



**UNIVERSIDADE ESTADUAL DE SANTA CRUZ  
PRO-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO**

**PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL  
EM CIÊNCIA E TECNOLOGIA**

**JORGE FARIAS HERCULANO**

**MÉTODO PARA AGRUPAMENTO DE GENES UTILIZANDO FATORES DE  
TRANSCRIÇÃO E EXPRESSÃO DIFERENCIAL, RELACIONANDO-OS COM A DOENÇA  
DE ALZHEIMER**

**ILHÉUS-BA  
2016**

**JORGE FARIAS HERCULANO**

**MÉTODO PARA AGRUPAMENTO DE GENES UTILIZANDO  
FATORES DE TRANSCRIÇÃO E EXPRESSÃO DIFERENCIAL,  
RELACIONANDO-OS COM A DOENÇA DE ALZHEIMER**

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional em Ciência e Tecnologia (PPGMC) da Universidade Estadual de Santa Cruz para obtenção do título de Mestre em Modelagem Computacional em Ciência e Tecnologia.

Orientador: Prof. Dr. Luciano Angelo de Souza Bernardes

Coorientador: Prof. Dr. Robson da Silva Magalhães

ILHÉUS-BA  
2016

H539

Herculano, Jorge Farias.

Método para agrupamento de genes utilizando fatores de transcrição e expressão diferencial, relacionando-os com a doença Alzheimer / Jorge Farias Herculano. – Ilhéus, BA: UESC, 2016.

74 f. : il.

Orientador: Luciano Angelo de Souza Bernardes.

Coorientador: Robson da Silva Magalhães.

Dissertação (Mestrado) – Universidade Estadual de Santa Cruz. Programa de Pós-Graduação em Modelagem Computacional em Ciência e Tecnologia.

Inclui referências e apêndices.

1. Alzheimer, Doença de. 2. Alzheimer, Doença de – Diagnóstico. 3. Algoritmo de agrupamento – Fuzzy C-Means. 4. Fatores de transcrição. I. Título.

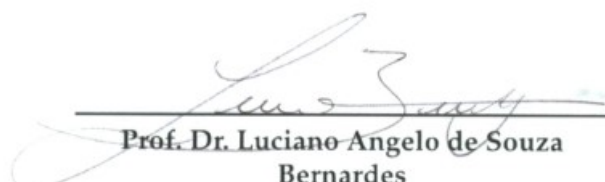
CDD 616.831

JORGE FARIAS HERCULANO

**MÉTODO PARA AGRUPAMENTO DE GENES UTILIZANDO  
FATORES DE TRANSCRIÇÃO E EXPRESSÃO DIFERENCIAL,  
RELACIONANDO-OS COM A DOENÇA DE ALZHEIMER**


Ilhéus-BA, 20/05/2016

Comissão Examinadora



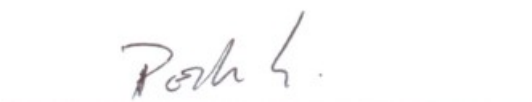
---

**Prof. Dr. Luciano Angelo de Souza  
Bernardes**  
UESC  
(Orientador)




---

**Prof. Dr. Robson da Silva Magalhães**  
UFSB  
(Coorientador)



---

**Prof. Dr. Paulo Eduardo Ambrósio**  
UESC



---

**Prof. Dr. Gesivaldo Santos**  
UESB

A Deus, por sua infinita bondade.

## **Agradecimentos**

- À UESC.
- Aos professores Luciano Bernardes e Robson Magalhães pelo apoio e orientação durante o desenvolvimento da pesquisa.
- Aos professores do PPGMC que contribuíram na formação através das atividades oferecidas pelo programa.
- Ao corpo técnico do programa pelo apoio durante o curso.
- Aos membros da banca examinadora por aceitar o convite e contribuir para o enriquecimento do trabalho.
- À minha família, por toda a dedicação e apoio.
- Aos amigos, em especial Neide Braga, pelo apoio e pelas inúmeras contribuições.
- Aos colegas do PPGMC com quem aprendi muito durante o mestrado.
- Aos colegas da Pró-reitoria de Tecnologia de Informação e Comunicação da UFSB, pelo convívio e aprendizado constante.

## Método para Agrupamento de Genes Utilizando Fatores de Transcrição e Expressão Diferencial, Relacionando-os com a Doença de Alzheimer

### Resumo

Pesquisas voltadas à causa e ao tratamento de doenças neurodegenerativas tem sido intensificadas nos últimos anos. Dentre estas, as pesquisas voltadas à causa da Doença de Alzheimer (DA) tem ganhado impulso em diversos países. No Brasil, não é diferente, pois com o envelhecimento da população, o aumento do número de casos dessa doença é mais perceptível. Com base na hipótese de que a ação conjunta de genes e fatores de transcrição (FT) contribui para o surgimento da doença, este trabalho teve por objetivo, propor um método para agrupamento de genes utilizando dados obtidos através de experimento com *microarray* e dados de fatores de transcrição, ambos relacionados à DA. No campo da bioinformática, geralmente, os dados apresentam incertezas. A utilização de um algoritmo que consiga englobar essas características se tornou imprescindível. Dessa forma, neste método foi utilizado o algoritmo *Fuzzy C-Means* para a realização do agrupamento. Após a geração dos *clusters*, foram feitas as seleções dos grupos que mais apresentaram concentração de genes e FT que possuem envolvimento com a DA já identificado. Ao final da aplicação do método foram selecionados 40 genes. Foram feitas pesquisas por publicações que associam estes genes com a Doença de Alzheimer, confirmando o envolvimento de alguns e indicando o possível envolvimento dos demais com a DA.

**Palavras-chave:** Doença de Alzheimer, Agrupamento, *Fuzzy C-Means*, Genes, Fatores de Transcrição.

# Method for Genes Cluster Using Transcription Factors and Differential Expression, Relating them with Alzheimer's Disease

## Abstract

Research focused on the causes and treatment of neurodegenerative diseases has been intensified in recent years. Among these, the research focused on the cause of Alzheimer's disease (AD) has gained incentive in several countries. In Brazil, it is no different, because with the aging population, the increasing number of cases of this disease is more noticeable. Based on the hypothesis that the joint action of genes and transcription factors (TF) contributes to the onset of the disease, this study aimed to propose a method for clustering genes using data obtained through experiment with microarray and data transcription factors, both related to AD. In the field of bioinformatics, generally, the data present uncertainty. The use of an algorithm that can include such characteristics became essential. Thus, this method was used Fuzzy C-Means algorithm for performing grouping. After the generation of clusters, selections were made of the groups that showed more concentration of genes and TF that have involvement with AD have identified. At the end of the method we selected 40 genes. Surveys were made by publications that these genes are associated with Alzheimer's disease, confirming the involvement of any and indicating the possible involvement of other genes with AD.

**Keywords:** Alzheimer's Disease, Clustering, Fuzzy C-Means, Genes, Transcription Factors.



## Lista de figuras

Figura 1 – Estrutura dos quatro nucleotídeos do DNA. . . . .	4
Figura 2 – Açúcar ribose do RNA e açúcar desoxirribose do DNA. . . . .	5
Figura 3 – A base Uracil do RNA, e a base Timina do DNA. . . . .	6
Figura 4 – Diferentes maneiras de agrupar o mesmo conjunto de dados. . . . .	13
Figura 5 – Fluxo do Método Proposto. . . . .	20
Figura 6 – $\log$ do módulo das diferenças para o conjunto de dados com maior valor de expressão. . . . .	24
Figura 7 – $\log$ do módulo das diferenças para o conjunto com a mediana do valor de expressão. . . . .	25
Figura 8 – $\log$ do módulo das diferenças para o conjunto com o menor valor de expressão. . . . .	25
Figura 9 – Histograma do <i>cluster</i> 01 do conjunto com o menor valor de expressão. . . . .	28
Figura 10 – Histograma do <i>cluster</i> 18 para o conjunto de dados com o maior valor de expressão. . . . .	29
Figura 11 – Histograma do <i>cluster</i> 18 identificando os FT relacionados aos genes envolvidos na DA. . . . .	30
Figura 12 – Percentual de genes e FT na faixa aproximada de 20% de menor pertinência nos <i>cluster</i> . . . . .	32
Figura 13 – Percentual de genes e FT na faixa aproximada de 20% de maior pertinência nos <i>cluster</i> . . . . .	33
Figura 14 – Percentual de genes e FT na faixa aproximada de 20% de menor pertinência nos <i>cluster</i> . . . . .	36
Figura 15 – Percentual de genes e FT na faixa aproximada de 20% de maior pertinência nos <i>cluster</i> . . . . .	37
Figura 16 – Percentual de genes e FT na faixa aproximada de 20% de menor pertinência nos <i>cluster</i> . . . . .	40
Figura 17 – Percentual de genes e FT na faixa aproximada de 20% de maior pertinência nos <i>cluster</i> . . . . .	41
Figura 18 – Intersecção dos <i>clusters</i> no Grupo 01. . . . .	43
Figura 19 – Intersecção dos <i>clusters</i> no Grupo 02. . . . .	44
Figura 20 – Intersecção dos <i>clusters</i> no Grupo 03. . . . .	44
Figura 21 – Intersecção dos <i>clusters</i> no Grupo 04. . . . .	45
Figura 22 – Níveis de expressão dos genes selecionados em relação à amostra controle (conjunto com menor valor de expressão). . . . .	49
Figura 23 – Níveis de expressão dos genes selecionados em relação à amostra controle (conjunto com a mediana do valor de expressão). . . . .	50

Figura 24 – Níveis de expressão dos genes selecionados em relação à amostra controle (conjunto com maior valor de expressão). . . . .	51
Figura 25 – Genes e FT na seleção de menor pertinência no <i>cluster</i> 08 do conjunto com maior valor de expressão. . . . .	64
Figura 26 – Genes e FT na seleção de menor pertinência no <i>cluster</i> 09 do conjunto com maior valor de expressão. . . . .	65
Figura 27 – Genes e FT na seleção de menor pertinência no <i>cluster</i> 16 do conjunto com maior valor de expressão. . . . .	65
Figura 28 – Genes e FT na seleção de maior pertinência no <i>cluster</i> 16 do conjunto com maior valor de expressão. . . . .	66
Figura 29 – Genes e FT na seleção de menor pertinência no <i>cluster</i> 05 do conjunto com a mediana valor de expressão. . . . .	66
Figura 30 – Genes e FT na seleção de maior pertinência no <i>cluster</i> 05 do conjunto com a mediana valor de expressão. . . . .	67
Figura 31 – Genes e FT na seleção de menor pertinência no <i>cluster</i> 07 do conjunto com a mediana valor de expressão. . . . .	67
Figura 32 – Genes e FT na seleção de maior pertinência no <i>cluster</i> 07 do conjunto com a mediana valor de expressão. . . . .	68
Figura 33 – Genes e FT na seleção de maior pertinência no <i>cluster</i> 09 do conjunto com a mediana valor de expressão. . . . .	68
Figura 34 – Genes e FT na seleção de maior pertinência no <i>cluster</i> 12 do conjunto com a mediana valor de expressão. . . . .	69
Figura 35 – Genes e FT na seleção de menor pertinência no <i>cluster</i> 17 do conjunto com a mediana valor de expressão. . . . .	69
Figura 36 – Genes e FT na seleção de maior pertinência no <i>cluster</i> 17 do conjunto com a mediana valor de expressão. . . . .	70
Figura 37 – Genes e FT na seleção de menor pertinência no <i>cluster</i> 18 do conjunto com a mediana valor de expressão. . . . .	70
Figura 38 – Genes e FT na seleção de menor pertinência no <i>cluster</i> 01 do conjunto com o menor valor de expressão. . . . .	71
Figura 39 – Genes e FT na seleção de menor pertinência no <i>cluster</i> 16 do conjunto com o menor valor de expressão. . . . .	71
Figura 40 – Genes e FT na seleção de maior pertinência no <i>cluster</i> 19 do conjunto com o menor valor de expressão. . . . .	72
Figura 41 – Genes e FT na seleção de maior pertinência no <i>cluster</i> 26 do conjunto com o menor valor de expressão. . . . .	72
Figura 42 – Genes e FT na seleção de menor pertinência no <i>cluster</i> 34 do conjunto com o menor valor de expressão. . . . .	73

Figura 43 – Genes e FT na seleção de maior pertinência no <i>cluster</i> 40 do conjunto com o menor valor de expressão. . . . .	73
Figura 44 – Genes e FT na seleção de maior pertinência no <i>cluster</i> 43 do conjunto com o menor valor de expressão. . . . .	74

## Lista de tabelas

Tabela 1 – Exemplo dos dados de <i>microarray</i> . . . . .	17
Tabela 2 – Exemplo dos dados de Fatores de Transcrição . . . . .	18
Tabela 3 – Exemplo dos cálculos para determinação do número ideal de <i>clusters</i> para o conjunto de dados considerando o maior valor de expressão . . . . .	24
Tabela 4 – Percentual da presença de genes e FT em aproximadamente 20% de maior e menor pertinência nos <i>clusters</i> do conjunto com maior valor de expressão . . . . .	31
Tabela 5 – Percentual da presença de genes e FT em aproximadamente 20% de maior e menor pertinência nos <i>clusters</i> do conjunto com a mediana do valor de expressão . . . . .	34
Tabela 6 – Percentual da presença de genes e FT em aproximadamente 20% de maior e menor pertinência nos <i>clusters</i> do conjunto com o menor valor de expressão . . . . .	38
Tabela 7 – <i>Clusters</i> por grupo selecionados para análise . . . . .	43
Tabela 8 – Percentual de presença Genes e FT relacionados ao Alzheimer nos grupos . . . . .	45
Tabela 9 – Publicações relacionadas ao genes selecionados do grupo 04 associadas com o termo <i>Alzheimer</i> . . . . .	46
Tabela 10 – Teste de Kruskal-Wallis nos 40 genes selecionados, considerando os níveis de expressão dos conjuntos com o maior, a mediana e o menor valor de expressão . . . . .	53
Tabela 11 – Principais doenças associadas aos genes selecionados . . . . .	55

## **Lista de quadros**

Quadro 1 – Genes Relacionados à Doença de Alzheimer. . . . .	19
--	----

## Lista de abreviaturas e siglas

ANOVA	Análise de Variância
DA	Doença de Alzheimer
DNA	Ácido desoxirribonucléico
FCM	<i>Fuzzy C-Means</i>
FT	Fatores de Transcrição
HTRIdb	Human Transcription Regulation Interactions Database
PHP	Hypertext Preprocessor
NCBI	National Center for Biotechnology Information
RNA	Ácido Ribonucléico

# Sumário

<b>1 – Introdução</b>	<b>1</b>
1.1 Problematização	1
1.2 Objetivos	2
1.2.1 Objetivo Geral	2
1.2.2 Objetivos Específicos	2
1.3 Justificativas	2
1.4 Organização do Trabalho	3
<b>2 – Referencial Teórico</b>	<b>4</b>
2.1 Princípios de Biologia Molecular	4
2.1.1 DNA e Genes	4
2.1.2 RNA	5
2.1.3 Proteínas	7
2.2 Doença de Alzheimer	8
2.3 <i>Microarray</i>	9
2.4 Fatores de Transcrição	10
2.5 Fuzzy	11
2.6 Agrupamento	12
2.6.1 Fuzzy c-Means	13
<b>3 – Aspectos Metodológicos, Materiais e Recursos Computacionais</b>	<b>16</b>
3.1 Aspectos Metodológicos	16
3.2 Materiais	17
3.3 Recursos Computacionais	19
<b>4 – Método Proposto e Resultados</b>	<b>20</b>
4.1 Método para Agrupamento de Genes Utilizando Fatores de Transcrição e Expressão Diferencial	20
4.2 Estudo de Caso: Aplicação do Método em Dados relacionados à Doença de Alzheimer	21
4.2.1 Obtenção dos Dados	21
4.2.2 Pré-processamento dos Dados	22
4.2.3 Determinação do Número de <i>Clusters</i>	23
4.2.4 <i>Fuzzy C-Means</i> (FCM)	26
4.2.5 Distribuição de Frequências	27
4.2.6 Seleção dos <i>Clusters</i>	31

4.2.7	Refinamento e Resultados . . . . .	42
4.3	Análise dos Resultados . . . . .	46
5	– Considerações Finais . . . . .	56
	Referências . . . . .	57
	<b>Apêndices</b>	<b>63</b>
	<b>APÊNDICE A – Grau de pertiência dos genes e fatores de transcrição por faixa de seleção nos <i>clusters</i> . . . . .</b>	<b>64</b>



# 1 Introdução

Durante os últimos anos, pesquisas voltadas à causa e ao tratamento de doenças neurodegenerativas tem sido intensificadas. A morte ou mau funcionamento de neurônios provoca alterações na memória, no comportamento e na capacidade de pensar com clareza, além de atrapalhar o armazenamento de informações recentes. Dentre os principais tipos de demências, a doença de Alzheimer (DA) é a que mais afeta a população mundial, sendo que as mudanças cerebrais provocadas pela DA prejudicam a capacidade de um indivíduo realizar funções corporais básicas como andar e engolir (ALZHEIMER'S ASSOCIATION, 2015).

O Brasil tem se despontado nessas pesquisas, contudo a causa da DA ainda é desconhecida, mas especialistas afirmam que a mesma se desenvolve em decorrência de diversos fatores, implicando em um diagnóstico que geralmente é tardio, ou seja, na maioria dos casos, acontece quando a DA já é moderada ou grave.

Neste contexto, uma pesquisa fundamentada na observação de dados de experimentos de microarranjo (do inglês *microarray*) pode contribuir para identificar, ou nos direcionar à uma melhor compreensão da origem do problema, e assim encontrar um tratamento mais efetivo ou até mesmo a cura. No entanto, desenvolver métodos computacionais eficientes para análise desses dados é um desafio para a bioinformática.

Os métodos de agrupamento, atualmente, consideram todos os pontos de um perfil de modulação dos genes, e quanto maior for a quantidade destes pontos melhor para a caracterização do agrupamento. O objetivo desta pesquisa foi desenvolver um método que seja capaz de agrupar os genes utilizando as proteínas da família de fatores de transcrição (FT) que demonstram interação com as suas regiões promotoras e os perfis de expressão.

## 1.1 Problematização

Para um efetivo tratamento na DA, é imprescindível que essa seja diagnosticada na sua fase inicial. No entanto, esse diagnóstico geralmente é tardio, ou seja, ocorre quando a DA se encontra de moderada a grave. Diversas pesquisas têm sido realizadas no intuito de identificar a raiz do problema, mas apesar dos avanços, a origem da DA ainda é desconhecida. Acredita-se que o estabelecimento da DA, ainda como em tantas outras doenças complexas, não ocorra devido a alterações em um único gene, mas seja resultante de um acúmulo de alterações, cada uma contribuindo com pequenos efeitos que resultam, em conjunto, no estabelecimento da doença em diferentes graus de severidade (FRIDMAN et al., 2004).

Diante disso, o problema que se segue é: como agrupar genes relacionados à Doença de Alzheimer de forma a possibilitar uma compreensão melhor da doença e assim contribuir para um diagnóstico precoce?

## 1.2 Objetivos

### 1.2.1 Objetivo Geral

Desenvolver um método para agrupar genes, utilizando proteínas da família de fatores de transcrição que demonstrem interação com suas regiões promotoras e com perfis de expressão, esses últimos, observados em experimentos com *microarrays*.

### 1.2.2 Objetivos Específicos

- (a) Extrair dados de sites de livre acesso e especializados (ex: GEO) em experimentos de *microarray*, relacionados à doença de Alzheimer;
- (b) Extrair dados de sites especializados em FT, comprovados por experimentos *in silico* ou de bancada, e buscar informações sobre relação FT/gene, ligadas à DA;
- (c) Estabelecer um método para criar subgrupos, a fim de garimpar genes e FT predominantes e atuantes;
- (d) Aplicar o método de agrupamentos aos dados selecionados;
- (e) Analisar os resultados obtidos pelo método de agrupamento;

## 1.3 Justificativas

A Doença de Alzheimer é um dos tipos de demência que mais afeta a população mundial. A cada 68 segundos alguém desenvolve a DA e os gastos públicos com a doença são na ordem de bilhões de dólares por ano (KLEIN LAB, 2015).

Pesquisas têm sido realizadas em todo o mundo no intuito de identificar a origem da DA. No entanto, apesar dos avanços, a origem da doença ainda é uma incógnita. O surgimento do *microarray* permitiu que as pesquisas ganhassem mais força. Com os dados obtidos através dos experimentos de *microarray*, agora é possível obter uma visão mais generalizada da atividade simultânea dos múltiplos caminhos celulares.

No entanto, analisar os dados de *microarray* pode não ser trivial. Segundo Cad-dick e Dobson (2007), o padrão de expressão não é necessariamente sincronizado. Agrupar genes com base apenas no seu perfil de expressão não diz como eles são regulados. Enquanto amplamente se utiliza o agrupamento por perfil de expressão

obtido em experimento de *microarray*, neste projeto foi proposto uma abordagem de agrupamento diferenciada (ver capítulo 4).

O Brasil tem despontado na busca pela causa e tratamento da Doença de Alzheimer. No entanto, o diagnóstico geralmente é tardio, ele ocorre quando a doença já é moderada ou grave. Uma solução seria a realização do diagnóstico precoce, utilizando *microarray*. Neste sentido, se faz necessário intensificar pesquisas em técnicas de análise de dados obtidos com *microarray* e dados de fatores de transcrição, possibilitando assim uma compreensão melhor da doença e contribuir para um diagnóstico precoce.

## 1.4 Organização do Trabalho

Este trabalho é organizado em cinco capítulos, com o capítulo 1 apresentando a introdução, a problematização, os objetivos e as justificativas. O capítulo 2 apresenta a fundamentação teórica que embasou a pesquisa. No capítulo 3 são apresentados os aspectos metodológicos, os materiais e os recursos computacionais utilizados. No capítulo 4 são apresentados e discutidos os resultados da pesquisa. Por fim, no capítulo 5 são feitas as considerações finais da pesquisa.

## 2 Referencial Teórico

### 2.1 Princípios de Biologia Molecular

#### 2.1.1 DNA e Genes

Segundo Griffiths et al. (2009), desde a descoberta da estrutura do Ácido Desoxirribonucleico (do inglês deoxyribonucleic acid, DNA) por Watson e Crick (1953), foram criados novos caminhos para a compreensão, a um nível molecular, da genética e da hereditariedade. A estrutura do DNA, proposta por eles, apresenta dois polinucleotídeos associados que se enrolam em conjunto para formar uma dupla hélice (LODISH et al., 2004).

Quimicamente, apesar de simples, os polinucleotídeos que compõem o DNA são longas cadeias formadas por quatro bases nitrogenadas: adenina (A), citosina (C), guanina (G) e timina (T), além do fosfato e do açúcar denominado de desoxirribose, como pode ser observado nos desenhos esquemáticos das estruturas para as quatro bases do DNA (Figura 1). Além disso, estes polinucleotídeos se ligam através de pontes de hidrogênio, formando os pares de bases A-T e G-C.

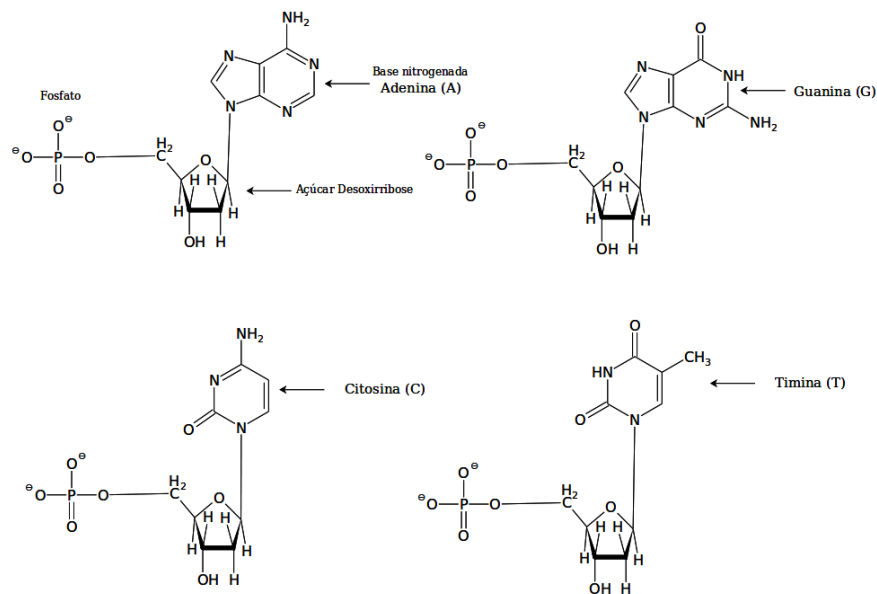


Figura 1 – Estrutura dos quatro nucleotídeos do DNA.

Fonte: Adaptado de Griffiths et al. (2009).

As informações armazenadas no DNA são organizadas em unidades hereditárias, denominadas genes, que determinam as características de um organismo. Um gene é, de acordo com Lodish et al. (2004), uma "unidade de DNA que contém a informação necessária para especificar a síntese de uma cadeia polipeptídica simples ou RNA funcional". Este processo é conhecido como transcrição, e consiste na cópia da informação armazenada no DNA para o Ácido Ribonucleico (do inglês *ribonucleic acid*, RNA). Este último, além de outras funções, está diretamente envolvido na síntese proteica.

O fluxo de informação genética nas células, de um modo geral, é de DNA → RNA → Proteína. Esse fluxo é denominado de Dogma Central da Biologia Molecular. Em uma representação geral do dogma central da biologia molecular, o fluxo da informação genética passa pelas seguintes fases: transcrição, processamento de RNA, tradução e replicação. No entanto, esta simples representação não reflete os papéis das proteínas na síntese de ácidos nucleicos (LODISH et al., 2004; ALBERTS et al., 2010).

### 2.1.2 RNA

O Ácido Ribonucleico (RNA) possui uma estrutura química semelhante ao DNA, sendo composto por quatro bases nitrogenadas, além do fosfato e açúcar. Uma das principais diferenças em relação ao DNA é a composição do açúcar presente em sua estrutura, que no caso do RNA é a ribose (Figura 2).

