



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

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

DIOGO PEREIRA SILVA DE NOVAIS

**RECONHECIMENTO DE PADRÕES EM DADOS DE EXPRESSÃO GÊNICA DE
PACIENTES PORTADORES DE OSTEOGÊNESE IMPERFEITA**

**ILHÉUS-BA
2016**

DIOGO PEREIRA SILVA DE NOVAIS

**RECONHECIMENTO DE PADRÕES EM DADOS DE
EXPRESSÃO GÊNICA DE PACIENTES PORTADORES DE
OSTEOGÊNESE IMPERFEITA**

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

Orientador: Prof. Dr. Paulo Eduardo Ambrósio

Coorientadora: Prof^a. Dr^a. Carla Martins Kaneto

ILHÉUS-BA
2016

N935

Novais, Diogo Pereira Silva de.

Reconhecimento de padrões em dados de expressão gênica de pacientes portadores de osteogênese imperfeita / Diogo Pereira Silva de Novais. – Ilhéus, BA: UESC, 2016.

61f. : il.

Orientador: Paulo Eduardo Ambrósio.

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

Inclui referências.

1. Osteogênese imperfeita. 2. Mapas auto-organizáveis. 3. Regulação de expressão gênica.
I. Título.

CDD 611.0182

DIOGO PEREIRA SILVA DE NOVAIS

**RECONHECIMENTO DE PADRÕES EM DADOS DE
EXPRESSÃO GÊNICA DE PACIENTES PORTADORES DE
OSTEOGÊNESE IMPERFEITA**

Ilhéus-BA, 15/02/2016

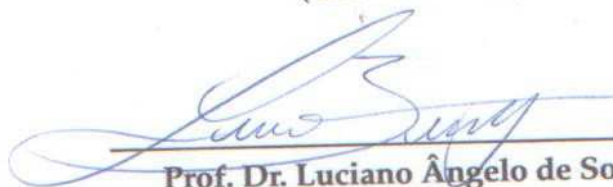
Comissão Examinadora



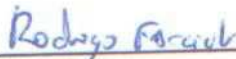
Prof. Dr. Paulo Eduardo Ambrósio
UESC
(Orientador)



Profª. Drª. Carla Martins Kaneto
UESC
(Coorientadora)



**Prof. Dr. Luciano Ângelo de Souza
Bernardes**
UESC



Prof. Dr. Rodrigo Antonio Faccioli
UniMauá

Agradecimentos

- À UESC através do PPGMC pela oferta do Mestrado em Modelagem Computacional em Ciência e Tecnologia
- Aos professores Paulo Ambrósio e Carla Kaneto pela orientação e oportunidade de trabalho com o tema desenvolvido neste trabalho.
- Ao quadro docente do PPGMC pelas contribuições na formação em todas atividades oferecidas pelo programa.
- Ao corpo técnico do programa e demais profissionais da UESC que possibilitam a manutenção do espaço onde o trabalho foi desenvolvido.
- Aos membros da banca examinadora pelo aceite do convite e pelas contribuições para o trabalho.
- A minha esposa Cátia, que além das contribuições pessoais, foi fonte de consultas constantes sobre conhecimentos de biologia molecular e revisora dos textos elaborados neste trabalho.
- Aos colegas do PPGMC pela parceria essencial no processo de formação do mestrado, pessoas com quem aprendi e espero ainda aprender muito.
- À comunidade do Campus Porto Seguro do IFBA pelo convívio acadêmico que propicia um ambiente continuado de formação acadêmica e científica.
- À minha família, a quem devo todas conquistas pessoais e profissionais.

Reconhecimento de Padrões em Dados de Expressão Gênica de Pacientes Portadores de Osteogênese Imperfeita

Resumo

Baseado na hipótese de que genes que apresentam perfis de expressão similares quando expostos a determinada condição podem estar envolvidos em processos funcionais relacionados ou possuem mecanismos de regulação em comum, a análise de agrupamento entre os padrões de expressão gênica pode revelar informações importantes acerca de genes envolvidos em determinado processo ou condição biológica. No entanto, a aplicação de algoritmos de agrupamento possui um conjunto de incertezas inerentes ao processo, uma vez que diferentes técnicas de normalização dos dados, diferentes algoritmos, ou até mesmo diferentes parâmetros podem evidenciar informações distintas. Em grande parte dos casos, definir o agrupamento que melhor representa os perfis existentes nos dados envolve novas análises experimentais ou buscas de relações entre os genes agrupados em bases de bioinformática e na literatura existente. Assim, este trabalho tem por objetivo analisar dados de expressão gênica de pacientes portadores de Osteogênese Imperfeita com diferentes algoritmos de agrupamento e apresentar uma discussão acerca dos perfis identificados por cada algoritmo e inferências possíveis acerca do envolvimento dos genes em processos biológicos relacionados com a patologia. Foram analisados dados de expressão gênica de aproximadamente quarenta mil genes de amostras de células tronco mesenquimais da medula óssea de três amostras de controle sadio, e cinco pacientes com Osteogênese Imperfeita, sendo dois portadores do tipo I e três do tipo III da patologia, durante o processo de osteogênese por um período de 21 dias. Desse grande conjunto de genes, após normalizados e pré-processados, foram selecionados os 100 genes com maior desvio padrão entre as amostras para construção dos agrupamentos. Os dados foram agrupados através de três algoritmos não hierárquicos: K-means, Mapas auto-organizáveis e o *Affinity Propagation*. Os três algoritmos encontraram dois grupos com o conjunto de genes bastante similares que se mostraram interessantes para compreensão da patologia. O primeiro grupo apresentou expressão gênica relativa elevada nas amostras de controle sadio e reduzida nas amostras dos pacientes com a patologia, o que sugere que alguns produtos gênicos podem ser gerados em quantidade menor que a necessária, prejudicando a osteogênese. O segundo grupo apresentou expressão gênica relativa reduzida nas amostras de controle sadio e elevada nas amostras dos pacientes com a patologia, o que pode sugerir desvio de osteogênese para outro processo ou alteração nas proporções das proteínas geradas, o que pode comprometer a composição do tecido. O *Affinity Propagation* revelou um terceiro grupo com expressão elevada apenas nas amostras de pacientes com Osteogênese Imperfeita Tipo III, que permite inferências análogas às do segundo grupo, mas direcionadas a

este tipo da patologia. Nas análises de bioinformática realizadas, foram encontrados alguns genes diretamente associados a características da patologia. Além disso, pode ser interessante a análise de genes que não possuem aparentemente relação com a patologia mas foram agrupados com genes relacionados. Estes genes podem estar relacionados a mecanismos de controle pós-transcricional ou em mecanismos associados à patologia ainda não conhecidos. Por fim, pode-se observar que o *Affinity Propagation* apresentou um agrupamento mais adequado ao tipo de análise realizada no trabalho. No entanto, a utilização de mais de um algoritmo de agrupamento e a comparação dos resultados mostrou-se interessante, tanto para reforçar a existência dos grupos encontrados quanto para complementar resultados obtidos pelos diferentes algoritmos.

Palavras-chave: Osteogênese Imperfeita, Agrupamento, K-means, Mapas auto-organizáveis, *Affinity Propagation*

Pattern Recognition with Gene Expression from Microarray Data from Patients with Osteogenesis Imperfecta

Abstract

Based on the hypothesis that genes which present similar expression profiles when exposed for some condition can be involved in related functional process or are regulated for similar cellular mechanisms, the cluster analysis between gene expression profiles can reveal important information about genes involved in some process or biological condition. However, the application of clustering algorithms has a set of uncertainties inherent to this process, once different techniques of preprocessing, different algorithms or even different parameters can reveal distinct information about the expressed genes. Thus, this work aims to analyse gene expression data from patients with Osteogenesis Imperfecta through different algorithms and to present a discussion about the identified profiles for each algorithm and possible inferences around genes involved in biological processes related to the pathology. There were analysed approximately forty thousands genes expressed in mesenchymal stem cells from the bone marrow of three healthy control samples and five patients with Osteogenesis Imperfecta from which two have the type I of the pathology and three have the type III, in the process of osteogenesis during 21 days. From this big set of genes, after normalizing and preprocessing, it were selected the 100 genes which have the biggest standard deviation between the samples to cluster analysis. The clusters had been built through three non-hierarchical algorithms: K-means, Self-Organizing Maps and Affinity Propagation. The three algorithms found two clusters with a set of very similar genes which seen to be interesting for understanding the pathology. The first cluster present high relative gene expression in the healthy control samples and low expression in the patients samples, what can suggest that some gene products can be generated in a amount smaller than the necessary, damaging the osteogenesis. The second cluster present low relative gene expression in the healthy control samples and high expression in the patients samples, what can suggest any deviation from the process of osteogenesis to other one or changing in the proportions of generated proteins, what can compromising the tissue composition. The Affinity Propagation revealed a third group with high expression only in the patients with Osteogenesis Imperfecta Type III samples what furnishes inferences analogs to the ones related to the second group, but directed to this kind of the pathology. In bioinformatics analysis it were found some genes directly associated to the pathology characteristics. Further, it can be interesting the analysis of genes which apparently are not related to the pathology that were clustered with related genes. These genes can be related with post transcriptional control mechanisms or in still unknown mechanisms associated to the pathology. Finally, it can be observed that the Affinity Propagation

presented a cluster more adequate to the analysis done in this work. However, the using of more than one cluster algorithm and the comparison of the results proved interesting, both to reinforce the existence of the found groups as to complement the results furnished from the used algorithms.

