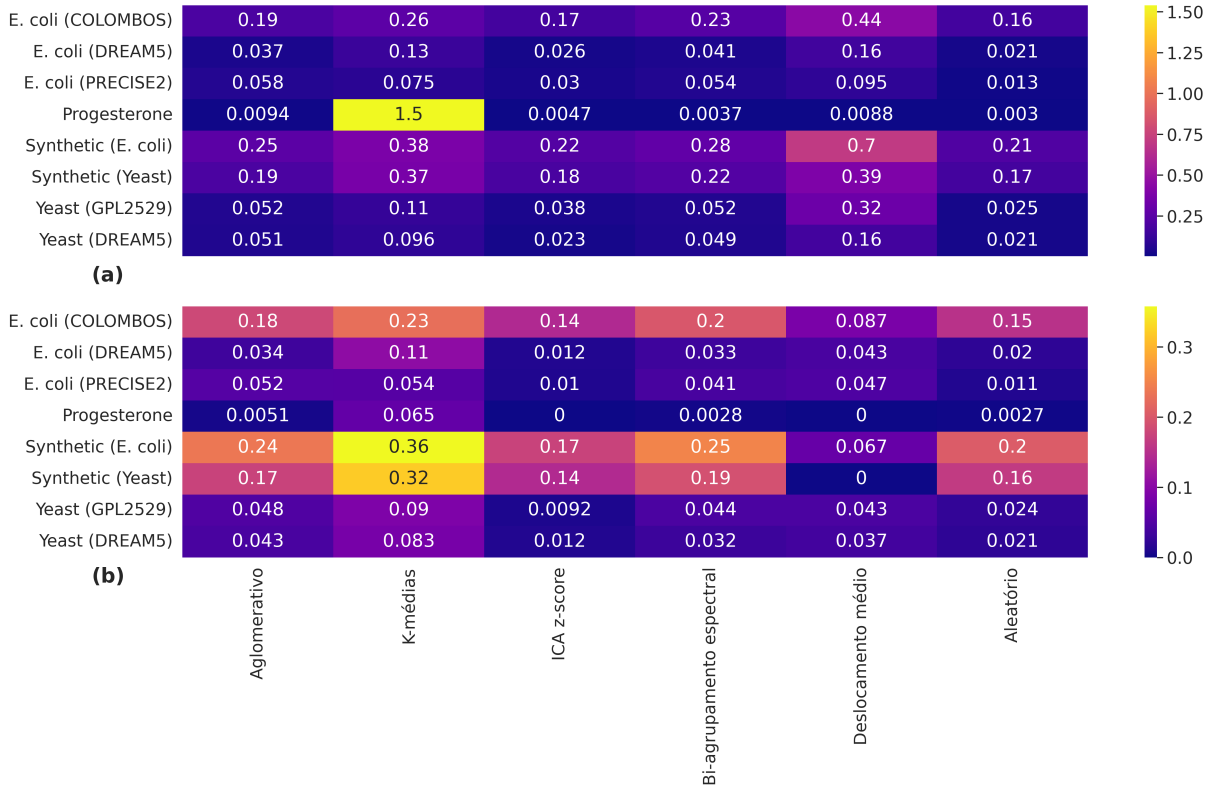
Uma representação das pontuações de treinamento e de teste para todos os conjuntos de dados pode ser visualizada na figura 22, enquanto os dados brutos utilizados no cálculo da pontuação podem ser conferidos nos apêndices E e F. Conforme a figura 22, a pontuação do desempenho dos métodos, em geral, foi maior quando avaliados com base nos conjuntos de dados *Synthetic (E. coli)* e *Synthetic (Yeast)*. Em partes, o fato da pontuação dos métodos, quando avaliados para o conjunto de dados relativos a progesterona, ter um comportamento especialmente diferente poderia ser justificado pelo fato deste ser um conjunto com dimensões diferentes dos demais, o que motivou a execução desses novos experimentos para ajuste de parâmetros.

Figura 22 – Escores segundo a validação externa ($f1rprr$) **(a)** de treinamento e **(b)** de teste por método e por conjuntos de dados



Para a avaliação segundo índices internos, o método com o melhor desempenho, quando avaliado pelo índice de Dunn, foi o algoritmo k-médias, conforme a figura 23. Esse resultado está de acordo com aquele obtido na etapa anterior, em que o k-médias também obteve o melhor desempenho, de acordo com o índice de Dunn, e, inclusive para o conjunto de dados de células beta-pancreáticas submetidas à progesterona, o método melhor avaliado é também o k-médias.

Figura 23 – Mapas de calor contendo os escores (a) de treinamento e (b) de teste por método e por conjunto de dados segundo o índice de Dunn, incluindo o conjunto de dados da progesterona



Segundo o índice da silhueta média, o método que obteve o melhor desempenho também foi o método k-médias, concordando novamente com os resultados obtidos na etapa anterior, em que se reproduziu o estudo de *benchmark*. Os valores obtidos para esta pontuação podem ser vistos na figura 24.

Figura 24 – Mapas de calor contendo os escores (a) de treinamento e (b) de teste por método e por conjunto de dados segundo o índice da silhueta média, incluindo o conjunto de dados da progesterona

Novamente, concordando com os resultados dos experimentos realizados na etapa anterior, segundo a pontuação calculada pelo índice de Davies-Bouldin, os métodos de deslocamento médio e k-médias foram melhor avaliados, conforme a figura 25. Por fim, na figura 26, é possível observar que o ICA z-scores obteve novamente o pior desempenho, superando apenas o agrupamento aleatório.

Figura 25 – Escores de treinamento e de teste segundo o índice de Davies-Bouldin para o conjunto de dados da progesterona.

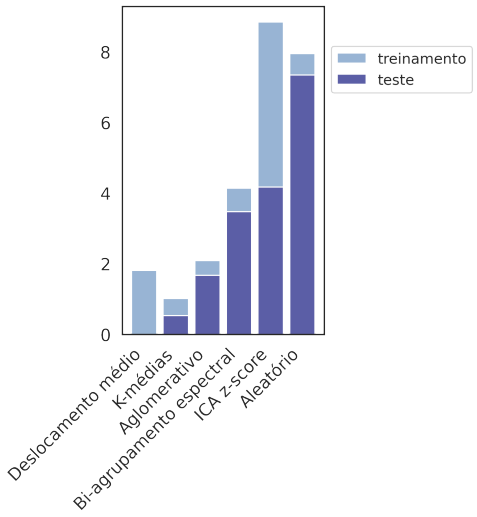
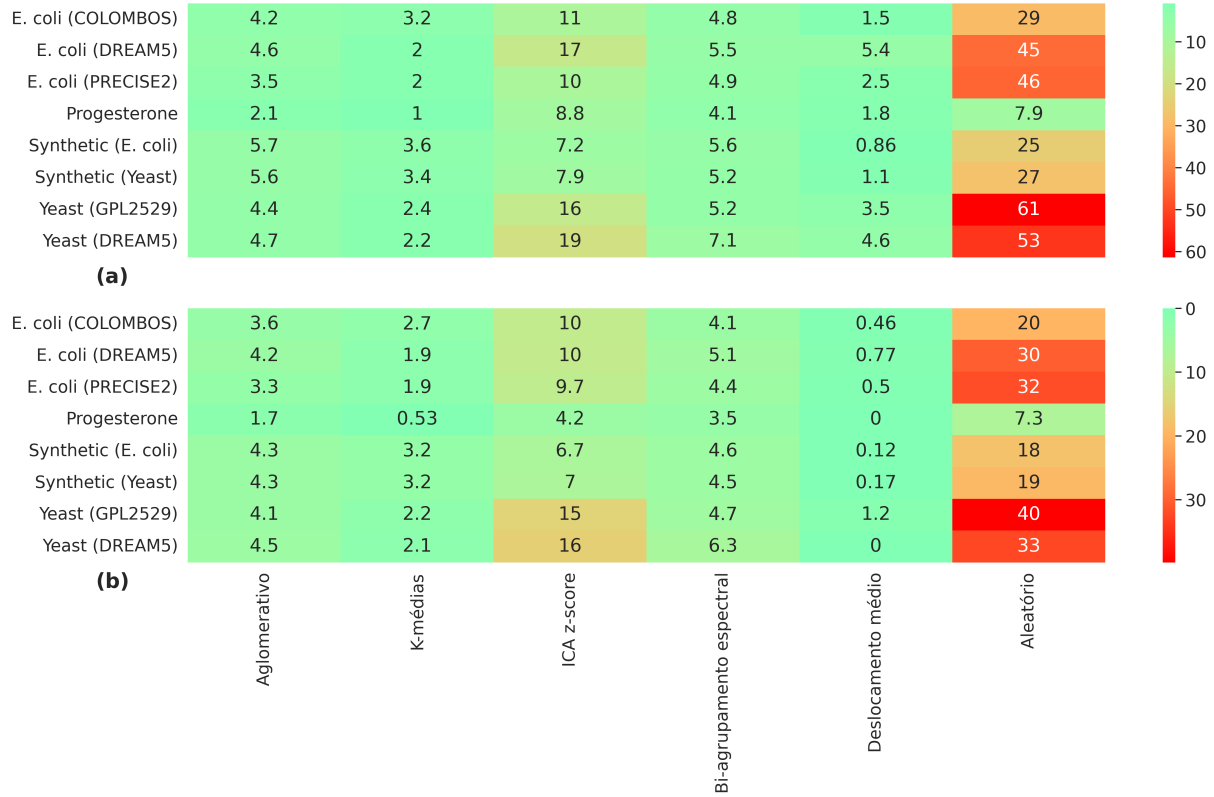


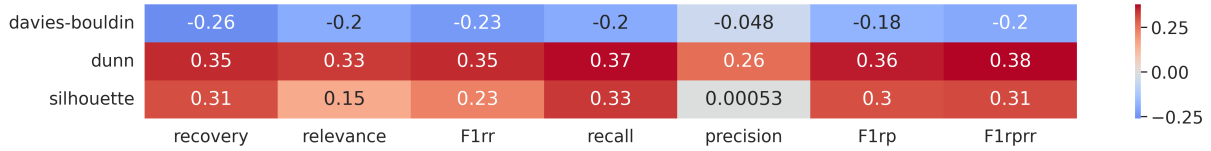
Figura 26 – Mapas de calor contendo os escores (a) de treinamento e (b) de teste por método e por conjunto de dados segundo o índice de Davies-Bouldin, incluindo o conjunto de dados da progesterona



De maneira análoga ao que foi feito na etapa II deste estudo, analisou-se a correlação de Pearson entre os índices internos e externos. Novamente, apesar de não haver fortes correlações entre esses índices, as correlações acompanham os sentidos da interpretação dos

índices, sendo negativas para o índice de Davies-Bouldin, para o qual os valores menores são os mais favoráveis. A matriz de correlação pode ser vista na figura 27.

Figura 27 – Matriz de correlação de Pearson entre índices internos e externos de todos os experimentos realizados conforme o estudo de *benchmark* considerando também o conjunto de dados de células beta-pancreáticas submetidas à progesterona.



Diante destes resultados, selecionou-se os métodos melhor avaliados segundo os índices internos e segundo os índices externos para prosseguir com a análise de enriquecimento funcional e com a interpretação dos resultados com o auxílio da especialista no domínio, a Profa. Dra. Anna Karenina Azevedo Martins. Na seção 4.2.4 serão discutidos esses resultados.

4.2.4 Análise de enriquecimento funcional dos grupos identificados no conjunto de dados de células beta-pancreáticas submetidas à progesterona

Aqui, os melhores resultados obtidos segundo cada uma das pontuações (internas e externa) foram enriquecidos funcionalmente e serão discutidos adiante.

Melhores resultados segundo índices externos ($f1rprr$)

De acordo com a pontuação baseada nos índices externos, o melhor resultado foi aquele obtido por meio do método de agrupamento hierárquico com o parâmetro k=6. Do ponto de vista da especialista do domínio, a análise de enriquecimento funcional dos grupos obtidos (ver apêndice G) não foi capaz de ampliar a compreensão do problema.

Na próxima seção serão discutidos os resultados mais relevantes do ponto de vista da especialista do domínio, a Profa. Dra. Anna Karenina Azevedo Martins, que foram obtidos com base nos índices de validação internos.

Melhores resultados segundo índices internos

Do ponto de vista dos índices internos, (i) os melhores resultados segundo as pontuações baseadas nos índices silhueta e Dunn; e (ii) o segundo melhor resultado segundo o índice de Davies-Bouldin (uma vez que, para o melhor resultado, todos os genes foram alocados em um único grupo), convergiram para o método k-médias com o parâmetro k=4.