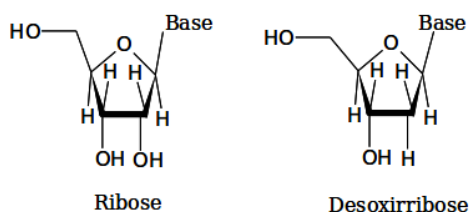


Figura 2 – Açúcar ribose do RNA e açúcar desoxirribose do DNA.

Fonte: Adaptado de Griffiths et al. (2009).

Outra diferença do RNA para o DNA é a substituição da base timina (T) pela uracil (U). A diferença química da timina em relação à base uracil é determinada pela presença da molécula CH<sub>3</sub> no carbono 5 de sua estrutura (Figura 3).

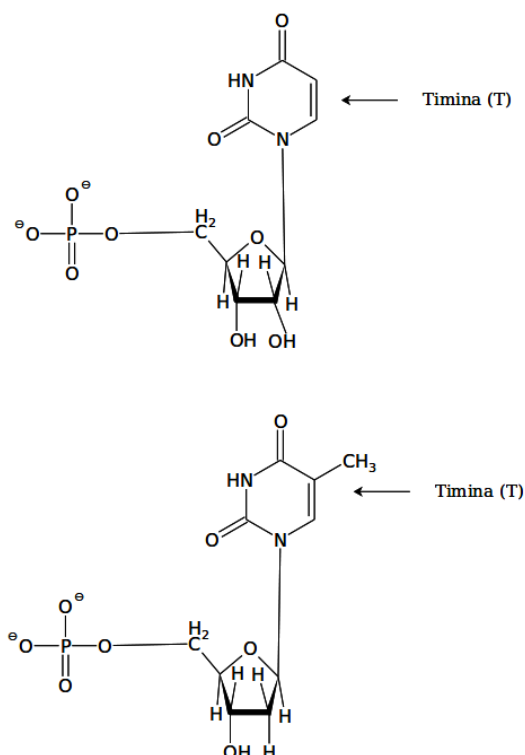


Figura 3 – A base Uracil do RNA, e a base Timina do DNA.

Fonte: Adaptado de Griffiths et al. (2009).

Segundo Lodish et al. (2004) o RNA pode assumir três funções distintas na síntese de proteínas, a saber:

- **RNA mensageiro (mRNA):** transporta a informação genética transcrita a partir do DNA sob a forma de uma série com três sequências de nucleótidos, chamados codões ou códons, cada um dos quais especifica um aminoácido.
- **RNA transportador (tRNA):** é a chave para decifrar os códons no mRNA. Cada tipo de aminoácido tem o seu próprio subconjunto de tRNA, cujos elementos se ligam a aminoácidos para transportá-los para a extremidade crescente de uma cadeia de polipeptídeos. Este transporte ocorrerá se o códon seguinte no mRNA o chamar. O tRNA correto, com o seu aminoácido ligado, é selecionado a cada passo, já que cada molécula de tRNA contém uma sequência de três nucleótidos, chamada de anticódon, que forma o par com o seu códon complementar no mRNA.
- **RNA ribossômico (rRNA):** se associa com um conjunto de proteínas para formar o ribossomo. São estruturas complexas, que se movem fisicamente ao longo

de uma molécula de mRNA catalisando a união dos aminoácidos em cadeias de polipeptídeos. Eles também se ligam a tRNA e a várias proteínas acessórias necessárias para a síntese de proteínas.

Desta forma, o mRNA é traduzido em proteína através da ação do tRNA, juntamente com o ribossomo.

### 2.1.3 Proteínas

As proteínas são, do ponto de vista de um químico, as moléculas com estruturas mais complexas e funcionalmente mais sofisticadas que se tem conhecimento (ALBERTS et al., 2010). Composta por uma longa cadeia de aminoácidos, as proteínas são o resultado do processo de tradução do mRNA.

As proteínas são as operárias das células, executam os programas de atividade codificados pelos genes. Sendo constituídas a partir de 20 tipos de aminoácidos, as proteínas desempenham grande variedade de tarefas (LODISH et al., 2004).

Os aminoácidos são os blocos de construção das proteínas. Segundo Lodish et al. (2004), eles têm uma estrutura característica que consiste em um átomo de carbono central  $\alpha$  ( $C_\alpha$ ) ligado a quatro grupos químicos diferentes: um grupo amino ( $NH_2$ ); um grupo carboxílico ( $COOH$ ); um átomo de hidrogênio ( $H$ ); e um grupo variável, chamado de uma cadeia lateral, ou grupo  $R$ . O grupo  $R$  é o que confere a cada aminoácido suas características únicas.

A função da proteína está relacionada à sua estrutura tridimensional, e essa estrutura é determinada pela distribuição dos aminoácidos ao longo da cadeia peptídica. Lodish et al. (2004) consideram a estrutura de proteínas em quatro níveis de organização, começando com os seus blocos de construção monoméricos, os aminoácidos:

- A estrutura primária de uma proteína é simplesmente o arranjo linear, ou sequência, dos resíduos de aminoácidos que compõem.
- O segundo nível na hierarquia da estrutura da proteína é composto dos vários arranjos espaciais resultantes da dobradura de partes localizadas de uma cadeia polipeptídica, esses arranjos são referidos como estruturas secundárias.
- A estrutura terciária refere-se à conformação geral de uma cadeia de polipéptido que é o arranjo tridimensional de todos os seus resíduos de aminoácidos.
- Proteínas multiméricas consistem de dois ou mais polipéptidos ou suas subunidades. A estrutura quaternária, descreve o número (estequimétrico) e as posições relativas das subunidades em proteínas multiméricas.

- O nível mais alto da estrutura proteica é a associação de proteínas em conjuntos macromoleculares.

As proteínas podem desempenhar diversas funções nas células, tais como: regulação, sinalização, estrutural, catalisadora, entre outras (LODISH et al., 2004).

As proteínas chamadas Fatores de Transcrição (FT), tem por objetivo a regulação da expressão gênica. Mais detalhes sobre o papel destas proteínas na regulação gênica serão tratados na seção 2.4.

## 2.2 Doença de Alzheimer

A Doença de Alzheimer, das doenças classificadas como demência, é a que representa a maioria dos casos. As últimas décadas trouxeram um progresso sem precedentes na compreensão da genética, fisiopatológica e histórico-natural dessa doença (GREEN, 2001). A DA é uma doença neurodegenerativa progressiva e fatal, que se manifesta: por deterioração cognitiva e da memória, pelo comprometimento progressivo das atividades da vida diária, e com uma variedade de sintomas neuropsiquiátricos e distúrbios comportamentais (CUMMINGS, 2004).

O termo demência é genericamente usado para descrever uma variedade de doenças e condições que se desenvolvem quando as células nervosas do cérebro (neurônios) morrem ou deixam de funcionar normalmente. O mau funcionamento e a morte dessas células provoca alterações na memória, no comportamento e na capacidade de pensar com clareza. Na DA, essas mudanças cerebrais, eventualmente, prejudicam a capacidade de um indivíduo realizar funções corporais básicas, como andar e engolir. A DA, é em última análise, fatal (ALZHEIMER'S ASSOCIATION, 2015).

A maioria dos especialistas concordam que a AD está associada à vários fatores ao invés de uma única causa. Esses fatores, incluem uma variedade de alterações cerebrais que antecedem, em até 20 anos, ao aparecimento dos sintomas. No início o cérebro do indivíduo funciona normalmente, mas com o aumento do dano neural o cérebro não consegue compensar essa falha e começa a apresentar problemas cognitivos (ALZHEIMER'S ASSOCIATION, 2015).

Segundo Fridman et al. (2004), o estabelecimento da DA deve-se ao acúmulo de eventos genéticos e ambientais. Cada um desses eventos contribui com pequenos efeitos que resultam, em conjunto, no estabelecimento da doença, com diferentes graus de severidade, sendo que, o fator genético é considerado como preponderante na etiopatogenia da DA (SERENIKI; VITAL, 2008).

A doença é caracterizada por acumulação extracelular do peptídeo beta-amilóide ( $A\beta$ ), placas senis, com a aparição de emaranhados neurofibrilares intracelulares e



perda sináptica e neuronal (FELICE et al., 2004; SELKOE, 2002). Tais placas contêm depósitos extracelulares de proteína beta-amiloide que ocorrem, principalmente, na forma filamentosa, ou seja, como massas em forma de estrela de fibrilas amilóides (SELKOE, 2002). Essas placas são umas das responsáveis por interromper a sinapse entre os neurônios ocasionando a morte ou mal funcionamento dos mesmos (ALZHEIMER'S ASSOCIATION, 2015).

Pesquisas tem indicado que as fibrilas não são as únicas formas tóxicas de ( $A\beta$ ), e talvez não sejam as neurotoxinas que são mais relevante para a DA. Pequenos oligômeros e protofibrilas também têm poderosa atividade neurológica (KLEIN, 2001).

As placas amilóides, oligômeros, fibrilas e a proteína beta-amiloide, presentes no tecido neural, são responsáveis pela ação danosa neste tecido.

Como pode ser percebido, a DA é uma patologia complexa, e portanto, ainda sem causa definida. Diante disso, o estudo para encontrar a origem da doença é um campo ainda em exploração (CAYTON et al., 2000). Os dados obtidos com *microarrays* são poderosas ferramentas para lidar com essa complexidade, porque permitem uma visão geral da atividade simultânea, considerando-se os múltiplos caminhos celulares (BLALOCK et al., 2004).

## 2.3 *Microarray*

O *microarray* de DNA, ou *DNA-chip*, é utilizado para analisar o conteúdo do mRNA em uma célula, de forma a revelar os padrões de expressão das proteínas ou para analisar o DNA genômico que revele os genes ausentes ou mutados (LESK, 2008). O *microarray* apresenta-se como uma importante ferramenta nas pesquisas genômicas. Os seres vivos possuem milhares de genes, e esta tecnologia permite uma visão geral dos níveis de modulação de cada um.

Os *microarrays* são produzidos com pequenos *chips* de vidro ou náilon, medindo cerca de  $2\text{cm}^2$ . Dispostos em um arranjo quadrado, oligonucleotídeos são fixados ao *chip* podendo variar entre 10.000 e 250.000 posições por  $\text{cm}^2$  (LESK, 2008).

Segundo Krishnamurthy (2006), dependendo do tipo de amostra imobilizada para a construção do *array* e as informações buscadas, os *microarrays* podem ser categorizados em três tipos:

- ***Microarray para análise de expressão:*** neste tipo de *microarray* o DNA complementar (cDNA) derivado do mRNA de genes previamente conhecidos é imobilizado no chip. São inseridas amostras de tecidos sadios e de tecidos doentes. Dessa forma, pontos com mais intensidade são obtidos para o gene do tecido doente, ou seja, se o gene está superexpresso na condição de doença. Este padrão de

expressão é então comparado com o padrão de expressão de um gene responsável por uma doença.

- **Microarray para análise de mutação:** para este tipo de *microarray*, os pesquisadores usam DNA genômico (gDNA). Os genes podem ser diferentes uns dos outros por apenas uma única base nucleotídica. A diferença de uma única base entre duas sequências é conhecido como Polimorfismo de Nucleotídeo Único.
- **Hibridização genômica comparativa:** é utilizado para a identificação do aumento ou diminuição dos fragmentos cromossômicos importantes que abrigam genes envolvidos em uma doença.

## 2.4 Fatores de Transcrição

O processo de transcrição é a primeira etapa da expressão gênica, resultando na produção de um transcrito primário de RNA a partir do DNA de um gene em particular (LATCHMAN, 1997). O controle da atividade do gene depende de proteínas de ligação ao DNA, denominadas Fatores de Transcrição, que atuam como interruptores, ativando ou reprimindo a transcrição de genes alvos (LODISH et al., 2004). Os fatores de transcrição são também conhecidos como proteínas de regulação gênica (ALBERTS et al., 2010).

A produção de RNA a partir de um molde de DNA é catalisada por enzimas chamadas RNA polimerase. Nas células eucarióticas existem três tipos destas enzimas: RNA polimerase I, que cataliza a produção de RNA ribossômico (rRNA); RNA polimerase III, que cataliza a produção do RNA transportador (tRNA); e o RNA polimerase II, que é responsável pela síntese de todo o RNA mensageiro (mRNA) (CREIGHTON, 1999).

A iniciação da transcrição é controlada através da interação de FT com os seus locais de ligação ao DNA, regiões promotoras e potenciadoras de genes (CREIGHTON, 1999). Os FT podem ser agrupados em famílias de acordo com os motivos (*motifs*) de ligação com o DNA. Os motivos estruturais, comuns encontrados nos fatores de transcrição das células eucarióticas, são segundo Philips e Hoopes (2008):

- *Basic helix-loop-helix*: nos organismos eucarióticos regula o desenvolvimento dos músculos, nervos, sangue e células pancreáticas.
- *Helix-turn-helix*: envolvidos em cenários de supressão de genes durante a diferenciação celular.
- *Zinc finger*: regula a remodelação da cromatina e da estrutura do DNA; também se liga a proteínas e lipídios.

- *Leucine zipper*: regula a divisão celular e a interação de proteínas.

As células de organismos superiores apresentam um número incrível de respostas gênicas ao seu ambiente. Isto é, em grande parte, resultado da ação dos FT que regem a forma como os genes são transcritos e como o RNA polimerase II é recrutado. Através destes mecanismos, os FT controlam aspectos importantes do desenvolvimento do organismo. As famílias dos FT aumentam ainda mais o nível de complexidade genética dos eucariotos, e muitos os FT, dentro da mesma família, frequentemente trabalham em conjunto na transcrição de um único gene (PHILIPS; HOOPES, 2008).

Os FT desempenham papel fundamental no controle da transcrição em células eucaróticas. Segundo Lodish et al. (2004), as proteínas ativadoras se ligam a elementos de controle específicos do DNA na cromatina, e interagem com máquinas de multiproteínas coativadoras, como mediador, para descondensar a cromatina e reunir a RNA polimerase e os FT gerais nos promotores. Genes inativos são reunidos em regiões da cromatina condensada, que inibem a RNA polimerase, e seus FT gerais associados de interação com os promotores. Alternativamente, as proteínas repressoras se ligam a outros elementos de controle para inibir a iniciação pelo RNA polimerase, e interagem com complexos de multiproteínas co-repressoras para condensar a cromatina.

## 2.5 Fuzzy

Um conjunto *fuzzy* é uma classe de objetos com uma continuidade de tipos de associações. Este conjunto é caracterizado pela presença da função de pertinência, que atribui a cada objeto um grau de adesão que varia entre zero e um (ZADEH, 1965). É uma alternativa à teoria de conjuntos clássica que não carrega qualquer imprecisão. Dessa forma, a teoria dos conjunto *fuzzy* é uma tentativa de encontrar uma aproximação de grupos imprecisos (NOVAK et al., 1999).

Na teoria dos conjuntos fuzzy a decisão se um objeto tem uma propriedade é equivalente à questão de saber se é verdade que o objeto tem essa propriedade. No entanto, tal questão não pode ser inequivocamente respondida. Uma solução razoável consiste em utilizar algum tipo de escala cujos elementos expressaria vários graus de verdade (NOVAK et al., 1999).

Zadeh (1965) define os conjuntos *fuzzy* da seguinte forma: Seja  $X$  um espaço de pontos (objetos), com um elemento genérico de  $X$  denominado  $x$ . Portanto,  $X = \{x\}$ . Um conjunto *fuzzy*  $A$  em  $X$  é caracterizado por uma função de pertinência  $f_A(x)$  que associa a cada ponto em  $X$  um número real no intervalo  $[0, 1]$ , com o valor de  $f_A(x)$  em  $x$  representando o grau de pertinência de  $x$  em  $A$ . Sendo assim, quanto mais próximo de um o valor de  $f_A(x)$ , maior o grau de pertinência de  $x$  em  $A$ .

Enquanto que, na teoria de conjuntos clássica os elementos pertencem ou não pertencem a um conjunto, na teoria fuzzy o elemento pertencerá a um conjunto com um grau de pertinência (verdade). Em um conjunto clássico, pode assumir somente dois valores 0 ou 1, ou seja, 1 caso  $x$  pertença a  $A$  ou 0 caso  $x$  não pertença a  $A$  (MALVEZZI, 2010). A noção de união, intersecção, complementar, relação, convexidade, etc, é estendida para tais conjuntos, e várias propriedades dessas noções no contexto dos conjuntos fuzzy estão estabelecidas nesta teoria (ZADEH, 1965; ZADEH, 1975).

Por exemplo, consideremos o conjunto dos números reais “próximos de 3”. O número 3 pertence a esse conjunto? E o número 10000? De acordo com a teoria dos conjuntos fuzzy, pode-se dizer que ambos os números pertencem ao conjunto, porém com diferentes graus de pertinência, de acordo com a propriedade que caracteriza o conjunto (CORCOLL-SPINA, 2010).

No campo da bioinformática em que estão presentes problemas que apresentam imprecisões, a aplicação da teoria dos conjuntos *fuzzy* se insere como uma ferramenta poderosa. A aplicação dessa teoria tem sido vista em diversas áreas do conhecimento, como: educação, engenharia elétrica e biomedicina (MALVEZZI, 2010; USIDA, 2007; ORTEGA, 2001). Os sistemas biológicos, médicos e epidêmicos apresentam vários tipos de incertezas inerentes aos seus processos. No entanto, essas áreas ainda carecem de estruturas matemáticas que possibilitem o tratamento das incertezas não-estatísticas típicas de alguns desses sistemas. Assim, devido às suas características, os sistemas *fuzzy* se apresentam como uma teoria adequada para tratar alguns desses problemas (ORTEGA, 2001).

## 2.6 Agrupamento

Segundo Camilo e Silva (2009), o agrupamento visa identificar e aproximar os registros baseado na similaridade entre eles. Um agrupamento (ou *clustering*) é definido como uma coleção de registros similares entre si e que diferem dos outros registros nos demais grupos. O agrupamento consiste em agrupar objetos de dados com base apenas em informações encontradas nos próprios dados que descrevem os objetos e seus relacionamentos (TAN et al., 2005).

No agrupamento, é a distribuição e composição dos dados que irão determinar os membros do grupo (MANNING et al., 2008). Dessa forma, um mesmo conjunto de dados pode ser disposto em grupos de maneiras diferentes (Figura 4).

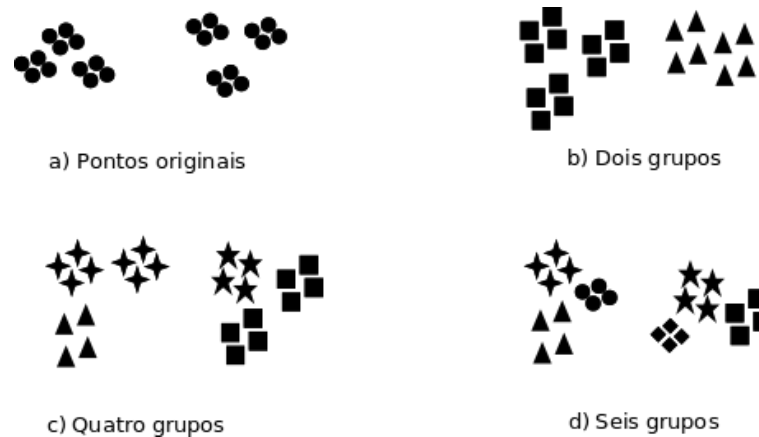


Figura 4 – Diferentes maneiras de agrupar o mesmo conjunto de dados.

Fonte: Adaptado de Tan et al. (2005).

De acordo com Camilo e Silva (2009), os métodos de agrupamento mais comuns são:

- **Métodos de Particionamento:** estes métodos tem como entradas um conjunto  $D$  de dados com  $n$  registros e  $k$  o número de agrupamentos desejados. A ideia por trás desta abordagem é organizar os objetos do conjunto em  $k$  grupos, tal que  $k \leq n$ . Os mais comuns são: *k-Means*, *k-Medoids* e *Fuzzy c-Means*.
- **Métodos Hierárquicos:** os métodos hierárquicos criam agrupamentos aglomerando ou dividindo os elementos do conjunto. Os métodos hierárquicos mais conhecidos são: Aglomerativos e Divisivos;
- **Outros métodos:** métodos baseados na densidade, métodos baseados em grade e métodos baseados em modelos.

### 2.6.1 Fuzzy c-Means

Em muitas aplicações práticas, um agrupamento clássico pode ser muito restritivo por causa da não completeza dos dados ou da imprecisão. Esse último pode surgir de diversas fontes (BEZDEK et al., 1984). A incerteza na classificação de elementos pode surgir também da sobreposição de várias classes, ou seja, pode existir características comuns entre as classes. Nas técnicas de classificação convencionais, comumente, assume-se que um dado pertença apenas a uma classe, o que nem sempre se verifica em domínios de dados reais, nos quais elementos pertencem a mais de uma classe, com diferentes graus (YONAMINE et al., 2002).

Os métodos científicos modernos fornecem aos pesquisadores ferramentas para sondar a natureza dos dados, reduzindo a complexidade e ajudando a expor a ordem oculta de padrões profundamente enterrados em dados (COX, 2005).

O algoritmo *Fuzzy c-Means* é um método de agrupamento que permite que um dado pertença a um ou mais de um grupo. Diferentemente dos métodos de agrupamentos clássicos em que os dados só pertencem a apenas um grupo. Este método foi desenvolvido por Dunn (1973) e melhorado por Bezdek et al. (1984), e é muito utilizado para reconhecimento de padrões. O método é baseado na minimização da seguinte função:

$$J_m(U, v) = \sum_{k=1}^N \sum_{i=1}^c (u_{ik})^m \|y_k - v_i\|_A^2 \quad (1)$$

Onde as variáveis são:

- $Y = \{y_1, y_2, \dots, y_N\} \subset \mathbb{R}^n$  = os dados
- $c$  = numeros de clusters em  $Y$ ;  $2 \leq c < n$
- $m$  = número real;  $1 \leq m < \infty$
- $U$  = c-partição fuzzy de  $Y$ ;  $U \in M_{fc}$
- $v = (v_1, v_2, \dots, v_c)$  = vetores de centros
- $v_i = (v_{i1}, v_{i2}, \dots, v_{in})$  = centro do cluster  $i$
- $\|\cdot\|_A$  = A-norma induzida em  $\mathbb{R}^n$
- $A$  = matriz (n x n) de pesos

O algoritmo *Fuzzy c-Means* (FMC) apresentado por Bezdek et al. (1984) é:

1. Inicialize a matrix  $U = [u_{ik}]$ ,  $U^{(0)}$
2. No passo  $k$ : calcule os vetores de centros  $v^{(k)}$ ,  $i = 1, 2, \dots, c$  com  $U^{(k)}$

$$v_i = \frac{\sum_{k=1}^N (u_{ik})^m y_k}{\sum_{k=1}^N (u_{ik})^m}; 1 \leq i \leq c$$

3. Atualize  $U^{(k)}$ ,  $U^{(k+1)}$

$$u_{ik} = \left( \sum_{j=1}^c \left( \frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)} \right)^{-1}; 1 \leq k \leq N; 1 \leq i \leq c$$

4. Se  $\|U^{(k+1)} - U^{(k)}\| < \epsilon$  então PARE; caso contrário retorne ao passo 2.

O *Fuzzy c-Means* é utilizado neste trabalho para realização dos agrupamentos e para determinação do número ideal de clusters.

### **3 Aspectos Metodológicos, Materiais e Recursos Computacionais**

#### **3.1 Aspectos Metodológicos**

Para o desenvolvimento da pesquisa fez-se a opção pelo método indutivo, o qual consiste numa interpretação generalizada da realidade partindo de algo particular. Considerado como um processo mental através do qual, se parte de dados particulares que sejam suficientes constatados e pode-se inferir uma verdade geral ou universal. O objetivo da indução é levar a uma conclusão generalizada com bases em premissas particulares (MARCONI; LAKATOS, 2003).

Marconi e Lakatos (2003), consideram três elementos fundamentais na indução: observação dos fenômenos, descoberta da relação entre os fenômenos e a generalização da relação. No primeiro é feita uma observação e análise dos fatos ou fenômenos com o objetivo de descobrir as causas dos mesmos. No segundo elemento, procura-se aproximar à realidade os fatos ou fenômenos com a finalidade de descobrir a relação entre eles através da comparação. Por fim, é feita a generalização da relação encontrada entre os fenômenos e fatos semelhantes.

Aplicando o método indutivo, realizou-se uma pesquisa exploratória na forma de estudo de caso, neste trabalho, sobre a Doença de Alzheimer. A pesquisa exploratória de acordo com Gil (2010) "tem como propósito proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a construir hipóteses". Por ser bastante flexível, este tipo de pesquisa, permite considerar os mais variados fatos e fenômenos acerca do objeto de estudo.

A pesquisa bibliográfica foi o marco inicial, pois objetiva-se um conhecimento prévio da situação em que se encontra o assunto na literatura da área.

As etapas do trabalho consistiram na busca, organização e pré-processamento dos dados de expressão gênica; escolha do algoritmo de agrupamento; desenvolvimento do método de agrupamento; aplicação do método; e por fim, a análise dos resultados. Nos resultados encontrados foram feitas análises de variância (CASTANHEIRA, 2010) para verificar se os genes selecionados estão com níveis de expressão aumentados ou reduzidos entre as amostras.



## 3.2 Materiais

Foi feita uma pesquisa por dados de *microarray* relacionados à Doença de Alzheimer na *National Center for Biotechnology Information* <sup>1</sup>. (NCBI), uma base de dados pública mantida pela *National Institutes of Health* <sup>2</sup> (NIH). A NCBI fornece mecanismos automatizados para armazenar e analisar conhecimento sobre a biologia molecular (NCBI, 2015). Também foram pesquisados dados de fatores de transcrição na base dados do HTRIdb (BOVOLenta et al., 2012) e genes já identificados com alguma relação com a Doença de Alzheimer na base de dados *PathCards* <sup>3</sup> (BELINKY et al., 2015).

Os dados de *microarray* utilizados nestes trabalhos se referem a 22 indivíduos (pós-morte) portadores da Doença de Alzheimer em vários estágios de severidade e 09 amostras de controle, e foram disponibilizados em um arquivo chamado *GDS810.soft*. Estes dados foram apresentados por (BLALOCK et al., 2004) no artigo: *Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses*. Dos 22 indivíduos, 07 se encontravam no estágio incipiente, 08 no estágio moderado e 07 no estágio grave.

A Tabela 1 apresenta uma amostra dos dados de *microarray* utilizados neste trabalho. Os dados são organizados no formato GEO Soft <sup>4</sup>.