Keywords: Osteogenesis Imperfecta, Clustering, K-means, Self Organizing Maps, *Affinity Propagation*

Lista de figuras

Figura 1 – Estrutura do DNA (GRIFFITHS et al., 2008).	7
Figura 2 – Replicação do DNA (GRIFFITHS et al., 2008).	8
Figura 3 – Fragmentos de Okazaki na replicação do DNA (ALBERTS et al., 2010)	9
Figura 4 – Estrutura básica de um aminoácido (ALBERTS et al., 2010).	10
Figura 5 – Dogma Central: transcrição e tradução (ALBERTS et al., 2010).	11
Figura 6 – Etapas da Transcrição (LODISH et al., 2003).	12
Figura 7 – Etapas da Tradução (ALBERTS et al., 2010).	14
Figura 8 – Exemplo de visualização de genes diferencialmente expressos no programa Genes (KANETO, 2011).	19
Figura 9 – Transformação não linear ϕ entre uma espaço $\alpha \subset R^n$ e um espaço discreto A implementada por um SOM. Adaptado de: (HAYKIN, 1999).	31
Figura 10 – Gráfico da função $h_{j,i}(x)$ em relação ao valor de σ . Adaptado de: (HAYKIN, 1999).	33
Figura 11 – Agrupamento dos 100 genes pré-selecionados utilizando o algoritmo <i>Kmeans</i> com número de grupos pré-estabelecido como cinco.	43
Figura 12 – Agrupamento dos 100 genes pré-selecionados utilizando o algoritmo SOM com número de grupos pré-estabelecido como cinco.	44
Figura 13 – Agrupamento dos 100 genes pré-selecionados utilizando o algoritmo <i>Affinity Propagation</i> com número de grupos pré-estabelecido como cinco.	45
Figura 14 – Percentual de similaridade entre os genes inseridos no grupo A nos diferentes algoritmos.	46
Figura 15 – Percentual de similaridade entre os genes inseridos no grupo B nos diferentes algoritmos.	48
Figura 16 – Expressão relativa dos genes do Grupo A em função do tempo.	50
Figura 17 – Expressão relativa do gene CCRL1, exemplar do Grupo A selecionado pelo <i>Affinity Propagation</i>	51
Figura 18 – Expressão relativa dos genes do Grupo B em função do tempo.	52
Figura 19 – Expressão relativa do gene MGC16291, exemplar do Grupo B selecionado pelo <i>Affinity Propagation</i>	52
Figura 20 – Expressão relativa dos genes do Grupo C em função do tempo.	53
Figura 21 – Expressão relativa do gene CR593560, exemplar do Grupo C selecionado pelo <i>Affinity Propagation</i>	54

Lista de tabelas

Tabela 1 – Código Genético	13
Tabela 2 – Ferramentas Blast	22
Tabela 3 – Genes Pertencentes a Grupos A e B por Algoritmo	47
Tabela 4 – Genes do Grupo C no <i>Affinity Propagation</i>	49

Lista de abreviaturas e siglas

BLAST	Basic Local Alignment Search Tool
cDNA	DNA complementar – do inglês <i>complementary DNA</i>
CTM	Célula Tronco Mesenquimal
DDBJ	DNA Databank of Japan
DNA	Ácido desoxirribonucleico
EMBL	European Molecular Biology Laboratory
miRNA	Micro RNA
mRNA	RNA mensageiro
NCBI	National Center for Biotechnology Information
ncRNA	RNA não codificante – do inglês <i>non coding RNA</i>
OI	Osteogênese Imperfeita
ORF	Fase de leitura aberta – do inglês <i>open read frame</i>
PCR	Reação em cadeia da polimerase – do inglês <i>Polymerase Chain Reaction</i>
PIR	Protein International Resource Database
RNA	Ácido ribonucleico
RNAi	RNA de interferência
rRNA	RNA ribossomal
snRNA	Pequenos RNA nucleares – do inglês <i>small nucleolar RNA</i>
SOM	Mapas auto-organizáveis – do inglês <i>Self Organizing Maps</i>
tRNA	RNA transportador

Sumário

1 – Introdução	1
1.1 Objetivo	2
1.1.1 Objetivos Específicos	2
1.2 Justificativa	3
1.3 Metodologia	4
1.4 Organização do Trabalho	5
2 – Fundamentos de Bioinformática	6
2.1 Princípios de Biologia Molecular	6
2.1.1 DNA e RNA	6
2.1.2 Proteínas	9
2.1.3 Do DNA à Proteína: o Dogma Central	10
2.1.4 RNAs Funcionais	15
2.2 Evolução das técnicas laboratoriais de análises biomoleculares	16
2.2.1 Sequenciamento de Sanger	16
2.2.2 Análise de expressão gênica – Microarranjos	17
2.3 Análise de dados de Bioinformática	19
2.3.1 Armazenamento e busca em base de dados	20
2.3.2 Alinhamento de Sequências	21
2.3.3 Análises de agrupamento em dados de microarranjo	22
2.3.3.1 Pré-processamento de Dados de Microarranjo	24
3 – Algoritmos de Agrupamento	25
3.1 K-means	27
3.1.1 Formalização do Algoritmo	28
3.2 Mapas auto-organizáveis	30
3.2.1 Formalização do Algoritmo	31
3.3 Affinity Propagation	35
3.3.1 Formalização do Algoritmo	36
4 – Análise dos Dados de Microarranjo	39
4.1 Caracterização da Osteogênese Imperfeita	39
4.2 Pré-processamento	41
4.3 Análise dos Agrupamentos	42
4.3.1 Consistência dos Agrupamentos	45
4.3.2 Análise dos Perfis Encontrados	49

5 – Considerações Finais	55
---	-----------

Referências	57
------------------------------	-----------

1 Introdução

A evolução das técnicas e equipamentos de análise biomolecular, juntos à construção colaborativa de grandes bases para armazenamento destes resultados, tem gerado massas de dados cada vez maiores, inviabilizando uma análise e comparação dos mesmos sem auxílio de ferramentas computacionais (BALDI; BRUNAK, 2001; SAMISH et al., 2015).

Além disso, os componentes celulares¹ envolvidos em processos biomoleculares podem apresentar comportamentos e funções distintas, nem sempre havendo conhecimento científico suficiente para se inferir deterministicamente o comportamento de cada componente em determinada situação (BALDI; BRUNAK, 2001).

Diante deste cenário, a análise computacional de dados biológicos tem ganhado representatividade entre novas pesquisas, onde por vezes os sistemas computacionais atuam como ferramenta para organização e análise dos dados e, em outros momentos, modelos matemáticos e algoritmos para tais demandas são objetos de estudo.

É importante que tais modelos de análise computacional sejam capazes de fornecer parâmetros para o auxílio à inferência e tomada de decisão diante do dinamismo e ausência de determinismo dos sistemas biológicos. Os próprios modelos de representação dos fenômenos biológicos possuem exceções conhecidas e catalogadas. Além disso, as análises laboratoriais muitas vezes fornecem um recorte estático de processos dinâmicos fornecendo uma visão limitada do problema estudado.

Duas técnicas comumente utilizadas para tratamento de dados desta natureza são modelos probabilísticos e técnicas de aprendizado computacional, pois se adéquam melhor a análises onde apenas parte das variáveis envolvidas no problema podem ser observadas e o sistema estudado possui estrutura complexa. Estes modelos são capazes de inferir relações entre conjuntos de dados, sem conhecimento completo dos relacionamentos e interações entre os mesmos (BALDI; BRUNAK, 2001).

Uma das aplicações da análise de dados biomoleculares é o estudo da relação de fatores genéticos com algumas doenças conhecidas, buscando mecanismos de controle, prognóstico, ou inferência sobre o estágio das mesmas através da aplicação de modelos estatísticos, probabilísticos ou de reconhecimento de padrões na análise de dados de DNA, RNA ou proteínas, evidenciando relações inicialmente não observáveis entre estes dados. Alguns trabalhos como (DESRIAC et al., 2013; CORTESI et al., 2014; ZHAO; LIN, 2014), podem exemplificar aplicações de técnicas desta natureza.

Neste contexto, o objetivo deste trabalho perpassa pela utilização de técnicas

¹No escopo da discussão principalmente DNA, RNA e Proteínas

computacionais de reconhecimento de padrões na análise de dados laboratoriais relacionados à Osteogênese Imperfeita (OI), um conjunto de patologias de ordem genética, relativamente raras, caracterizadas pela produção defeituosa ou reduzida de colágeno. A produção deficiente de colágeno causa nos pacientes problemas principalmente na formação óssea, resultando em um quadro de osteopenia generalizada fazendo com que os pacientes tenham problemas de crescimento, ossos quebradiços e em variações mais severas da patologia, podem resultar na morte dos pacientes.

Duas pesquisas que precedem este trabalho constituíram uma base teórica e material que possibilitaram a execução do mesmo. No primeiro deles, Kulterer et al. (2007) fornecem evidências experimentais que as células tronco mesenquimais (CTMs) não sofrem alterações significativas no processo de cultura e expansão *ex vivo*, mostrando viabilidade do uso de técnicas deste tipo para terapias gênicas e alguns estudos de perfis de expressão gênica. Além disso, no mesmo trabalho é realizada uma análise *ex vivo* de perfis de expressão gênica durante a osteogênese (diferenciação de CTMs em células do tecido ósseo), no qual foram identificadas três fases distintas para a osteogênese: proliferação celular, maturação e mineralização da matriz óssea, elencando genes que caracterizam cada uma das fases.

Em outro trabalho, Kaneto (2011) realiza um estudo da expressão gênica durante a diferenciação osteogênica *in vitro* de CTMs em pacientes com Osteogênese Imperfeita, evidenciando perfis de expressão diferencial entre pacientes com a patologia e amostras de controle, onde são elencados genes relacionados à osteogênese que apresentam perfis de expressão diferentes das amostras de controle de indivíduos saudáveis. Os dados de expressão gênica utilizados neste trabalho foram provenientes das análises laboratoriais discutidas em (KANETO, 2011).