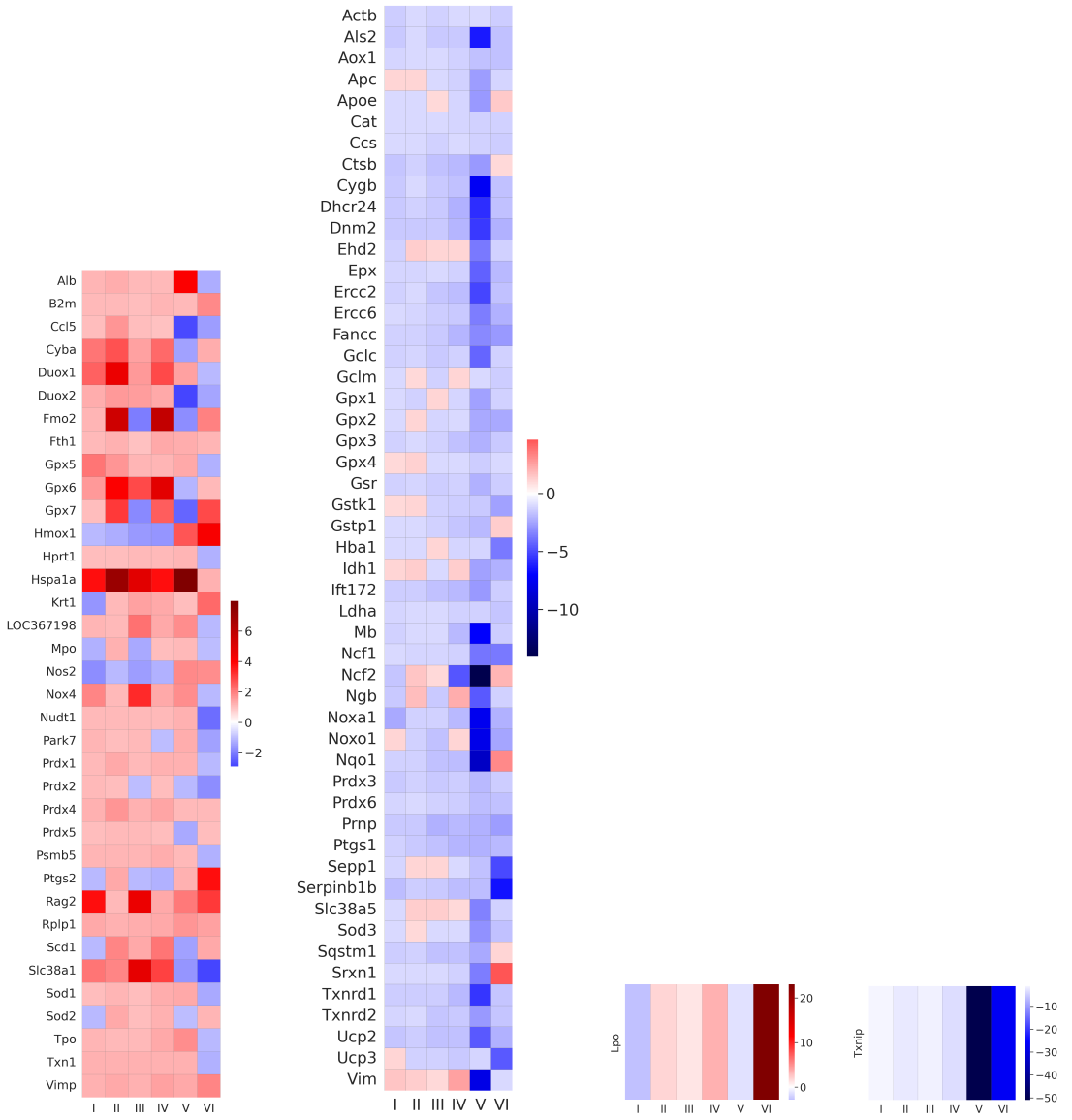
Os mapas de calor que ilustram a expressão dos genes dentro de cada um dos grupos obtidos podem ser visualizados na figura 28. Em um primeiro olhar, pode-se notar que, no geral, dois dos grupos foram predominantemente superexpressos (figura 28(a) e figura 28(c)), estando representados em tons de vermelho, e dois dos grupos foram predominantemente subexpressos (figura 28(b) e figura 28(d)), representados em tons de azul.

A interpretação desses resultados se deu por meio da combinação entre: (i) os mapas de calor; (ii) os resultado da análise de enriquecimento funcional dos genes; (iii) a opinião da especialista no domínio e; (iv) do apoio em pesquisas na literatura; e será discutida a seguir.

A análise de enriquecimento funcional foi realizada por meio de vários conjuntos de dados disponíveis no *Enrichr* e os resultados mais relevantes para o problema são discutidos aqui. Para fins de compreensão, nas próximas seções os módulos são referidos como: (i) módulo A (figura 28(a)); (ii) módulo B (figura 28(b)); (iii) módulo C (figura 28(c)); e (i) módulo D (figura 28(d)).

- **Módulo A**

Este módulo (grupo de genes), quando enriquecido funcionalmente por meio da biblioteca *DisGeNET*, apresenta termos relacionados aos diabetes tipo I e tipo II (figura 29). Este fato está de acordo com a relação do estresse oxidativo em células beta-pancreáticas e diabetes. Por outro lado, o enriquecimento funcional com a Ontologia de Genes (GO) retornou termos como *resposta celular ao estresse oxidativo* e resposta com *via de sinalização mediada por citocinas*. Por meio do princípio conhecido por *guilt by association*, pode-se inferir que os genes deste módulo estão desempenhando um papel na resposta ao estímulo da progesterona, ou um papel na defesa antioxidante.



(a) (b) (c) (d)

Figura 28 – Visão geral dos módulos obtidos por meio da melhor configuração dos experimentos. As figuras (a), (b), (c) e (d) ilustram a alocação dos genes ao longo dos quatro módulos detectados. Os experimentos *0.1 μM 6h*,*0.1 μM 24h*,*1 μM 6h*,*1 μM 24h*,*100 μM 6h*,*100 μM 6h* são referenciados por I, II, III, IV, V e V, respectivamente. As cores representam a expressão dos genes durante a condição experimental em relação ao grupo controle.

- **Módulo B**

A análise de enriquecimento deste módulo não retornou termos relacionados à doença estudada quando enriquecido por meio da biblioteca *DisGeNET*. No entanto, o enriquecimento funcional por meio da Ontologia de Genes (GO) mostra termos como *regulação negativa do processo apoptótico* (ou anti-apoptose), que por definição é qualquer processo que pare, previna ou reduza a frequência, taxa ou extensão da morte celular por

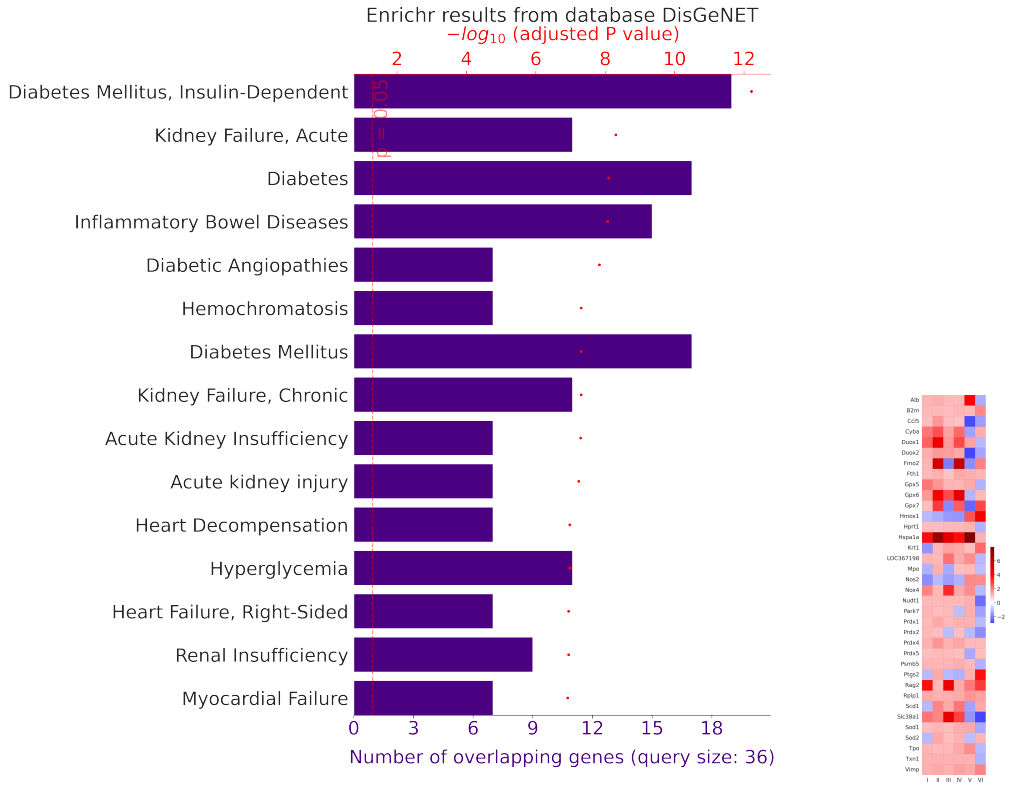
processo apoptótico. Por meio do princípio conhecido por *guilt by association*, pode-se inferir que os genes deste módulo, apesar de capazes de desempenhar um papel na defesa antioxidante, não puderam desempenhar tal papel de defesa da célula, uma vez que este módulo teve sua expressão suprimida. A supressão desses genes pode estar contribuindo para a morte celular induzida por progesterona.

- **Módulo C**

Neste módulo, um único gene foi alocado (gene LPO). A análise de enriquecimento não retornou termos relacionados à doença estudada com base na biblioteca *DisGeNET*. No entanto, os termos enriquecidos por meio da Ontologia de Genes (GO) mostram que o módulo está relacionado à regulação negativa da divisão celular, que está relacionada a processos que interrompem, impedem ou reduzem a frequência da divisão celular. Como esse gene é superexpresso, esse processo pode contribuir para uma redução da massa de células beta-pancreáticas, fator que pode contribuir para o diabetes gestacional.

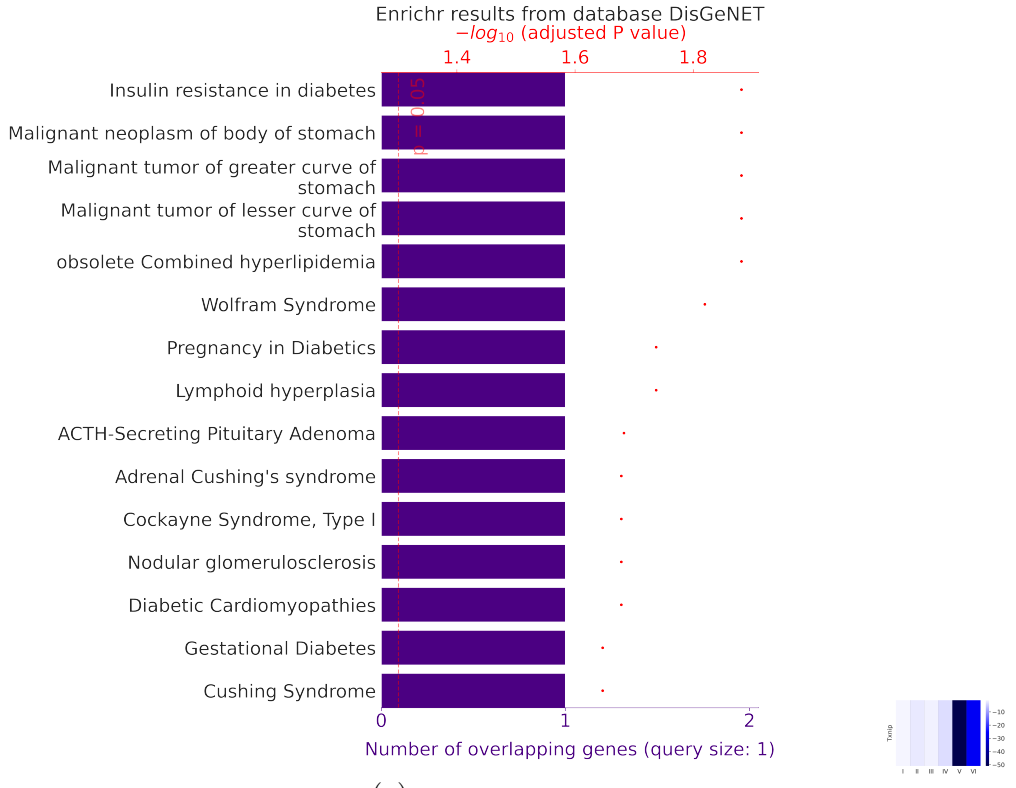
- **Módulo D**

O enriquecimento *DisGeNET* (figura 30), retornou termos como "diabetes na gravidez" e "diabetes gestacional", enquanto o enriquecimento por meio da Ontologia de Genes (GO) demonstrou que este módulo também está associado à regulação negativa da divisão celular. Isso pode contribuir para a preservação da massa de células beta-pancreáticas, que são células que se reproduzem por divisão celular. Este fato é curioso, pois ao mesmo tempo que a literatura relata que a superexpressão desse gene está relacionada aos diabetes tipo I e II (WONDAFRASH *et al.*, 2020) (LEI *et al.*, 2022) e que o estudo de Chen *et al.* (2008), aponta que a deficiência de TXNIP foi capaz de resgatar completamente camundongos do diabetes, a deficiência de TXNIP induzida pela progesterona não foi capaz de prevenir a morte das células pancreáticas. Portanto, este gene pode ser um fator chave na compreensão do problema do diabetes gestacional e necessita de mais estudos.



(a)

Figura 29 – Resultado da análise de enriquecimento funcional utilizando o DisGeNET para o módulo A



(a)

Figura 30 – Resultado da análise de enriquecimento funcional utilizando o DisGeNET para o módulo D

5 Conclusão e trabalhos futuros

Este trabalho partiu do problema de que, devido ao crescente uso farmacológico de progestógenos ao longo da gravidez para a prevenção do parto prematuro, a relação entre esses hormônios e o diabetes gestacional requer atenção. A morte de células beta-pancreáticas está associada aos diabetes tipo I e tipo II , mas ainda precisa ser melhor compreendida no contexto do diabetes gestacional.

Com a intenção de investigar este problema, experimentos de *microarray* foram conduzidos estudando-se células da linhagem RINm5F submetidas à progesterona em três doses (0,1 μM, 1 μM e 100 μM) e dois tempos (6h e 24h), resultando em um conjunto de dados de expressão gênica que foi objeto da análise deste trabalho, que se propôs a auxiliar na compreensão do fenômeno realizando a identificação e a aplicação das técnicas de agrupamento mais adequadas ao problema e a interpretação baseada em análises de enriquecimento funcional dos resultados no sentido de contribuir para a compreensão da doença do diabetes gestacional.

Para tanto, reproduziu-se o procedimento consolidado no estudo de *benchmark* proposto por Saelens, Cannoodt e Saeys, em 2018, para as principais técnicas utilizadas na literatura, incluindo aquelas mais bem avaliadas no esquema de pontuação proposto, a saber, agrupamento hierárquico aglomerativo, k-médias, agrupamento de deslocamento médio, análise de componentes independentes e bi-agrupamento espectral e conforme era esperado, os resultados obtidos foram semelhantes aos descritos no estudo de referência.