Tabela 1 – Exemplo dos dados de *microarray*

ID_REF	IDENTIFIER	GSM21215	GSM21217	GSM21218	GSM21219
1007_s_at	DDR1	2735	3746	3317	5689
1053_at	RFC2	42.7	26.1	138.9	77.6
117_at	HSPA6	161.3	153.6	381.2	248.8
121_at	PAX8	1272.3	979.6	917.4	1291.9

Fonte: Blalock et al. (2004)

A primeira coluna apresenta um identificador de referência para o gene, genes repetidos tem identificadores de referência distintos. Na segunda coluna tem-se o nome do gene. As demais colunas representam os valores de modulação de cada gene para cada indivíduo.

Os dados de fatores de transcrição foram obtidos através do HTRIdb <sup>5</sup>. O HTRIdb (do inglês, Human Transcriptional Regulation Interaction Database) é um repositório

<sup>1</sup><http://www.ncbi.nlm.nih.gov>

<sup>2</sup><http://www.nih.gov>

<sup>3</sup><http://pathcards.genecards.org>

<sup>4</sup><http://www.ncbi.nlm.nih.gov/geo/info/soft.html>

<sup>5</sup><http://www.lbbc.ibb.unesp.br/htri>

que mantém dados obtidos através de experimentos de fatores de transcrição humano relacionados com os seus genes alvos (BOVOLENTA et al., 2012). O arquivo contendo as informações dos FT e os genes alvos está disponível nos repositórios do HTRIdb. Nesse arquivo o fator de transcrição é representado pelo gene que o origina.

A Tabela 2 apresenta um exemplo dos FT disponíveis no HTRIdb que são utilizados nesta pesquisa.

Tabela 2 – Exemplo dos dados de Fatores de Transcrição

OID	GENEID_TF	SYMBOL_TF	GENEID_TG	SYMBOL_TG	PUBMED_ID
111793	142	PARP1	675	BRCA2	18990703
111793	142	PARP1	675	BRCA2	18990703
114604	196	AHR	1543	CYP1A1	9698073
114602	196	AHR	1543	CYP1A1	17785579

Fonte: Bovolenta et al. (2012)

Para completar os dados necessários, foi realizada uma pesquisa de genes que possuem alguma relação já identificada com a Doença de Alzheimer na base de dados chamada *PathCards*.

*PathCards* é uma base de dados integrada de processos biológicos humanos e suas anotações. Estes processos são clusterizados dentro de *SuperPaths* baseados na similaridade de genes. Cada *PathCard* provém informações de um *SuperPath* no qual representa um ou mais processos biológicos humanos (BELINKY et al., 2015). Ao pesquisar pela Doença de Alzheimer, foram fornecidas informações de 51 genes que estão de alguma forma relacionados com a DA (Quadro 1).

Foram utilizadas informações da base dados AlzGene <sup>6</sup>, que segundo Bertram et al. (2007), é uma ferramenta poderosa para decifrar a genética da DA, além de servir como modelo para pesquisa de genes candidatos em outras doenças genéticas.

<sup>6</sup><http://www.alzgene.org>

Quadro 1 – Genes Relacionados à Doença de Alzheimer.

51 Genes								
A2M	BACE2	EGFR	FGFR2	INSR	NTRK1	PRKCB	PRKCQ	SRC
ADAM10	C1D	EPOR	FGFR3	LCN2	NTRK2	PRKCD	PRKCZ	TMED9
ADAM17	CAPN1	ERBB2	FGFR4	LIFR	NTRK3	PRKCE	PRKD1	TPPP
APOE	CDK5	ERBB3	GSK3B	LRP1	PDGFRA	PRKCG	PRKD3	
APP	CDK5R1	ERBB4	HRES1	MAPK1	PDGFRB	PRKCH	PSEN1	
BACE1	DCTN5	FGFR1	IGF1R	MAPT	PRKCA	PRKCI	PSEN2	

Fonte: WEIZMANN INSTITUTE OF SCIENCE (2015)

### 3.3 Recursos Computacionais

O algoritmo *Fuzzy c-Means* utilizado neste trabalho foi desenvolvido na linguagem de programação R e está disponível através do pacote Mfuzz (KUMAR; FUTSCHIK, 2007) do Biocondutor. O Biocondutor é um pacote *open source* que fornece ferramentas para análise e compreensão de dados genômicos e da biologia molecular (HUBER et al., 2015).

Foram utilizados três pacotes do Biocondutor: o pacote GEOquery (DAVIS; MELTZER, 2007), para fazer a leitura dos dados de expressão diferencial que estão no formato SOFT; o pacote Mfuzz, para gerar os *clusters*; e o pacote Limma (RITCHIE et al., 2015). Foi utilizada a função *vennDiagram* do pacote Limma para geração dos diagramas de Venn. No entanto, esta função limita para o número ( $n$ ) máximo de cinco conjuntos a geração dos diagramas. Isto se deve à quantidade de combinações feitas entre os conjuntos que é de  $n^2$ . Dessa forma, o número máximo de combinações é  $2^5 = 32$ . Números acima de cinco conjuntos aumentam a quantidade de combinações, o que implica no prejuízo da visualização dos conjuntos.

Para manipulação de arquivos foram desenvolvidos *scripts* tanto na linguagem PHP<sup>7</sup> (*Hypertext Preprocessor*) quanto na linguagem R<sup>8</sup>.

O computador utilizado no processamento dos dados foi um Notebook HP com processador Core i3 2.40 GHz Intel, memória RAM de 4 GB e HD de 320 GB, Sistema Operacional *Linux Mint Debian Edition 2*.

<sup>7</sup><http://www.php.net>

<sup>8</sup><http://www.r-project.org>

## 4 Método Proposto e Resultados

Com o objetivo de agrupar os genes considerando as suas modulações em uma determinada doença, neste trabalho foi proposto um método de agrupamento utilizando fatores de transcrição e expressão diferencial. Neste capítulo são apresentadas as etapas do método e suas aplicações nos dados da Doença de Alzheimer. A justificativa para a escolha desta patologia foi apresentada na seção 1.3, e os dados sobre a doença são explicados na seção 3.2.

### 4.1 Método para Agrupamento de Genes Utilizando Fatores de Transcrição e Expressão Diferencial

O método proposto neste trabalho foi desenvolvido baseado na utilização do algoritmo de agrupamento *Fuzzy c-Means* com informações da expressão dos genes obtidos através de experimentos de *microarray*, dados de genes identificados com alguma relação com a patologia escolhida e dados de fatores de transcrição (vide seção 3.2). Este método possui 07 etapas em um fluxo sequencial (Figura 5).

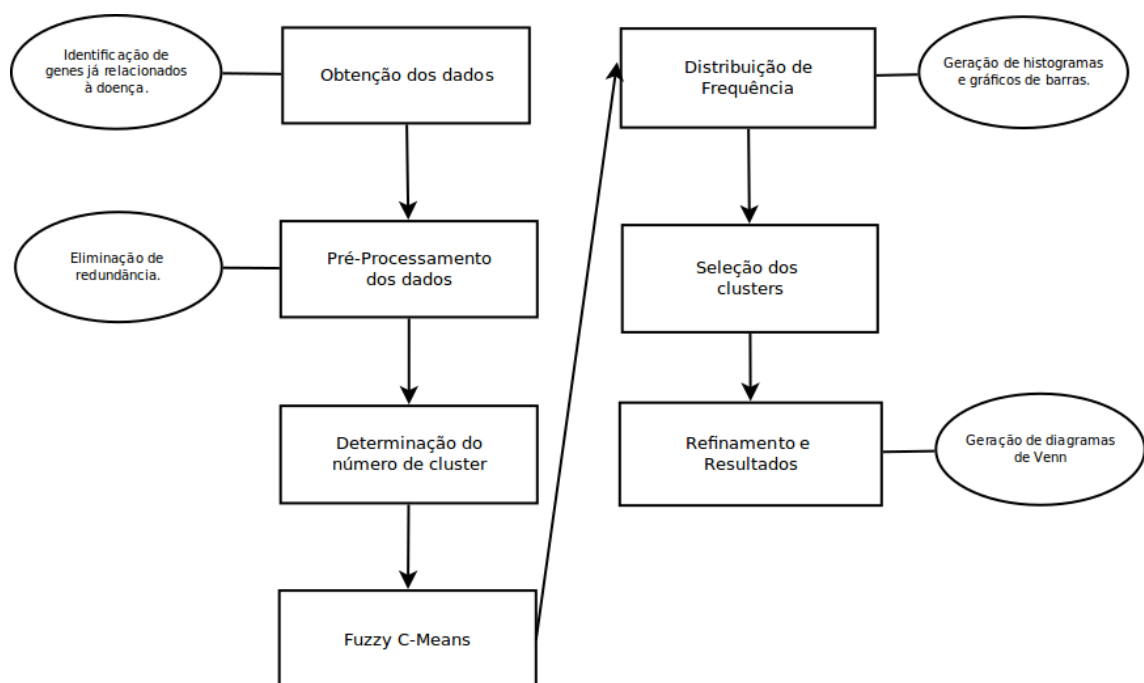


Figura 5 – Fluxo do Método Proposto.

Fonte: Autoria própria.

As etapas do método são:

- **Obtenção dos dados:** nesta etapa foram pesquisados os dados de *microarray* relacionados à patologia escolhida, neste caso a Doença de Alzheimer, bem como os genes já conhecidos para a doença e os dados de fatores de transcrição;
- **Pré-processamento dos dados:** foi feita a eliminação da redundância nos dados de *microarray*, bem como a separação dos genes e fatores de transcrição que se tem conhecimento da sua relação com a doença;
- **Determinação do número de *clusters*:** nesta etapa foi determinada a quantidade ideal de *clusters* para o agrupamento e posterior aplicação da ferramenta *Fuzzy C-Means*;
- ***Fuzzy C-Means*:** nesta etapa foi feita a aplicação do algoritmo *Fuzzy C-Means* no agrupamento dos genes com o número de *clusters* determinado na etapa anterior;
- **Distribuição de Frequências:** foi feita a distribuição de frequências e foram gerados histogramas com a identificação dos genes e dos fatores de transcrição relacionados à patologia;
- **Seleção dos *clusters*:** como no agrupamento feito pelo *Fuzzy c-Means* todos os genes pertencem a todos os *clusters*, com certo grau de pertinência, foi necessário escolher quais dos *clusters* revelaram a maior concentração de genes com os fatores de transcrição relacionados à patologia;
- **Refinamento e resultados:** foi realizado um refinamento nos dados dos *clusters* selecionados e apresentados os resultados.

## 4.2 Estudo de Caso: Aplicação do Método em Dados relacionados à Doença de Alzheimer

Nesta seção é apresentada a aplicação do método para agrupamento de genes utilizando fatores de transcrição e expressão diferencial. O método é aplicado aos dados relacionados à DA. Nas subseções seguintes são descritas as etapas e os resultados.

### 4.2.1 Obtenção dos Dados

A primeira etapa do método consiste na obtenção dos dados para realização do agrupamento. Maiores detalhes sobre a obtenção dos dados pode ser obtida na seção 3.2 (Materiais). Os dados utilizados na aplicação do método foram:

- Dados de expressão gênica de 22.283 genes, obtidos através de experimentos com *microarray* na DA. Este conjunto de dados possuem o nível de expressão dos genes de 31 amostras de indivíduos (pós-morte), dentre os quais 09 foram amostras de controle, 07 no estágio inicial, 08 no estágio moderado e 07 no estágio grave da doença;
- 285 fatores de transcrição e seus genes alvos;
- 51 genes com relação à DA já estabelecida.

#### 4.2.2 Pré-processamento dos Dados

A segunda etapa foi a de pré-processamento e preparação dos dados para as etapas seguintes. Foram executadas as seguintes ações:

- Remoção de 68 genes que possuíam ID\_RF iniciados em "AFFX-" do arquivo *GDS810.soft*, por serem dados de controle usados durante a preparação do *microarray*. A permanência destes genes poderiam interferir no agrupamento;
- Eliminação de redundância de genes no arquivo *GDS810.soft*, considerando o menor, a mediana e o maior valor de expressão dos genes redundantes para cada amostra. Dessa forma, o trabalho passou a ser realizado com três conjuntos de dados de *microarray*, e ao invés dos 22.283 genes iniciais, com a retirada dos genes redundantes, cada conjunto de dados passou a ter 14.092 genes. Estes conjuntos de dados são referenciados neste trabalho como: **conjunto com o menor valor de expressão, conjunto com a mediana do valor de expressão e conjunto com o maior valor de expressão**. Os genes que não apresentavam redundância foram mantidos com os valores de expressão original, sendo inserido como mais um dos elementos em todos os três conjuntos citados. A eliminação da redundância dos genes também foi necessária, pois a permanência dos genes redundantes poderia interferir no agrupamento;
- Pesquisa da disponibilidade dos genes relacionados à DA no conjunto de dados de *microarray*. Dos 51 genes conhecidos, 50 estavam disponíveis no conjunto de dados;
- A mesma pesquisa da disponibilidade foi realizada com os fatores de transcrição para identificar os que estavam presentes. Dos 285 FT conhecidos, 258 foram identificados;
- Foi criado um arquivo onde se relaciona os fatores de transcrição com os genes envolvidos na DA. Foram identificados 32 FT que possuem relação com os genes envolvidos na Doença de Alzheimer.

### 4.2.3 Determinação do Número de *Clusters*

Um dos mais difíceis problemas no agrupamento é identificar o número de *clusters* para o conjunto de dados (SUGAR; JAMES, 2003). Na maioria dos algoritmos de agrupamento temos que especificar o número ideal de grupos antes da sua execução (SUBBALAKSHMI et al., 2015). Muitas abordagens para este problema têm sido sugeridas ao longo dos anos. No entanto, muitas dessas abordagens são para problemas específicos e que requerem muitos parâmetros (SUGAR; JAMES, 2003).

Nesta pesquisa, a escolha do numero ideal de *clusters* para a aplicação do *Fuzzy C-Means* foi feita de forma experimental. Para cada conjunto de dados (menor, mediana e maior valor de expressão) foi aplicado o algoritmo *Fuzzy C-Means* com  $k$  clusters, no intervalo  $[2, 70]$ . A cada saída do algoritmo foram armazenados os valores de pertinência dos genes em cada *cluster*. Devido à necessidade de determinar um número que representasse todo o conjunto de dados, foi calculado o desvio padrão da pertinência de cada gene nos *clusters*.

Segundo Castanheira (2010), o desvio padrão é uma medida de dispersão que considera os desvios em relação à média, dessa forma verificando o quanto os valores estão dispersos em relação à média. Foram feitos os cálculos do desvio padrão amostral obedecendo a Fórmula 2. O desvio padrão amostral é representado neste trabalho pela letra  $s$ , onde

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}. \quad (2)$$

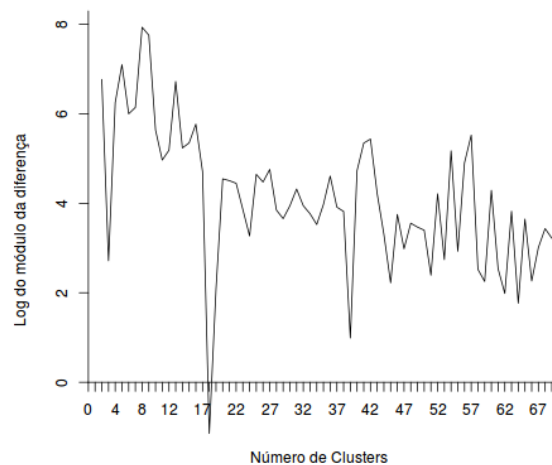
Depois de calculado o desvio padrão da pertinência de cada gene nos *clusters*, foi feita a soma desses desvios, para todos os genes. Em seguida, foi calculada a diferença dessas somas que foram obtidas para um número de *cluster* posterior e o seu anterior. E, por fim, foi aplicada a função logarítmica ( $\log$ ) na base 10 no módulo das diferenças, para normalização dos valores, como exemplificado na Tabela 3.

Tabela 3 – Exemplo dos cálculos para determinação do número ideal de *cluster* para o conjunto de dados considerando o maior valor de expressão

Número de Clusters	$\sum_1^{14092} s$	Diferença	$\log$ do módulo da dif.
2	5733,7431473756	871,0268500356	2,9400315427
3	6604,7699974112	-15,1065322323	1,1791647817
4	6589,6634651789	-518,7678191274	2,7149730275
5	6070,8956460515	-1210,556111021	3,0829849244
$\vdots$	$\vdots$	$\vdots$	$\vdots$
67	1219,423997392	-20,3193506	1,3079098239
68	1199,1046467919	31,0280793745	1,4917548938
69	1230,1327261664	-25,0271858945	1,3984120194
70	1205,1055402719	-25,0271858945	1,3984120194

Fonte: Autoria própria.

Foram plotados gráficos com os valores obtidos aplicando a função logarítmica no módulo das diferenças. A utilização do  $\log$  deve-se à necessidade de normalização dos dados. Ao analisar o gráfico do  $\log$  do módulo das diferenças para o conjunto com maior valor de expressão (Figura 6), pode ser percebido que nele é apresentado um mínimo global localizado no número de *clusters* 18.

Figura 6 –  $\log$  do módulo das diferenças para o conjunto de dados com maior valor de expressão.

Fonte: Autoria própria.



O mesmo ocorre com o gráfico do *log* do módulo das diferenças para o conjunto com a mediana do valor de expressão (Figura 8). Há um mínimo global localizado no número de *clusters* 53.

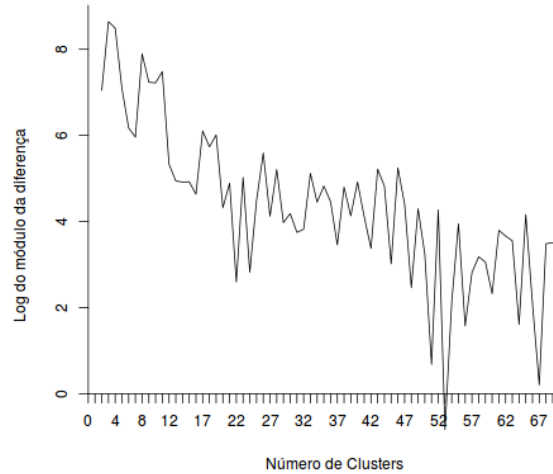


Figura 7 – *log* do módulo das diferenças para o conjunto com a mediana do valor de expressão.

Fonte: Autoria própria.

Para o conjunto considerando o menor valor de expressão também foi gerado um gráfico do *log* do módulo das diferenças (Figura 7). Nele também é apresentado um mínimo global, neste caso, para o número de clusters 54.

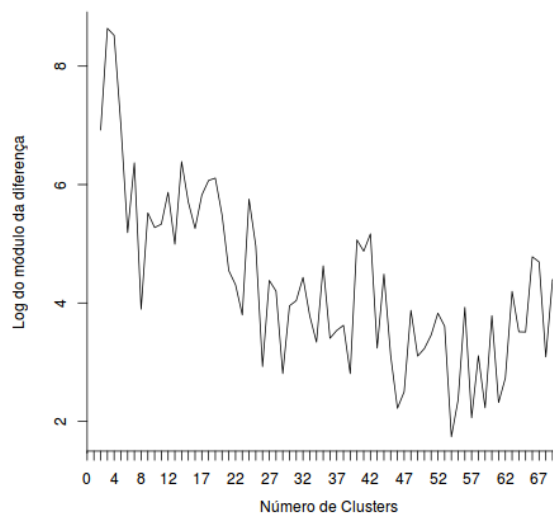


Figura 8 – *log* do módulo das diferenças para o conjunto com o menor valor de expressão.

Fonte: Autoria própria.

Uma vez calculadas as somas dos desvios padrão e aplicado o *log* no módulo das diferenças dessas somas, ficou evidenciado que 18, 53 e 54, dentre os números no intervalo  $[2, 70]$ , foram os números inteiros que expressam a quantidade de *clusters* que minimiza a variação da soma do desvio padrão, portanto, foram usados para o agrupamento nos conjuntos de dados. Como pode ser observado, comparando-se os três gráficos, a quantidade de *clusters* adotada para a mediana é bem próximo da quantidade de *clusters* adotada para o conjunto com o menor valor de expressão. Isto justifica-se pelo fato de que no conjunto de dados a mediana do valor de expressão estar mais próxima dos menores valores do que dos maiores. Portanto, os números mais representativos para o conjunto com o maior, mediana e o menor valor de expressão são respectivamente 18, 53 e 54.

#### 4.2.4 Fuzzy C-Means (FCM)

Como explicado anteriormente, o *Fuzzy C-Means* foi utilizado para identificação do número ideal de *clusters* em cada conjunto de dados. O algoritmo FCM (DUNN, 1973; BEZDEK et al., 1984) utilizado neste trabalho foi implementado na linguagem R e está disponível através da função *mfuzz* do pacote *Mfuzz* (KUMAR; FUTSCHIK, 2007).

Segundo Kumar e Futschik (2007), para o agrupamento *fuzzy*, os centros dos *clusters* resultam da soma ponderada de todos os membros do *cluster*, e mostram o padrão de expressão global de cada *cluster*. O valor de pertinência indica o quanto um determinado gene está representado pelo *cluster*. O agrupamento é implementado na função *mfuzz* e utiliza o algoritmo *fuzzy c-means* (do pacote 'e1071').

Ainda de acordo com Kumar e Futschik (2007), para aplicação do algoritmo, é necessário informar, além do conjunto de dados, o número de *clusters* e o parâmetro de fuzzificação  $m$ . Neste trabalho, foram utilizados os números 18, 53 e 54, que representam a quantidade de *clusters* para os conjuntos de dados com o maior, mediana e o menor valor de expressão. Para o parâmetro de fuzzificação foi utilizado o valor padrão apresentado pelo pacote, ou seja,  $m = 1,25$ .

Durante sua execução apresentou uma boa performance para os conjuntos de dados (14092 genes de 32 amostras em cada conjunto). Ao aplicar o FCM nos dados, foram gerados arquivos contendo o grau de pertinência dos genes, nos *clusters*. As próximas etapas consideram essas informações como fundamentais para análise dos grupos.

### 4.2.5 Distribuição de Frequências

Para análise dos *clusters*, foram feitas as distribuições de frequências utilizando o grau de pertinência dos genes em cada *cluster*. Segundo Correa (2003), não existe um método exato para determinar o número de classes de uma distribuição de frequência.

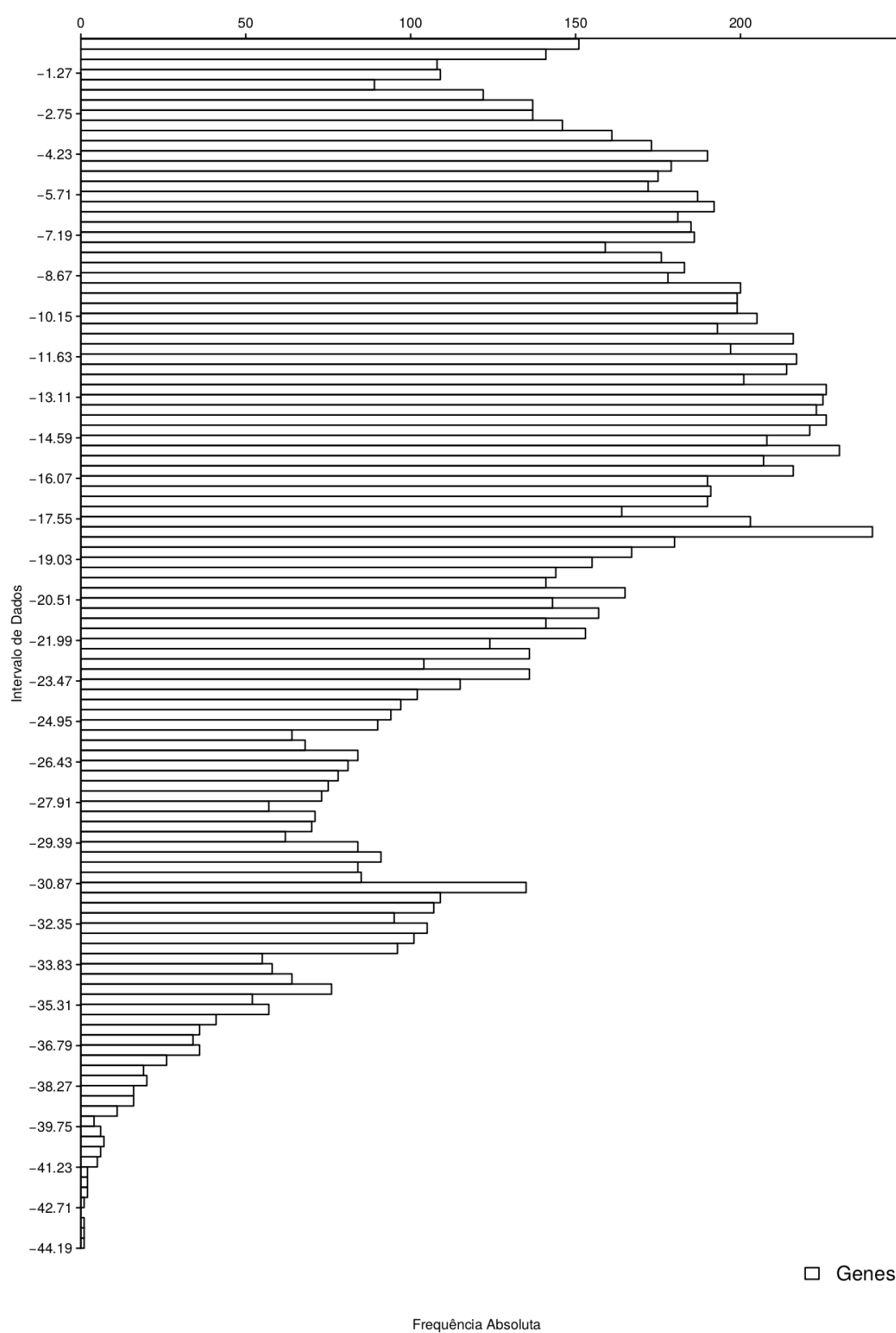
No entanto, há algumas regras adotadas, como a regra da raiz quadrada usada neste trabalho (CORREA, 2003). Dessa forma, temos o número de classes  $c$  sendo igual a  $\sqrt{n}$ , neste caso,  $n$  é o número de elementos no conjunto de dados. Ao aplicar a raiz quadrada nos 14092 genes, foi obtido como resultado  $c \cong 118,709$ , que foi aproximado para o inteiro 119, sendo este o número de classes utilizadas para construção dos histogramas.