1.1 Objetivo

Este trabalho objetiva analisar comparativamente a aplicação de diferentes algoritmos de agrupamento para visualização e extração de informação utilizando dados de expressão gênica de pacientes portadores de Osteogênese Imperfeita, possivelmente evidenciando padrões de expressão diferencial que sugiram genes associados à patologia que direcionem novas análises laboratoriais para confirmação de tais relações.

1.1.1 Objetivos Específicos

- Discutir a relevância biológica de agrupamentos obtidos por diferentes algoritmos;
- Revelar padrões de expressão diferencial que sugiram genes associados à patologia;

- Ampliar conhecimentos acerca da aplicação de algoritmos de agrupamento em dados de expressão gênica;
- Contribuir com a compreensão de padrões de expressão gênica diferencial em pacientes portadores de Osteogênese Imperfeita.

1.2 Justificativa

Uma vez que a Osteogênese Imperfeita possui causas genéticas, o que dificulta sua cura, várias pesquisas nas últimas décadas são direcionadas para o desenvolvimento de terapias gênicas que atenuem os efeitos da patologia. No entanto, conforme apontado por Kaneto (2011), e ainda válido nos dias atuais, o conhecimento genético acerca das diferentes variações da patologia ainda é insuficiente para subsidiar o desenvolvimento de terapias gênicas eficazes para todos os tipos conhecidos da patologia. Apesar de existirem caracterizações moleculares para as variações catalogadas da patologia, apontando genes característicos envolvidos, as causas da doença são poligênicas e podem envolver diferentes genes para cada tipo, dificultando a caracterização completa dos genes envolvidos.

Além disso, por ser uma doença rara e ainda pouco explorada, não existem bases de dados amplas disponíveis para análise se comparada a doenças mais comuns como a doença de Alzheimer e Huntington², por exemplo. Assim, a exploração dos dados existentes com diferentes técnicas computacionais amplia sua importância, permitindo a inferência de novas relações entre genes envolvidos com a patologia, inicialmente não encontrados, e como direcionador de novas pesquisas laboratoriais, possibilitando a redução de custos através da seleção mais criteriosa de genes para análises mais complexas, como o PCR em tempo real.

Do ponto de vista computacional, o trabalho contribui com a aplicação do algoritmo de agrupamento *Affinity Propagation* na análise de perfis de expressão gênica em dados de microarranjo de cDNA buscando grupos de genes relacionados com a patologia, abordagem ainda pouco explorada, principalmente por conta da publicação do algoritmo ter menos de uma década. A comparação dos resultados deste algoritmo com dois algoritmos amplamente utilizados como o K-means e os Mapas auto-organizáveis (SOMs) tende a contribuir para a tomada de decisão em novas pesquisas sobre o uso do *Affinity Propagation* para análises similares.

Além da utilização de um algoritmo relativamente novo, a análise de agrupamentos por diferentes algoritmos que partem de princípios matemáticos distintos e

²Do ponto de vista do número de indivíduos portadores da patologia, a doença de Huntington atinge um número expressivamente menor de casos. No entanto, várias pesquisas nas últimas décadas se dedicam ao estudo molecular da doença e existem bases de dados especializadas com uma quantidade significativa de dados sobre pacientes com Huntington.

avaliação da consistência dos grupos de interesse através da similaridade dos grupos formados pelos diferentes algoritmos, apesar de sugerida na literatura sobre reconhecimento de padrões, é prática pouco comum em trabalhos na área de bioinformática, onde por vezes a escolha do algoritmo se dá pela afinidade do grupo de pesquisa pelo uso de determinada técnica ou pela adaptação de metodologia encontrada em trabalho semelhante. Assim, espera-se contribuir com a proposta de análise do mesmo conjunto de dados por diferentes algoritmos, mostrando que os mesmos podem evidenciar relações complementares, não perceptíveis pela observação direta dos dados.

1.3 Metodologia

As análises realizadas no presente trabalho foram desenvolvidas com dados laboratoriais obtidos por Kaneto (2011) (compartilhados em comunicação pessoal), limitando-se a análises computacionais, não sendo parte deste, experimentos laboratoriais, coleta de amostras ou atividades semelhantes.

Foram obtidos dados de expressão gênica de amostras de células tronco mesenquimais (CTMs) da medula óssea de três amostras de controle sadio, e cinco pacientes com Osteogênese Imperfeita, sendo dois portadores de OI Tipo I e três portadores de OI tipo III.

As células coletadas foram expandidas através de cultivo *in vitro* e induzidas para diferenciação em células do tecido ósseo, processo que tem duração de 21 dias. Os níveis de expressão gênica das células durante a osteogênese foram analisados antes da indução e com 2, 7, 12, 17 e 21 dias após a indução utilizando experimentos de microarranjo com o *Whole Human Genome Microarray Kit* (Agilent).

Os níveis de expressão gênica relativa das amostras foram obtidos através de análise das imagens das lâminas de microarranjo com o software *Agilent Feature Extraction* (versão 9.5.3.1) e exportados para planilhas no formato do *Microsoft Excel* (.xls) de modo que cada arquivo contém os níveis de expressão gênica de aproximadamente 44.000 genes da amostra de controle e de pacientes com OI tipo 1 e OI tipo 3, em 6 tempos cada.

Os níveis de expressão gênica contidos nos diferentes arquivos foram reunidos numa única matriz, na qual as linhas representam cada gene e as colunas cada amostra em determinado dia (exemplo: Controle Sadio - dia 2, OI Tipo III - dia 2).

Uma vez reunidos os dados, o pré-processamento e agrupamento dos genes foram realizados utilizando o ambiente computacional R (R Core Team, 2015) e os pacotes *Limma* (RITCHIE et al., 2015), para normalização, *kohonen* (WEHRENS; BUYDENS, 2007) e *apcluster* (BODENHOFER et al., 2011) para agrupamento com o Mapas

auto-organizáveis e *Affinity Propagation*, respectivamente.

Após as análises de agrupamento e escolha de grupos que apresentam padrões interessantes para compreensão da patologia foram utilizadas a base de genes do NCBI³ e a literatura disponível na busca de relações entre os genes destes grupos e características da patologia.

1.4 Organização do Trabalho

Este trabalho é organizado em cinco capítulos, sendo o primeiro a introdução do trabalho e o último considerações finais acerca do mesmo. Os capítulos 2 e 3 apresentam uma fundamentação teórica que suplementa a compreensão das análises realizadas e sustenta a metodologia adotada no trabalho apresentando. Estes capítulos apresentam, respectivamente, conhecimentos básicos de bioinformática e técnicas computacionais utilizadas nas análises de agrupamento. O quarto capítulo apresenta os resultados das análises de agrupamento com os dados de expressões gênica e discute relações entre os perfis observados que podem ser inferidas a partir destes resultados.

³<http://www.ncbi.nlm.nih.gov/gene>

2 Fundamentos de Bioinformática

A aplicação de técnicas de reconhecimento computacional de padrões em um conjunto de dados requer conhecimentos na área de domínio do problema estudado, para que seja possível avaliar a relevância e dar significado aos padrões encontrados com a aplicação de algoritmos desta natureza. É importante ainda que se conheça a natureza dos dados obtidos e que tipo de inferências podem ser realizadas acerca dos mesmos, para que seja possível formular e refutar novas hipóteses a partir da análise destes dados.

Alguns princípios de biologia molecular fundamentam as hipóteses assumidas nas análises realizadas neste trabalho, sendo essencial a compreensão dos mesmos para realização de inferências acerca dos padrões apontados pelas ferramentas computacionais utilizadas. Adicionalmente, o conhecimento básico das ferramentas e propostas da bioinformática, permite compreender esta área do conhecimento que instrumentaliza as análises computacionais realizadas neste trabalho.

2.1 Princípios de Biologia Molecular

2.1.1 DNA e RNA

Em meados do Século XX, um conjunto de descobertas científicas, revelou importantes características da estrutura molecular da célula, aproximando ainda mais os estudos de Biologia Molecular à discussão dos “fatores hereditários” propostos por Mendel, processo que resultou no estudo da genética a partir da análise de estruturas simples do ponto de vista molecular, hoje conhecidos por genes.

Experimentos iniciados por Griffith (1928), com diferentes linhagens da bactéria *Streptococcus pneumoniae*, evidenciaram que o Ácido Desoxirribonucleico (DNA) é o princípio transformante responsável por transferir a característica de virulência (capacidade de provocar doença e morte) entre as linhagens da bactéria. Esta é conhecida como primeira evidência que os genes (material hereditário) são compostos por DNA (GRIFFITHS et al., 2008).

Posteriormente, experimentos com o vírus fago T2 e a *Escherichia coli*, conduzidos por Hershey e Chase (1952), mostraram que o material transferido pelo vírus para as células bacterianas que levava à propagação do vírus era composto essencialmente por DNA, ratificando que DNA é o material hereditário, carregando portanto a informação genética dos organismos (GRIFFITHS et al., 2008).

Um dos motivos do entusiasmo da comunidade científica com as descobertas

acerca do DNA enquanto material hereditário, se deu por sua simplicidade do ponto de vista químico, sendo composto por fosfato, um açúcar, chamado desoxirribose e quatro bases nitrogenadas: adenina, citosina, guanina e timina (GRIFFITHS et al., 2008). Estes componentes se dispõem em grupos contendo: um composto de um grupo de fosfato, uma molécula de desoxirribose e uma das quatro bases nitrogenadas. Estes grupos são chamados de nucleotídeos e é comum na literatura se identificar um nucleotídeo pela primeira letra da base nitrogenada presente no nucleotídeo: A, C, G ou T.