No entanto, Saelens, Cannoodt e Saeys atribuiram a pontuação aos métodos de agrupamento segundo apenas índices de validação externos, fato que nem sempre reflete a realidade, que conta com cenários em que o pesquisador não tem à sua disposição os agrupamentos de referência. Por esse motivo, a discussão do estudo de *benchmark* foi estendida por meio do cálculo e da discussão da pontuação com base em índices de validação internos, a saber, índice de Dunn, índice da silhueta média e índice de Davies-Bouldin. Essa discussão mostrou-se muito interessante pois a pontuação dos métodos, quando calculada para os mesmos resultados da reprodução do estudo de *benchmark*, em muitos casos, contrariou a avaliação segundo os índices externos.

Apesar de estender-se a discussão do estudo de *benchmark* aos índices de validação internos, esses insumos ainda não foram suficientes para selecionar as melhores técnicas

para os dados de células beta-pancreáticas, uma vez que este conjunto tem dimensões muito diferentes (em termos de quantidades de genes e quantidade de condições experimentais) daqueles originalmente utilizados. Para resolver esse problema, reproduziu-se o procedimento do estudo de *benchmark* e toda a avaliação descrita até aqui incluindo-se o conjunto de células beta-pancreáticas. Para que isso fosse possível, uma vez que não se tem disponível agrupamento de referência para o conjunto de dados células beta-pancreáticas submetidas à progesterona, se fez necessário realizar o levantamento de módulos (grupos de genes) de referência. Para que os agrupamentos de referência fossem aproximados da melhor maneira possível, isso foi feito tanto com base nos dados, disponíveis na ferramenta *Cytoscape*, do banco de dados denominado *Comparative Toxicogenomics Database* quanto aplicando-se a técnica de agrupamento baseada em análise de rede de co-expressão de genes ponderada.

Uma vez calculadas as pontuações segundo índices internos e externos para cada método, realizou-se a análise de enriquecimento funcional dos melhores resultados obtidos para o conjunto de células beta-pancreáticas submetidas à progesterona segundo cada pontuação. A análise de enriquecimento foi então apresentada à especialista do domínio, a Profa. Dra. Anna Karenina Azevedo Martins, que auxiliou com a interpretação e seleção dos resultados relevantes para o problema. Essa interpretação resultou em *insights* relevantes para a compreensão do diabetes gestacional, como a compreensão de cada gene para a morte celular, com base nos grupos em que foram atribuídos. Desta análise, o principal *insight* é que o comportamento do gene TXNIP nesse contexto corroborou com alguns resultados identificados na literatura, que já direcionam esse gene como tendo um grande potencial para ser estudado compreensão do diabetes gestacional.

Deste trabalho de mestrado decorreram as seguintes contribuições e trabalhos futuros.

5.1 Contribuições do trabalho

Enumeram-se as seguintes contribuições deste trabalho:

- Extensão da discussão do estudo proposto por Saelens, Cannoodt e Saeys, em 2018 aos índices de validação internos.

- Constatação de que a pontuação proposta, segundo os índices de validação internos, foi capaz de apresentar os melhores insights do ponto de vista da especialista do domínio, para compreensão do problema do diabetes gestacional e da sua relação com a progesterona.
- Santos, M. D. L., Oliveira, P. R., Martins, A. K. A. (2022). Cluster analysis indicates genes involved in progesterone-induced oxidative stress in pancreatic beta cells: insights to understanding gestational diabetes.
  Artigo apresentado na conferência intitulada Brazilian Symposium on Bioinformatics (BSB). O trabalho apresenta desde o procedimento realizado neste trabalho de mestrado para a identificação da melhor técnica de agrupamento para a análise dos dados de célula beta-pancreáticas submetidas à progesterona até os *insights* obtidos a partir da análise de enriquecimento funcional dessa técnica para a compreensão do diabetes gestacional.

5.2 Trabalhos futuros

Das limitações desta pesquisa, decorrem os seguintes trabalhos futuros:

- Pelo fato de que as técnicas aqui estudadas foram restritas àquelas melhor avaliadas e mais utilizadas na literatura, sugere-se ampliar a discussão do estudo comparativo para avaliar uma maior quantidade de técnicas de agrupamento.
- Os índices de validação interna de agrupamento utilizados, por levarem em consideração a coesão e separação como critérios de qualidade, podem favorecer determinados tipos de agrupamento, como por exemplo o k-médias. Devido a tal limitação, em trabalhos futuros pretende-se estender a análise em uma busca por outros índices.
- Dado que a análise de enriquecimento funcional é fundamental para a compreensão de estudos de expressão gênica, há a intenção de disponibilizar os resultados decorrentes deste trabalho de mestrado como contribuição para futuras análises de enriquecimento funcional, submetendo-os, por exemplo, à ferramenta *Enrichr*.
- Uma vez que esse gene apresenta um grande potencial para a compreensão do fenômeno, investigar o papel do gene TXNIP, por meio da realização de mais experimentos e estudos, para a compreensão da doença do diabetes gestacional.

Referências

ABU-JAMOUS, B.; FA, R.; ROBERTS, D. J.; NANDI, A. K. Method for the identification of the subsets of genes specifically consistently co-expressed in a set of datasets. In: IEEE. *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*. [S.l.], 2013. p. 1–6. Citado na página 46.

AHMED, F. E. Molecular techniques for studying gene expression in carcinogenesis. *Journal of Environmental Science and Health, Part C*, Taylor & Francis, v. 20, n. 2, p. 77–116, 2002. Citado na página 20.

ALBERTS, B. *Molecular biology of the cell*. [S.l.]: Garland science, 2017. Citado na página 19.

BALDI, P.; BRUNAK, S. *Bioinformatics: the machine learning approach*. [S.l.]: MIT press, 2001. Citado 2 vezes nas páginas 20 e 25.

BERGER, B.; PENG, J.; SINGH, M. Computational solutions for omics data. *Nature reviews genetics*, Nature Publishing Group, v. 14, n. 5, p. 333–346, 2013. Citado na página 20.

BHANDARI, N.; WALAMBE, R.; KOTECH, K.; KHARE, S. Comprehensive survey of computational learning methods for analysis of gene expression data in genomics. *arXiv preprint arXiv:2202.02958*, 2022. Citado na página 56.

BINDER, J. X.; PLETSCHER-FRANKILD, S.; TSAFOU, K.; STOLTE, C.; O'DONOGHUE, S. I.; SCHNEIDER, R.; JENSEN, L. J. Compartments: unification and visualization of protein subcellular localization evidence. *Database*, Oxford Academic, v. 2014, 2014. Citado na página 64.

BOLSHAKOVA, N.; AZUAJE, F. Cluster validation techniques for genome expression data. *Signal processing*, Elsevier, v. 83, n. 4, p. 825–833, 2003. Citado na página 55.

CHEN, J.; HUI, S. T.; COUTO, F. M.; MUNGRUE, I. N.; DAVIS, D. B.; ATTIE, A. D.; LUSIS, A. J.; DAVIS, R. A.; SHALEV, A. Thioredoxin-interacting protein deficiency induces akt/bcl-xl signaling and pancreatic beta-cell mass and protects against diabetes. *The FASEB Journal*, Wiley Online Library, v. 22, n. 10, p. 3581–3594, 2008. Citado na página 83.

CHEN, R. Y.; KUNG, V. L.; DAS, S.; HOSSAIN, M. S.; HIBBERD, M. C.; GURUGE, J.; MAHFUZ, M.; BEGUM, S. K. N.; RAHMAN, M. M.; FAHIM, S. M. *et al.* Duodenal microbiota in stunted undernourished children with enteropathy. *New England Journal of Medicine*, Mass Medical Soc, v. 383, n. 4, p. 321–333, 2020. Citado 2 vezes nas páginas 22 e 55.

COMANICIU, D.; MEER, P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 24, n. 5, p. 603–619, 2002. Citado 2 vezes nas páginas 28 e 33.

CONSORTIUM, G.; ARDLIE, K. G.; DELUCA, D. S.; SEGRÈ, A. V.; SULLIVAN, T. J.; YOUNG, T. R.; GELFAND, E. T.; TROWBRIDGE, C. A.; MALLER, J. B.; TUKIAINEN, T. *et al.* The genotype-tissue expression (gtex) pilot analysis: multitissue

gene regulation in humans. *Science*, American Association for the Advancement of Science, v. 348, n. 6235, p. 648–660, 2015. Citado na página 60.

CONSORTIUM, G. O. Gene ontology consortium: going forward. *Nucleic acids research*, Oxford University Press, v. 43, n. D1, p. D1049–D1056, 2015. Citado na página 64.

CONSORTIUM, G. O. The gene ontology resource: enriching a gold mine. *Nucleic acids research*, Oxford University Press, v. 49, n. D1, p. D325–D334, 2021. Citado na página 64.

CONSORTIUM, I. H. G. S. *et al.* Finishing the euchromatic sequence of the human genome. *Nature*, Nature Publishing Group, v. 431, n. 7011, p. 931, 2004. Citado na página 19.

DALTON, L.; BALLARIN, V.; BRUN, M. Clustering algorithms: on learning, validation, performance, and applications to genomics. *Current genomics*, Bentham Science Publishers, v. 10, n. 6, p. 430–445, 2009. Citado 2 vezes nas páginas 21 e 57.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, n. 2, p. 224–227, 1979. Citado na página 37.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, 1979. Citado na página 37.

DAVIS, A. P.; GRONDIN, C. J.; JOHNSON, R. J.; SCIAKY, D.; WIEGERS, J.; WIEGERS, T. C.; MATTINGLY, C. J. Comparative toxicogenomics database (ctd): update 2021. *Nucleic acids research*, Oxford University Press, v. 49, n. D1, p. D1138–D1143, 2021. Citado na página 63.

DEZA, M. M.; DEZA, E. Encyclopedia of distances. In: *Encyclopedia of Distances*. [S.l.]: Springer, 2009. p. 1–583. Citado na página 35.

DUNN, J. C. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, Taylor & Francis, v. 4, n. 1, p. 95–104, 1974. Citado na página 37.

EISEN, M. B.; SPELLMAN, P. T.; BROWN, P. O.; BOTSTEIN, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 95, n. 25, p. 14863–14868, 1998. Citado 6 vezes nas páginas 21, 44, 46, 52, 53 e 114.

FRATELLO, M.; CATTELANI, L.; FEDERICO, A.; PAVEL, A.; SCALA, G.; SERRA, A.; GRECO, D. Unsupervised algorithms for microarray sample stratification. In: *Microarray Data Analysis*. [S.l.]: Springer, 2022. p. 121–146. Citado na página 56.

FRIENDLY, M.; DENIS, D. J. Milestones in the history of thematic cartography, statistical graphics, and data visualization. *URL http://www. datavis. ca/milestones*, Citeseer, v. 32, p. 13, 2001. Citado na página 44.

FUKUNAGA, K.; HOSTETLER, L. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, IEEE, v. 21, n. 1, p. 32–40, 1975. Citado na página 33.

GAZI, V. P.; KAYIS, E. Comparing clustering techniques for real microarray data. In: IEEE. *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. [S.l.], 2012. p. 788–791. Citado na página 53.

GEHLENBORG, N.; WONG, B. Heat maps. *Nature Methods*, v. 9, n. 3, p. 213, 2012. Citado 2 vezes nas páginas 43 e 44.

GONçALVES, A.; ONG, I.; LEWIS, J. A.; COSTA, V. S. Discovering differentially expressed genes in yeast stress data. In: IEEE. *Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on*. [S.l.], 2014. p. 537–538. Citado na página 54.

HENNIG, S.; GROTH, D.; LEHRACH, H. Automated gene ontology annotation for anonymous sequence data. *Nucleic Acids Research*, Oxford University Press, v. 31, n. 13, p. 3712–3715, 2003. Citado na página 19.

HUANG, D. W.; SHERMAN, B. T.; LEMPICKI, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, Oxford University Press, v. 37, n. 1, p. 1–13, 2009. Citado na página 46.

HYVÄRINEN, A.; OJA, E. Independent component analysis: algorithms and applications. *Neural networks*, Elsevier, v. 13, n. 4-5, p. 411–430, 2000. Citado 2 vezes nas páginas 28 e 33.

JAIN, A. K.; DUBES, R. C. Algorithms for clustering data. Prentice-Hall, Inc., 1988. Citado 2 vezes nas páginas 28 e 30.

JIANG, D.; TANG, C.; ZHANG, A. Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, v. 16, n. 11, p. 1370–1386, 2004. Citado na página 21.

KANEHISA, M.; SATO, Y.; KAWASHIMA, M.; FURUMICHI, M.; TANABE, M. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, Oxford University Press, v. 44, n. D1, p. D457–D462, 2016. Citado na página 64.

KLUGER, Y.; BASRI, R.; CHANG, J. T.; GERSTEIN, M. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research*, Cold Spring Harbor Lab, v. 13, n. 4, p. 703–716, 2003. Citado 2 vezes nas páginas 28 e 34.

KULESHOV, M. V.; JONES, M. R.; ROUILLARD, A. D.; FERNANDEZ, N. F.; DUAN, Q.; WANG, Z.; KOPLEV, S.; JENKINS, S. L.; JAGODNIK, K. M.; LACHMANN, A. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, Oxford University Press, v. 44, n. W1, p. W90–W97, 2016. Citado 2 vezes nas páginas 47 e 64.

LANGFELDER, P.; HORVATH, S. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, BioMed Central, v. 9, n. 1, p. 1–13, 2008. Citado na página 62.

LAWLOR, M. A.; CAO, W.; ELLISON, C. E. A transposon expression burst accompanies the activation of y-chromosome fertility genes during drosophila spermatogenesis. *Nature communications*, Nature Publishing Group, v. 12, n. 1, p. 1–12, 2021. Citado 2 vezes nas páginas 22 e 55.