Para a análise da distribuição de frequências foram gerados histogramas para cada um dos *clusters*. O histograma da distribuição da frequência dos genes no *cluster* 01 do conjunto com menor valor de expressão pode ser observado na Figura 9.

Na distribuição de frequência foi utilizada a função *log* na base 2 para a normalização do grau de pertinência .

Foram gerados os histogramas com a identificação, nas classes, dos genes envolvidos na DA, e dos FT relacionados aos genes envolvidos na DA. O histograma do *cluster* 18, do conjunto com o maior valor de expressão, identificando os genes envolvidos na DA, pode ser observado na Figura 10.

Os histogramas foram marcadas com linhas horizontais. A linha mais acima (vermelha) separa as classes com aproximadamente 20% dos genes com maior grau de pertinência (MPG), e a linha mais abaixo (azul) separa as classes com aproximadamente 20% dos genes com o menor grau de pertinência (mPG).

Figura 9 – Histograma do *cluster* 01 do conjunto com o menor valor de expressão.

Fonte: Autoria própria.

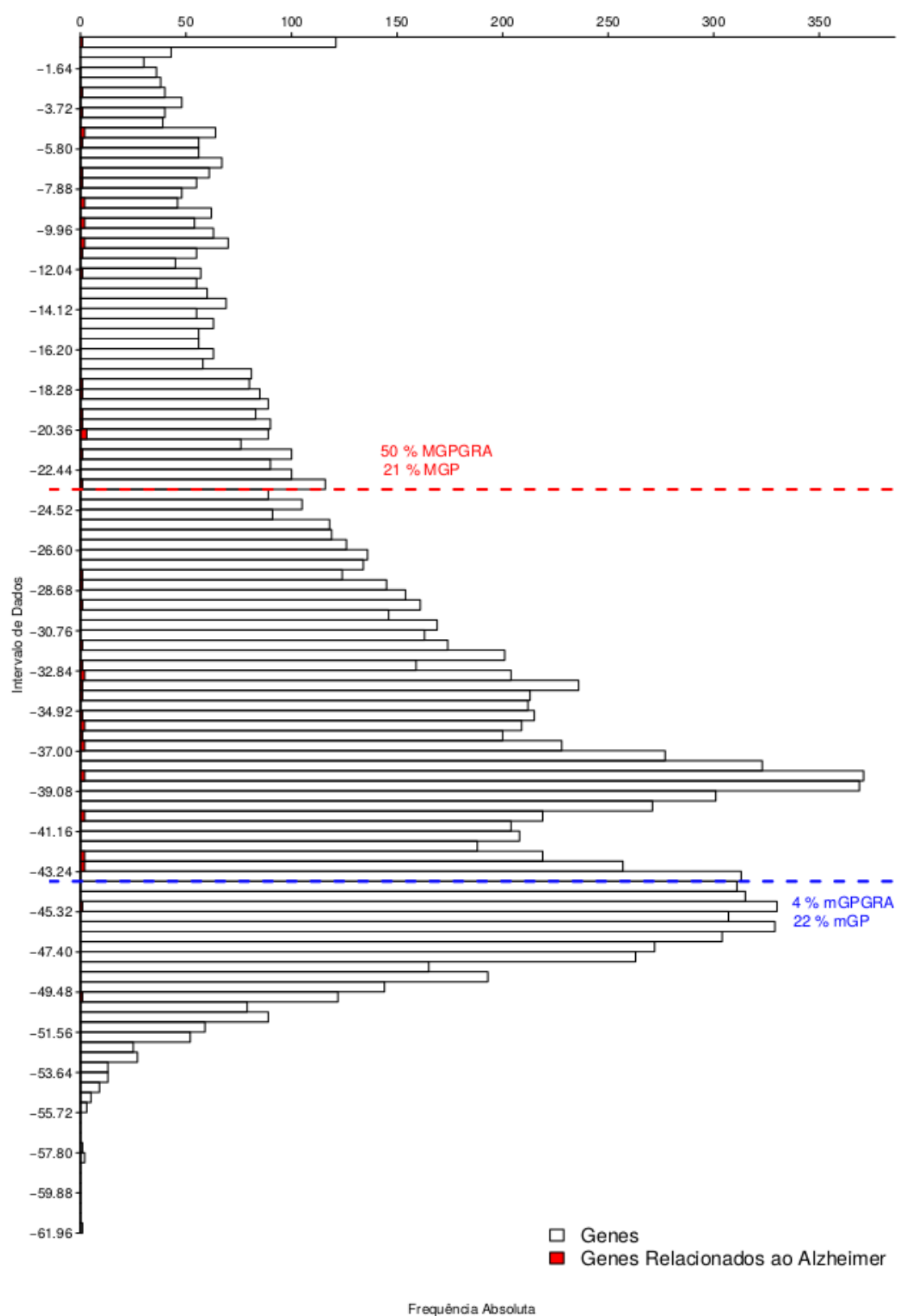


Figura 10 – Histograma do *cluster* 18 para o conjunto de dados com o maior valor de expressão.

Fonte: Autoria própria.

O histograma do *cluster* 18, para o conjunto com o maior valor de expressão identificando os fatores de transcrição dos genes envolvidos na DA, é apresentado na Figura 11.

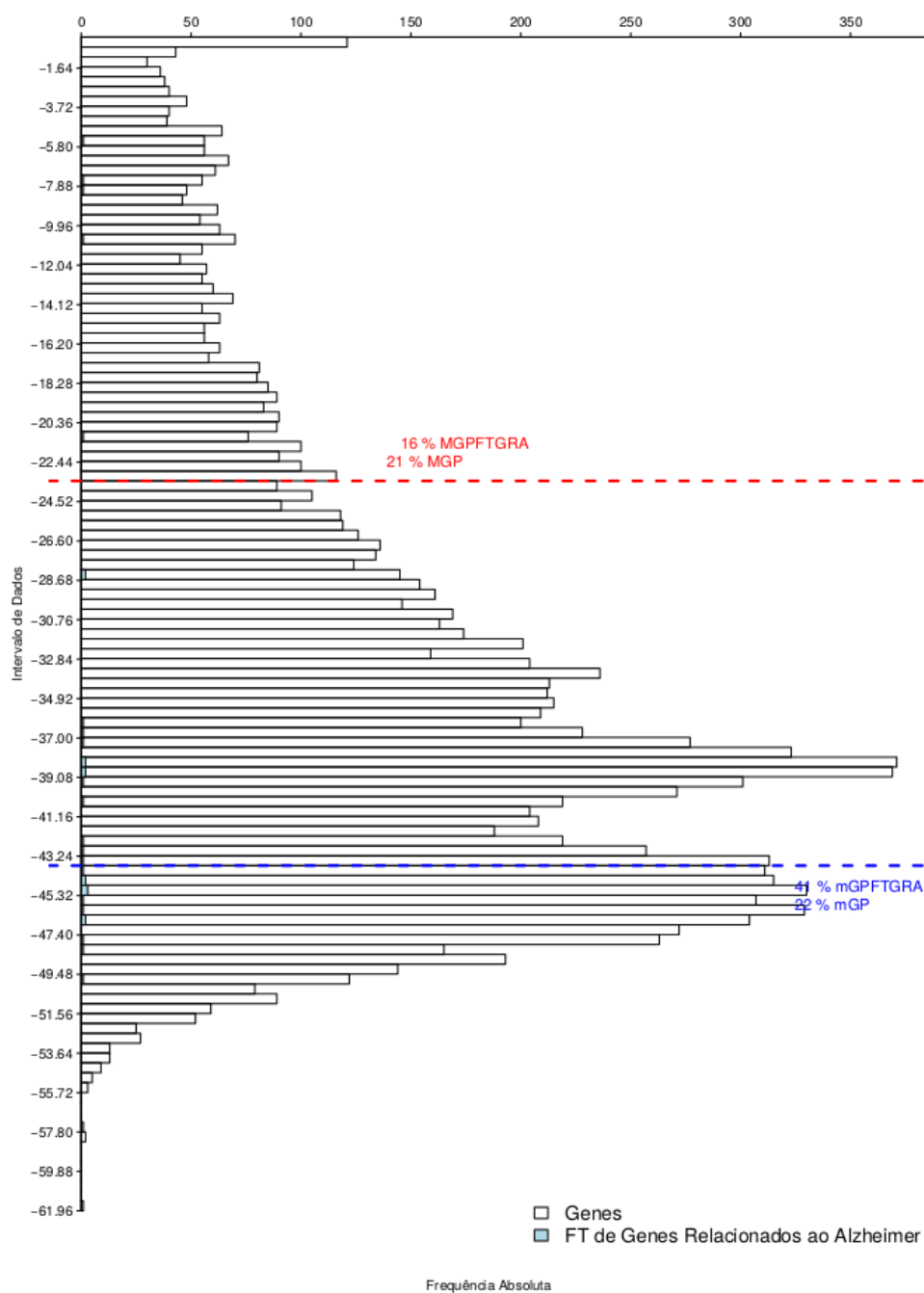


Figura 11 – Histograma do *cluster* 18 identificando os FT relacionados aos genes envolvidos na DA.

Fonte: Autoria própria.

### 4.2.6 Seleção dos *Clusters*

Feita a distribuição de frequências e a geração dos histogramas para os três conjuntos de dados, foi contabilizado o percentual da presença de genes e fatores de transcrição relacionados à DA; na faixa inferior aproximada de 20%, para os genes de menor pertinência, e na faixa superior, aproximada de 20%, para os de maior pertinência, em cada *cluster*.

Para a escolha do percentual de 20%, utilizamos o Princípio de Pareto, também conhecido com Princípio 80/20. Segundo Koch (1998), o Princípio de Pareto afirma que uma minoria de causas, entradas ou esforço, geralmente levam a maioria dos resultados, produtos ou recompensas. Por exemplo, oitenta por cento do que é conseguido em um trabalho vem de vinte por cento do tempo gasto.

A Tabela 4 apresenta o percentual da presença de genes e FT em aproximadamente 20% de maior e menor pertinência nos *clusters* do conjunto com maior valor de expressão.

Tabela 4 – Percentual da presença de genes e FT em aproximadamente 20% de maior e menor pertinência nos *clusters* do conjunto com maior valor de expressão

Cluster	Genes Relacionados AD		FT Relacionados à AD		Grupo
	Menor	Maior	Menor	Maior	
1	4,00	18,00	37,50	12,50	
2	4,00	50,00	37,50	15,63	
3	4,00	50,00	37,50	15,63	
4	4,00	50,00	40,63	15,63	
5	4,00	50,00	37,50	15,63	
6	4,00	50,00	37,50	15,63	
7	4,00	50,00	37,50	15,63	
8	<b>4,00</b>	<b>50,00</b>	<b>43,75</b>	<b>15,63</b>	<b>03</b>
9	<b>4,00</b>	<b>50,00</b>	<b>43,75</b>	<b>15,63</b>	<b>03</b>
10	4,00	36,00	37,50	15,63	
11	8,00	20,00	25,00	21,88	
12	4,00	50,00	37,50	15,63	
13	4,00	26,00	34,38	9,38	
14	4,00	50,00	40,63	15,63	
15	4,00	44,00	37,50	12,50	
16	<b>34,00</b>	<b>16,00</b>	<b>21,88</b>	<b>50,00</b>	<b>01</b>
17	4,00	50,00	37,50	15,63	
18	4,00	50,00	40,63	15,63	

Fonte: Autoria própria.

Utilizando os dados presentes na Tabela 4, foram elaborados dois gráficos de barras para análise dos percentuais apresentados. O primeiro gráfico apresenta o percentual de presença de genes para a faixa de aproximadamente 20% de menor pertinência nos *clusters*, para o conjunto com maior valor de expressão (Figura 12).

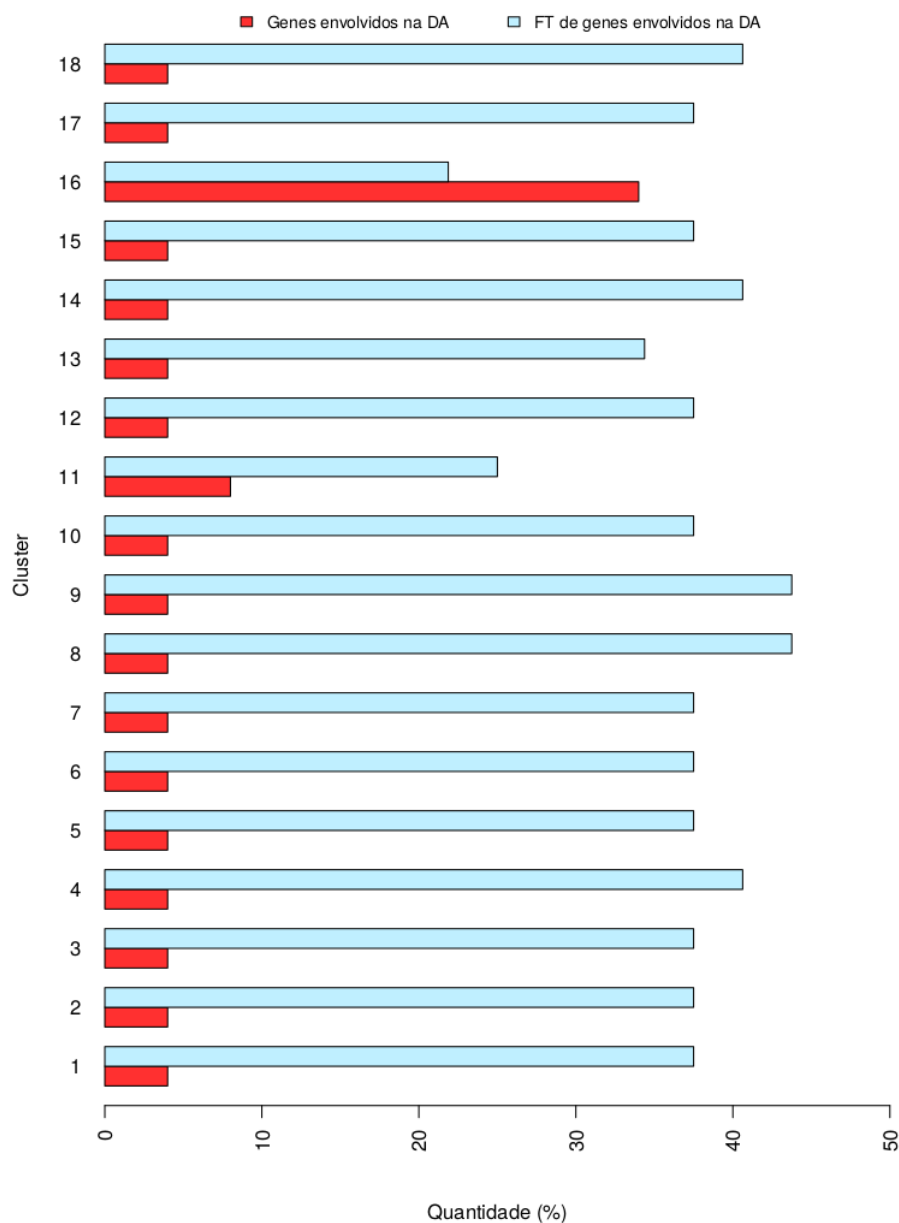


Figura 12 – Percentual de genes e FT na faixa aproximada de 20% de menor pertinência nos *cluster*.

Fonte: Autoria própria.

No segundo gráfico, é apresentado o percentual para a faixa de aproximadamente 20% de maior pertinência nos *clusters*, para o conjunto com o maior valor de expressão (Figura 13).



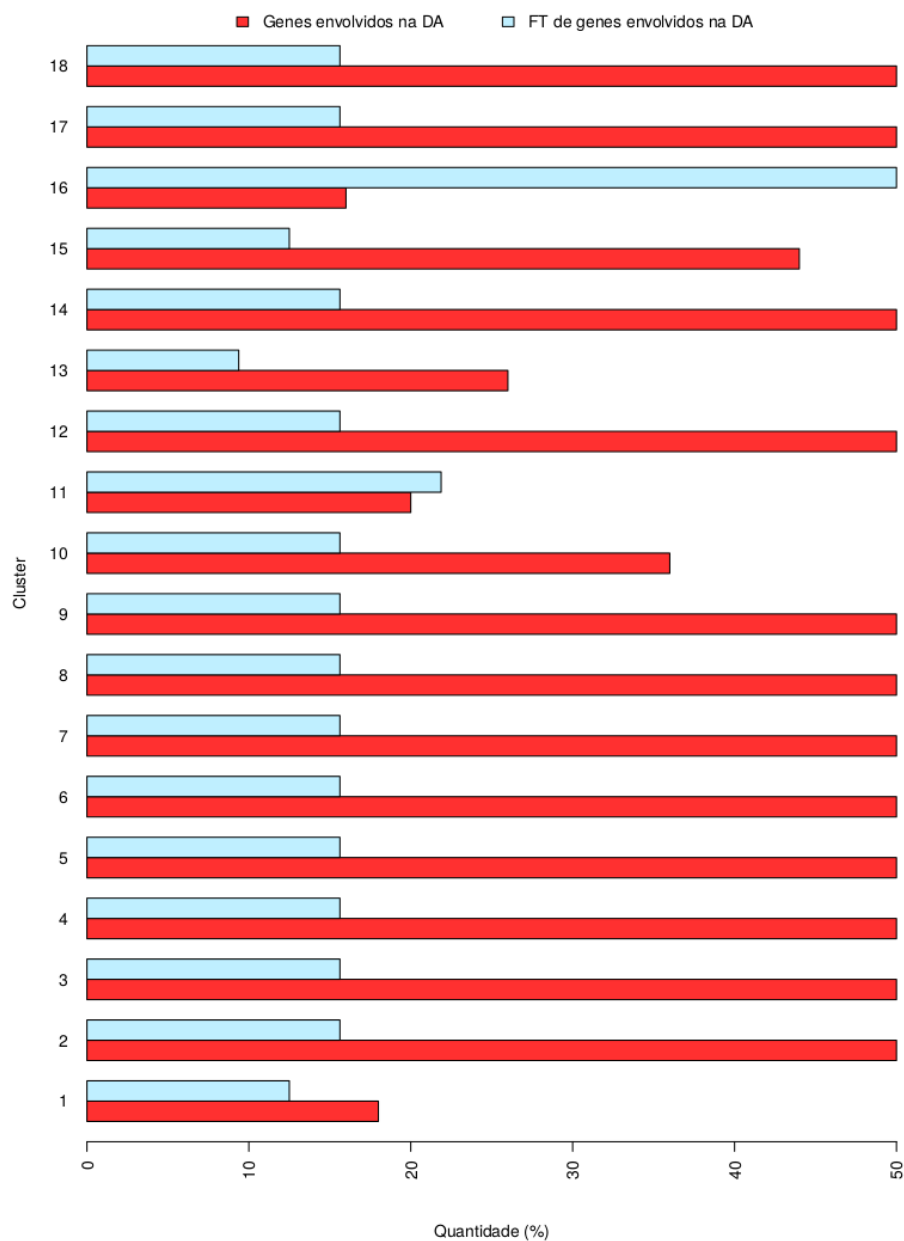


Figura 13 – Percentual de genes e FT na faixa aproximada de 20% de maior pertinência nos *cluster*.

Fonte: Autoria própria.

A mesma estratégia foi utilizada para o conjunto considerando a mediana e o menor valor de expressão dos genes. A Tabela 5 apresenta o percentual de genes e FT para os 53 *clusters* do conjunto com a mediana do valor de expressão.

Tabela 5 – Percentual da presença de genes e FT em aproximadamente 20% de maior e menor pertinência nos *clusters* do conjunto com a mediana do valor de expressão

Cluster	(%) Genes Relacionados AD		(%)FT Relacionados à AD		Grupo
	Menor	Maior	Menor	Maior	
1	8,00	34,00	43,75	9,38	
2	8,00	26,00	37,50	9,38	
3	8,00	22,00	43,75	12,50	
4	8,00	28,00	43,75	9,38	
5	<b>16,00</b>	<b>36,00</b>	<b>46,88</b>	<b>12,50</b>	<b>02 e 03</b>
6	10,00	34,00	46,88	12,50	
7	<b>30,00</b>	<b>12,00</b>	<b>15,63</b>	<b>40,63</b>	<b>01 e 04</b>
8	10,00	32,00	43,75	12,50	
9	<b>12,00</b>	<b>36,00</b>	<b>46,88</b>	<b>12,50</b>	<b>02</b>
10	14,00	34,00	46,88	12,50	
11	8,00	28,00	43,75	12,50	
12	<b>14,00</b>	<b>36,00</b>	<b>46,88</b>	<b>12,50</b>	<b>02</b>
13	10,00	32,00	43,75	12,50	
14	12,00	32,00	46,88	12,50	
15	16,00	34,00	46,88	12,50	
16	8,00	28,00	43,75	6,25	
17	<b>30,00</b>	<b>12,00</b>	<b>12,50</b>	<b>43,75</b>	<b>01</b>
18	<b>20,00</b>	<b>16,00</b>	<b>25,00</b>	<b>21,88</b>	<b>01</b>
19	10,00	34,00	46,88	12,50	
20	10,00	34,00	46,88	12,50	
21	10,00	24,00	46,88	9,38	
22	8,00	30,00	43,75	9,38	
23	10,00	32,00	43,75	12,50	
24	12,00	34,00	46,88	12,50	
25	16,00	36,00	46,88	12,50	
26	10,00	34,00	46,88	12,50	
27	10,00	34,00	46,88	12,50	
28	18,00	22,00	31,25	9,38	
29	10,00	32,00	46,88	12,50	
30	14,00	34,00	46,88	12,50	
31	10,00	32,00	43,75	12,50	
32	8,00	24,00	43,75	12,50	
33	8,00	24,00	37,50	12,50	
34	10,00	34,00	46,88	12,50	
35	16,00	36,00	46,88	12,50	
36	10,00	32,00	46,88	12,50	
37	10,00	32,00	43,75	9,38	
38	10,00	34,00	46,88	12,50	
39	10,00	34,00	46,88	12,50	
40	16,00	36,00	46,88	12,50	

Continua na página seguinte.

Tabela 5 Continuação da página anterior.

Cluster	(%) Genes Relacionados AD		(%)FT Relacionados à AD		Grupo
	Menor	Maior	Menor	Maior	
41	12,00	34,00	46,88	12,50	
42	10,00	32,00	46,88	12,50	
43	8,00	24,00	43,75	12,50	
44	8,00	30,00	40,63	9,38	
45	16,00	32,00	46,88	12,50	
46	16,00	36,00	46,88	12,50	
47	8,00	26,00	43,75	12,50	
48	10,00	28,00	46,88	9,38	
49	16,00	34,00	46,88	12,50	
50	10,00	32,00	46,88	12,50	
51	10,00	30,00	43,75	12,50	
52	10,00	32,00	46,88	12,50	
53	10,00	32,00	46,88	12,50	
					Concluída

Fonte: Autoria própria.

O gráfico de barras com o percentual para a faixa de aproximadamente 20% de menor pertinência nos *clusters* é apresentado na Figura 14, para o conjunto com a mediana do valor de expressão.

Para a faixa de aproximadamente 20% de maior pertinência nos *clusters*, o gráfico de barras é apresentado na Figura 15, para o conjunto com a mediana do valor de expressão.

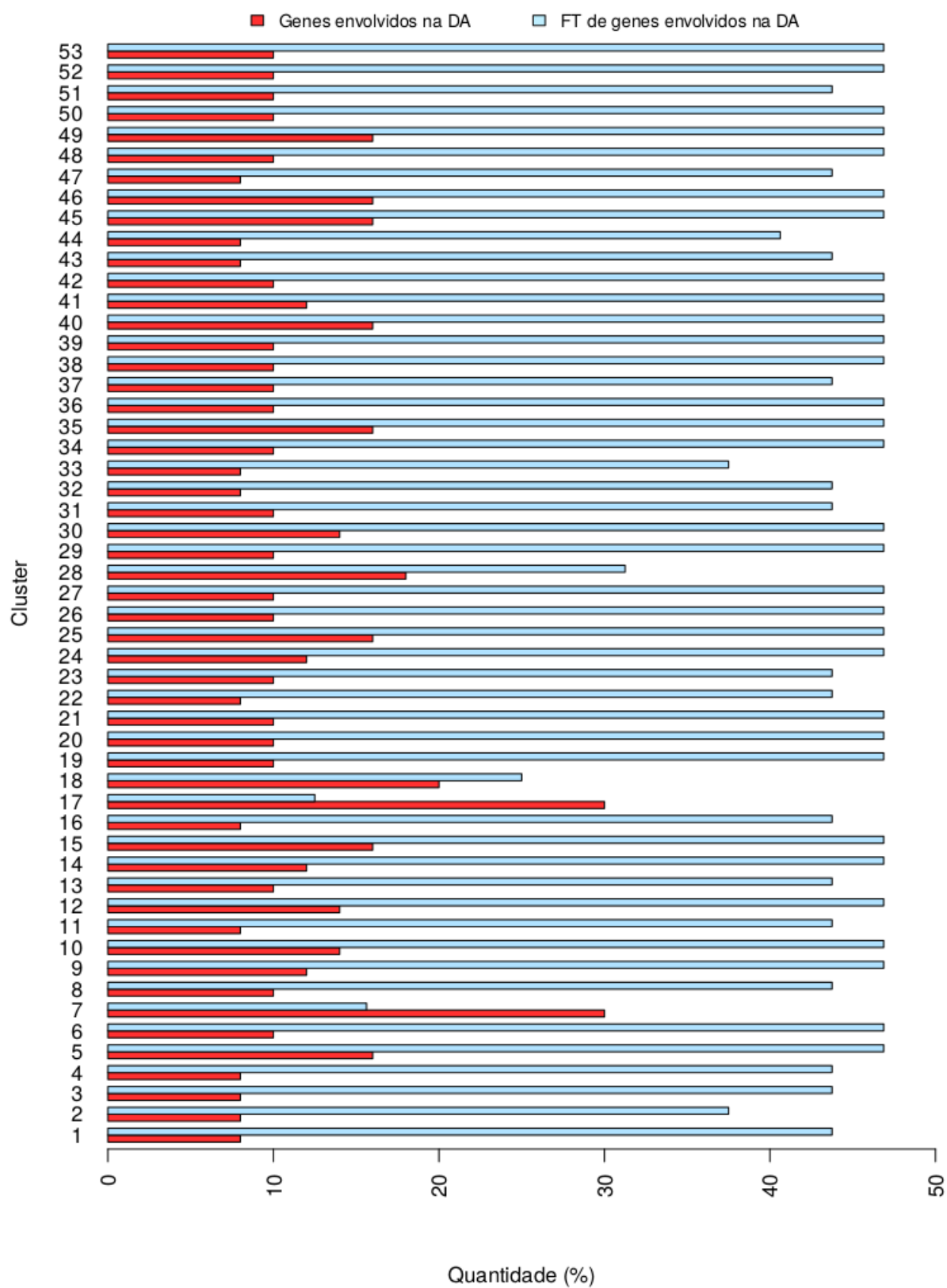


Figura 14 – Percentual de genes e FT na faixa aproximada de 20% de menor pertinência nos *cluster*.