Pesquisas desenvolvidas por Watson e Crick (1953) revelaram a estrutura tri-dimensional da molécula de DNA, conhecida como **dupla hélice** (ilustrada na Figura 1), que é formada por dois filamentos (ou fitas) de nucleotídeos ligados por pontes de hidrogênio entre suas bases nitrogenadas, torcidos como uma escada em espiral. Além disso, Watson e Crick (1953) concluíram que as fitas que compõem a dupla hélice são formadas de maneira complementar, de modo que um nucleotídeo A é ligado apenas com T e C, apenas com G. Com base nesta informação, o conhecimento sobre os nucleotídeos de uma das fitas permite a inferência da fita complementar (GRIFFITHS et al., 2008).

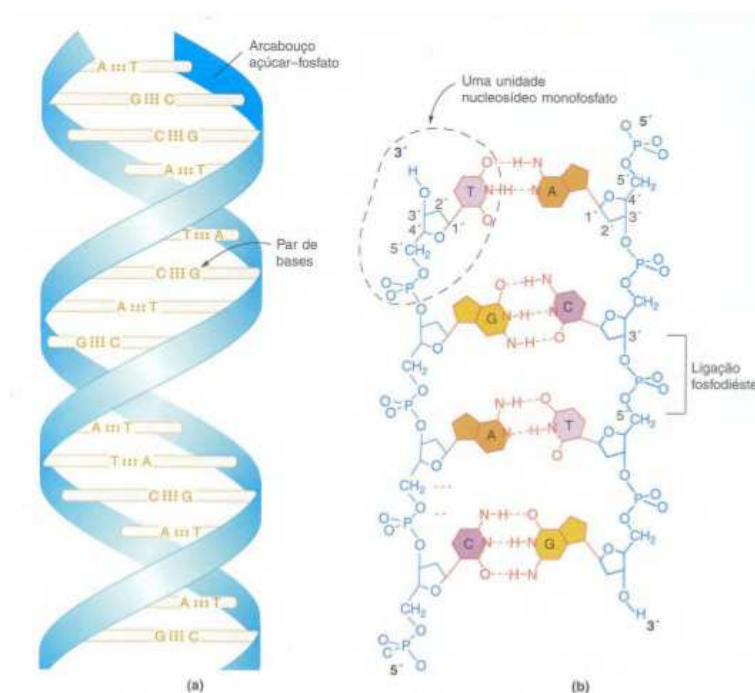


Figura 1 – Estrutura do DNA (GRIFFITHS et al., 2008).

O arcabouço das fitas que compõem a dupla hélice é formado por unidades alternadas de fosfato e desoxirribose conectados por ligações fosfodiéster. Como pode ser observado na Figura 1, as ligações fosfodiéster unem o átomo de carbono 5' ¹ ao

¹Os carbonos do grupamento açúcar são numerados de 1' até 5' (lê-se "um linha"). Cada fita é formada através da ligação do carbono 5' com o carbono 3' do grupamento açúcar adjacente.

átomo de carbono 3' da desoxirribose adjacente, dando à ligação uma polaridade (ou sentido) 5' – 3' ou 3' – 5', sendo que as duas fitas devem possuir polaridades inversas para que se forme a estrutura de dupla hélice (GRIFFITHS et al., 2008).

Um importante processo celular compreendido através do entendimento da estrutura celular é a replicação do DNA, que garante a preservação da informação genética no processo de reprodução celular. Para que a partir de uma molécula de DNA possa ser gerada uma nova molécula com a mesma sequência de nucleotídeos, uma enzima conhecida por DNA polimerase separa as fitas da dupla hélice, de modo que cada fita possa servir de molde para uma nova molécula de DNA. Uma vez expostas, as bases nitrogenadas em ambiente contendo nucleotídeos livres, uma nova fita é gerada para cada uma das fitas moldes, através da ligação de nucleotídeos complementares por pontes de hidrogênio (ALBERTS et al., 2010).

Como pode ser visto na Figura 2, as novas moléculas são formadas pela fita molde e uma fita complementar, de forma que cada uma das moléculas contém uma fita da molécula antiga e outra recém sintetizada. Este processo é conhecido como replicação semiconservativa (ALBERTS et al., 2010; GRIFFITHS et al., 2008).

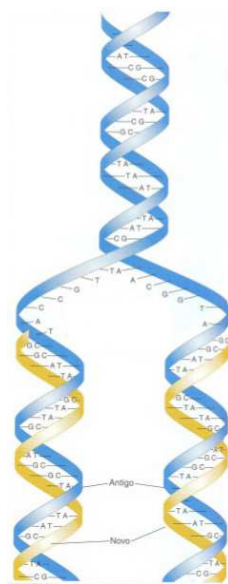


Figura 2 – Replicação do DNA (GRIFFITHS et al., 2008).

As novas fitas de DNA são sintetizadas a medida que a DNA-polimerase vai ampliando a região em que as pontes de hidrogênio foram quebradas, desfazendo a estrutura de dupla hélice. Essa região, mostrada na Figura 2, é conhecida como forquilha de replicação.

Apesar das mesmas regiões das fitas moldes estarem na forquilha de replicação ao mesmo tempo, as fitas não são polimerizadas na mesma velocidade e na mesma direção, uma vez que a DNA polimerase comumente é capaz de polimerizar apenas

no sentido 5' – 3' e as fitas são antiparalelas. Dessa forma, uma fita é polimerizada no sentido 5' – 3' de forma contínua e a outra é polimerizada em pequenos fragmentos contendo 1000 a 2000 nucleotídeos, aproximadamente, assim que está disponível um fragmento onde é possível polimerizar a nova fita no sentido 5' – 3'. Estes fragmentos são conhecidos por fragmentos de Okazaki. As fitas que são polimerizadas de forma contínua e descontínua são conhecidas como fita-líder (do inglês – *leading*) e fita-retardada (do inglês - *lagging*), respectivamente (ALBERTS et al., 2010; GRIFFITHS et al., 2008). A polimerização das fitas antiparalelas pode ser vista na Figura 3.

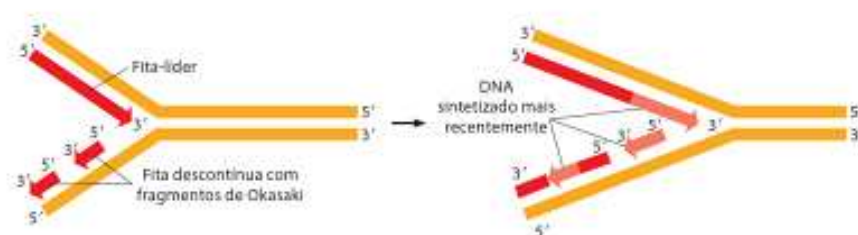


Figura 3 – Fragmentos de Okazaki na replicação do DNA (ALBERTS et al., 2010)

Um outro tipo de ácido nucleico existente nos organismos celulares envolvido no transporte da informação gênica é o Ácido Ribonucleico (RNA), que possui composição química similar ao DNA, diferenciando-se, em sua composição química apenas pelo tipo de açúcar, que é a ribose, ao invés da desoxirribose e a base nitrogenada uracila ao invés de timina. Similarmente ao DNA, nucleotídeos que compõem uma molécula de RNA, são geralmente referenciados pela primeira letra da base nitrogenada: A, C, G ou U (BERG et al., 2002).

O RNA, diferencia-se ainda do DNA pela sua estrutura, sendo formado por uma fita simples, ao invés de fita dupla, o que permite que o mesmo assuma diferentes estruturas tridimensionais, fazendo com que moléculas de RNA possam assumir diferentes papéis na maquinaria celular. Ainda diferente do DNA, os RNAs não são gerados por replicação, mas sim por um processo conhecido como transcrição, que é descrito na seção 2.1.3.

2.1.2 Proteínas

As proteínas são as macromoléculas responsáveis pelas mais diversas funções celulares, como controle de passagem de moléculas menores para dentro e fora da célula, transporte de mensagens entre células ou transmissão de sinais da membrana plasmática para o núcleo celular. Além disso, proteínas mais especializadas podem atuar como anticorpos, toxinas, hormônios, fibras elásticas etc. A compreensão do

funcionamento de um organismo a nível molecular, portanto, perpassa pelo estudo de suas proteínas, seus níveis de expressão e suas funções (ALBERTS et al., 2010).

Uma proteína, do ponto de vista químico, pode ser definida como um composto de monômeros conhecidos como aminoácidos, conectados por ligações peptídicas. A junção de um conjunto de aminoácidos também é chamada de polipeptídeo (LEHNINGER et al., 2005).

Os aminoácidos possuem composição química bastante semelhante, contendo um grupo carboxila e um grupo amino, ligados pelo mesmo átomo de carbono, diferenciando-se apenas por suas cadeias laterais, que possuem diferentes estruturas, tamanhos e carga elétrica. A estrutura geral de um aminoácido pode ser vista na Figura 4 (LEHNINGER et al., 2005).

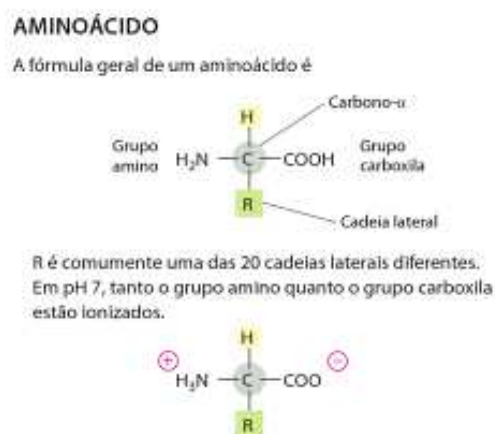


Figura 4 – Estrutura básica de um aminoácido (ALBERTS et al., 2010).

Existem 20 aminoácidos conhecidos que fazem parte da composição de proteínas que possuem propriedades químicas que fazem com que combinados em sequência definam a conformação tridimensional da proteína. Como é de se esperar pela complexidade dos processos nos quais estão envolvidas e na variedade de proteínas existentes com base na combinação de 20 aminoácidos, o estudo das funções das proteínas é relativamente complexo, comumente tendo sua função associada à sua estrutura tridimensional e sendo, por vezes, necessário analisar as interações entre conjuntos de proteínas para a compreensão da função de determinada proteína em um organismo.