LEI, Z.; CHEN, Y.; WANG, J.; ZHANG, Y.; SHI, W.; WANG, X.; XING, D.; LI, D.; JIAO, X. Txnip deficiency promotes β-cell proliferation in the hfd-induced obesity mouse model. *Endocrine connections*, Bioscientifica Ltd, v. 11, n. 4, 2022. Citado na página 83.

LEWIS, S.; ASHBURNER, M.; REESE, M. G. Annotating eukaryote genomes. *Current opinion in structural biology*, Elsevier, v. 10, n. 3, p. 349–354, 2000. Citado na página 45.

LLOYD, S. Least squares quantization in pcm. *IEEE transactions on information theory*, IEEE, v. 28, n. 2, p. 129–137, 1982. Citado na página 28.

LOUA, T. *Atlas statistique de la population de Paris*. [S.l.]: J. Dejey & cie, 1873. Citado 2 vezes nas páginas 43 e 115.

LUEBBERT, L.; PACHTER, L. Efficient querying of genomic reference databases with gget. *bioRxiv*, 2022. Citado na página 64.

MAJI, P.; SHAH, E. Significance and functional similarity for identification of disease genes. *IEEE/ACM transactions on computational biology and bioinformatics*, IEEE, v. 14, n. 6, p. 1419–1433, 2017. Citado 3 vezes nas páginas 20, 46 e 55.

MARBACH, D.; COSTELLO, J. C.; KÜFFNER, R.; VEGA, N. M.; PRILL, R. J.; CAMACHO, D. M.; ALLISON, K. R.; KELLIS, M.; COLLINS, J. J.; STOLOVITZKY, G. Wisdom of crowds for robust gene network inference. *Nature methods*, Nature Publishing Group, v. 9, n. 8, p. 796–804, 2012. Citado na página 59.

MCPHERSON, J. D.; MARRA, M.; HILLIER, L. D.; WATERSTON, R. H.; CHINWALLA, A.; WALLIS, J.; SEKHON, M.; WYLIE, K.; MARDIS, E. R.; WILSON, R. K. *et al.* A physical map of the human genome. *Nature*, Nature Publishing Group, v. 409, n. 6822, p. 934–941, 2001. Citado na página 19.

MEYSMAN, P.; SONEGO, P.; BIANCO, L.; FU, Q.; LEDEZMA-TEJEIDA, D.; GAMA-CASTRO, S.; LIEBENS, V.; MICHIELS, J.; LAUKENS, K.; MARCHAL, K. *et al.* Colombos v2. 0: an ever expanding collection of bacterial expression compendia. *Nucleic acids research*, Oxford University Press, v. 42, n. D1, p. D649–D653, 2014. Citado na página 59.

MOCELLIN, S.; ROSSI, C. R. Principles of gene microarray data analysis. In: *Microarray Technology and Cancer Gene Profiling*. [S.l.]: Springer, 2007. p. 19–30. Citado na página 20.

NAGI, S.; BHATTACHARYYA, D. K.; KALITA, J. K. Gene expression data clustering analysis: A survey. In: IEEE. *2011 2nd National Conference on Emerging Trends and Applications in Computer Science*. [S.l.], 2011. p. 1–12. Citado 2 vezes nas páginas 25 e 27.

NUNES, V. A.; PORTIOLI-SANCHES, E. P.; ROSIM, M.; ARAUJO, M.; PRAXEDES-GARCIA, P.; VALLE, M.; ROMA, L.; HAHN, C.; GURGUL-CONVEY, E.; LENZEN, S. *et al.* Progesterone induces apoptosis of insulin-secreting cells: insights into the molecular mechanism. *Journal of Endocrinology*, Soc Endocrinology, v. 221, n. 2, p. 273–284, 2014. Citado 2 vezes nas páginas 22 e 57.

OCHIENG, P. J.; TARIGAN, S. I.; DIDIK, H. A clustering model for identification of time course gene expression patterns. In: IEEE. *Biomedical Engineering (IBIOMED), International Conference on*. [S.l.], 2016. p. 1–6. Citado na página 54.

OLIVER, S. Guilt-by-association goes global. *Nature*, Nature Publishing Group, v. 403, n. 6770, p. 601–602, 2000. Citado 2 vezes nas páginas 21 e 46.

OYELADE, J.; ISEWON, I.; OLADIPUPO, F.; AROMOLARAN, O.; UWOGHIREN, E.; AMEH, F.; ACHAS, M.; ADEBIYI, E. Clustering algorithms: their application to gene expression data. *Bioinformatics and Biology insights*, SAGE Publications Sage UK: London, England, v. 10, p. BBI–S38316, 2016. Citado na página 21.

PALASCA, O.; SANTOS, A.; STOLTE, C.; GORODKIN, J.; JENSEN, L. J. Tissues 2.0: an integrative web resource on mammalian tissue expression. *Database*, Oxford Academic, v. 2018, 2018. Citado na página 64.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 2 vezes nas páginas 33 e 34.

PERGIALIOTIS, V.; BELLOS, I.; HATZIAGELAKI, E.; ANTSAKLIS, A.; LOUTRADIS, D.; DASKALAKIS, G. Progestogens for the prevention of preterm birth and risk of developing gestational diabetes mellitus: a meta-analysis. *American journal of obstetrics and gynecology*, Elsevier, v. 221, n. 5, p. 429–436, 2019. Citado 2 vezes nas páginas 22 e 57.

PIÑERO, J.; RAMÍREZ-ANGUITA, J. M.; SAÜCH-PITARCH, J.; RONZANO, F.; CENTENO, E.; SANZ, F.; FURLONG, L. I. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, Oxford University Press, v. 48, n. D1, p. D845–D855, 2020. Citado na página 64.

PLETSCHER-FRANKILD, S.; PALLEJÀ, A.; TSAFOU, K.; BINDER, J. X.; JENSEN, L. J. Diseases: Text mining and data integration of disease–gene associations. *Methods*, Elsevier, v. 74, p. 83–89, 2015. Citado na página 64.

POUDEL, S.; TSUNEMOTO, H.; SEIF, Y.; SASTRY, A. V.; SZUBIN, R.; XU, S.; MACHADO, H.; OLSON, C. A.; ANAND, A.; POGLIANO, J. *et al.* Revealing 29 sets of independently modulated genes in staphylococcus aureus, their regulators, and role in key physiological response. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 117, n. 29, p. 17228–17239, 2020. Citado 2 vezes nas páginas 22 e 55.

REED, J. L.; FAMILI, I.; THIELE, I.; PALSSON, B. O. Towards multidimensional genome annotation. *Nature Reviews Genetics*, Nature Publishing Group, v. 7, n. 2, p. 130, 2006. Citado na página 45.

ROJAS, J.; BERMUDEZ, V.; PALMAR, J.; MARTÍNEZ, M. S.; OLIVAR, L. C.; NAVA, M.; TOMEY, D.; ROJAS, M.; SALAZAR, J.; GARICANO, C. *et al.* Pancreatic beta cell death: novel potential mechanisms in diabetes therapy. *Journal of Diabetes Research*, Hindawi, v. 2018, 2018. Citado 2 vezes nas páginas 22 e 57.

ROTIVAL, M.; ZELLER, T.; WILD, P. S.; MAOUCHE, S.; SZYMCZAK, S.; SCHILLERT, A.; CASTAGNÉ, R.; DEISEROTH, A.; PROUST, C.; BROCHETON, J. *et al.* Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. *PLoS genetics*, Public Library of Science San Francisco, USA, v. 7, n. 12, p. e1002367, 2011. Citado na página 33.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987. Citado 2 vezes nas páginas 37 e 38.

SAELENS, W.; CANNOODT, R.; SAEYS, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nature communications*, Nature Publishing Group, v. 9, n. 1, p. 1–12, 2018. Citado 23 vezes nas páginas 17, 21, 22, 23, 28, 33, 39, 49, 50, 51, 55, 56, 57, 58, 59, 61, 62, 64, 65, 66, 67, 68 e 73.

SASTRY, A. V.; HU, A.; HECKMANN, D.; POUDEL, S.; KAVVAS, E.; PALSSON, B. O. Independent component analysis recovers consistent regulatory signals from disparate datasets. *PLoS computational biology*, Public Library of Science San Francisco, CA USA, v. 17, n. 2, p. e1008647, 2021. Citado 2 vezes nas páginas 22 e 55.

SCHAFFTER, T.; MARBACH, D.; FLOREANO, D. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, Oxford University Press, v. 27, n. 16, p. 2263–2270, 2011. Citado na página 60.

SHANNON, P.; MARKIEL, A.; OZIER, O.; BALIGA, N. S.; WANG, J. T.; RAMAGE, D.; AMIN, N.; SCHWIKOWSKI, B.; IDEKER, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, Cold Spring Harbor Lab, v. 13, n. 11, p. 2498–2504, 2003. Citado na página 62.

SMYTH, G. K. Limma: linear models for microarray data. In: *Bioinformatics and computational biology solutions using R and Bioconductor*. [S.l.]: Springer, 2005. p. 397–420. Citado na página 54.

SOLLERO, B. P.; GRYNBERG, P. Tutorial para análise funcional a partir de estudos de associação genômica ampla e transcriptômicos utilizando o banco de dados mesh (medical subject headings) no programa r. *Embrapa Pecuária Sul-Documentos (INFOTECA-E)*, Bagé: Embrapa Pecuária Sul, 2020., 2020. Citado 2 vezes nas páginas 47 e 64.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: principles and methods. *Data & knowledge engineering*, Elsevier, v. 25, n. 1-2, p. 161–197, 1998. Citado 2 vezes nas páginas 19 e 46.

TAN, J.; SASTRY, A. V.; FREMMING, K. S.; BJØRN, S. P.; HOFFMEYER, A.; SEO, S.; VOLDBORG, B. G.; PALSSON, B. O. Independent component analysis of e. coli's transcriptome reveals the cellular processes that respond to heterologous gene expression. *Metabolic Engineering*, Elsevier, v. 61, p. 360–368, 2020. Citado 2 vezes nas páginas 22 e 55.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to data mining. 1st*. [S.l.]: Boston: Pearson Addison Wesley. xxi, 2005. 487–515, 532–540, 546–550 p. Citado 4 vezes nas páginas 29, 31, 32 e 36.

TOMCZAK, A.; MORTENSEN, J. M.; WINNENBURG, R.; LIU, C.; ALESSI, D. T.; SWAMY, V.; VALLANIA, F.; LOFGREN, S.; HAYNES, W.; SHAH, N. H. *et al.* Interpretation of biological experiments changes with evolution of the gene ontology and its annotations. *Scientific reports*, Nature Publishing Group, v. 8, n. 1, p. 1–10, 2018. Citado na página 47.

TUIMALA, J.; LAINE, M. M. DNA microarray data analysis. *Espoo, Finland: Finnish IT Center for Science*, 2003. Citado 3 vezes nas páginas 25, 26 e 27.

VANICHAYOBON, S.; SIRIPHAN, W.; WIPHADA, W. Microarray gene selection using self-organizing map. In: *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*. [S.l.: s.n.], 2007. Citado na página 52.

VOEHRINGER, D.; HIRSCHBERG, D.; XIAO, J.; LU, Q.; ROEDERER, M.; LOCK, C.; HERZENBERG, L.; STEINMAN, L. Gene microarray identification of redox and mitochondrial elements that control resistance or sensitivity to apoptosis. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 97, n. 6, p. 2680–2685, 2000. Citado na página 52.

WIWIE, C.; BAUMBACH, J.; RÖTTGER, R. Comparing the performance of biomedical clustering methods. *Nature methods*, Nature Publishing Group, v. 12, n. 11, p. 1033–1038, 2015. Citado 6 vezes nas páginas 22, 39, 56, 57, 68 e 72.

WONDAFRASH, D. Z.; NIRE'A, A. T.; TAFERE, G. G.; DESTA, D. M.; BERHE, D. A.; ZEWDIE, K. A. Thioredoxin-interacting protein as a novel potential therapeutic target in diabetes mellitus and its underlying complications. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, Dove Press, v. 13, p. 43, 2020. Citado na página 83.

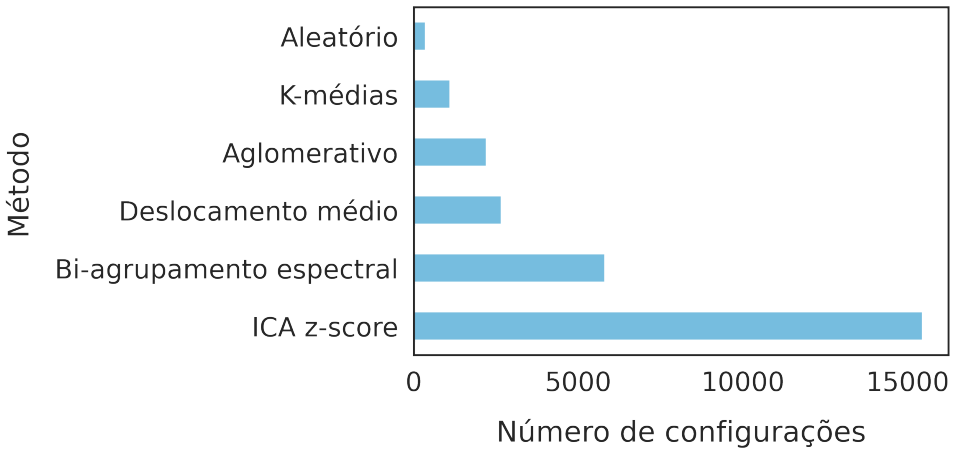
YANG, C.; WAN, B.; GAO, X. Effectivity of internal validation techniques for gene clustering. In: SPRINGER. *International Symposium on Biological and Medical Data Analysis*. [S.l.], 2006. p. 49–59. Citado na página 56.