Fonte: Autoria própria.

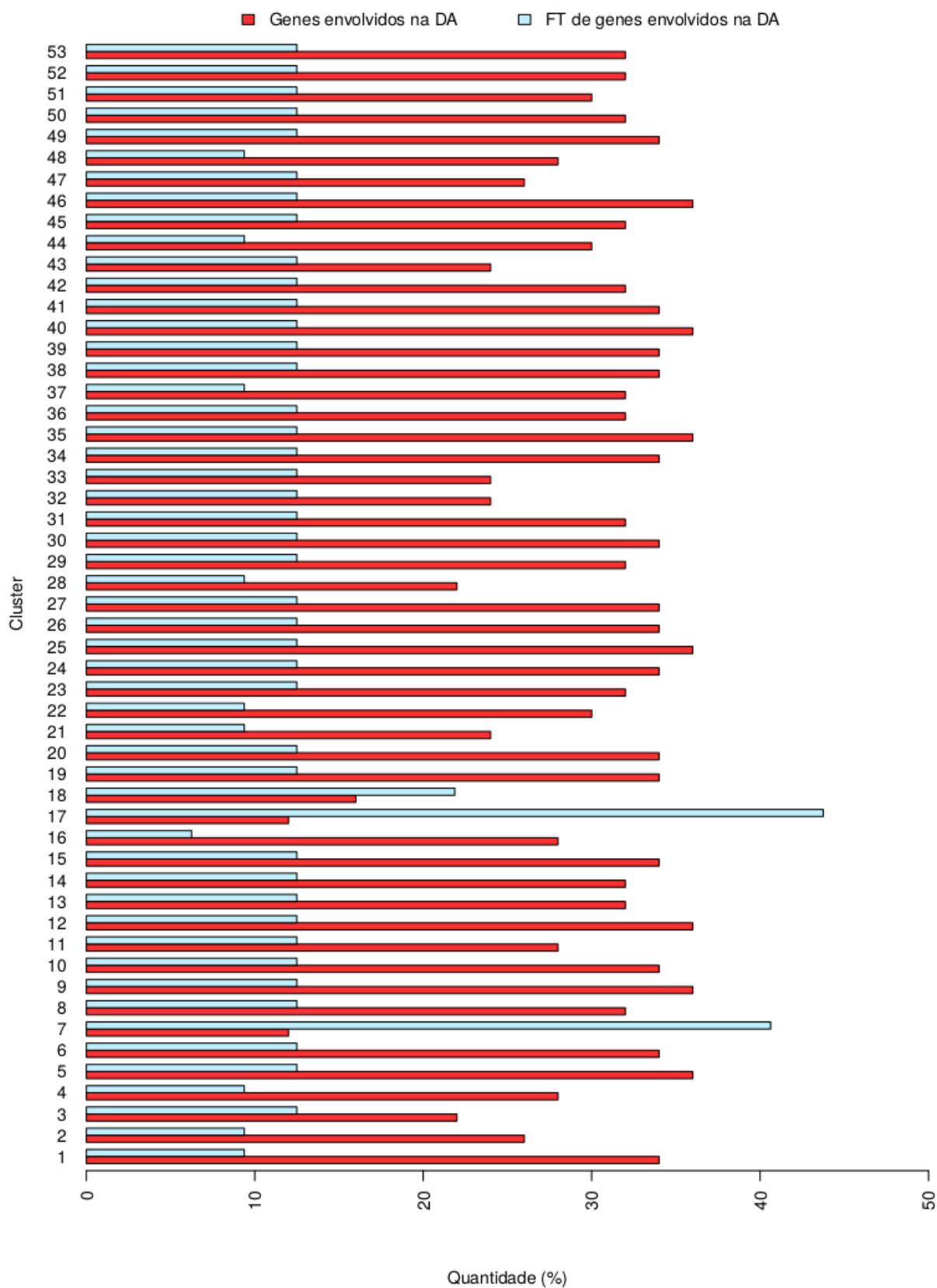


Figura 15 – Percentual de genes e FT na faixa aproximada de 20% de maior pertinência nos *cluster*.

Fonte: Autoria própria.

Para o conjunto considerando o menor valor de expressão, a Tabela 6 apresenta o percentual de genes e FT nos 54 *clusters*.

Tabela 6 – Percentual da presença de genes e FT em aproximadamente 20% de maior e menor pertinência nos *clusters* do conjunto com o menor valor de expressão

Cluster	(%) Genes Relacionados AD		(%)FT Relacionados à AD		Grupo
	Menor	Maior	Menor	Maior	
<b>1</b>	<b>22,00</b>	<b>22,00</b>	<b>37,50</b>	<b>6,25</b>	<b>01</b>
2	22,00	20,00	40,63	9,38	
3	18,00	22,00	37,50	25,00	
4	22,00	16,00	40,63	9,38	
5	22,00	18,00	40,63	9,38	
6	22,00	22,00	37,50	3,13	
7	24,00	20,00	37,50	9,38	
8	22,00	22,00	37,50	12,50	
9	22,00	18,00	37,50	9,38	
10	22,00	22,00	37,50	12,50	
11	22,00	18,00	43,75	9,38	
12	22,00	16,00	40,63	9,38	
13	22,00	16,00	40,63	9,38	
14	22,00	16,00	37,50	9,38	
15	22,00	18,00	40,63	9,38	
<b>16</b>	<b>22,00</b>	<b>18,00</b>	<b>46,88</b>	<b>9,38</b>	<b>03</b>
17	22,00	14,00	37,50	9,38	
18	22,00	18,00	37,50	9,38	
<b>19</b>	<b>22,00</b>	<b>34,00</b>	<b>37,50</b>	<b>9,38</b>	<b>02</b>
20	22,00	14,00	37,50	9,38	
21	22,00	18,00	37,50	9,38	
22	22,00	24,00	40,63	9,38	
23	22,00	16,00	37,50	9,38	
24	22,00	18,00	40,63	9,38	
25	22,00	20,00	40,63	9,38	
<b>26</b>	<b>18,00</b>	<b>18,00</b>	<b>18,75</b>	<b>28,13</b>	<b>04</b>
27	22,00	16,00	37,50	9,38	
28	22,00	16,00	37,50	9,38	
29	22,00	18,00	37,50	12,50	
30	22,00	20,00	40,63	12,50	
31	22,00	18,00	40,63	9,38	
32	22,00	26,00	37,50	9,38	
33	22,00	16,00	37,50	9,38	
<b>34</b>	<b>22,00</b>	<b>18,00</b>	<b>43,75</b>	<b>9,38</b>	
35	22,00	18,00	37,50	9,38	<b>03</b>
36	22,00	14,00	37,50	9,38	
37	22,00	18,00	40,63	9,38	

Continua na página seguinte.

Tabela 6 Continuação da página anterior.

Cluster	(%) Genes Relacionados AD		(%)FT Relacionados à AD		Grupo
	Menor	Maior	Menor	Maior	
38	22,00	16,00	37,50	9,38	<b>04</b>
39	22,00	18,00	37,50	9,38	
<b>40</b>	<b>20,00</b>	<b>22,00</b>	<b>12,50</b>	<b>43,75</b>	
41	22,00	18,00	40,63	9,38	
42	22,00	14,00	40,63	9,38	<b>02</b>
<b>43</b>	<b>22,00</b>	<b>30,00</b>	<b>37,50</b>	<b>15,63</b>	
44	22,00	14,00	37,50	9,38	
45	22,00	18,00	37,50	9,38	
46	22,00	16,00	40,63	9,38	
47	22,00	18,00	40,63	9,38	
48	22,00	18,00	40,63	9,38	
49	22,00	18,00	40,63	9,38	
50	22,00	16,00	37,50	9,38	
51	16,00	24,00	37,50	18,75	
52	22,00	14,00	40,63	9,38	
53	22,00	24,00	40,63	6,25	
54	22,00	16,00	37,50	9,38	
					Concluída

Fonte: Autoria própria.

O gráfico de barras com o percentual para a faixa de aproximadamente 20% de menor pertinência nos *clusters* é apresentado na Figura 16, para o conjunto com o menor valor de expressão.

O percentual para aproximadamente 20% de maior pertinência nos *clusters* é mostrado no gráfico de barras na Figura 17, para o conjunto com o menor valor de expressão.

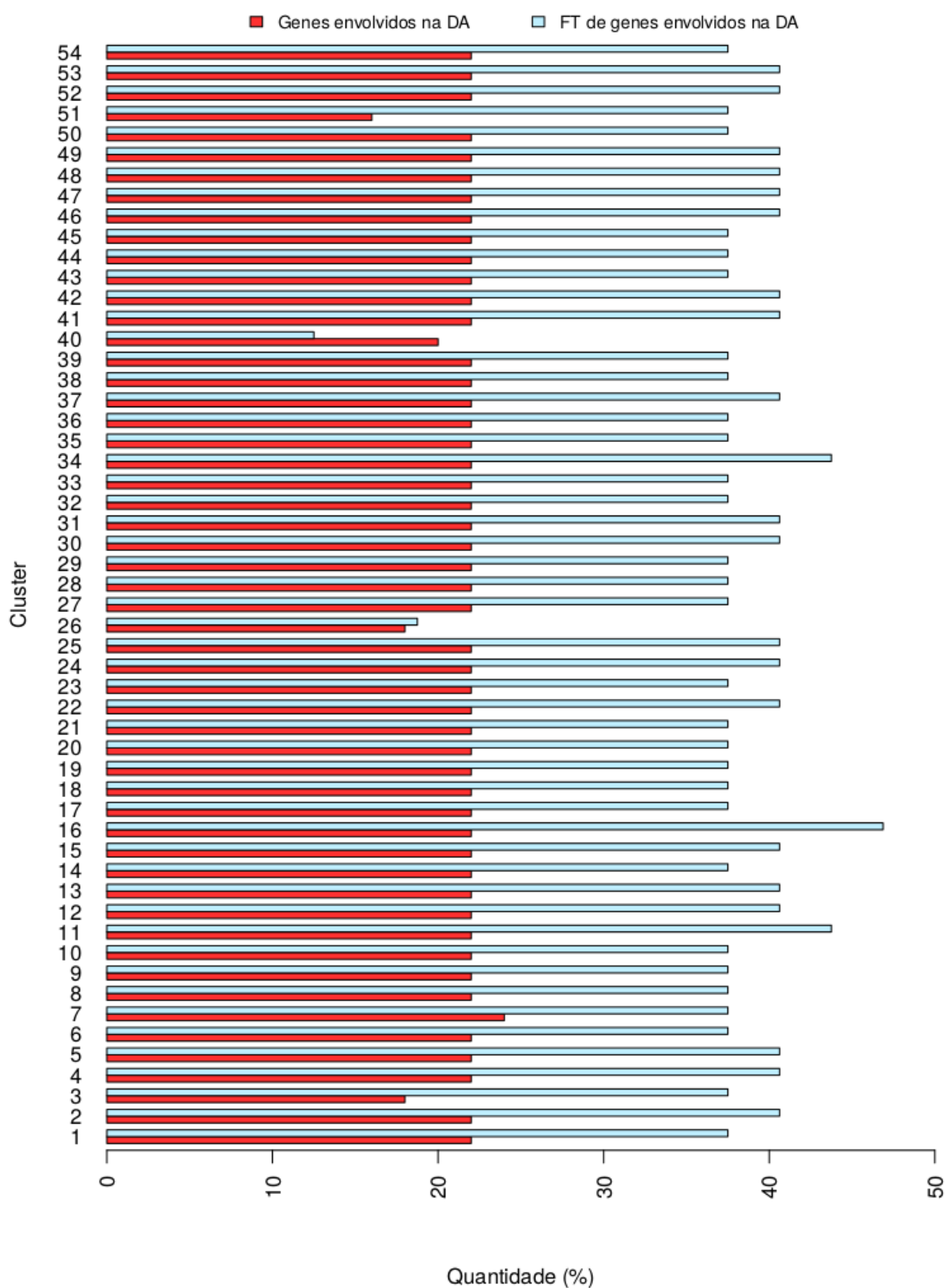


Figura 16 – Percentual de genes e FT na faixa aproximada de 20% de menor pertinência nos *cluster*.

Fonte: Autoria própria.



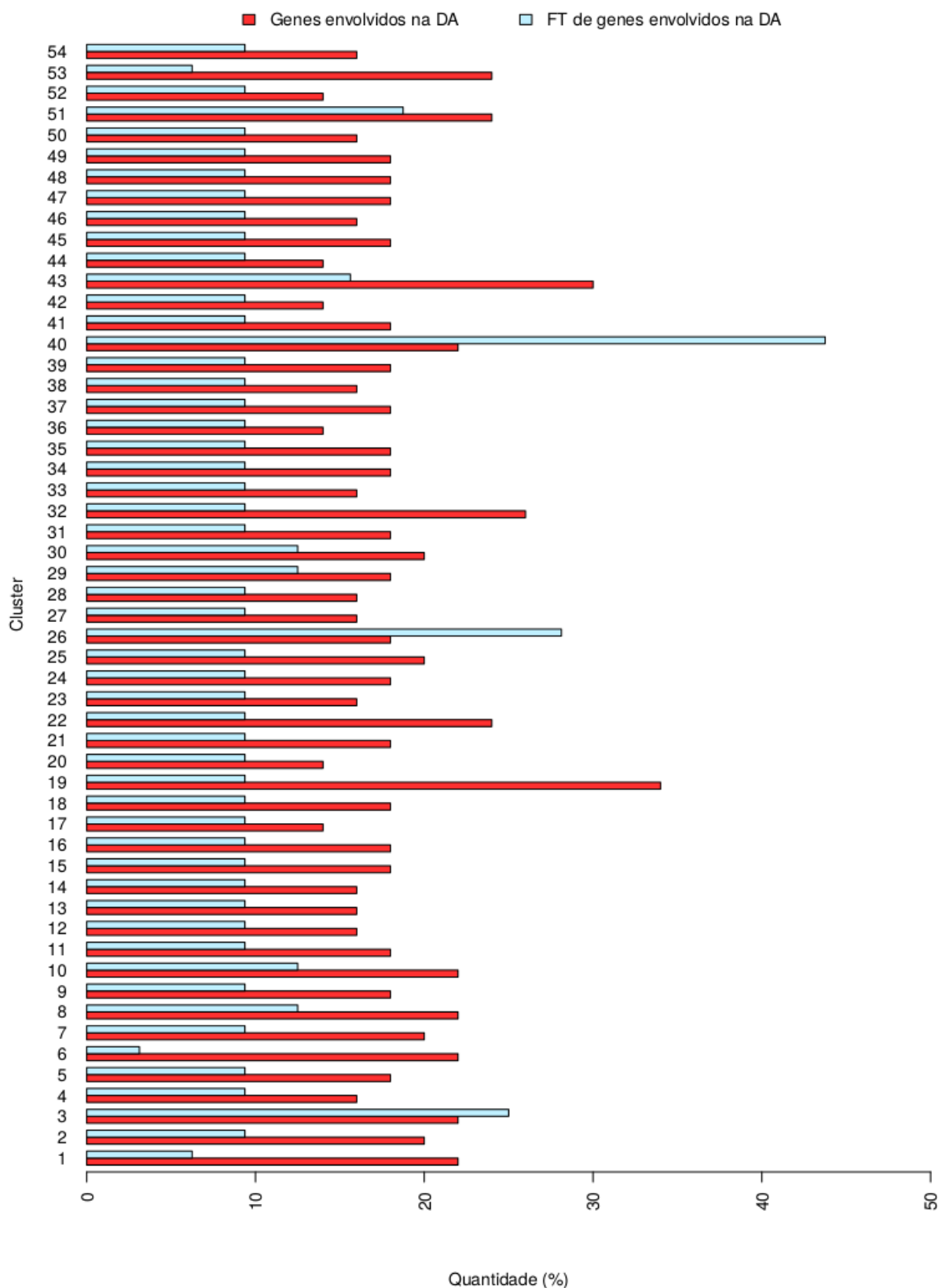


Figura 17 – Percentual de genes e FT na faixa aproximada de 20% de maior pertinência nos *cluster*.

Fonte: Autoria própria.

### 4.2.7 Refinamento e Resultados

Baseado nas informações contidas nas tabelas de percentual de presença e nos gráficos de barras, foram observados o comportamento da distribuição dos genes ao longo dos *clusters*. Fundamentado no conceito do agrupamento realizado pelo *Fuzzy C-Means*, em que todos os genes pertencem a todos os *clusters* com certo grau de pertinência, foram selecionados os *clusters* por faixa de menor e maior pertinência. Esta seleção leva em consideração os genes e os fatores de transcrição relacionados à DA, bem como seu grau de pertinência nos *clusters*.

Nos gráficos das Figuras 25 à 44 do Apêndice A são mostrados o grau de pertinência dos genes nas referidas faixas, com identificação dos genes e fatores de transcrição relacionados à DA dos *clusters* escolhidos. Como podem ser observados nessas figuras, os genes e FT relacionados à DA tendem a apresentar concentração nestas áreas.

Reunindo as informações dos gráficos do grau de pertinência, foram criados quatro grupos formados por cinco *clusters* cada, conforme Tabela 7. Os grupos foram formados obedecendo os seguintes critérios:

- **Grupo 01:** concentração do maior percentual da presença dos genes relacionados ao Alzheimer na faixa com menor grau de pertinência;
- **Grupo 02:** concentração do maior percentual da presença dos genes relacionados ao Alzheimer na faixa com maior grau de pertinência;
- **Grupo 03:** concentração do maior percentual da presença dos fatores de transcrição relacionados ao Alzheimer na faixa com menor grau de pertinência;
- **Grupo 04:** concentração do maior percentual da presença dos fatores de transcrição relacionados ao Alzheimer na faixa com maior grau de pertinência.

Tabela 7 – *Clusters* por grupo selecionados para análise

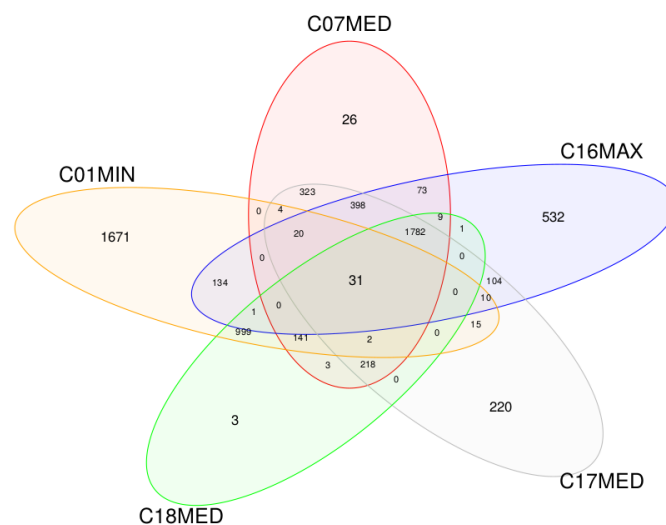
Grupo	Conjunto de dados (valor de expressão)	Cluster
01	Maior	16
	Menor	01
	Mediana	07, 17 e 18
02	Maior	-
	Menor	19 e 43
	Mediana	05, 09 e 12
03	Maior	08 e 09
	Menor	16 e 34
	Mediana	05
04	Maior	16
	Menor	26 e 40
	Mediana	07 e 17

Fonte: Autoria própria.

Os genes presentes nos grupos formados foram utilizados para geração de diagramas de Venn com o objetivo de identificar a recorrência dos genes nos cinco *clusters* de cada grupo.

A escolha de utilizar a quantidade de cinco *clusters* por grupo se deve à uma limitação da função do pacote Limma (ver seção 3.3) que gera o diagrama para o número máximo de cinco conjuntos.

Foram encontrados 31 genes recorrentes no grupo 01 (Figura 18).

Figura 18 – Intersecção dos *clusters* no Grupo 01.

Fonte: Autoria própria.

No grupo 02 foram encontrados 946 genes recorrentes (Figura 19).

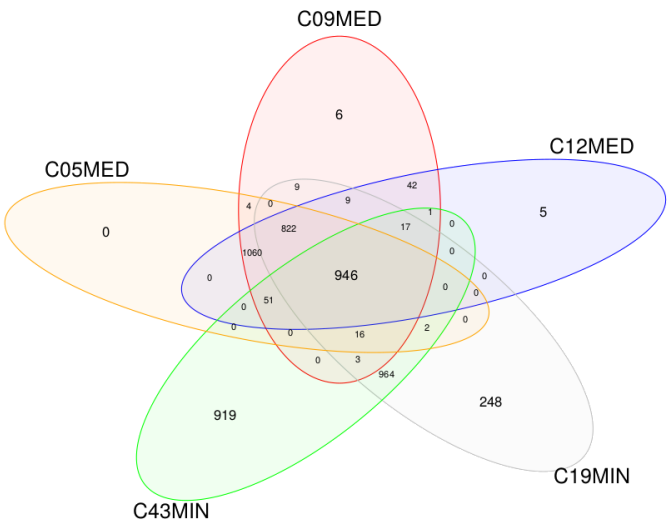


Figura 19 – Intersecção dos *clusters* no Grupo 02.

Fonte: Autoria própria.

O grupo 03 apresentou 1981 genes recorrentes (Figura 20).

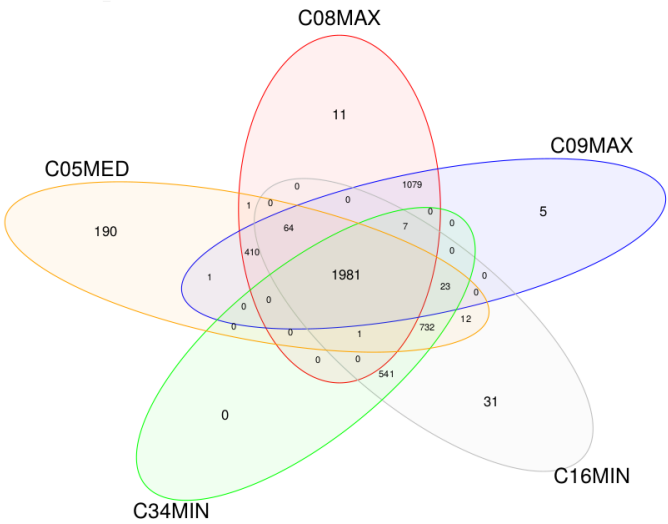


Figura 20 – Intersecção dos *clusters* no Grupo 03.

Fonte: Autoria própria.

No grupo 04, considerando o maior percentual da presença dos fatores de transcrição relacionados à DA na faixa com maior grau de pertinência, foram encontrados 40 genes recorrentes (Figura 21).

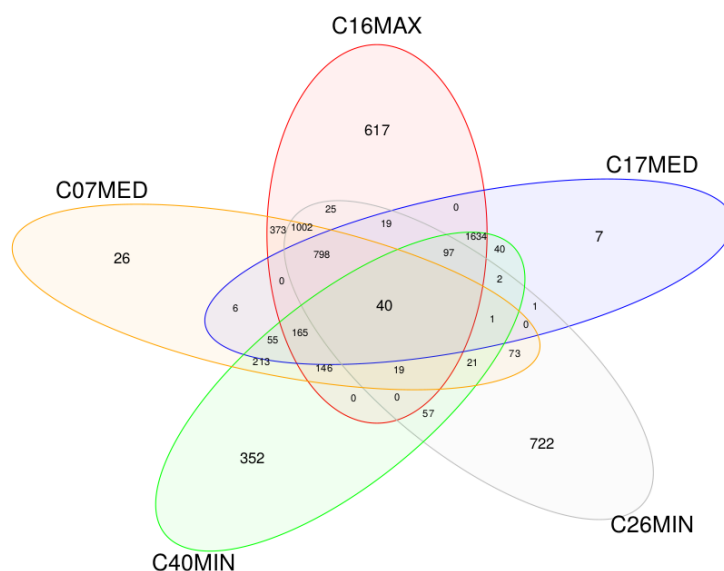


Figura 21 – Intersecção dos *clusters* no Grupo 04.

Fonte: Autoria própria.

Com base nas intersecção dos cinco *clusters* nos grupos formados, foi feita uma comparação na proporção de genes e fatores de transcrição relacionados à DA (Tabela 8) entre os quatro grupos. Ao comparar essa proporção, pôde-se observar que o grupo 04 apresentou maior proporção de genes e FT relacionados à DA na intersecção dos 05 *clusters* em comparação com os outros grupos.

Tabela 8 – Percentual de presença Genes e FT relacionados ao Alzheimer nos grupos

Grupo	Total	Genes e FT's no total	Intersecção	Genes e FT's na Intersecção
01	6720	0,71%	31	0,00%
02	5124	0,60%	946	0,95%
03	5089	0,65%	1981	0,40%
<b>04</b>	<b>6511</b>	<b>0,65%</b>	<b>40</b>	<b>2,50%</b>

Fonte: Autoria própria.

Dessa forma, considerando a maior concentração dos genes e FT relacionados à DA nas intersecções, foram selecionados para a análise, os 40 genes da intersecção do grupo 04.

### 4.3 Análise dos Resultados

Nesta seção é realizada uma análise dos 40 genes selecionados do Grupo 04. Após a seleção dos genes foi realizada uma pesquisa no PubMed (NCBI, 2016) utilizando como palavras-chave os identificadores dos genes, apresentados na Tabela 9, e o termo "Alzheimer". A Tabela 9 apresenta os números de publicações retornados nas buscas. Cinco genes, como por exemplo o *AW451806*, não foram identificados nas bases do PubMed. Dos 35 genes identificados, oito tiveram publicações associadas ao Alzheimer.