2.1.3 Do DNA à Proteína: o Dogma Central

Conforme já foi dito na seção 2.1.1, uma série de descobertas mostrou que o DNA é responsável pelo armazenamento da informação genética. Por outro lado, grande parte das funcionalidades internas e externas a uma célula são realizadas por proteínas. Uma importante peça para o quebra-cabeças da biologia molecular é processo através

do qual a informação genética é transmitida do DNA para outras estruturas moleculares de modo a gerar moléculas funcionais como proteínas, determinando o fenótipo de um organismo.

A sequência de nucleotídeos presentes no genoma de um organismo define que proteínas ele será capaz de gerar. No entanto, as cadeias de aminoácidos não são sintetizadas diretamente através do DNA. Um conjunto de mecanismo de regulação da célula definem em que condições determinado trecho do DNA será expresso. Esse trecho de DNA é transcrito inicialmente em uma molécula de RNA (um RNA mensageiro ou mRNA), que é posteriormente traduzida em uma proteína. Este princípio que define o fluxo de informação genética nos organismos, ilustrado na Figura 5, é conhecido por **dogma central** (ALBERTS et al., 2010).

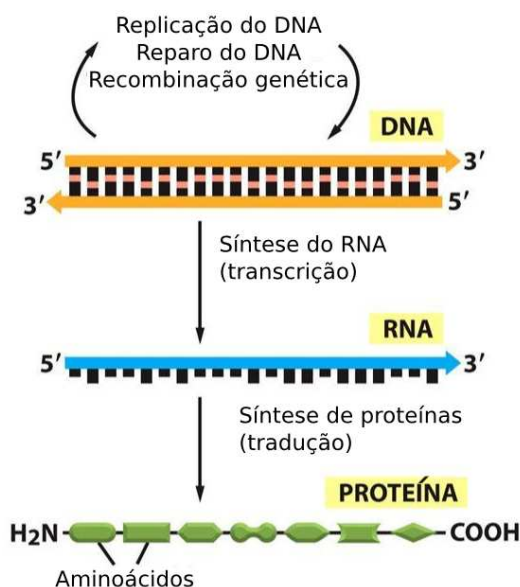


Figura 5 – Dogma Central: transcrição e tradução (ALBERTS et al., 2010).

Cabe ressaltar, que existem exceções ao dogma central, principalmente por conta de alguns tipos de RNA que podem assumir estruturas tridimensionais mais complexas e executar funções nos organismos (ex: RNA ribossomais, RNA de interferência etc.), mesmo assim ele continua válido enquanto princípio norteador no estudo da biologia molecular.

O processo de transcrição do DNA em RNA ocorre de maneira muito similar à replicação do DNA. A RNA-polimerase se fixa a um sítio específico na molécula de DNA, conhecido por região promotora, e inicia a separação da dupla hélice, de modo a manter uma região da fita de DNA exposta para servir de fita molde. A nova fita de RNA é polimerizada através da adição de ribonucleotídeos, de forma complementar à região exposta do DNA, ligando A com U, G com C, T com A e C com G. Ao chegar no final da

sequência, o transcrito de RNA é desligado da molécula de DNA e a RNA-polimerase também se dissocia do DNA molde (LODISH et al., 2003).

A transcrição, conforme ilustrado na Figura 6 pode ser conceitualmente dividida em três etapas, iniciação, alongamento e terminação, que são: a fixação da região promotora e catalisação de primeiras ligações fosfodiéster, inserção das demais bases nitrogenadas por complementaridade, separação do RNA da fita molde e dissociação da RNA-polimerase do DNA molde.

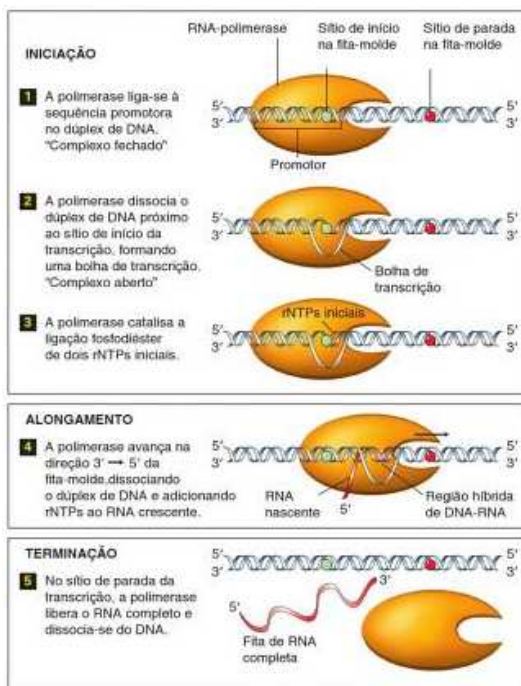


Figura 6 – Etapas da Transcrição (LODISH et al., 2003).

Uma vez transcrita a informação genética no mRNA, o próximo passo para que possa ser sintetizada a proteína é a tradução. Este processo ocorre no ribossomo, que é formado por um conjunto de proteínas e RNAs específicos conhecidos como RNAs ribossomais ou rRNA. Outro tipo de RNA importante na tradução é o RNA transportador ou tRNA, que possui a função de se associar a um aminoácido específico e possui uma região de três bases nitrogenadas, conhecida como anticódon, complementares a um trio de bases nitrogenadas do mRNA, conhecido como códon, ao qual se ligará, permitindo a inserção de um novo aminoácido na cadeia (LODISH et al., 2003).

Dessa forma, a sequência de aminoácidos em uma proteína é definida pela sequência de códons existentes no mRNA. Cada códon no mRNA é associado a um tRNA específico, que se associa a um aminoácido específico, permitindo a inferência da sequência de aminoácidos a ser gerada, a partir da sequência de códons no mRNA. Esta relação entre códons e aminoácidos é conhecida como código genético (ALBERTS et al., 2010).

Apesar de existirem 4 tipos de nucleotídeos, consequentemente 64 possibilidades de arranjos lineares de 3 em 3, existem somente 20 aminoácidos conhecidos que formam proteínas. Isto se explica pelo fato de mais de um códon poder gerar o mesmo aminoácido, inserindo redundâncias no código genético. Além de aminoácidos, existem códons específicos que atuam como indicadores de início e término da região de tradução (ALBERTS et al., 2010). A tabela 1 mostra os aminoácidos gerados por cada códon. Cabe ressaltar que apesar deste código ser válido para a maioria dos organismos conhecidos, já existem resultados que mostram exceções ao código genético.

Tabela 1 – Código Genético - Adaptado de (LODISH et al., 2003)

1ª Posição		2ª Posição				3ª Posição
	U	C	A	G		
U	Fenilamina	Serina	Tirosina	Cisteína	U	
	Fenilalanina	Serina	Tirosina	Cisteína	C	
	Leucina	Serina	Stop	Stop	A	
	Leucina	Serina	Stop	Triptofano	G	
C	Leucina	Prolina	Histidina	Arg	U	
	Leucina	Prolina	Histidina	Arginina	C	
	Leucina	Prolina	Glutamina	Arginina	A	
	Leucina	Prolina	Glutamina	Arginina	G	
A	Isoleucina	Treonina	Asparagina	Serina	U	
	Isoleucina	Treonina	Asparagina	Serina	C	
	Isoleucina	Treonina	Lisina	Arginina	A	
	Metionina (Start)	Treonina	Lisina	Arginina	G	
G	Valina	Alanina	Ácido aspártico	Glicina	U	
	Valina	Alanina	Ácido aspártico	Glicina	C	
	Valina	Alanina	Ácido glutâmico	Glicina	G	
	Valina	Alanina	Ácido glutâmico	Glicina	A	

De modo análogo à transcrição, a tradução possui fases de início e término da sequência. Nem toda sequência transcrita em um mRNA é decodificada em proteína. O ribossomo inicia o processo de tradução assim que encontra um códon específico, o AUG. Após iniciar a tradução, os tRNA ligam-se com os códons do mRNA, adicionando aminoácidos à nova proteína. A tradução é encerrada assim que é encontrado um dos códons de terminação UAA, UAG ou UGA. A região presente entre o códon de iniciação (do inglês - *start codon*) e um dos códons de terminação (do inglês - *stop*

condon) é conhecida como região de leitura. Os códons de terminação não adicionam aminoácido à sequência polipeptídica, já o códon de iniciação (AUG), pode significar a inserção de uma Metionina (Met), inclusive no interior na sequência (ALBERTS et al., 2010; LODISH et al., 2003). A Figura 7 mostra o funcionamento da tradução.

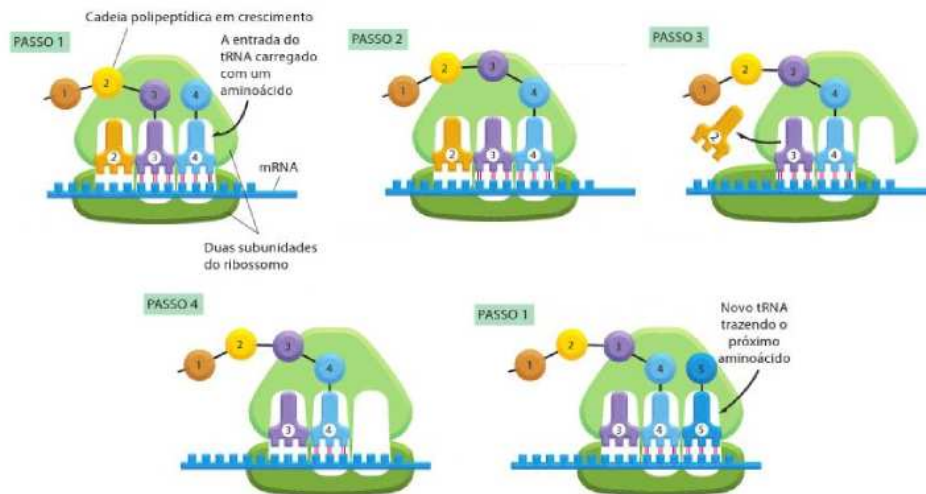


Figura 7 – Etapas da Tradução (ALBERTS et al., 2010).