YUVARAJ, K.; MANJULA, D. A performance analysis of clustering based algorithms for the microarray gene expression data. *International Journal of Engineering and Technology (UAE)*, v. 7, n. 2, p. 201–205, 2018. Citado na página 26.

Apêndice A – Distribuição das configurações de avaliação para a primeira fase do experimento (reprodução do *benchmark*)

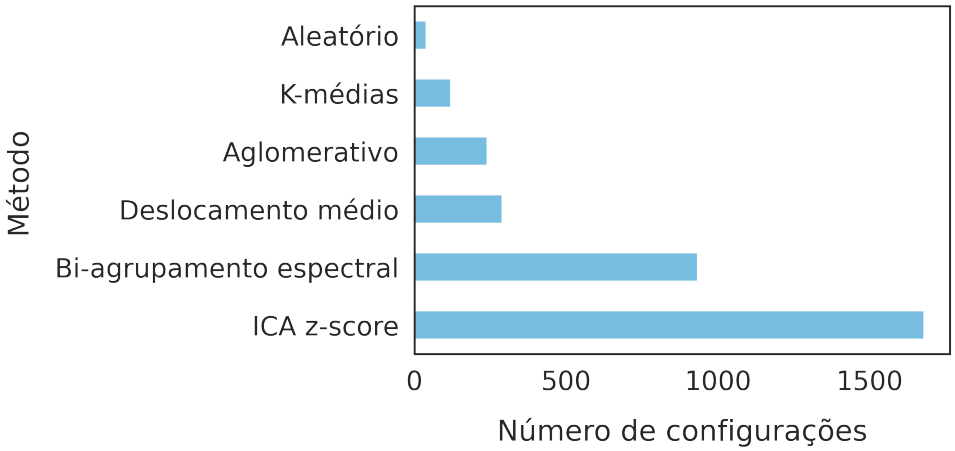
Do ponto de vista da avaliação baseada em módulos conhecidos, os experimentos resultaram em 26.789 combinações distintas de conjunto de dados, métodos, parâmetros e referência de avaliação, distribuídas por método, conforme a figura 35.

Figura 31 – Número de experimentos utilizando índices externos



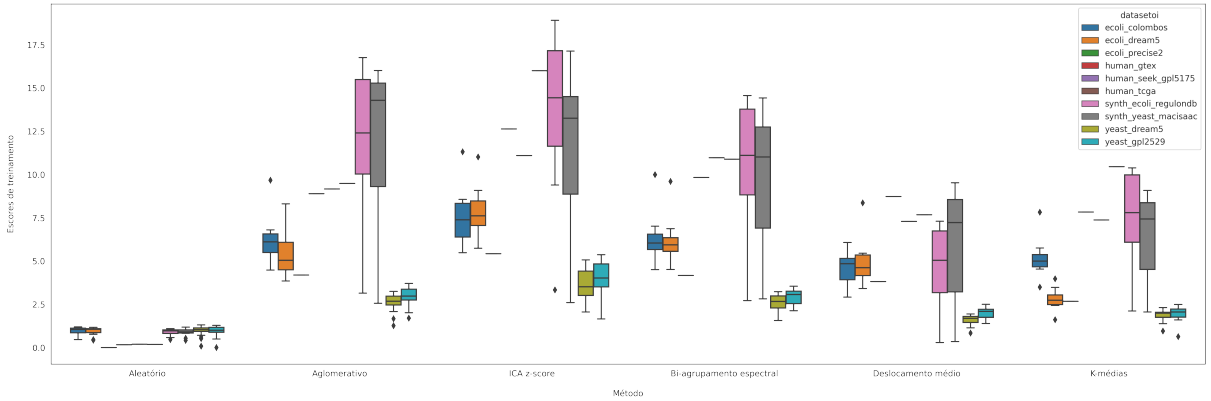
Adicionalmente, foram calculados os índices internos para cada configuração de experimento. Se considerarmos a quantidade de índices internos distintos provindos das combinações de métodos, conjuntos de dados e parâmetros, totalizaram-se 2.869 configurações, distribuidas por método seguindo a figura 36.

Figura 32 – Número de experimentos utilizando índices internos



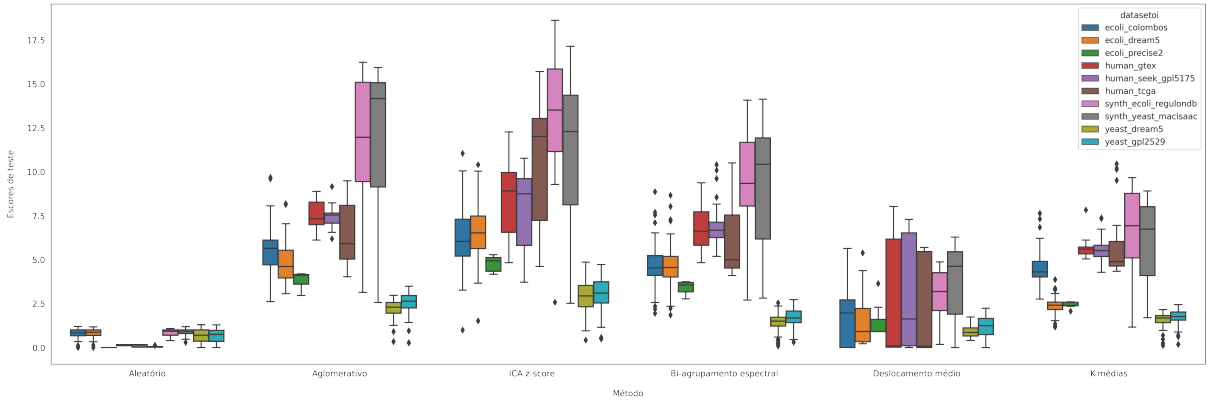
Apêndice B – Distribuição dos escores de treinamento para a reprodução do estudo de *benchmark*

Figura 33 – Escores de treinamento



Apêndice C – Distribuição dos escores de teste obtidos da etapa de reprodução do *benchmark*

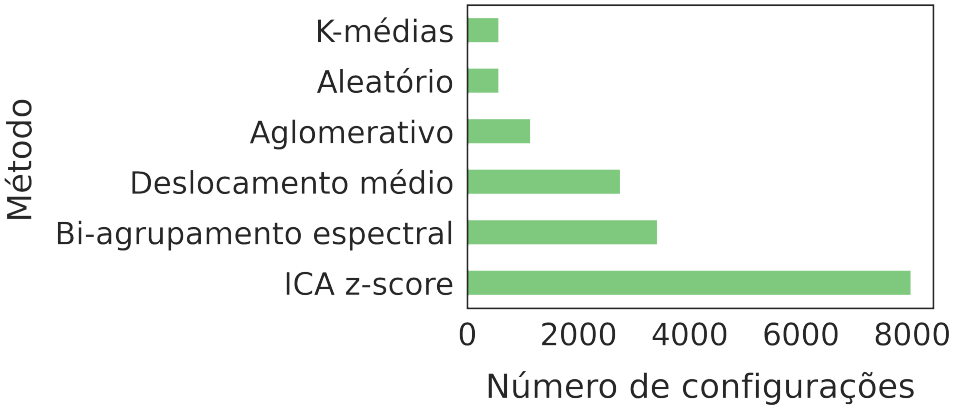
Figura 34 – Escores de teste.



Apêndice D – Distribuição das configurações de avaliação para a terceira fase do experimento (reprodução do *benchmark* incluindo o conjunto de dados de células beta)

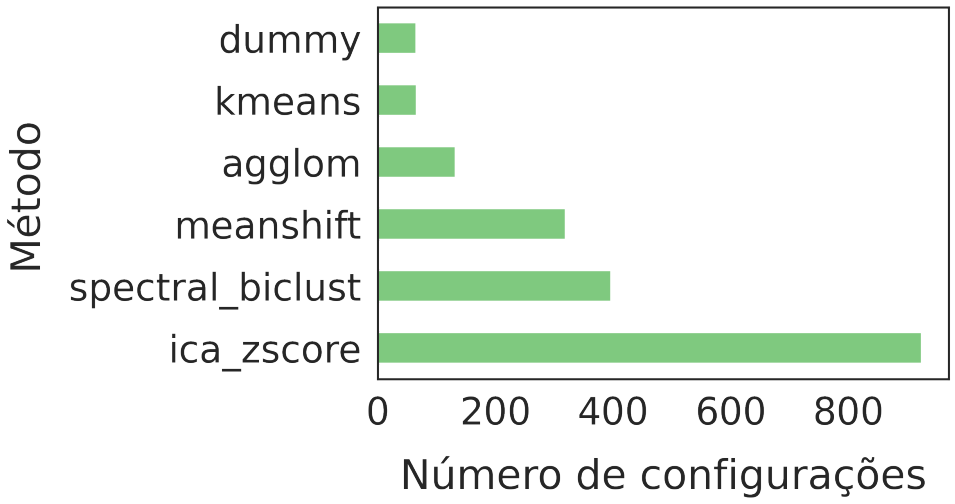
Do ponto de vista da avaliação baseada em módulos conhecidos, os experimentos resultaram em 15.913 combinações distintas de conjunto de dados, métodos, parâmetros e referência de avaliação, distribuídas por método, conforme a figura 35.

Figura 35 – Número de experimentos utilizando índices externos



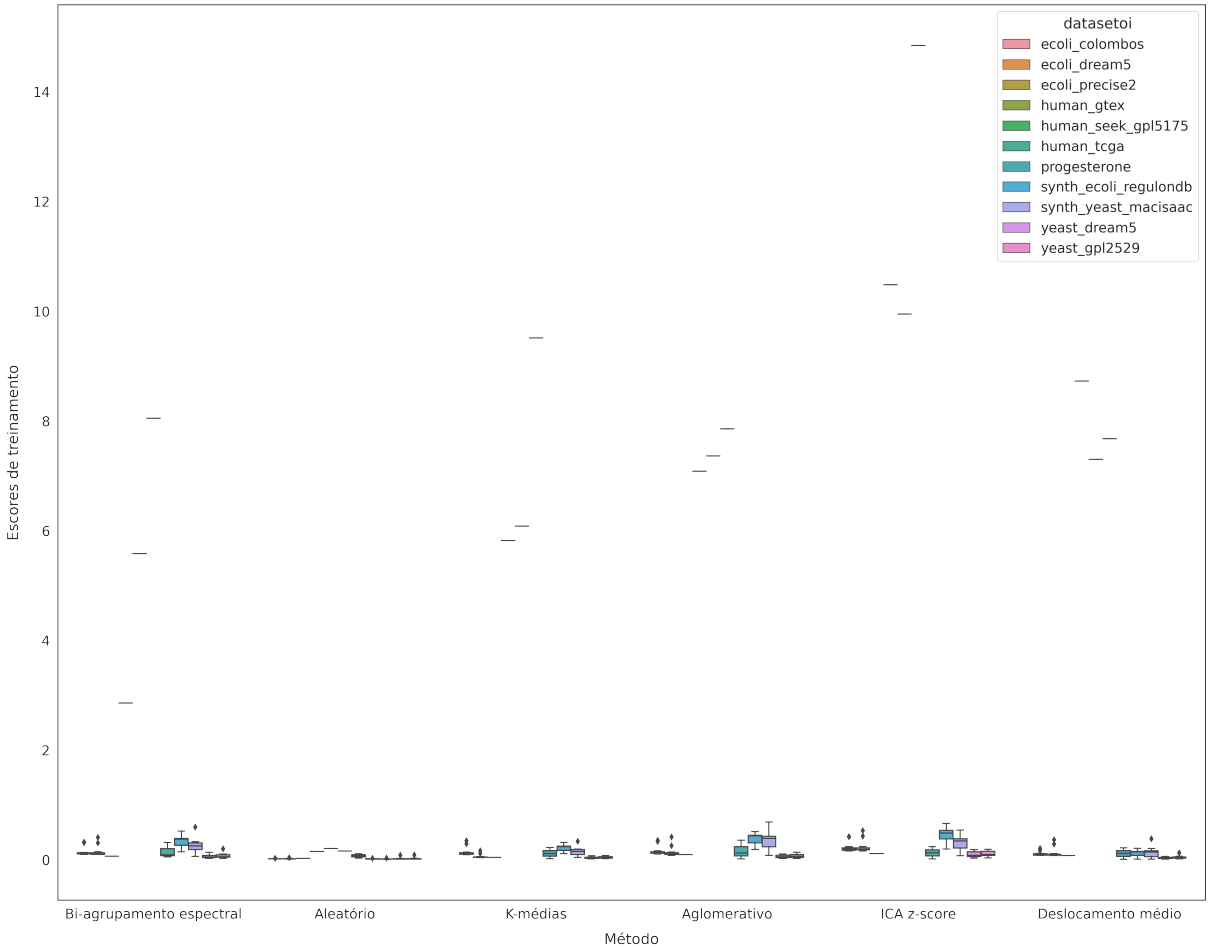
Adicionalmente, foram calculados os índices internos para cada configuração de experimento. Se considerarmos a quantidade de índices internos distintos provindos das combinações de métodos, conjuntos de dados e parâmetros, totalizaram-se 1.902 configurações, distribuidas por método seguindo a figura 36.