Tabela 9 – Publicações relacionadas ao genes selecionados do grupo 04 associadas com o termo *Alzheimer*

Gene	Publicações PubMed	Gene	Publicações PubMed
216498_at	Gene não identificado	MYL9	-
221275_s_at	Gene não identificado	NEK2	-
AF130071	Gene não identificado	NR1I2	-
AP1M2	-	OPN1SW	-
APOL1	1	PAX5	1
AW451806	Gene não identificado	PTGFR	-
CCL24	-	PYCRL	-
CEL	7	SCGB1D2	-
CEP152	-	SLC30A4	4
CLPB	-	SLC6A5	1
CXCL11	1	SPTLC3	-
FAHD2CP	Gene não identificado	SSX5	-
FAM135A	-	SULT2A1	1
GPR18	-	TEAD4	-
HIST1H3H	-	TEP1	-
HOXA3	-	TERT	73
LILRB3	-	TNFRSF10D	-
MAP1LC3C	-	TRAIP	-
MEP1A	-	ZNF208	-
MTNR1A	-	ZNF613	-

Fonte: Autoria própria.

Também foi feita uma busca pelos genes da Tabela 9 na base de dados AlzGene (BERTRAM et al., 2007), no entanto, o único gene que retornou resultados foi o *TEP1*. Foram encontradas seis publicações no AlzGene, no entanto, cinco apontam negativamente para a DA e uma não indica se há ou não evidência com a doença. De

acordo com Safran et al. (2010), o produto do *TEP1* é um componente do complexo ribonucleoproteico responsável pela atividade de telomerase, que cataliza a adição de novos telômeros nas extremidades do cromossomo.

Ao analisar as publicações encontradas para os oito genes, pôde ser observado que:

- *APOL1*: a publicação de Okamoto (2004) não apresenta informações que possa identificar a influência deste gene na Doença de Alzheimer;
- *CEL*: Não foi possível identificar nas publicações encontradas a relação deste gene com a DA. As publicações associadas ao termo "CEL", dizem respeito ao *N<sub>ε</sub> – carboxy – ethil – lysine*, substância que é considerada um fator de envelhecimento e que possui influência em doenças degenerativas (XUE et al., 2011; PEREZ-GRACIA et al., 2009);
- *CXCL11*: este gene é regulado pelos fatores de transcrição *STAT3* e *AML1a* (SAFRAN et al., 2010). A publicação de Hashioka et al. (2012) indica que o fator de transcrição *STAT3* contribui para a neurotoxicidade quando está superexpresso;
- *PAX5*: a publicação de Hsu et al. (1996) não apresenta informações que possam identificar o papel deste gene na DA;
- *SLC30A4*: este gene está envolvido no transporte de Zinco para fora do citoplasma (SAFRAN et al., 2010). As observações de níveis elevados de Zinco na DA, indicam que proteínas reguladoras de Zinco podem ser fatores-chave na fisiopatologia da DA (LYUBARTSEVA et al., 2010; ZHANG et al., 2008; SMITH et al., 2006);
- *SLC6A5*: este gene codifica um transportador para os neurotransmissores de glicina dependente de cloreto de sódio. Segundo Espay e Chen (2013), este gene apresenta uma relação com a mioclonia, uma síndrome reconhecida em pacientes com doenças neurodegenerativas, principalmente na Doença de Alzheimer;
- *SULT2A1*: segundo Vankova et al. (2015), sua pesquisa representa a primeira tentativa de avaliar o papel de *SULT2A1* (esteróide sulfotransferase) na fisiopatologia da doença de Alzheimer com base no metaboloma de esteróides na circulação, mostrando-se positiva na relação do *SULT2A1* com a DA. De acordo com os mesmos autores, para este efeito, foram selecionados os esteróides não conjugados e os seus homólogos conjugados em que a sulfatação catalisada por *SULT2A1* domina sobre glucuronidação;
- *TERT*: segundo Spilsbury et al. (2015), é sugerido que a proteína telomerase (TERT) parece ser um fator protetor no cérebro e pode oferecer resistência neuronal contra

*tau* patológica, reduzindo a produção de espécies oxidantes e melhorar a função mitocondrial. Segundo Alonso et al. (1996), a proteína *tau* associada a microtúbulos anormalmente, torna-se hiperfosforilada na DA e acumula-se como filamentos emaranhados emparelhados helicoidais, em neurônios que sofrem degeneração.

Para análise dos níveis de expressão nas amostras dos genes selecionados, foi feita nos níveis de expressão uma correção de *background* utilizando a função *background-Correct* e uma normalização entre *arrays* utilizando a função *normalizeBetweenArrays*, ambas funções do pacote *limma* do Biocondutor. Em seguida foram gerados três gráficos, um para cada conjunto de dados (Figuras 24, 23 e 22), com a média ponderada dos níveis de expressão dos genes por estágio da doença. As médias ponderadas das amostras de controle estão representadas em linha nos gráficos, enquanto que as amostras doentes estão representadas em barras.



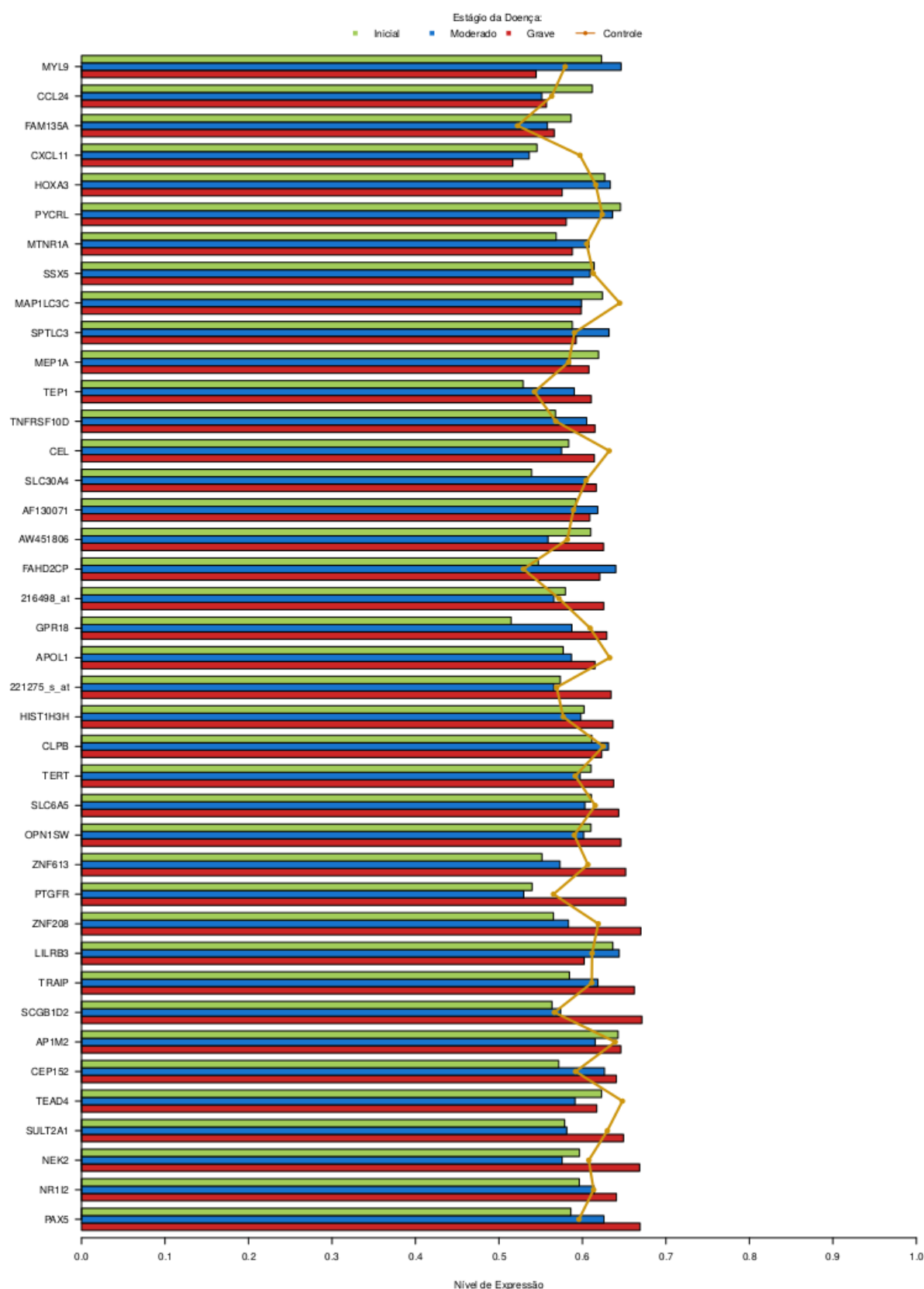


Figura 22 – Níveis de expressão dos genes selecionados em relação à amostra controle (conjunto com menor valor de expressão).

Fonte: Autoria própria.

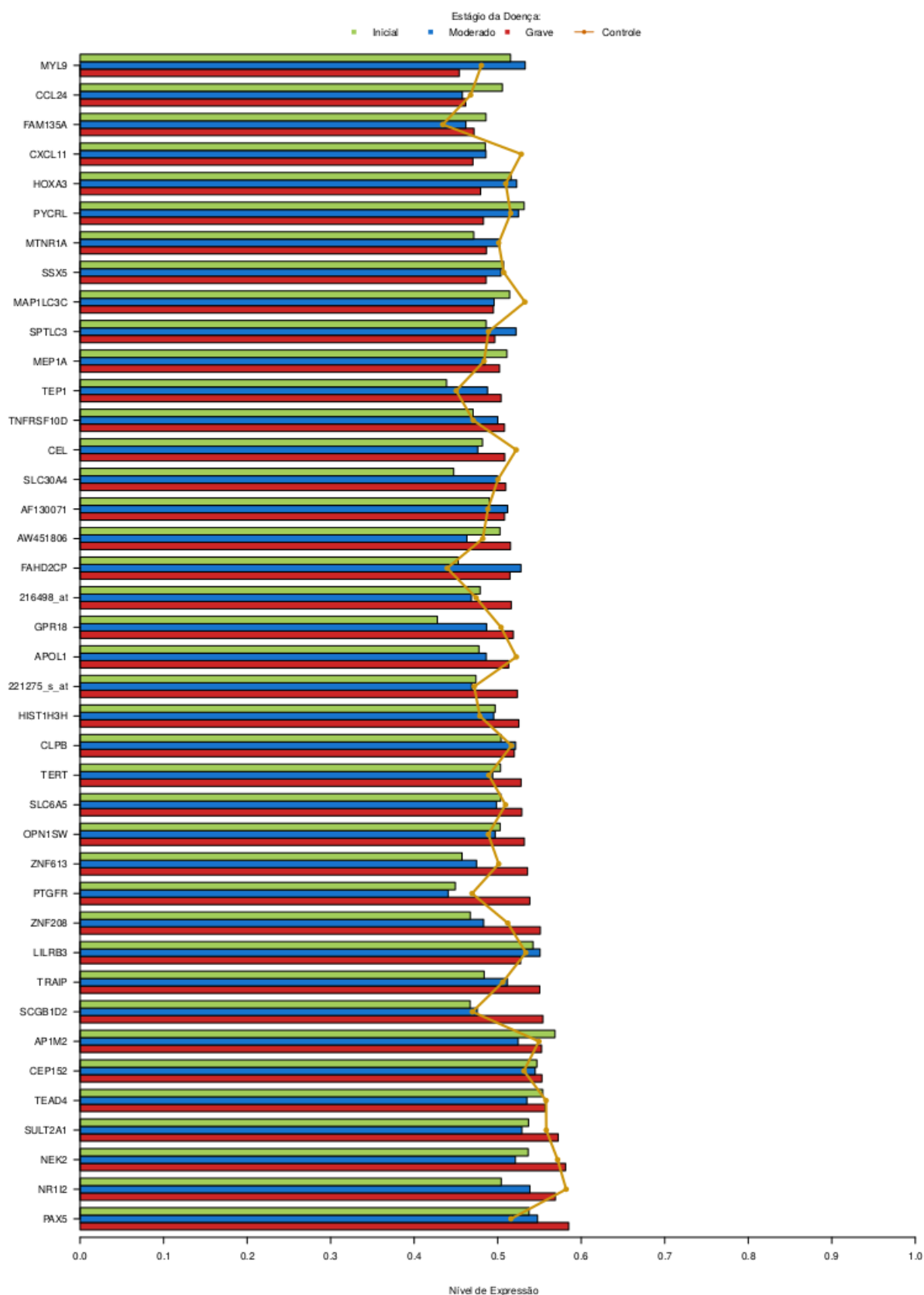


Figura 23 – Níveis de expressão dos genes selecionados em relação à amostra controle (conjunto com a mediana do valor de expressão).

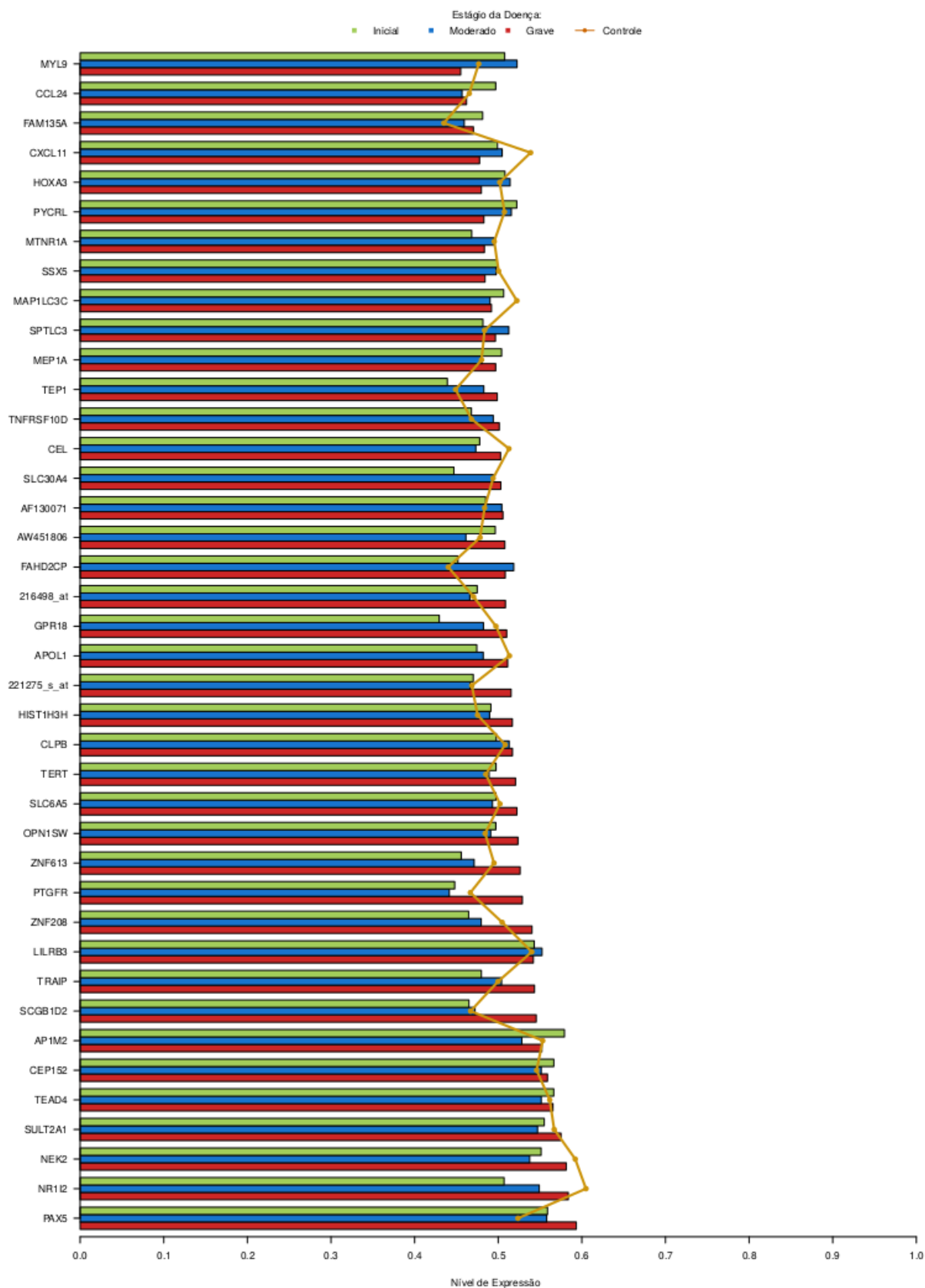


Figura 24 – Níveis de expressão dos genes selecionados em relação à amostra controle (conjunto com maior valor de expressão).

Fonte: Autoria própria.

Analisar graficamente, as médias dos níveis de expressão dos genes selecionados, não é suficiente para determinar se estão diferencialmente expressos nas amostras dos grupos da doença (inicial, moderado e grave) em relação às amostras de controle. Para este tipo de análise, geralmente, é feita uma Análise de Variância (ANOVA).

Segundo Castanheira (2010), a análise de variância permite que as médias de duas ou mais grupos sejam comparadas simultaneamente. São testadas duas hipóteses: a hipótese nula ( $H_0$ ), que afirma a igualdade das médias dos grupos; e a hipótese alternativa ( $H_1$ ), que afirma uma desigualdade entre as médias. No entanto, três condições são colocadas (CASTANHEIRA, 2010):

1. As amostras dos grupos devem estar em uma distribuição normal;
2. Os grupos são aleatórios independente;
3. As variâncias dos grupos possuem o mesmo valor.

Aplicando um teste de normalidade em cada um dos grupo para cada gene, utilizando a função *shapiro.test* do pacote *Limma* do Biocondutor, foi verificado que nem todos os grupos estão em uma distribuição normal. Dessa forma, se fez necessária a adoção de outra ferramenta para analisar os níveis de expressão das amostras dos grupos.

O Teste de Kruskal-Wallis (KRUSKAL; WALLIS, 1952) é uma alternativa não-paramétrica à análise de variância, ou seja, não é exigida as três condições impostas na ANOVA. Segundo Kruskal e Wallis (1952), o teste estatístico para a hipótese nula é calculado pela seguinte equação:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^C \frac{Ri^2}{ni} - 3(N+1). \quad (3)$$

Em que,

- $C$  = número de grupos;
- $ni$  = número de observações dos grupos;
- $N = \sum ni$ , número de observações de todos os grupos combinados;
- $Ri$  =, a soma das fileiras no  $i$ -ésimo grupo.

Por fim, o p-valor é aproximado pela distribuição qui-quadrado ( $\chi^2$ ), sendo  $Pr(\chi^2(C-1) \geq H)$ . A hipótese nula ( $H_0$ ) é rejeitada caso p-valor  $< 0,05$ .

Tabela 10 – Teste de Kruskal-Wallis nos 40 genes selecionados, considerando os níveis de expressão dos conjuntos com o maior, a mediana e o menor valor de expressão

Gene	p-valor(maior)	p-valor(mediana)	p-valor(menor)
<b>NR1I2</b>	<b>0,020489</b>	<b>0,062844</b>	<b>0,598380</b>
<b>GPR18</b>	<b>0,025429</b>	<b>0,023066</b>	<b>0,019450</b>
ZNF208	0,053611	0,056013	0,062540
TEP1	0,111248	0,111248	0,088240
MTNR1A	0,112089	0,112089	0,182750
OPN1SW	0,114739	0,101255	0,101050
CEL	0,149571	0,148318	0,208580
AP1M2	0,240280	0,185176	0,639160
PAX5	0,254258	0,449694	0,312060
CCL24	0,301029	0,296865	0,343400
SLC6A5	0,303554	0,339728	0,354580
PYCRL	0,309585	0,309585	0,383760
SCGB1D2	0,383192	0,360665	0,250670
FAHD2CP	0,385308	0,385308	0,359880
CXCL11	0,421592	0,472089	0,363860
FAM135A	0,427579	0,417733	0,386460
CLPB	0,433405	0,448036	0,400970
CEP152	0,472388	0,561485	0,493250
SSX5	0,473667	0,504823	0,552030
SULT2A1	0,518317	0,373191	0,484470
NEK2	0,520942	0,275913	0,569800
TRAIP	0,523007	0,523007	0,557730
MYL9	0,525425	0,509466	0,540570
PTGFR	0,528763	0,529592	0,524480
SPTLC3	0,566139	0,536581	0,718610
MEP1A	0,571053	0,583363	0,568190
APOL1	0,581306	0,591515	0,537210
HOXA3	0,611911	0,621765	0,535340
SLC30A4	0,628498	0,623518	0,641450
AF130071	0,682027	0,665907	0,758150
TNFRSF10D	0,711727	0,707570	0,690470
216498_at	0,751979	0,740770	0,650410
MAP1LC3C	0,798209	0,819108	0,841210
ZNF613	0,809893	0,787787	0,779270
TERT	0,820291	0,836316	0,919660
221275_s_at	0,857276	0,847010	0,812870
TEAD4	0,876433	0,565284	0,234240
HIST1H3H	0,883214	0,884229	0,843320
AW451806	0,956618	0,965800	0,975270
LILRB3	0,993298	0,992697	0,925610

Fonte: Autoria própria.

O teste foi aplicado aos 40 genes nos três conjuntos de dados. Como pôde ser observado na Tabela 10, no conjunto com o maior valor de expressão os genes que apresentaram p-valor menor que 0.05 foram os genes *NR1I2* e *GPR18*. Nos conjuntos com o menor e a mediana do valor de expressão o gene que apresentou p-valor menor que 0,05 foi apenas o *GPR18*.

O gene *NR1I2*, também chamado de *PXR* (Receptor Pregnano X), tem como produto pertencente à superfamília de receptores nucleares (PRUITT et al., 2014). Segundo (JAIN et al., 2014), o acúmulo de *A $\beta$*  acompanhada de inflamação, é um dos principais fatores que contribuem para a neurodegeneração associada à DA. O *PXR*, como um receptor nuclear, mostra resultados promissores na regulação da transcrição de transportadores de efluxo, principalmente, a glicoproteína-p nas barreiras sanguíneas do cérebro. A superexpressão dos transportadores de efluxo nas barreiras sanguíneas do cérebro é mediada pelo *PXR*, que reduz a carga *A $\beta$*  pela via intracerebral. Além disso, *PXR* controla a *ApoE* e a inflamação que oferece uma vantagem adicional para o seu papel na AD Jain et al. (2014),.

Gantz et al. (1997), informa que transcritos de *GPR18*, além de serem encontrados no baço, são detectados em vários outros tecidos associados com o sistema imunológico. O padrão de expressão observados pelos autores, sugere que este gene possa ter uma relação na regulação do sistema imunológico.

Foi feita uma busca nas bases de dados *GeneCards* (SAFRAN et al., 2010) por doenças associadas aos 40 genes selecionados. A Tabela 11 apresenta as principais doenças relacionadas a estes genes. Como pode ser observado, apesar desta busca não apontar uma relação direta destes genes com a DA, não exclui o possível papel que estes genes possam desempenhar na DA. Dessa forma, mais pesquisas voltadas na determinação da relação destes genes com a DA são necessárias.

Tabela 11 – Principais doenças associadas aos genes selecionados

Gene	Doenças Associadas
216498_at	-
221275_s_at	-
AF130071	-
AP1M2	Agnosia
APOL1	Glomerulosclerose segmentar e Síndrome nefrótica idiopática córtico-resistente esporádica com hialinose segmentar e focal
AW451806	-
CCL24	Eosinofilia e Asma
CEL	Diabetes da maturidade-início dos jovens
CEP152	Microcefalia Autossômica Recessiva e Síndrome de Seckel
CLPB	3-Methylglutaconic Aciduria tipo 7 com catarata, envolvimento neurológico e neutropenia e Tularemia pneumônica
CXCL11	Adenocarcinoma da cavidade nasal e Câncer da cavidade nasal
FAHD2CP	-
FAM135A	Retinite Pigmentosa
GPR18	-
HIST1H3H	-
HOXA3	Hemiagenesia da tiroide
LILRB3	-
MAP1LC3C	-
MEP1A	-
MTNR1A	Escoliose idiopática and Escoliose idiopática do adolescente
MYL9	-
NEK2	Retinite Pigmentosa e Neurofibrona Plexiforme
NR1I2	Frimartinismo e Xantomatose cerebrotendínea.
OPN1SW	Daltonismo Tritan e Daltonismo Azul
PAX5	Leucemia linfóide aguda de precursores das células B e Leucemia linfóide aguda
PTGFR	Hipertensão Ocular e Glaucoma limítrofe
PYCRL	-
SCGB1D2	Câncer de Mama
SLC30A4	Acrodermatite enteropática
SLC6A5	Hiperecplexia
SPTLC3	-
SSX5	Sarcoma sinovial
SULT2A1	Defeito seletivo de células T e Doenças da diferenciação do sexo
TEAD4	Narcolepsia
TEP1	-
TERT	Disceratose congênita e Fibrose pulmonar
TNFRSF10D	Câncer de Próstata
TRAIP	-
ZNF208	-
ZNF613	-

## 5 Considerações Finais

Este trabalho teve por objetivo propor um método para agrupamento de genes utilizando dados de fatores de transcrição e expressão diferencial. Foi escolhida como patologia para aplicação do método e como estudo de caso, a Doença de Alzheimer, que apesar dos avanços nas pesquisas ainda possui causa não determinada.

Após apresentar as etapas do método, foram feitas pesquisas em bases de dados públicas por dados de expressão diferencial obtidos através de *microarray* relacionados à DA, e por fatores de transcrição e seus genes alvos.

Para aplicação do método aos dados obtidos, foi necessário preparar os dados de *microarray* e de fatores de transcrição gerando três novos conjuntos de dados. Além disso, foi determinado o número de *clusters* para os agrupamentos, foram feitas distribuições de frequências com os graus de pertinência dos genes nos *clusters* e gerados histogramas para análise.

Com a criação de grupos formados por cinco *clusters* e identificado os genes recorrentes, o método nos permitiu selecionar um total de 40 genes (Tabela 9). Foram feitas pesquisas por publicações na base de dados PubMed associando cada gene ao termo "Alzheimer". Dos 40, oito retornaram publicações com estudos relacionados à DA, sendo que a maioria destas publicações indicavam o envolvimento direto ou indireto dos genes com a DA.

Foi utilizado o Teste de Kruskal-Wallis para encontrar genes que se apresentam diferencialmente expressos nas amostras com a DA em relação às amostras de controle (Tabela 10). Dos 40 genes, dois se apresentaram diferencialmente expressos: o *NR1I2* e o *GPR18*.