Apesar da evolução do conhecimento em torno do código genético e dos processos de transcrição e tradução, inferir que tipo de proteína é gerada a partir de um fragmento de DNA ainda é uma atividade complexa, dado a inexatidão e dinâmica de fenômenos biológicos. Além das exceções conhecidas do código genético, a existência de sequências no DNA que não codificam proteínas, somado a outras incertezas inerentes ao funcionamento dos organismos, dificulta ainda mais análises deste tipo.

Através do conhecimento do dogma central somado ao código genético, para grande parte dos procariotos é possível inferir que tipo proteína é gerada a partir de sequências de DNA e, inclusive construir hipóteses sobre suas funções através de comparações com outras proteínas similares conhecidas. Já nos eucariotos, a transcrição e a tradução não acontecem de forma concorrente, uma vez que a transcrição acontece no núcleo celular e a tradução no citoplasma. Neste caso, a transcrição gera um precursor de mRNA, chamado de pré-mRNA, que passa por uma etapa de pré-processamento antes de ser traduzido em proteína (ALBERTS et al., 2010).

No primeiro passo do pré-processamento, o pré-mRNA passa por um capeamento no qual são adicionados componentes químicos (chamados de quepe) na extremidade 5', que permitirá que a célula diferencie o mRNA de outros tipos de RNA fora do núcleo celular (ALBERTS et al., 2010). Também é adicionado um complemento na extremidade 3', conhecido por calda de poli-A (LODISH et al., 2003).

Uma importante etapa do pré-processamento, é a separação de regiões codifi-

cantes e não codificantes. Esta descoberta se deu como consequência da percepção que nos eucariotos, as porções de DNA que serão codificadas em proteínas (chamadas de éxons), são transcritas intercaladas por porções muito maiores de sequências que não serão codificadas (íntrons). A remoção dos íntrons de um pré-mRNA é conhecida como *splicing* do RNA. Além da existência de sequências não codificantes, pesquisas posteriores revelaram em eucariotos a capacidade de realizar em diferentes situações *splicing* em regiões diferentes, fazendo com que um mesmo pré-mRNA possa gerar diferentes mRNAs, consequentemente diferentes proteínas. Esta propriedade é apresentada na literatura como *splicing* alternativo (ALBERTS et al., 2010).

2.1.4 RNAs Funcionais

Como já foi discutido nesta seção, apenas os mRNAs podem ser traduzidos em proteínas. Além disso, diferente do DNA que possui estrutura bifilamentar, a estrutura unifilamentar do RNA permite que ele assuma estruturas tridimensionais e faça parte de mecanismos funcionais na maquinaria celular, como é caso dos tRNAs e dos rRNA. Estes e outros tipos de RNA que atuam diretamente em algum mecanismo funcional da célula, sem serem codificados em proteínas, são conhecidos como RNAs Funcionais ou RNAs não codificantes (ALBERTS et al., 2010).

Um outro tipo de RNA funcional conhecido são os pequenos RNA nucleares ou snRNA (do inglês - *small nucleolar RNA*) que juntamente com complexos de proteínas específicas formam o spliceossomo, estrutura molecular responsável pelo *splicing* do pré-mRNA.

Outros dois tipos de RNA funcionais conhecidos por atuar na repressão da expressão gênica são os RNA de interferência (RNAi) e os micro-RNAs (miRNA), ambos compostos por pequenas sequências de nucleotídeos (geralmente menos de 100 bases), e atuam na repressão pós-transcricional, fazendo com que o mRNA não seja codificado em proteína.

Tendo sido descobertos a pouco mais de duas décadas, os miRNA ganharam um espaço significativo nas pesquisas em torno de marcadores biomoleculares para alterações fisiológicas. Alguns estudos apresentam indícios de um mecanismo central de redes de miRNAs, atuando na regulação da expressão gênica em condições de stress patofisiológico (MENDELL; OLSON, 2012).

Com bases nestes princípios funcionais dos miRNAs várias pesquisas os apontam como bons marcadores biomoleculares de alterações fisiológicas, principalmente relacionados a doenças cardíacas e alguns tipos de câncer como em (LI et al., 2012; REID et al., 2011; WITTMANN; JäCK, 2010; DI STEFANO et al., 2011). Apesar do uso comum para alterações com origem patológicas, níveis de expressão de miRNAs também po-

dem indicar outras condições, a exemplo dos estudos de alteração de alguns miRNA específicos em períodos de gravidez em humanos apontados por Reid et al. (2011).

2.2 Evolução das técnicas laboratoriais de análises biomoleculares

A análise e simulação computacional de dados biológicos teve como origem de suas demandas a evolução das técnicas de análises biomoleculares laboratoriais, que resultaram na produção de grandes massas de dados que atualmente são compartilhadas em bancos de dados especializados, mantidos e compartilhados pela comunidade científica (LIEW et al., 2005; LIN et al., 2006; DOUGHERTY, 2005). Um grande exemplo de projeto possibilitado por essas novas técnicas laboratoriais é o sequenciamento completo do genoma humano, realizado e publicado por dois grupos diferentes de cientistas (CONSORTIUM, 2002; VENTER, 2001).

Nesta seção são apresentadas duas técnicas que revolucionaram a análise de dados biomoleculares e foram precursores de uma geração de equipamentos que produzem grandes massas de dados biológicos. Uma vez obtidos dados em quantidades não analisáveis de forma eficiente apenas com o olhar do pesquisador, surge a demanda de aplicações de modelos computacionais para armazenamento, busca de dados e descoberta de informação nestas bases de dados.

2.2.1 Sequenciamento de Sanger

Um passo que marcou a evolução do sequenciamento de DNA, foi uma técnica desenvolvida por Sanger e Nicklen (1977), que permitiu de maneira eficiente o sequenciamento de cadeias relativamente grandes de DNA.

A técnica de sequenciamento, que passou a ser conhecida por sequenciamento de Sanger, é baseada na reprodução da replicação do DNA, com alterações que permitem localizar a posição de nucleotídeos que foram quimicamente alterados para marcar sua posição na sequência (SANGER; NICKLEN, 1977). Segue um resumo da técnica:

- O trecho de DNA que se deseja sequenciar é inserido em solução contendo DNA-polimerase e bases nitrogenadas livres para que novas sequências possam ser geradas;
- Algumas dessas bases, que formam um nucleotídeo específico (A, C, G ou T), são quimicamente modificados para que a DNA-polimerase interrompa a geração da cadeia após a inserção deste nucleotídeo. A replicação repetitiva do trecho faz com que a polimerização seja interrompida em diferentes pontos da sequência;

- As sequências de diferentes tamanho, são separadas em gel de acrilamida por eletroforese, fornecendo informação sobre a posição do nucleotídeo gerado pela base modificada;
- A execução da eletroforese em paralelo com material dos 4 nucleotídeos revela a sequência do trecho analisado.

A técnica de Sanger e suas variações permitiram o sequenciamento de trechos e do genoma completo de algumas espécies. Tais pesquisas, somadas à corrida em busca do sequenciamento completo do genoma humano, impulsionaram evoluções nas tecnologias de sequenciamento, gerando o que hoje são conhecidos como sequenciadores de nova geração.

Os sequenciadores de nova geração diferenciam-se da técnica de Sanger, principalmente pela clonagem do DNA que é feita *in vitro*, ao invés da utilização de bactérias; a sequência é determinada pela síntese de um novo DNA por complementaridade sem terminação química de cadeia; e várias amostras podem ser sequenciadas simultaneamente com a utilização de técnicas massivamente paralelas (ANDERSON; SCHRIJVER, 2010). Em (METZKER, 2010) pode ser encontrada uma discussão mais detalhada sobre as tecnologias de sequenciamento de nova geração e comparativos de desempenho.

2.2.2 Análise de expressão gênica – Microarranjos

O conhecimento do genoma completo de um organismo por si só, revela poucas informações sobre seu fenótipo. Mesmo que se possa inferir a sequência de aminoácidos a ser gerada por um gene específico, apenas através da sequência de nucleotídeos, não se pode determinar em que condições e em que quantidade este gene será expresso e esse produto gênico gerado. Exemplo desta complexidade é o fato do DNA de um organismo multicelular ser o mesmo em todos os tipos de células, apesar da maioria delas não expressar mais que metade dos genes em seu funcionamento (ALBERTS et al., 2010).

Um organismo multicelular, possui genes que codificam proteínas essenciais ao funcionamento de processos comuns a todas as células, sendo estes expressos em todos tipos de células. Os demais genes são expressos como respostas a sinais internos e externos à célula, em diferentes proporções. Além disso, o fato de um gene ser transcrito, não garante que será traduzido, existem vários mecanismo de controle da célula, como os miRNA e os RNAi, que podem fazer com que o mRNA não seja traduzido.

Assim, para compreensão da função de proteínas codificadas por determinados genes, ou a inferência de genes associados a determinadas propriedades do organismo, por exemplo resistência a doenças ou outras condições estresse, é importante se analisar

de que maneira os genes presentes no genoma deste organismo são expressos nestas condições, ou que modificações no fenótipo são geradas por variação na expressão de determinados genes.