Figura 36 – Número de experimentos utilizando índices internos



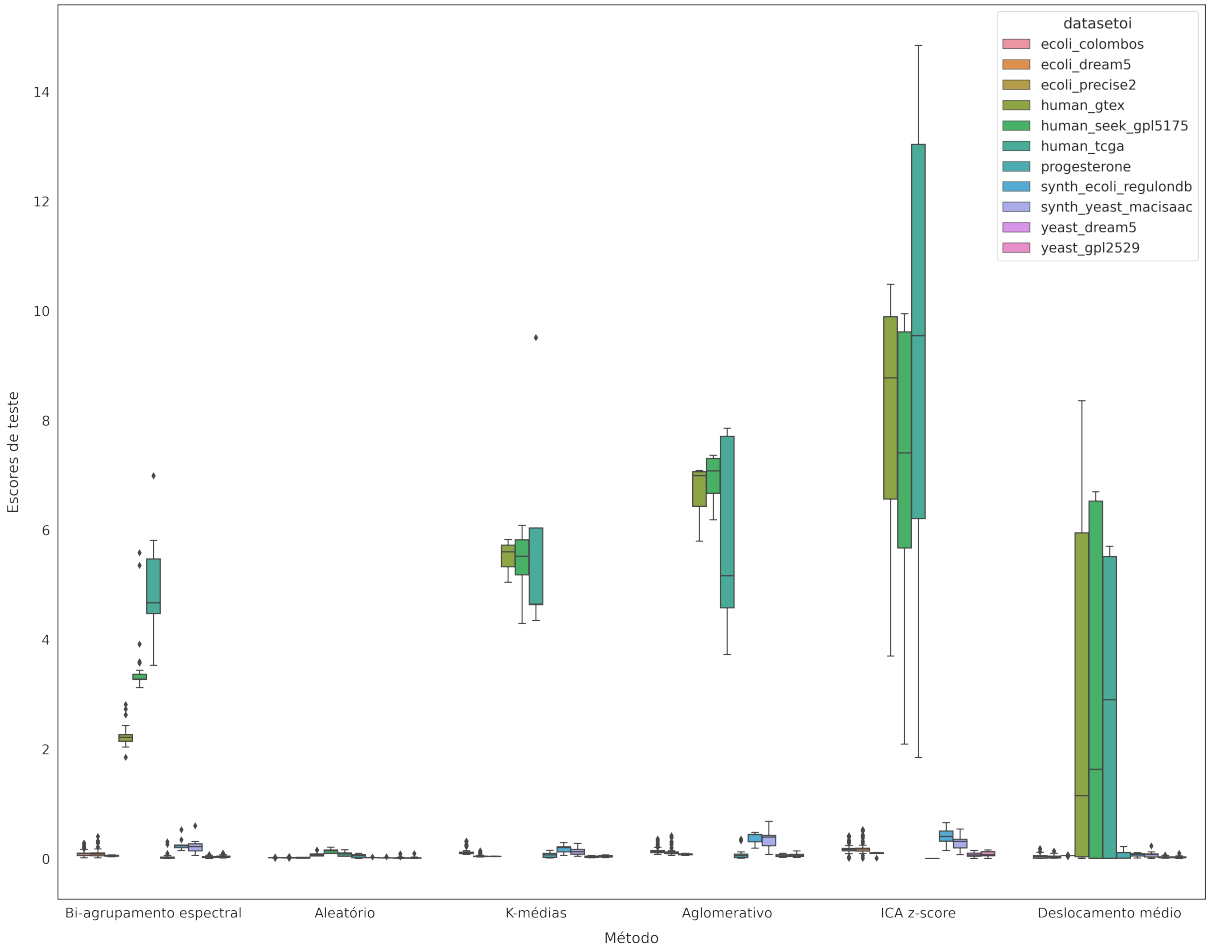
Apêndice E – Distribuição dos escores de treinamento para a reprodução do estudo de *benchmark* incluindo o conjunto de dados de células beta-pancreáticas

Figura 37 – Escores de treinamento



Apêndice F – Distribuição dos escores de teste obtidos da etapa de reprodução do *benchmark*

Figura 38 – Escores de teste.



Apêndice G – Resultados do agrupamento hierárquico obtido para o conjunto de dados com células beta-pancreáticas submetidas à progesterona (método melhor pontuado, segundo avaliação externa)

Figura 39 – Visualização do agrupamento hierárquico com k=6. As cores à esquerda da figura representam os módulos identificados por meio da técnica.

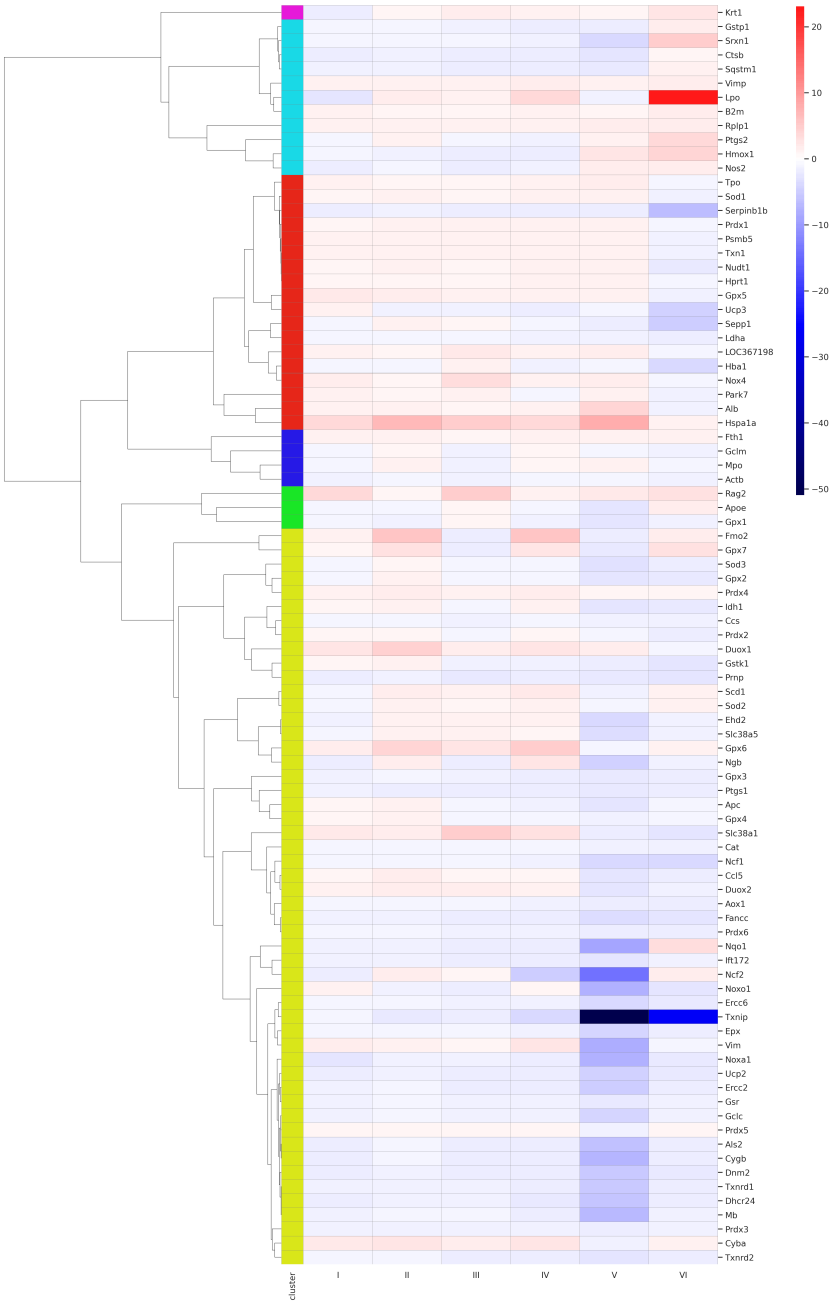


Figura 40 – Enriquecimento funcional para o módulo com o gene Krt1, o primeiro representado no dendrograma da figura 39 (de cima para baixo).

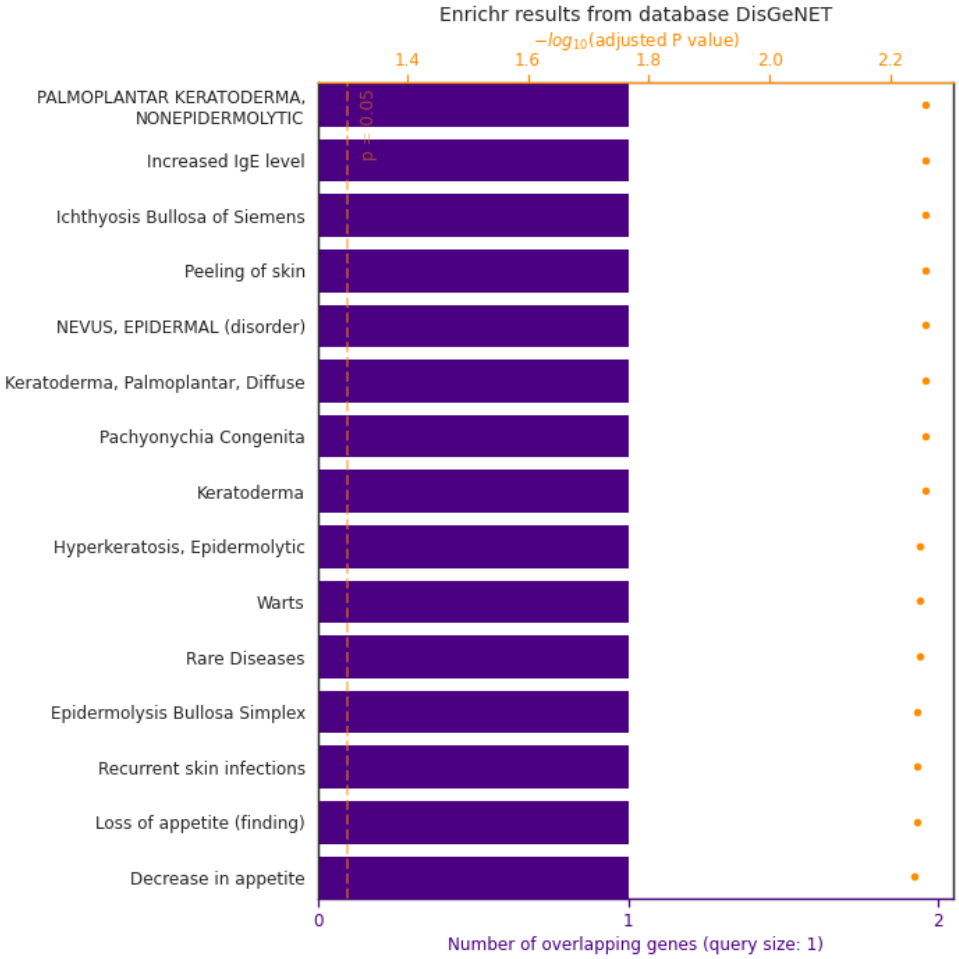


Figura 41 – Enriquecimento funcional para o módulo com os genes B2m, Ctsb, Gstp1, Hmox1, Lpo, Nos2, Ptgs2, Rplp1, Sqstm1, Srxn1 e Vimp, o segundo representado no dendrograma da figura 39 (de cima para baixo).

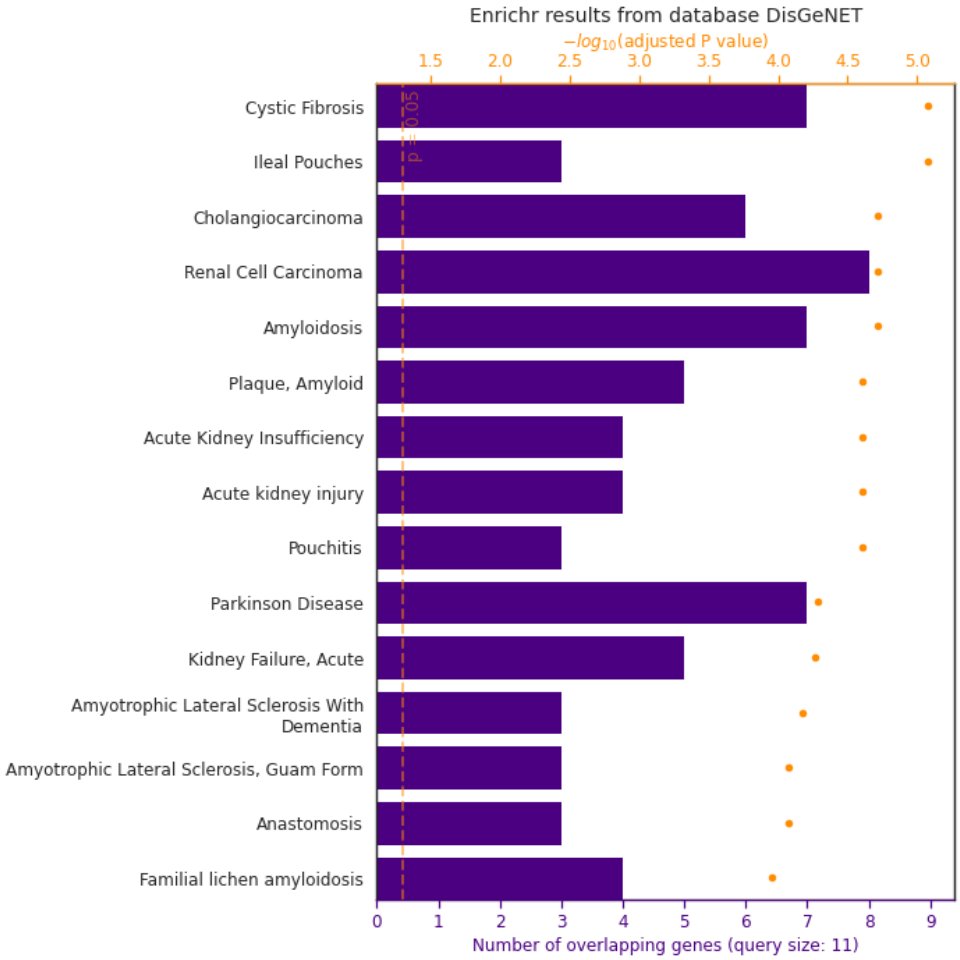


Figura 42 – Enriquecimento funcional para o módulo com os genes Alb, Gpx5, Hba1, Hprt1, Hspa1a, LOC367198, Ldha, Nox4, Nudt1, Park7, Prdx1, Psmb5, Sepp1, Serpinb1b, Sod1, Tpo, Txn1 e Ucp3, o terceiro representado no dendrograma da figura 39 (de cima para baixo).