Foram pesquisadas nas bases do *GeneCards* doenças já associadas aos genes selecionados (Tabela 11). Apesar dos resultados das buscas não apresentarem associação destes genes com a DA, isto não exclui as possíveis participações destes na DA.

Dessa forma, são indicados como trabalhos futuros: a pesquisa das funções biológicas destes genes e de suas interações com a DA; e o desenvolvimento de uma ferramenta automatizando o método proposto, facilitando assim a utilização do método.



## Referências

- ALBERTS, B.; ALEXANDER, J.; JULIAN, L.; RAFF, M.; ROBERTS, K.; WALTER, P. **Biologia Molecular da Célula**. 5. ed. Porto Alegre: Artmed, 2010. ISBN 978-85-363-2066-3.
- ALONSO, A. C.; GRUNDKE-IQBAL, I.; IQBAL, K. Alzheimer's disease hyperphosphorylated tau sequesters normal tau into tangles of filaments and disassembles microtubules. **Nature medicine**, v. 2, n. 7, p. 783–7, jul 1996. ISSN 1078-8956. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/8673924>>.
- ALZHEIMER'S ASSOCIATION. 2015 Alzheimer's Disease Facts and Figures. **Alzheimer's & Dementia: The Journal of the Alzheimer's Association**, v. 11, n. 2, p. 332–384, 2015.
- BELINKY, F.; NATIV, N.; STELZER, G.; ZIMMERMAN, S.; Iny Stein, T.; SAFRAN, M.; LANCET, D. PathCards: multi-source consolidation of human biological pathways. **Database**, v. 2015, p. bav006–bav006, feb 2015. ISSN 1758-0463. Disponível em: <<http://database.oxfordjournals.org/cgi/doi/10.1093/database/bav006>>.
- BERTRAM, L.; MCQUEEN, M. B.; MULLIN, K.; BLACKER, D.; TANZI, R. E. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. **Nature Genetics**, v. 39, n. 1, p. 17–23, jan 2007. ISSN 1061-4036. Disponível em: <<http://www.nature.com/doi/10.1038/ng1934>>.
- BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM: The fuzzy c-means clustering algorithm. **Computers & Geosciences**, v. 10, n. 2-3, p. 191–203, jan 1984. ISSN 00983004. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/0098300484900207>>.
- BLALOCK, E. M.; GEDDES, J. W.; CHEN, K. C.; PORTER, N. M.; MARKESBERY, W. R.; LANDFIELD, P. W. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. **Proceedings of the National Academy of Sciences of the United States of America**, v. 101, p. 2173–2178, 2004. ISSN 0027-8424.
- BOVOLENTA, L. A.; ACENCIO, M. L.; LEMKE, N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. **BMC Genomics**, v. 13, n. 1, p. 405, 2012. ISSN 1471-2164. Disponível em: <<http://www.biomedcentral.com/1471-2164/13/405>>.
- CADDICK, M.; DOBSON, C. Gene regulation. In: GOLDMAN, G. H.; OSMANI, S. A. (Ed.). **The Aspergilli: Genomics, Medical Applications, Biotechnology, and Research Methods**. [S.l.]: CRC Press, 2007. p. 153–161. ISBN 9780849390807 - CAT# 9080.
- CAMILO, C. O.; SILVA, J. C. da. **Mineração de Dados: Conceitos, tarefas, métodos e ferramentas**. [S.l.], 2009. 29 p. Disponível em: <[http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF{\\\_}001-09](http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF{\_}001-09)>.
- CASTANHEIRA, N. P. **Estatística Aplicada a Todos os Níveis**. 5. ed. Curitiba: Ibpx, 2010. 255 p.

CAYTON, H.; WARNER, J.; NORIGRAHAM. **Tudo Sobre Doença de Alzheimer**. São Paulo: Andrei, 2000. 161 p. ISBN 8574762601.

CORCOLL-SPINA, C. d. O. **Lógica fuzzy : reflexões que contribuem para a questão da subjetividade na construção do conhecimento matemático**. São Paulo: Faculdade de Educação da Universidade de São Paulo, 2010.

CORREA, S. M. B. B. **Probabilidade e Estatística**. 2. ed. Belo Horizonte: PUC Minas Virtual, 2003. 116 p.

COX, E. Principal Model Types. In: **Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration**. Elsevier, 2005. p. 540. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/B9780121942755500049>>.

CREIGHTON, T. E. **Encyclopedia of Molecular Biology**. New York, Chichester, Weinheim, Brisbane, Singapore, Toronto: John Wiley & Sons, Inc., 1999. ISBN 0-471-15302-8.

CUMMINGS, J. L. Alzheimer's Disease. **The New England Journal of Medicine**, v. 351, p. 56–67, 2004.

DAVIS, S.; MELTZER, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. **Bioinformatics (Oxford, England)**, v. 23, n. 14, p. 1846–7, jul 2007. ISSN 1367-4811. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/17496320>>.

DUNN, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. **Journal of Cybernetics**, v. 3, n. 3, p. 32–57, jan 1973. ISSN 0022-0280. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/01969727308546046>>.

ESPAY, A. J.; CHEN, R. Myoclonus. **CONTINUUM: Lifelong Learning in Neurology**, v. 19, p. 1264–1286, oct 2013. ISSN 1080-2371. Disponível em: <<http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00132979-201310000-00>>.

FELICE, F. G. D.; VIEIRA, M. N. N.; SARAIVA, L. M.; FIGUEROA-VILLAR, J. D.; GARCIA-ABREU, J.; LIU, R.; CHANG, L.; KLEIN, W. L.; FERREIRA, S. T. Targeting the neurotoxic species in Alzheimer's disease: inhibitors of Abeta oligomerization. **The FASEB journal : official publication of the Federation of American Societies for Experimental Biology**, v. 18, n. 1, p. 1366–1372, 2004. ISSN 0892-6638.

FRIDMAN, C.; GREGÓRIO, S. P.; Dias Neto, E.; Benquique Ojopi, É. P. Alterações genéticas na doença de Alzheimer. **Revista de Psiquiatria Clínica**, v. 31, n. 1, p. 19–25, 2004. ISSN 01016083.

GANTZ, I.; MURAOKA, A.; YANG, Y. K.; SAMUELSON, L. C.; ZIMMERMAN, E. M.; COOK, H.; YAMADA, T. Cloning and chromosomal localization of a gene (GPR18) encoding a novel seven transmembrane receptor highly expressed in spleen and testis. **Genomics**, v. 42, n. 3, p. 462–6, jun 1997. ISSN 0888-7543. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/9205118>>.

GIL, A. C. **Como elaborar projetos de pesquisa**. [S.l.]: Atlas, 2010. ISBN 9788522458233.

GREEN, R. C. **Diagnóstico e tratamento da doença de Alzheimer e outras demências**. Rio de Janeiro: [s.n.], 2001. 223 p. ISBN 8587403494.

GRIFFITHS, A. J. F.; LEWONTIN, R. C.; CARROLL, S. B.; WESSLER, S. R. **Introdução à Genética**. 9. ed. Rio de Janeiro: Guanabara Koogan, 2009.

HASHIOKA, S.; KLEGERIS, A.; MCGEER, P. L. Inhibition of human astrocyte and microglia neurotoxicity by calcium channel blockers. **Neuropharmacology**, v. 63, n. 4, p. 685–91, sep 2012. ISSN 1873-7064. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/22659089>>.

HSU, P. L.; XIE, S. S.; HSU, S. M. Absence of T-cell- and B-cell-specific transcription factors TCF-1, GATA-3, and BSAP in Hodgkin's Reed-Sternberg cells. **Laboratory investigation; a journal of technical methods and pathology**, v. 74, n. 2, p. 395–405, feb 1996. ISSN 0023-6837. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/8780159>>.

HUBER, W.; CAREY, V. J.; GENTLEMAN, R.; ANDERS, S.; CARLSON, M.; CARVALHO, B. S.; BRAVO, H. C.; DAVIS, S.; GATTO, L.; GIRKE, T.; GOTTARDO, R.; HAHNE, F.; HANSEN, K. D.; IRIZARRY, R. A.; LAWRENCE, M.; LOVE, M. I.; MACDONALD, J.; OBENCHAIN, V.; OLEŚ, A. K.; PAGÈS, H.; REYES, A.; SHANNON, P.; SMYTH, G. K.; TENENBAUM, D.; WALDRON, L.; MORGAN, M. Orchestrating high-throughput genomic analysis with Bioconductor. **Nature Methods**, v. 12, n. 2, p. 115–121, jan 2015. ISSN 1548-7091. Disponível em: <<http://www.nature.com/doifinder/10.1038/nmeth.3252>>.

JAIN, S.; RATHOD, V.; PRAJAPATI, R.; NANDEKAR, P. P.; SANGAMWAR, A. T. Pregnane X Receptor and P-glycoprotein: a connexion for Alzheimer's disease management. **Molecular Diversity**, v. 18, n. 4, p. 895–909, nov 2014. ISSN 1381-1991. Disponível em: <<http://link.springer.com/10.1007/s11030-014-9550-6>>.

KLEIN LAB. **About Alzheimer's**. 2015. Disponível em: <[http://www.kleinlab.org/about/{\\\_}alzheimers](http://www.kleinlab.org/about/{\_}alzheimers)>.

KLEIN, W. Targeting small A $\beta$  oligomers: the solution to an Alzheimer's disease conundrum? **Trends in Neurosciences**, v. 24, n. 4, p. 219–224, apr 2001. ISSN 01662236. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0166223600017495>>.

KOCH, R. **The 80-20 Principle-The secret of achieving more with less**. London: NICHOLAS BREALEY PUBLISHING, 1998. 313 p.

KRISHNAMURTHY, P. **Approaches to clustering gene expression time course data**. Tese (Doutorado) — The Faculty of the Graduate School of The State University of New York at Buffalo, 2006.

KRUSKAL, W. H.; WALLIS, W. A. Use of Ranks in One-Criterion Variance Analysis. **Journal of the American Statistical Association**, v. 47, n. 260, p. 583, dec 1952. ISSN 01621459. Disponível em: <<http://www.jstor.org/stable/2280779?origin=crossref>>.

KUMAR, L.; FUTSCHIK, M. Mfuzz: a software package for soft clustering of microarray data. **Bioinformatics**, v. 2, n. 1, p. 5–7, 2007. ISSN 0973-2063. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/18084642>><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2139991>>.

LATCHMAN, D. S. **Transcription factors: An overview**. 1997. 1305–1312 p.

LESK, A. M. **Introdução à Bioinformática**. 2. ed. Porto Alegre: Artmed, 2008. 384 p. ISBN 978-85-363-1104-3.

LODISH, H.; BERK, A.; MATSUDAIRA, P.; KAISER, C. A.; KRIEGER, M.; SCOTT, M. P.; ZIPURSKY, L.; DARNELL, J. **Molecular Cell Biology**. 5. ed. [S.l.]: W.H. Freeman and Company, 2004. 973 p.

LYUBARTSEVA, G.; SMITH, J. L.; MARKESBERY, W. R.; LOVELL, M. A. Alterations of Zinc Transporter Proteins ZnT-1, ZnT-4 and ZnT-6 in Preclinical Alzheimer's Disease Brain. **Brain Pathology**, v. 20, n. 2, p. 343–350, mar 2010. ISSN 10156305. Disponível em: <<http://doi.wiley.com/10.1111/j.1750-3639.2009.00283.x>>.

MALVEZZI, W. R. **Uma ferramenta baseada em teoria fuzzy para o acompanhamento de alunos aplicado ao modelo de educação presencial mediado por tecnologia**. São Paulo: Escola Politécnica da Universidade de São Paulo, 2010.

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. **Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2008. ISBN 9780511809071. Disponível em: <<http://ebooks.cambridge.org/ref/id/CBO9780511809071>>.

MARCONI, M.; LAKATOS, E. **Fundamentos de metodologia científica**. [S.l.: s.n.], 2003. 310 p. ISSN 9788522457588. ISBN 8522433976.

NCBI. **About NCBI: Our Mission**. 2015. Disponível em: <<http://www.ncbi.nlm.nih.gov/home/about/mission.shtml>>.

NCBI. **PubMed**. 2016. <http://www.ncbi.nlm.nih.gov/pubmed> p.

NOVAK, V.; PERFILIEVA, I.; MOCKOR, J. **Mathematical Principles of Fuzzy Logic**. Boston, MA: Springer US, 1999. 320 p. ISBN 978-1-4613-7377-3. Disponível em: <<http://link.springer.com/10.1007/978-1-4615-5217-8>>.

OKAMOTO, Y. [Other apolipoproteins (apolipoprotein F, G, H, L, M and so on)]. **Nihon rinsho. Japanese journal of clinical medicine**, v. 62 Suppl 1, p. 123–6, dec 2004. ISSN 0047-1852. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/15658279>>.

ORTEGA, N. R. S. **Aplicação da Teoria de Conjuntos Fuzzy a Problemas da Biomedicina**. 166 p. Tese (Tese de Doutorado) — Universidade de São Paulo, 2001. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/43/43134/tde-04122013-133237/>>.

PEREZ-GRACIA, E.; BLANCO, R.; CARMONA, M.; CARRO, E.; FERRER, I. Oxidative stress damage and oxidative stress responses in the choroid plexus in Alzheimer's disease. **Acta Neuropathologica**, v. 118, n. 4, p. 497–504, oct 2009. ISSN 0001-6322. Disponível em: <<http://link.springer.com/10.1007/s00401-009-0574-4>>.

PHILIPS, T.; HOOPES, L. Transcription Factors and Transcriptional Control in Eukaryotic Cells. **Nature Education**, v. 1, n. 1, p. 119, 2008.

PRUITT, K. D.; BROWN, G. R.; HIATT, S. M.; THIBAUD-NISSEN, F.; ASTASHYN, A.; ERMOLAEVA, O.; FARRELL, C. M.; HART, J.; LANDRUM, M. J.; MCGARVEY, K. M.; MURPHY, M. R.; O'LEARY, N. A.; PUJAR, S.; RAJPUT, B.; RANGWALA, S. H.; RIDDICK, L. D.; SHKEDA, A.; SUN, H.; TAMEZ, P.; TULLY, R. E.; WALLIN, C.; WEBB, D.; WEBER, J.; WU, W.; DICUCCIO, M.; KITTS, P.; MAGLOTT, D. R.; MURPHY, T. D.;

OSTELL, J. M. RefSeq: an update on mammalian reference sequences. **Nucleic acids research**, v. 42, n. Database issue, p. D756–63, jan 2014. ISSN 1362-4962. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/24259432><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3965018>>.

RITCHIE, M. E.; PHIPSON, B.; WU, D.; HU, Y.; LAW, C. W.; SHI, W.; SMYTH, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. **Nucleic Acids Research**, v. 43, n. 7, p. e47–e47, apr 2015. ISSN 0305-1048. Disponível em: <<http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv007>>.

SAFRAN, M.; DALAH, I.; ALEXANDER, J.; ROSEN, N.; Iny Stein, T.; SHMOISH, M.; NATIV, N.; BAHIR, I.; DONIGER, T.; KRUG, H.; SIROTA-MADI, A.; OLENDER, T.; GOLAN, Y.; STELZER, G.; HAREL, A.; LANCET, D. GeneCards Version 3: the human gene integrator. **Database**, v. 2010, p. baq020–baq020, aug 2010. ISSN 1758-0463. Disponível em: <<http://database.oxfordjournals.org/cgi/doi/10.1093/database/baq020>>.

SELKOE, D. J. Alzheimer's Disease Is a Synaptic Failure. **Science**, v. 298, n. 5594, p. 789–791, oct 2002. ISSN 00368075. Disponível em: <<http://www.sciencemag.org/cgi/doi/10.1126/science.1074069>>.

SERENIKI, A.; VITAL, M. A. B. F. **A doença de Alzheimer: aspectos fisiopatológicos e farmacológicos**. [S.l.]: scielo, 2008. 0 p.

SMITH, J.; XIONG, S.; MARKESBERY, W.; LOVELL, M. Altered expression of zinc transporters-4 and -6 in mild cognitive impairment, early and late Alzheimer's disease brain. **Neuroscience**, v. 140, n. 3, p. 879–888, jan 2006. ISSN 03064522. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0306452206003022>>.

SPILSBURY, A.; MIWA, S.; ATTEMS, J.; SARETZKI, G. The role of telomerase protein TERT in Alzheimer's disease and in tau-related pathology in vitro. **The Journal of neuroscience : the official journal of the Society for Neuroscience**, v. 35, n. 4, p. 1659–74, jan 2015. ISSN 1529-2401. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/25632141><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4308607>>.

SUBBALAKSHMI, C.; KRISHNA, G. R.; RAO, S. K. M.; RAO, P. V. A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set. **Procedia Computer Science**, v. 46, p. 346–353, 2015. ISSN 18770509. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S1877050915000940>>.

SUGAR, C. A.; JAMES, G. M. Finding the Number of Clusters in a Dataset. **Journal of the American Statistical Association**, v. 98, n. 463, p. 750–763, sep 2003. ISSN 0162-1459. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1198/016214503000000666>>.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., 2005. 769 p. ISBN 0321321367.

USIDA, W. F. **Controle Fuzzy para Melhoria do Perfil de Tensão em Sistemas de Distribuição de Energia Elétrica**. São Carlos: Escola de Engenharia de São Carlos, 2007.

VANKOVA, M.; HILL, M.; VELIKOVA, M.; VCELAK, J.; VACINOVA, G.; LUKASOVA, P.; VEJRAZKOVA, D.; DVORAKOVA, K.; RUSINA, R.; HOLMEROVA, I.; JAROLIMOVA, E.; VANKOVA, H.; BENDLOVA, B. Reduced sulfotransferase SULT2A1 activity in patients with Alzheimer's disease. **Physiological research / Academia Scientiarum Bohemoslovaca**, v. 64 Suppl 2, p. S265–73, 2015. ISSN 1802-9973. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/26680489>>.

WATSON, J. D.; CRICK, F. H. C. **Molecular structure of nucleic acids**. 1953. 737–738 p. Disponível em: <[http://www.nature.com/physics/looking-back/crick/\\$\delimiter"026E30F\\$nhhttp://www.ncbi.nlm.nih.gov/pubmed/13054692](http://www.nature.com/physics/looking-back/crick/$\delimiter)>.

WEIZMANN INSTITUTE OF SCIENCE. **Alzheimers Disease Pathway**. 2015. Disponível em: <[http://pathcards.genecards.org/card/alzheimers{\\\_}disease{\\\_}p](http://pathcards.genecards.org/card/alzheimers{\_}disease{\_}p)>.

XUE, J.; RAI, V.; SINGER, D.; CHABIERSKI, S.; XIE, J.; REVERDATTO, S.; BURZ, D. S.; SCHMIDT, A. M.; HOFFMANN, R.; SHEKHTMAN, A. Advanced Glycation End Product Recognition by the Receptor for AGEs. **Structure**, v. 19, n. 5, p. 722–732, may 2011. ISSN 09692126. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0969212611001055>>.

YONAMINE, F. S.; SPECIA, L.; CARVALHO, V. O. de; NICOLETTI, M. d. C. **Aprendizado não supervisionado em domínios fuzzy - algoritmo fuzzy c-means**. São Carlos: [s.n.], 2002. 20 p.

ZADEH, L. Fuzzy sets. **Information and Control**, v. 8, n. 3, p. 338–353, jun 1965. ISSN 00199958. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S001999586590241X>>.

ZADEH, L. A. The concept of a linguistic variable and its application to approximate reasoning. **Information Sciences**, v. 1, p. 119–249, 1975.

ZHANG, L.-H.; WANG, X.; STOLTENBERG, M.; DANSCHER, G.; HUANG, L.; WANG, Z.-Y. Abundant expression of zinc transporters in the amyloid plaques of Alzheimer's disease brain. **Brain research bulletin**, v. 77, n. 1, p. 55–60, sep 2008. ISSN 1873-2747. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/18639746>>.

## Apêndices

## APÊNDICE A – Grau de pertinência dos genes e fatores de transcrição por faixa de seleção nos *clusters*

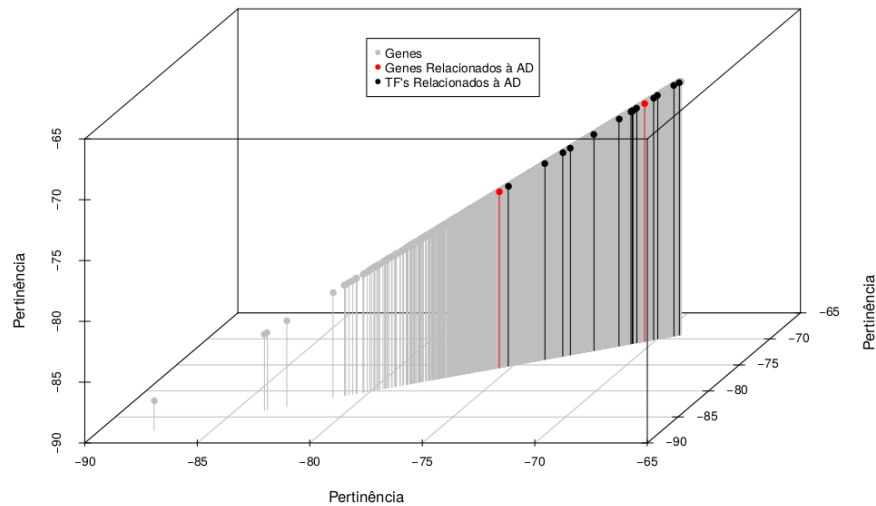


Figura 25 – Genes e FT na seleção de menor pertinência no *cluster* 08 do conjunto com maior valor de expressão.

Fonte: Autoria própria.



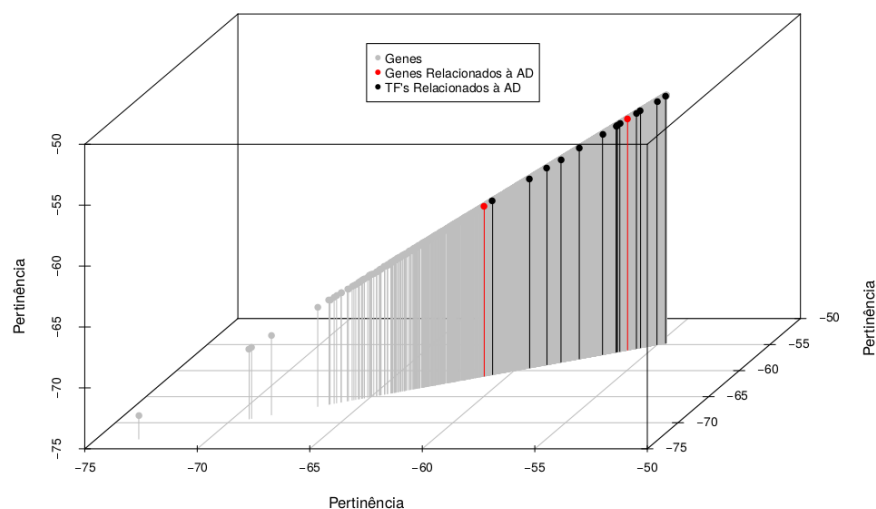


Figura 26 – Genes e FT na seleção de menor pertinência no *cluster* 09 do conjunto com maior valor de expressão.

Fonte: Autoria própria.

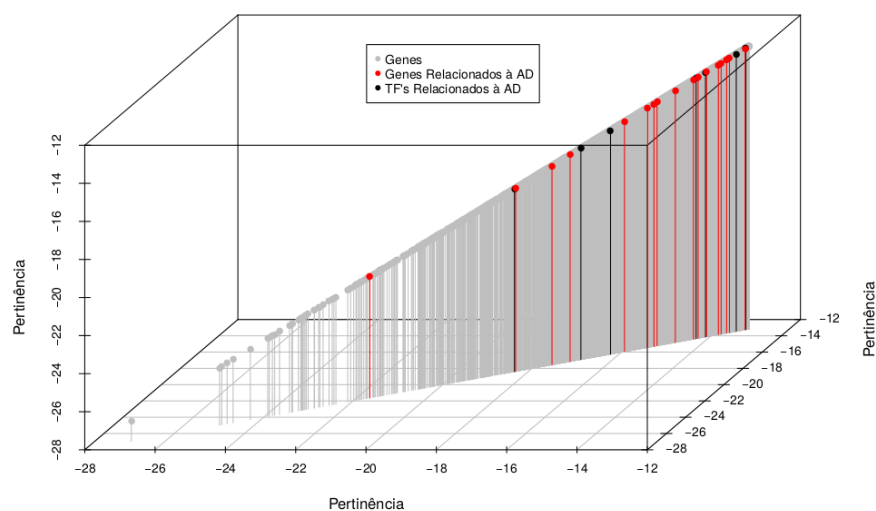


Figura 27 – Genes e FT na seleção de menor pertinência no *cluster* 16 do conjunto com maior valor de expressão.

Fonte: Autoria própria.

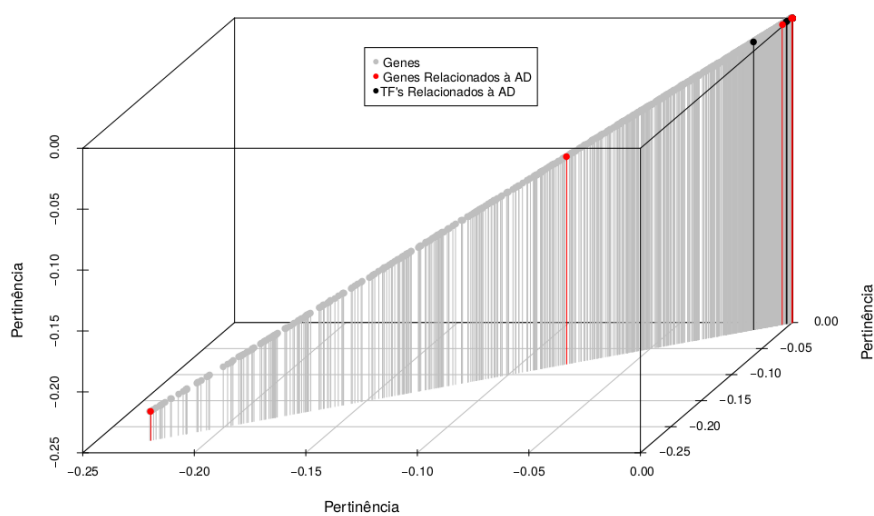


Figura 28 – Genes e FT na seleção de maior pertinência no *cluster* 16 do conjunto com maior valor de expressão.