Uma técnica de suma importância no estudo de expressão gênica é a Reação em Cadeia de Polimerase ou PCR (do inglês – *Polymerase Chain Reaction*), que permite a amplificação de uma sequência de DNA ou cDNA, de modo que se possa verificar a existência de uma sequência específica ou quantificar a expressão de determinado gene em diferentes indivíduos ou diferentes condições ambientais.

A amplificação de DNA via PCR se dá através da mistura do DNA que, possivelmente contém a sequência que se deseja amplificar, com uma DNA-polimerase resistente a altas temperaturas e iniciadores (primers), que se ligam de forma complementar no início da região que se deseja amplificar. A solução é aquecida para desnaturar as pontes de hidrogênio que ligam as bases nitrogenadas, deixando expostas as fitas do DNA, nas quais os iniciadores são flanqueados e as novas sequências são polimerizadas no sentido 5' – 3'. O processo pode ser repetido até que se obtenha quantidade suficiente da sequência que se deseja amplificar, para que torne visível para revelação em gel por eletroforese (LEHNINGER et al., 2005).

Esta técnica também pode ser aplicada em sequências de mRNA, para se estudar perfis de expressão gênica de organismos em diferentes condições. Neste caso, é utilizada a enzima transcriptase reversa, que polimeriza DNA a partir de mRNA por complementaridade de bases nitrogenadas. O DNA gerado é conhecido como DNA complementar ou cDNA.

Com a evolução das tecnologias de análise de biologia molecular surgiram os Microarranjos ou Chips de DNA, que possibilitam a análise de perfis de expressão automática de centenas, ou até milhares genes simultaneamente. A análise é feita através circuitos controlados computacionalmente, o que permite a análise paralela de várias amostras, com uma boa precisão. O resultado de uma análise de microarranjo é uma tabela contendo o nível de expressão de cada gene para cada uma das amostras (LODISH et al., 2003). A Figura 8 mostra um modo de visualização de um microarranjo, onde intensidades de cores são utilizadas para representar de maneira visualmente mais perceptível os níveis de expressão dos genes.

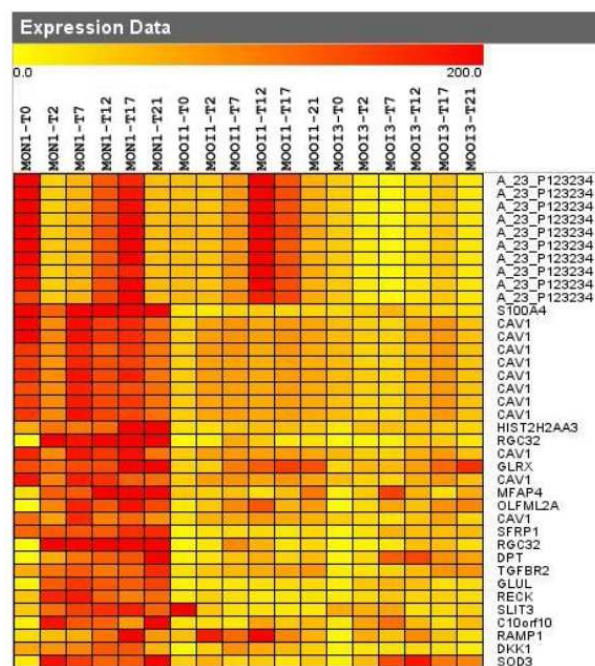


Figura 8 – Exemplo de visualização de genes diferencialmente expressos no programa Genes (KANETO, 2011).

2.3 Análise de dados de Bioinformática

Além da grande quantidade de dados produzidos com a evolução das técnicas e equipamentos laboratoriais de biologia molecular, Kasabov (2007) ressalta a complexidade e a dinâmica envolvida nas interações biomoleculares, entre genes, RNAs, proteínas e os diversos componentes reguladores da célula. Prever, por exemplo, que proteína será gerada a partir de um pré-mRNA em um eucarioto pode ser uma atividade difícil, uma vez que o *splicing* alternativo pode fazer com que diferentes regiões ora sejam tratadas como éxons, ora como íntrons.

Assim, os desafios da bioinformática devem ir além de armazenar e compartilhar este grande volume de dados produzidos. As respostas das questões biológicas dependem da descoberta de informações a partir dos dados obtidos, tarefa que ganha complexidade com o crescimento da quantidade de dados disponíveis. Kasabov (2007) define como maior objetivo atual da área de Bioinformática a modelagem das interações biomoleculares e a extração de padrões significativos a partir dos modelos obtidos. Destaca ainda os seguintes problemas de pesquisa relacionados ao reconhecimento de padrões e bioinformática:

- Reconhecimento de padrões em dados de DNA, como reconhecimento de região promotora, identificação de ORF, etc.;

- Reconhecimento de padrões em dados de RNA, como predição de regiões de *splicing*, predição de éxons e íntrons;
- Modelagem de perfis de expressão gênica associado a diferentes patologias;
- Predição de estruturas de proteínas, atividade de grande importância no projeto de fármacos;
- Criação de sistemas de apoio a decisões médicas baseados em informações biomoleculares e clínicas, possibilitando melhores diagnósticos e prognósticos.

As subseções que seguem apresentam algumas importantes áreas de pesquisa atual da bioinformática que contribuem com a compreensão deste trabalho e o contexto no qual se insere.

2.3.1 Armazenamento e busca em base de dados

Uma vez ampliada a capacidade de produção de dados, apêndices de publicações científicas passam a não comportar o volume de dados relacionados a estas pesquisas. Várias bases de dados biológicos foram criadas, com o propósito de utilizar técnicas de armazenamento e busca de dados convencionais na computação, para compartilhamento de dados biológicos pela comunidade científica.

Algumas das bases de dados biológicos inicialmente propostas, se tornaram populares e recebem diariamente dados oriundos de pesquisas em diversas áreas e diversas regiões do mundo, ao tempo que disponibilizam mecanismos eficientes de busca e comparação de novos dados com os preexistentes.

Entre as bases de dados de sequências de nucleotídeos, merecem um destaque especial o GenBank, mantido pelo *National Center for Biotechnology Information* (NCBI²), o *European Molecular Biology Laboratory* (EMBL)/*EBI Nucleotide Sequence Database*³ e o *DNA Databank of Japan* (DDBJ)⁴, as quais, seus mantenedores juntos formam o *International Nucleotide Sequence Database Collaboration*⁵, que diariamente compartilham as alterações e registros em suas bases de dados (MOUNT, 2004).

Além das bases de dados existentes para nucleotídeos, existem bases especializadas em dados relacionados a proteínas, em todos os níveis de estrutura, entre elas, a base de dados suíça, SwissProt⁶, o *Protein International Resource Database* (PIR)⁷ do National Biomedical Research Foundation in Washington. Existem ainda, bases

²<http://www.ncbi.nlm.nih.gov/>

³<http://www.embl.de/>

⁴<http://www.ddbj.nig.ac.jp/>

⁵<http://www.insdc.org/>

⁶<http://www.uniprot.org/>

⁷<http://www.uniprot.org/>

especializadas em dados de RNAs não codificantes como o NONCODE⁸, uma base de dados especializada em ncRNAs e a *LncRNADisease Database*⁹, uma base de dados especializada em RNAs não codificantes longos. Em (MOUNT, 2004) pode ser encontrada uma lista com diversas bases de dados biológicos populares, que compartilham dados de sequências de DNA, RNA e proteínas.

2.3.2 Alinhamento de Sequências

O alinhamento de sequências tem por objetivo as dispor de maneira que permita uma análise comparativa das sequências, de modo a revelar relações estruturais, funcionais ou evolutivas através da similaridade entre as sequências (SCHULER, 2001). Tais sequências podem ser resultantes de análises de DNA ou RNA (nucleotídeos) ou de proteínas (aminoácidos).

Os primeiros métodos de comparação de sequências, baseavam-se na comparação das sequências em sua totalidade, buscando um alinhamento que refletisse sua similaridade. Esta estratégia é conhecida como **alinhamento global** (SCHULER, 2001). A proposta inicial de alinhamento global foi apresentada em (NEEDLEMAN; WUNSCH, 1970), para alinhamento de sequências de aminoácidos.

A evolução das pesquisas em biologia molecular revelou a existência de sub-regiões de sequências de DNA e proteínas, conservadas durante a evolução das espécies, responsáveis pela funcionalidade do produto gênico resultante. As regiões alteradas por inserções ou deleções no processo evolutivo são comumente pouco informativas (MOUNT, 2004). Especificamente em relação a sequências de DNA, como apenas os éxons são transcritos e possivelmente traduzidos, divergências nos íntrons podem não ser significativas se o objetivo for avaliar a similaridade das proteínas resultantes (SCHULER, 2001).

Com bases nestes pressupostos que justificam a busca de sub-regiões que fornecem bons alinhamentos entre as sequências, Smith e Waterman (1981) propuseram alterações nos algoritmos de Needleman para comportar a busca de sub-regiões com bons alinhamentos entre duas sequências (MOUNT, 2004).

Muitas vezes ao analisar uma nova sequência, principalmente de aminoácidos, a informação de proteínas similares em outros organismos, sua estrutura e funcionalidade, podem permitir a inferência de informações importantes sobre essa proteína. Além da comparação entre duas sequências, conhecido como alinhamento simples ou par a par, outra técnica de comparação muito utilizada é o alinhamento múltiplo, que permite a comparação simultânea de um conjunto de sequências (BARTON, 2001).

⁸<http://www.noncode.org/>

⁹<http://cmbi.bjmu.edu.cn/lncrnadisease>

Especificamente em relação ao DNA, a existência de uma região altamente conservada em diferentes espécies pode fundamentar a suposição desta região ser codificante.