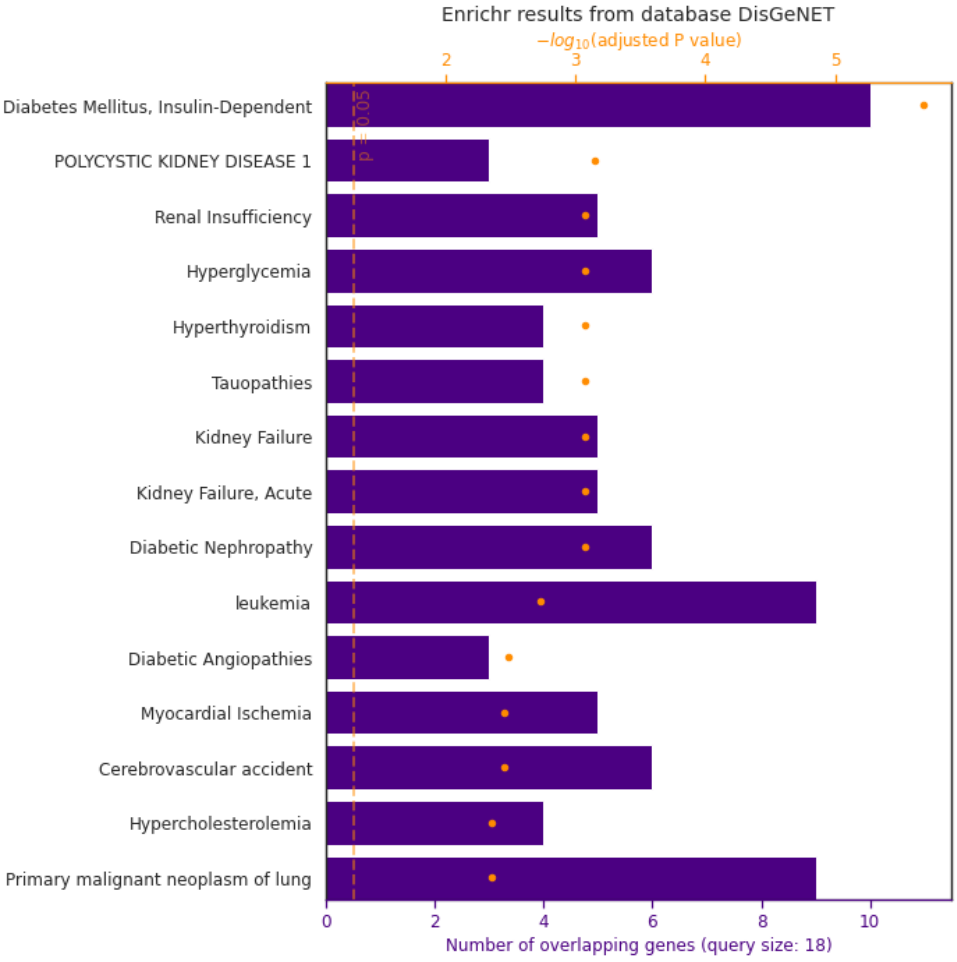


Figura 43 – Enriquecimento funcional para o módulo com os genes Actb, Fth1, Gclm e Mpo, o quarto representado no dendrograma da figura 39 (de cima para baixo).

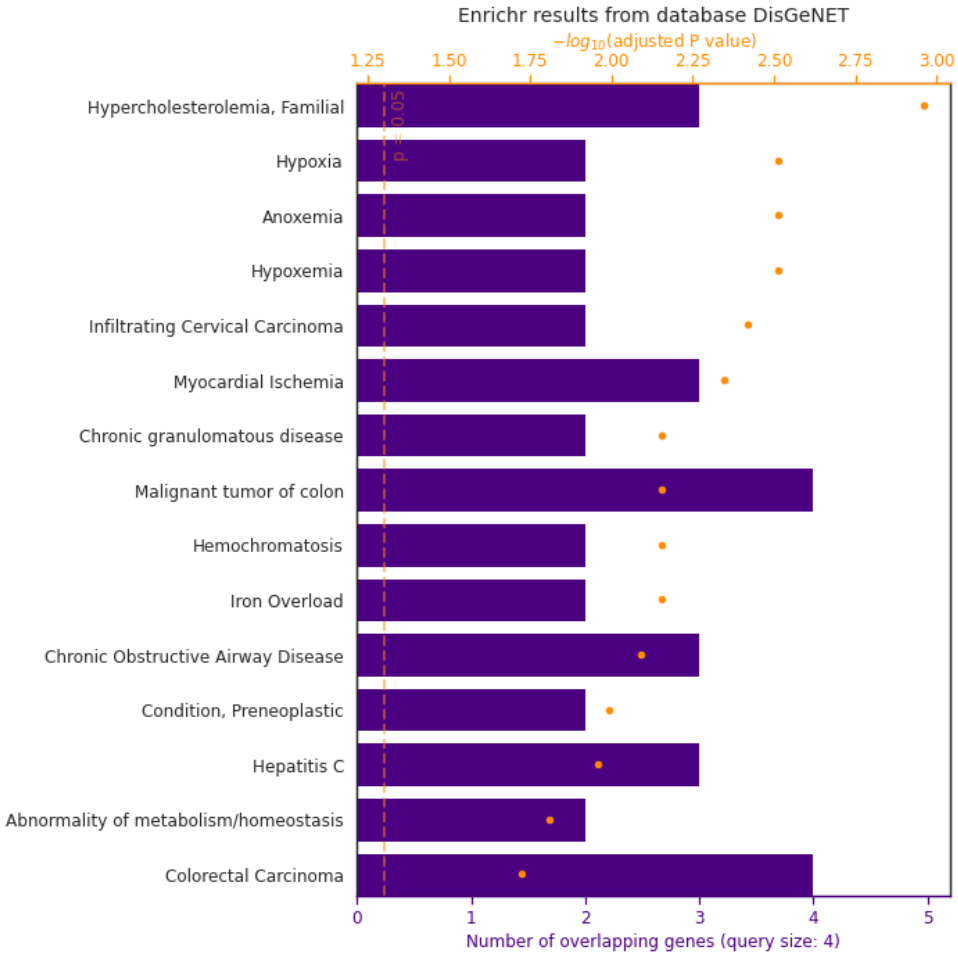


Figura 44 – Enriquecimento funcional para o módulo com os genes Apoe, Gpx1 e Rag2, o quinto módulo representado no dendrograma da figura 39 (de cima para baixo).

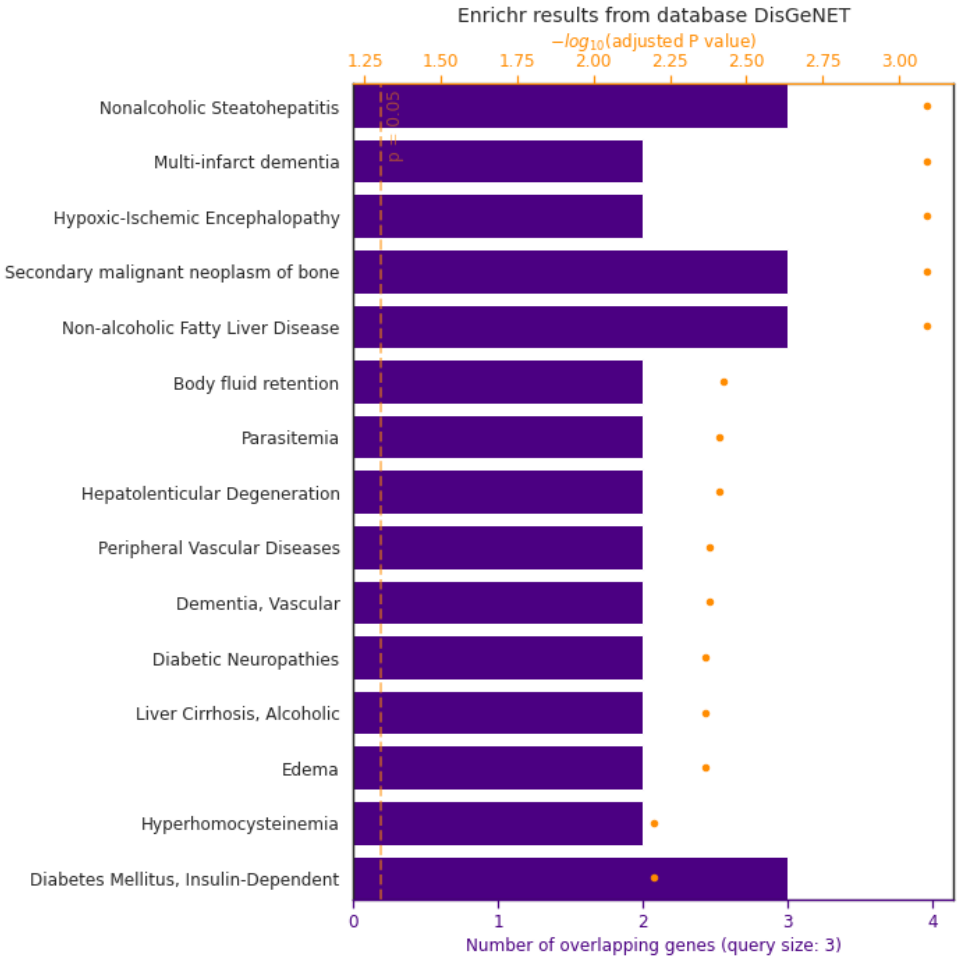
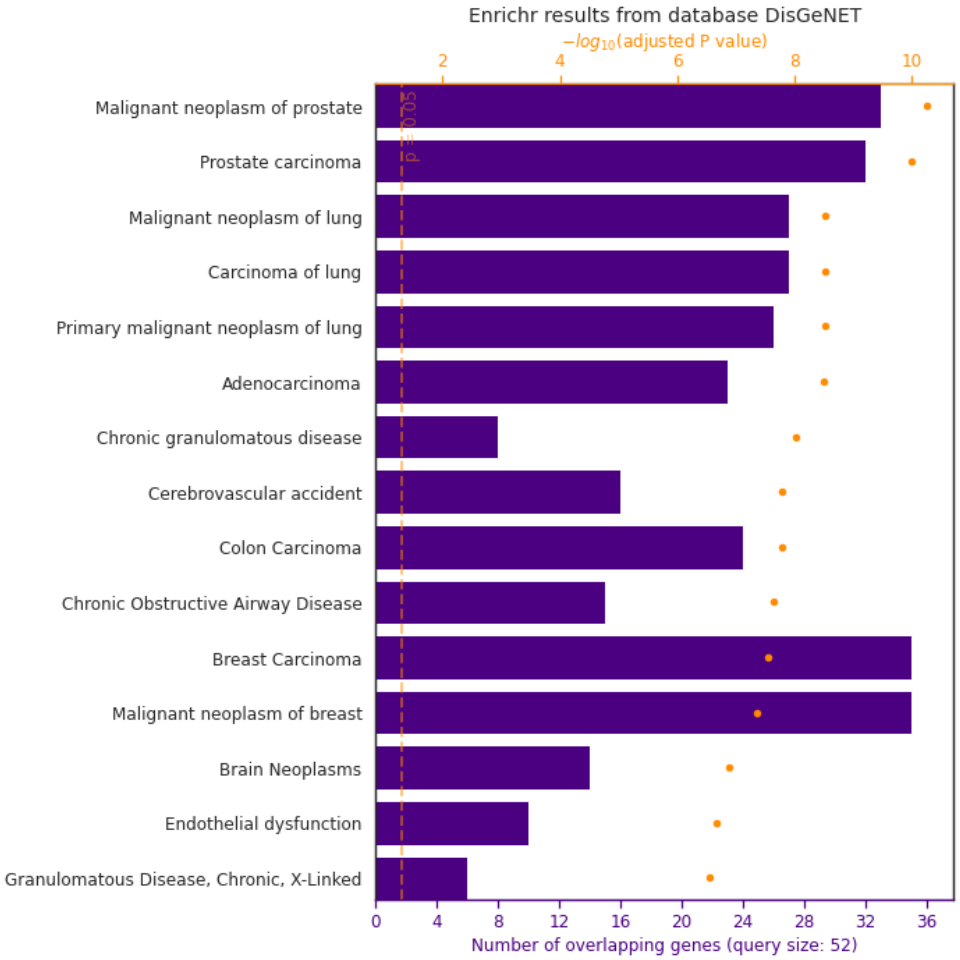


Figura 45 – Enriquecimento funcional para o módulo com os genes Als2, Aox1, Apc, Cat, Ccl5, Ccs, Cyba, Cygb, Dhcr24, Dnm2, Duox1, Duox2, Ehd2, Epx, Ercc2, Ercc6, Fancc, Fmo2, Gclc, Gpx2, Gpx3, Gpx4, Gpx6, Gpx7, Gsr, Gstk1, Idh1, Ift172, Mb, Ncf1, Ncf2, Ngb, Noxa1, Noxo1, Nqo1, Prdx2, Prdx3, Prdx4, Prdx5, Prdx6, Prnp, Ptgs1, Scd1, Slc38a1, Slc38a5, Sod2, Sod3, Txnip, Txnrd1, Txnrd2, Ucp2 e Vim, o sexto módulo representado no dendrograma da figura 39 (de cima para baixo).



Apêndice H – Conjunto de dados

Tabela 8 – Conjunto de dados de *microarray*.