Fonte: Autoria própria.

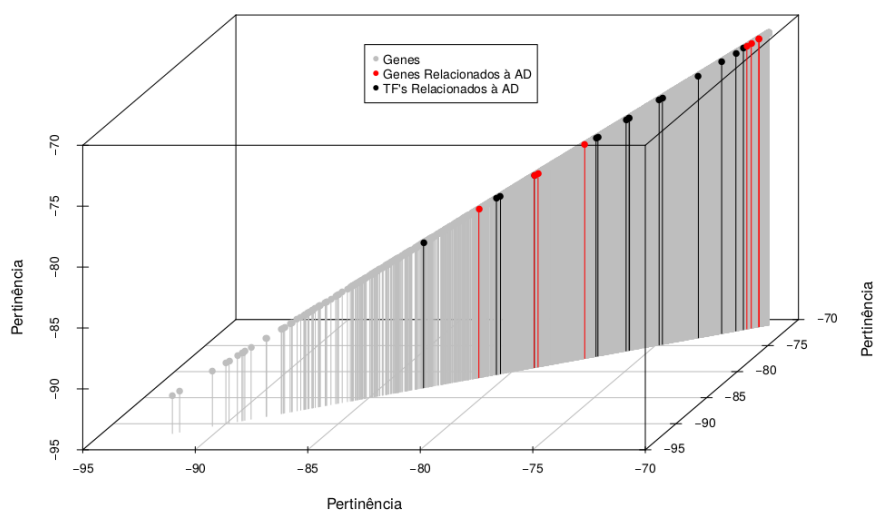


Figura 29 – Genes e FT na seleção de menor pertinência no *cluster* 05 do conjunto com a mediana valor de expressão.

Fonte: Autoria própria.

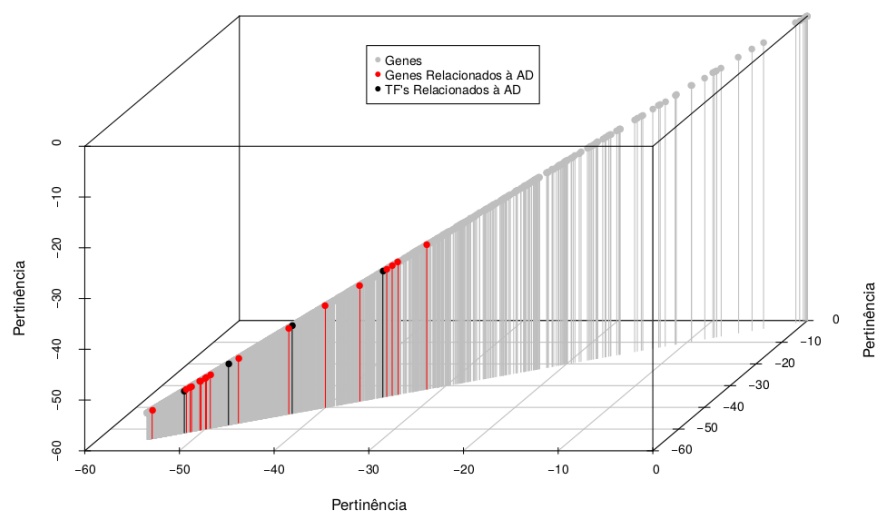


Figura 30 – Genes e FT na seleção de maior pertinência no *cluster* 05 do conjunto com a mediana valor de expressão.

Fonte: Autoria própria.

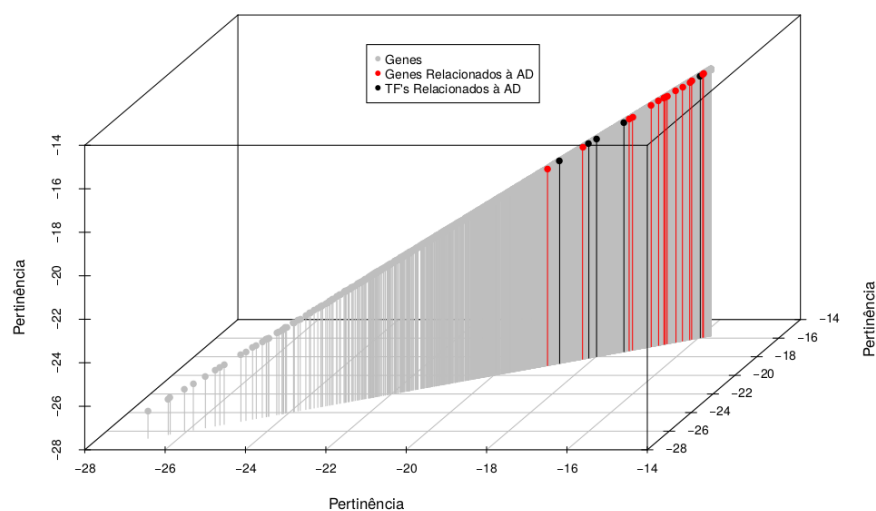


Figura 31 – Genes e FT na seleção de menor pertinência no *cluster* 07 do conjunto com a mediana valor de expressão.

Fonte: Autoria própria.

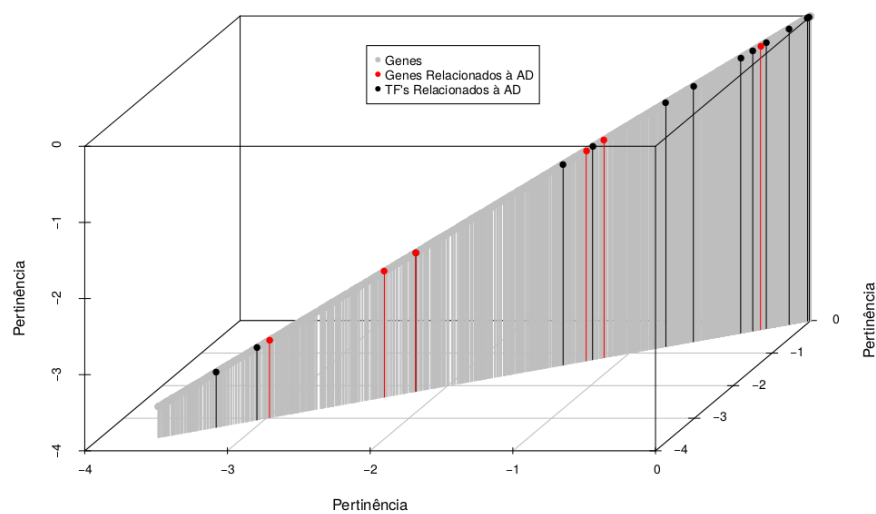


Figura 32 – Genes e FT na seleção de maior pertinência no *cluster* 07 do conjunto com a mediana valor de expressão.

Fonte: Autoria própria.

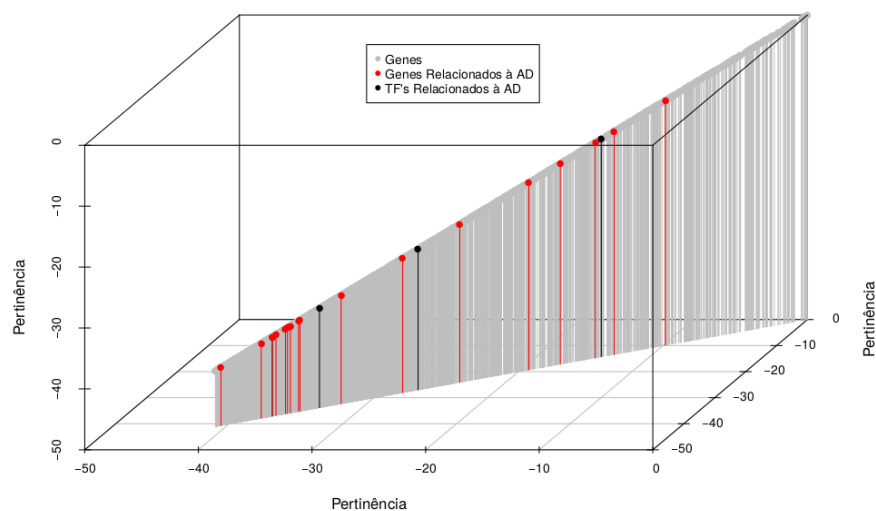


Figura 33 – Genes e FT na seleção de maior pertinência no *cluster* 09 do conjunto com a mediana valor de expressão.

Fonte: Autoria própria.

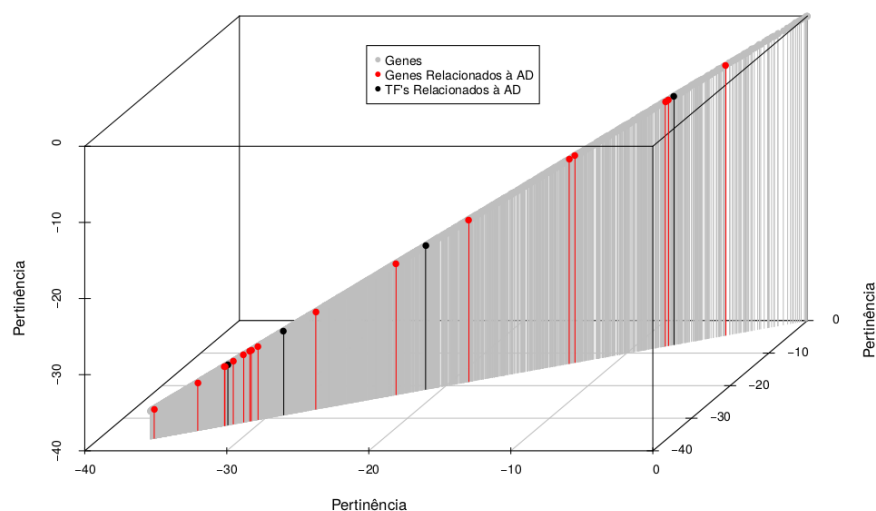


Figura 34 – Genes e FT na seleção de maior pertinência no *cluster* 12 do conjunto com a mediana valor de expressão.

Fonte: Autoria própria.

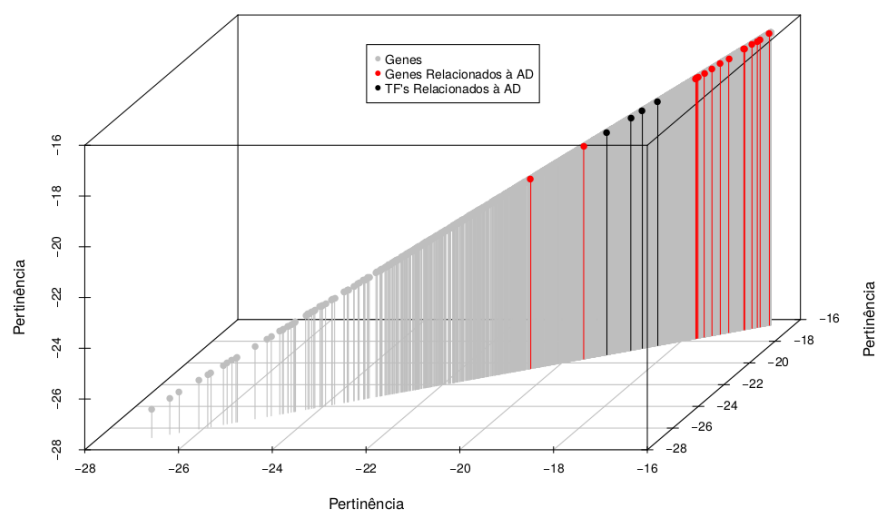


Figura 35 – Genes e FT na seleção de menor pertinência no *cluster* 17 do conjunto com a mediana valor de expressão.

Fonte: Autoria própria.

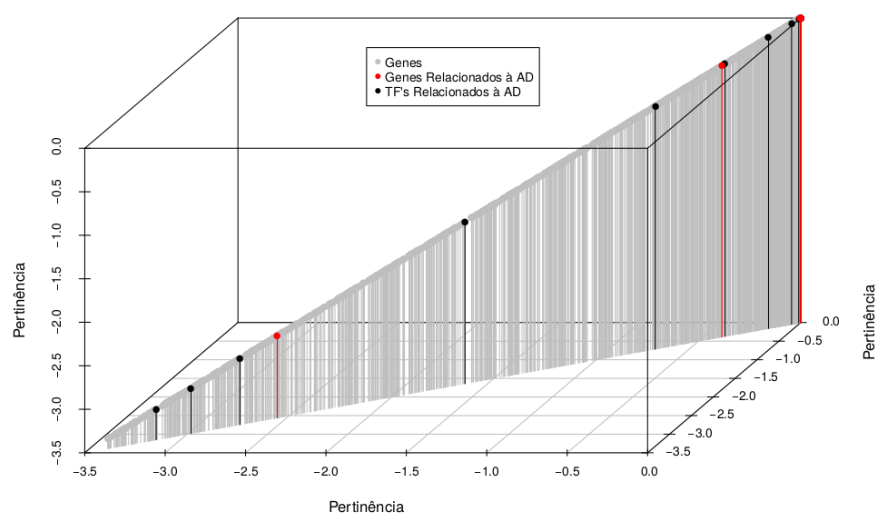


Figura 36 – Genes e FT na seleção de maior pertinência no *cluster* 17 do conjunto com a mediana valor de expressão.

Fonte: Autoria própria.

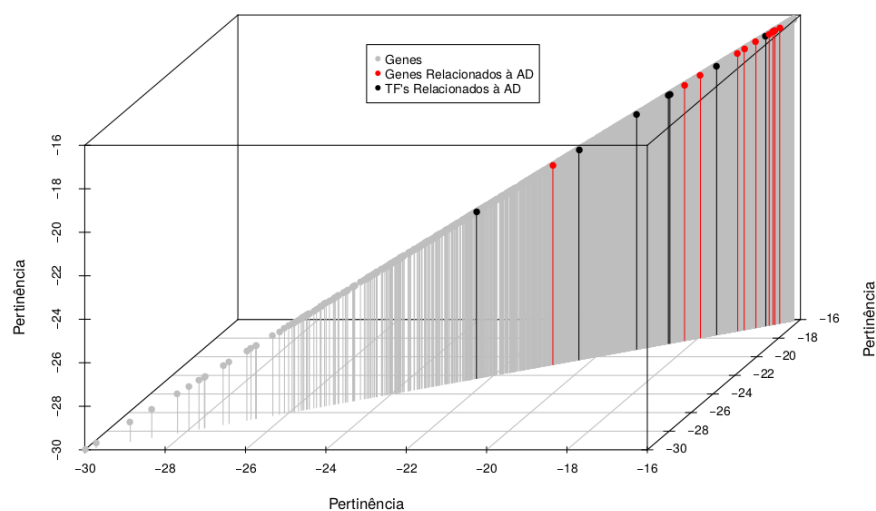


Figura 37 – Genes e FT na seleção de menor pertinência no *cluster* 18 do conjunto com a mediana valor de expressão.

Fonte: Autoria própria.

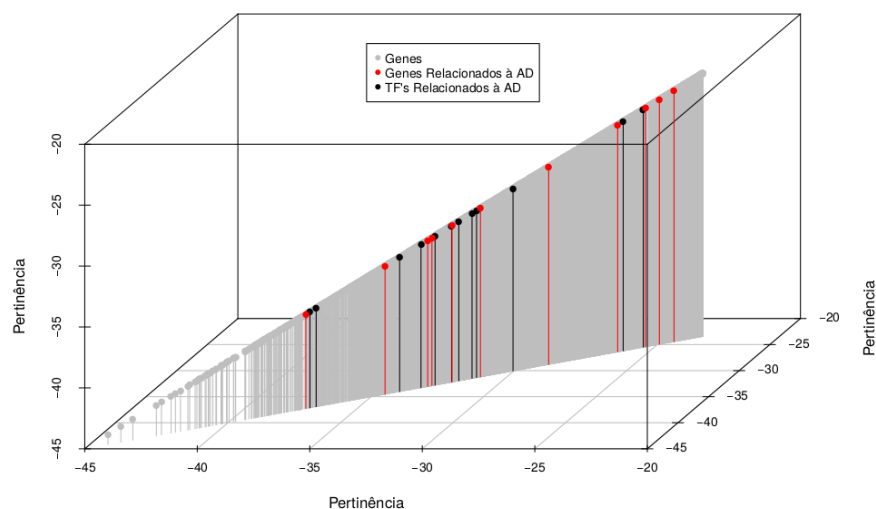


Figura 38 – Genes e FT na seleção de menor pertinência no *cluster* 01 do conjunto com o menor valor de expressão.

Fonte: Autoria própria.

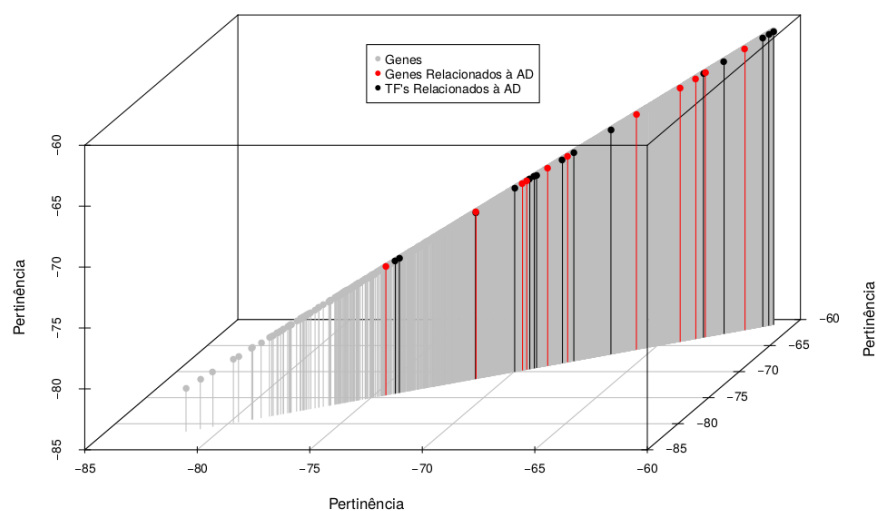


Figura 39 – Genes e FT na seleção de menor pertinência no *cluster* 16 do conjunto com o menor valor de expressão.

Fonte: Autoria própria.

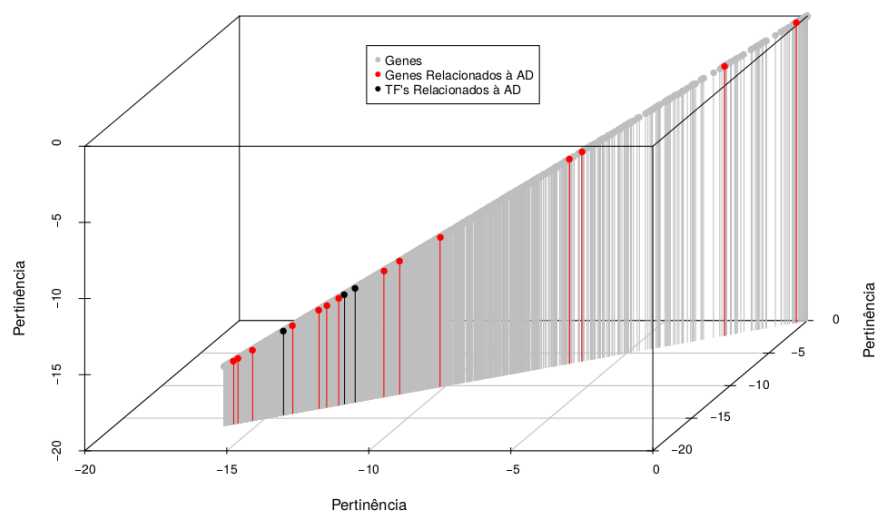


Figura 40 – Genes e FT na seleção de maior pertinência no *cluster* 19 do conjunto com o menor valor de expressão.

Fonte: Autoria própria.

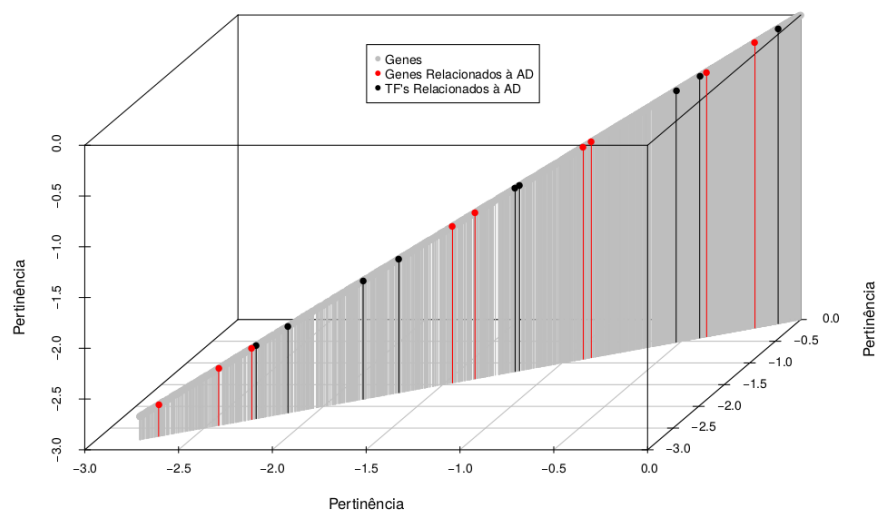


Figura 41 – Genes e FT na seleção de maior pertinência no *cluster* 26 do conjunto com o menor valor de expressão.

Fonte: Autoria própria.



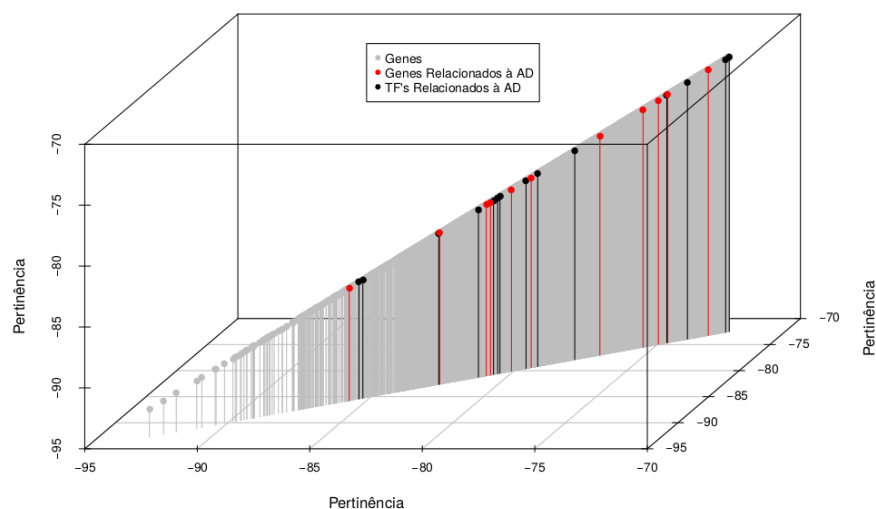


Figura 42 – Genes e FT na seleção de menor pertinência no *cluster* 34 do conjunto com o menor valor de expressão.

Fonte: Autoria própria.

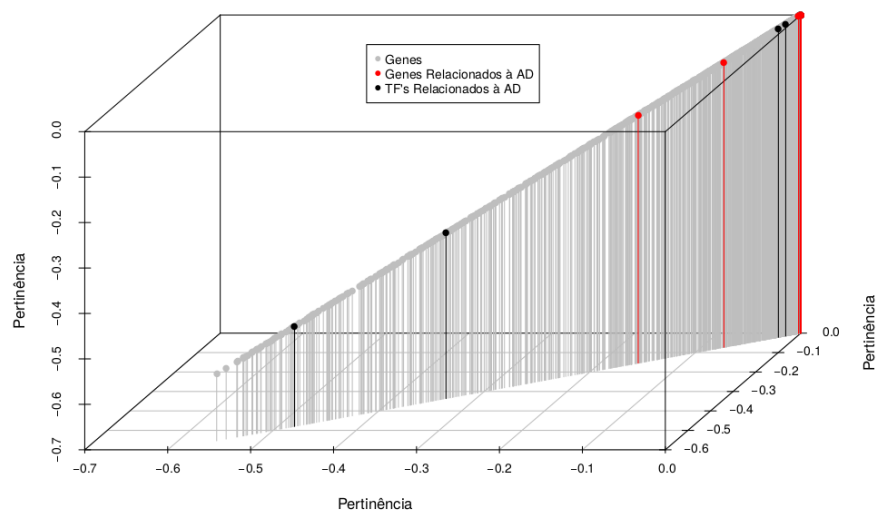


Figura 43 – Genes e FT na seleção de maior pertinência no *cluster* 40 do conjunto com o menor valor de expressão.

Fonte: Autoria própria.

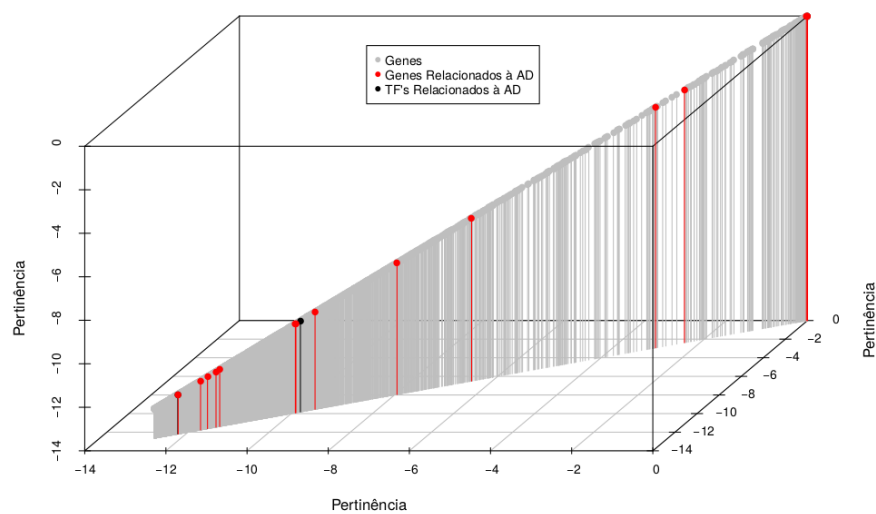


Figura 44 – Genes e FT na seleção de maior pertinência no *cluster* 43 do conjunto com o menor valor de expressão.

Fonte: Autoria própria.