Gene	PG 100 nM 6h	PG 100 nM 24h	PG 1 μM 6h	PG 1 μM 24h	PG 100 μM 6h	PG 100 μM 24h
Alb	1.14	1.3	1.1	1.13	4.04	-1.25
Als2	-1.51	-1.13	-1.5	-1.55	-6.4	-1.74
Aox1	-1.15	-1.03	-1.09	-1.33	-1.74	-1.8
Apc	1.07	1.16	-1.06	-1.24	-2.75	-1.17
Apoe	-1.03	-1.06	1	-1.16	-2.85	1.47
Cat	-1.11	-1.11	-1.06	-1.17	-1.34	-1.24
Ccl5	1.03	1.62	1.01	1	-2.84	-1.53
Ccs	-1.08	-1.12	-1.25	-1.12	-1.27	-1.39
Ctsb	-1.59	-1.27	-1.73	-1.9	-2.81	1
Cyba	2.17	2.69	1.5	2.33	-1.46	1.27
Cygb	-1.58	-1.1	-1.55	-1.77	-7.41	-1.67
Dhcr24	-1.55	-1.36	-1.45	-2.13	-5.77	-1.72
Dnm2	-1.51	-1.48	-1.56	-2.04	-5.46	-2.1
Duox1	2.45	4.54	1.59	2.84	1.45	-1.04
Duox2	1.28	1.57	1.55	1.34	-2.9	-1.4
Ehd2	-1.35	1.32	1.12	1.19	-3.69	-1.36

Tabela 8 - continuação da página anterior.

Gene	PG 100 nM 6h	PG 100 nM 24h	PG 1 μM 6h	PG 1 μM 24h	PG 100 μM 6h	PG 100 μM 24h
Epx	-1.2	-1.15	-1.06	-1.26	-4.34	-1.9
Ercc2	-1.31	-1.14	-1.6	-1.81	-5.16	-1.7
Ercc6	-1.11	-1.2	-1.38	-1.48	-3.61	-2.14
Fancc	-1.27	-1.34	-1.53	-2.04	-3.31	-2.86
Fmo2	1.15	5.55	-2.01	5.89	-1.76	1.96
Fth1	1.13	1.19	1	1.33	1.28	1.14
Gclc	-1.31	-1.18	-1.48	-1.28	-4.27	-1.33
Gclm	-1.11	1.02	-1.29	1.05	-1.02	-1.43
Gpx1	-1.06	-1.36	1.08	-1.21	-2.58	-1.31
Gpx2	-1.13	1.15	-1.17	-1.11	-2.39	-2.33
Gpx3	-1.22	-1.05	-1.33	-1.68	-2.11	-1.53
Gpx4	1.01	1.21	-1.06	-1.12	-1.37	-1.09
Gpx5	2.16	1.63	1.15	1.14	1.31	-1.23
Gpx6	1.57	4.06	2.82	4.82	-1.15	1.12
Gpx7	1.03	3.07	-1.83	2.53	-2.35	2.87
Gsr	-1.25	-1.19	-1.39	-1.35	-2.17	-1.38
Gstk1	1.03	1.13	-1.27	-1.4	-1.51	-2.55
Gstp1	-1.02	-1.12	-1.31	-1.64	-1.97	1.4

Tabela 8 - continuação da página anterior.

Gene	PG 100 nM 6h	PG 100 nM 24h	PG 1 μM 6h	PG 1 μM 24h	PG 100 μM 6h	PG 100 μM 24h
Hba1	-1.06	-1.1	1.14	-1.21	-1.18	-3.68
Hmox1	-1.09	-1.26	-1.57	-1.63	2.66	4.18
Hspa1a	3.81	7.14	4.85	3.8	7.98	1.24
Idh1	1.05	1.28	-1.03	1.34	-2.56	-2.2
Ift172	-1.39	-1.41	-1.66	-1.95	-2.8	-1.41
Krt1	-1.66	1.1	1.47	1.35	1.01	2.33
LOC367198	1.14	1.1	2.23	1.35	1.78	-1.04
Lpo	-2.81	1.9	1.2	3.62	-1.33	23.08
Mb	-1.32	-1.1	-1.1	-1.91	-7.13	-1.4
Mpo	-1.19	1.23	-1.31	1.05	1.13	-1.02
Ncf1	-1.03	-1.19	-1.16	-1.33	-3.81	-3.68
Ncf2	-1.6	1.53	1.04	-4.72	-14.08	1.95
Ngb	-1.51	1.76	-1.56	2.29	-4.64	-1.28
Nos2	-1.76	-1.03	-1.52	-1.22	1.86	1.78
Nox4	1.87	1.1	3.34	1.35	1.78	-1.04
Noxa1	-2.37	-1.26	-1.22	-1.9	-7.6	-2.15
Noxo1	1.17	-1.36	-1.71	1.09	-7.83	-2.47
Nqo1	-1.28	-1.22	-1.7	-1.82	-9.26	3.18

Tabela 8 - continuação da página anterior.

Gene	PG 100 nM 6h	PG 100 nM 24h	PG 1 μM 6h	PG 1 μM 24h	PG 100 μM 6h	PG 100 μM 24h
Nudt1	1.06	1.11	1.07	1.13	1.2	-2.29
Park7	1.14	1.05	1.11	-1.01	1.3	-1.48
Prdx1	1.1	1.36	1.13	1.26	1.2	-1.05
Prdx2	1.06	1.04	-1	1.01	-1.03	-1.78
Prdx3	-1.44	-1.22	-1.46	-1.4	-1.94	-1.4
Prdx4	1.25	1.66	1.22	1.42	1.07	1.06
Prdx5	1.05	1.07	1.08	1.02	-1.33	1.04
Prdx6	-1.19	-1.14	-1.26	-1.39	-1.81	-1.69
Prnp	-1.56	-1.46	-2.23	-2.01	-2.1	-2.7
Psmb5	1.16	1.16	1.17	1.28	1.13	-1.24
Ptgs1	-1.33	-1.55	-1.71	-2.05	-2.15	-1.96
Ptgs2	-1.08	1.38	-1.09	-1.21	1.24	3.77
Rag2	3.73	1.1	4.6	1.35	2.09	3.08
Scd1	-1.11	1.93	1.39	2.15	-1.42	1.38
Vimp	1.25	1.33	1.18	1.47	1.36	1.92
Sepp1	-1.19	1.15	1.1	-1.11	-1.71	-5.07
Serpinb1b	-1.82	-1.43	-1.51	-1.85	-1.86	-6.55
Slc38a1	2.2	1.92	4.75	2.96	-1.66	-2.9

Tabela 8 - continuação da página anterior.

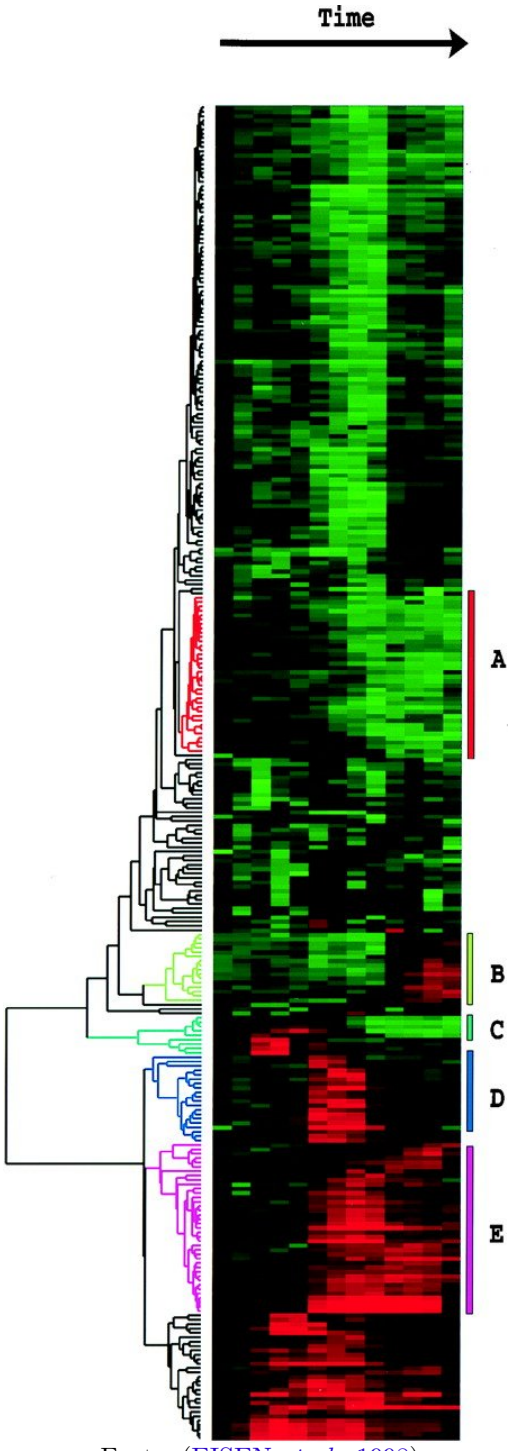
Gene	PG 100 nM 6h	PG 100 nM 24h	PG 1 μM 6h	PG 1 μM 24h	PG 100 μM 6h	PG 100 μM 24h
Slc38a5	-1.08	1.32	1.39	1.01	-3.45	-1.29
Sod1	1.05	1.14	1.04	1.29	1.39	-1.31
Sod2	-1.04	1.35	1.01	1.24	-1.01	1.14
Sod3	-1.03	1	-1.08	-1.05	-3.07	-1.76
Sqstm1	-1.42	-1.24	-1.78	-1.66	-2.33	1.11
Srxn1	-1.11	-1.09	-1.13	-1.23	-3.61	4.63
Tpo	1.14	1.1	1.1	1.35	1.78	-1.04
Txn1	1.25	1.23	1.2	1.24	1.25	-1.22
Txnip	-1.16	-2.19	-1.56	-3.53	-50.94	-26.65
Txnrd1	-1.39	-1.43	-1.37	-1.98	-5.54	-1.62
Txnrd2	-1.2	-1.1	-1.61	-1.57	-2.89	-1.59
Ucp2	-1.62	-1.4	-1.79	-1.73	-4.51	-2.1
Ucp3	1.18	-1.23	-1.31	-1.53	-1.16	-4.51
Vim	1.58	1.29	1.01	2.55	-7.99	-1.14
Actb	-1.41	-1.02	-1.22	-1.04	-1.08	-1.38
B2m	1.11	1.08	1.03	1.14	1.1	1.85
Hprt1	1.05	1.05	1.09	1.12	1.17	-1.16
Ldha	-1.16	-1.13	-1.12	-1.27	-1.28	-1.6

Tabela 8 - continuação da página anterior.

Gene	PG 100 nM 6h	PG 100 nM 24h	PG 1 μM 6h	PG 1 μM 24h	PG 100 μM 6h	PG 100 μM 24h
Rplp1	1.39	1.22	1.3	1.33	1.64	1.47

Anexo A – Dendrograma ampliado

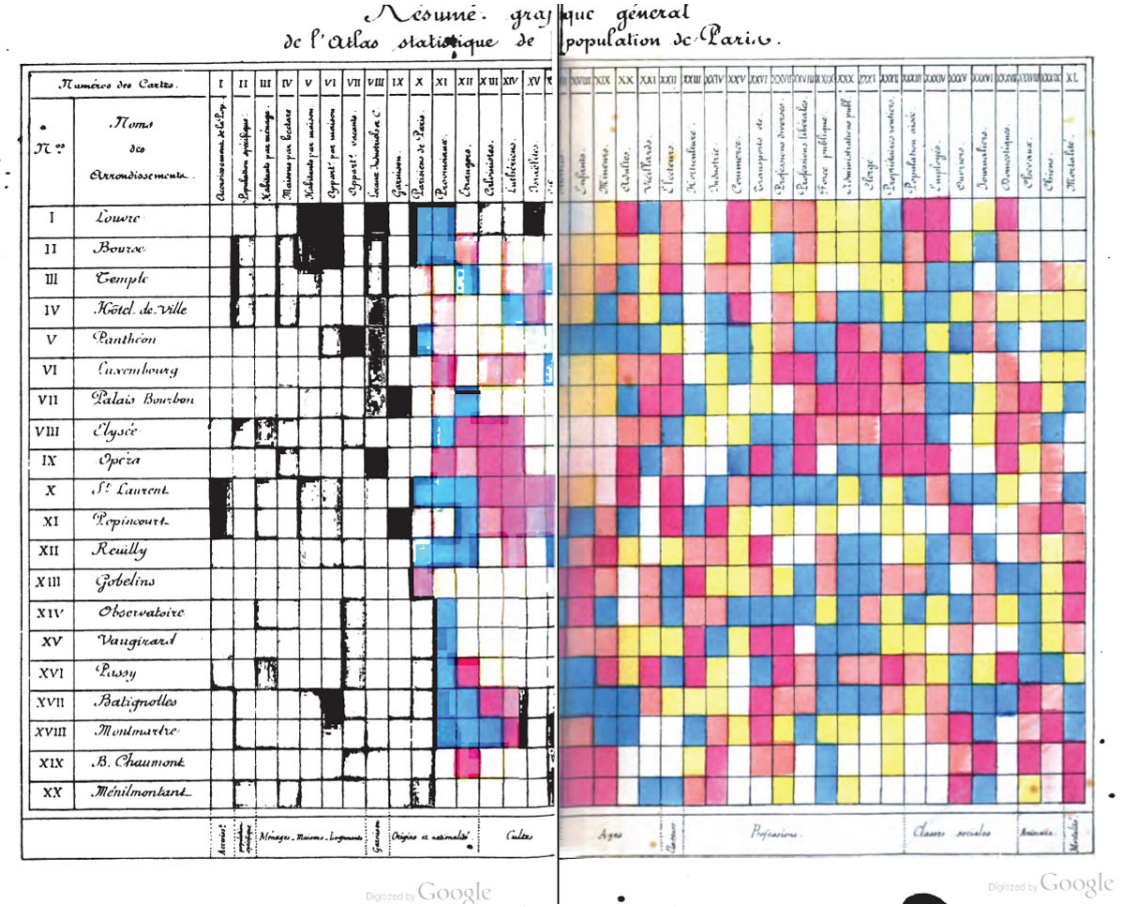
Figura 46 – Mapa de calor de dados de expressão gênica



Fonte: (EISEN *et al.*, 1998)

Anexo B – Figura ampliada de visualização utilizando esquema de cores

Figura 47 – Exemplo de tabela colorida, datada de 1873, semelhante aos mapas de calor.



Fonte: (LOUA, 1873)