Atualmente, várias ferramentas computacionais, muitas delas disponíveis na Web, fornecem recursos para alinhamento simples e múltiplo de sequências, principalmente com buscas em bases de dados preexistentes. Um conjunto de ferramentas importantes para comparação de sequências disponibilizado pelo NCBI é o BLAST (*Basic Local Alignment Search Tool*), que permite inclusive o alinhamento simples de uma sequência com várias outras existentes na base de dados. A tabela 2 mostra as principais ferramentas de alinhamento disponíveis no BLAST. Já para o alinhamento múltiplo, existem ferramentas como o *Muscle* e o *ClustalW*, que permitem além do alinhamento múltiplo, algumas análises filogenéticas.

Tabela 2 – Principais Ferramentas Disponíveis no BLAST.

Ferramenta	Descrição
Nucleotide BLAST	Busca em uma base de nucleotídeos através de uma sequência de nucleotídeos de entrada
Protein BLAST	Busca em uma base de proteínas através de uma sequência de aminoácidos de entrada
BLASTX	Busca em uma base de proteínas através de uma sequência de nucleotídeos de entrada traduzidos
TBLASTN	Busca em uma base de nucleotídeos traduzidos através de uma sequência de aminoácidos de entrada
TBLASTX	Busca em uma base de nucleotídeos traduzidos através de uma sequência de nucleotídeos de entrada traduzidos.

2.3.3 Análises de agrupamento em dados de microarranjo

Um dos principais objetivos da análises de dados de expressão gênica é a identificação de genes que têm alterações significativas em seus níveis de expressão sobre determinada condição experimental. Baseado na hipótese que genes que tem sua expressão aumenta ou reduzida de forma similar (co-expressão) quando expostos a diferentes condições podem estar envolvidos em processos funcionais semelhantes ou serem regulados pelos mesmos mecanismo de controle, a análise da relação entre os padrões de expressões dos genes pode revelar informações acerca de genes envolvidos em determinado processo ou condição biológica (Qizheng Sheng, Yves Moreau, Frank De Smet, Kathleen Marchal, 2005).

Alguns exemplos de aplicações de algoritmos de agrupamento na análise de

perfis de expressão gênica são:

- Kaneto (2011): Analisou perfis de expressão gênica com dados de microarranjo de pacientes portadores de Osteogênese Imperfeita;
- Van 't Veer et al. (2002): Analisou perfis de expressão de 5000 genes em grupos de controle e pacientes com diferentes estágios de câncer de mama, mostrando que existem padrões similares de expressão que podem auxiliar no prognóstico destes pacientes;
- Slavkov et al. (2006): Utilizou algoritmos de agrupamento (*Predictive Clustering Trees*) para classificação de pacientes com diferentes estágios da doença de Huntington a partir de dados de expressão gênica obtidos com microarranjos;
- Covell et al. (2003): Utilizou Mapas auto-organizáveis para agrupamento de dados de microarranjo de expressão gênica de 14 classes de tumor e um grupo de controle.

Mesmo a partir da suposição que genes co-expressos estão correlacionados funcionalmente, a tarefa de obter informações a partir de perfis de expressão gênica é cercada de incertezas, pois diferentes técnicas de normalização dos dados, diferentes algoritmos, ou até mesmo diferentes parâmetros podem revelar informações distintas sobre os genes expressos (QUACKENBUSH, 2001). O capítulo 3 apresenta e discute conceitos gerais de algoritmos de agrupamento, apresentando a formalização matemática dos algoritmos utilizados neste trabalho, ao tempo que apresenta o contexto atual de pesquisas que utilizam técnicas de agrupamento para análise de dados de expressão gênica, fornecendo uma visão mais aprofundada deste tipo de análise de bioinformática.

Por fim, é importante ressaltar que análise de microarranjos é apenas uma das etapas na descoberta de genes associados a determinada condição estudada, uma vez que fornecem um retrato estático de um processo dinâmico: a expressão gênica. Após selecionados genes de interesse, técnicas mais precisas como PCR de tempo real podem ser utilizadas para confirmação da expressão diferencial dos genes selecionados. Além disso, a simples expressão de um gene não garante que o produto gênico será produzido. Vários mecanismos de controle pós-transcricional, miRNA por exemplo, podem atuar silenciar a tradução do mRNA em proteína.

Neste contexto, a aplicação de técnicas de agrupamento em dados de microarranjo tem por principal finalidade fornecer uma melhor visualização dos perfis de expressão diferencial e associada a técnicas estatísticas permite a redução da quantidade de amostras para confirmações experimentais posteriores.

Uma etapa essencial que deve anteceder a análise de agrupamento é o pré-processamento dos dados obtidos, de modo a evitar que os ruídos inerentes às atividades

experimentais e a digitalização para análise computacional não levem a conclusões equivocadas acerca das informações obtidas. Segue uma pequena discussão sobre o pré-processamento padrão de dados de microarranjo.

2.3.3.1 Pré-processamento de Dados de Microarranjo

A análise de microarranjos verifica o nível de expressão de um gene em uma amostra de acordo com sua intensidade de fluorescência relativa a uma amostra de referência. No fim do experimento, é registrada uma imagem contendo estes pontos com diferentes intensidades de luminosidade (*spots*), a partir dos quais o nível de expressão do gene é calculado através de técnicas de processamento de imagens.

Este processo de obtenção do nível de expressão possui imprecisões comuns a diferentes aparelhos, de modo que já existe na literatura técnicas de pré-processamento aplicáveis à análise de dados de microarranjo que deve ser realizada antes da submissão dos dados a rotinas de agrupamento. Duas etapas de pré-processamento se destacam neste contexto: a correção de *background* e a normalização entre as amostras.

A correção de *background* tem por finalidade corrigir erros que podem resultar da identificação de reflexos dos *spots* no chip de DNA (ou cDNA) e outros ruídos como parte do nível de expressão das amostras (SILVER et al., 2009). Em (RITCHIE et al., 2007) é apresentado o algoritmo *normexp* que é capaz de realizar de forma eficiente a correção de *background* de dados de microarranjo obtidos em imagens multicanais, que posteriormente foi melhorado por Silver et al. (2009). A versão melhorada do algoritmo é atualmente utilizada no pacote de bioinformática Limma (RITCHIE et al., 2015) do ambiente estatístico R (R Core Team, 2015).

A normalização das amostras, por sua vez, tem por objetivo eliminar variações entre os experimentos, uma vez que as amostras podem ser analisadas em diferentes experimentos de microarranjo, fazendo com que condições ambientais, alterações nos reagentes utilizados e outras interferências externas ao experimento gerem ruídos na captura da intensidade dos *spots* de uma análise para outra.

Um algoritmo muito utilizado para normalização entre as amostras é a normalização quantílica, proposta por Bolstad et al. (2003), que tem por objetivo uniformizar a distribuição de intensidades entre as amostras, evitando que alterações nos experimentos levem à percepção equivocada que amostras se expressaram de maneiras diferentes. Alguns trabalhos propõem aprimoramentos à normalização quantílica, exemplo destes é (HU; HE, 2007), que combina a decomposição em valores singulares após a normalização quantílica para recuperar parte da informação perdida na normalização.

3 Algoritmos de Agrupamento

Os algoritmos de agrupamento tem por finalidade a descoberta de padrões em conjuntos de dados, de modo que estes possam ser agrupados, possibilitando a percepção de similaridades entre os padrões e a inferência de conhecimento acerca dos dados. No estudo de reconhecimento de padrões, estes algoritmos são também chamados de aprendizagem não supervisionada (THEODORIDIS; KOUTROUMBAS, 2009).

O objetivo da aplicação de um algoritmo de agrupamento é a organização dos dados de entrada em grupos, de forma que os dados sejam mais similares dos que estão no mesmo grupo que os de outros grupos (RUSPINI, 1969). Matematicamente, Muthukalathi et al. (2014) definem a tarefa de agrupamento de padrões de um conjunto X em K grupos como a definição de uma partição P de X com $P = \{P_1, P_2, \dots, P_K\}$ de forma que:

$$\bigcup_{k=1}^K P_k = X \text{ e } P_i \cap P_j = \emptyset, \forall i, j : i \neq j. \quad (1)$$

Esta formulação permite uma definição genérica que atende a grande parte dos algoritmos, não se aplicando exatamente nesta forma aos algoritmos fuzzy.

Entre as diversas possíveis aplicações de análise de agrupamentos, podem ser destacados: a redução de dados, onde a análise de um conjunto muito grande pode ser simplificada pela análise dos grupos; geração de hipóteses, através da inferência de relações entre dados a partir dos agrupamentos que podem ser confirmados experimentalmente ou através de novas análises; confirmação de hipóteses teóricas através da análise de um conjunto de dados e; predição baseada em grupos, na qual após aplicado um algoritmo de agrupamento em conjunto um de dados, conhecidos e identificados os agrupamentos, informações sobre um novo padrão podem ser inferidas, apenas observando em que grupo o algoritmo o adicionará (THEODORIDIS; KOUTROUMBAS, 2009).

Apesar da existência de diferentes algoritmos, um agrupamento passa pelos seguintes passos (THEODORIDIS; KOUTROUMBAS, 2009):

- Seleção de Características: os dados devem ser analisados e codificados, de forma a evitar redundância e otimizar a análise pelo algoritmo escolhido. Técnicas de pré-processamento podem ser utilizadas nesta etapa;
- Medida de Proximidade: deve ser definida ou selecionada uma medida que expresse quão similares ou diferentes são dois vetores. Durante o pré-processamento

dos dados, deve-se evitar que o peso de uma característica seja muito grande em relação às demais na medida escolhida;

- Critério de agrupamento: uma parcela considerável dos algoritmos de agrupamento atuam de maneira repetitiva, refinando os grupos encontrados entre os dados. Nestes casos é necessário estabelecer um critério de parada que indique que o agrupamento encontrado satisfaz a demanda do agrupamento;
- Algoritmo de agrupamento: de acordo com o padrão dos dados e natureza da análise, um ou mais algoritmos de agrupamento podem ser escolhidos;
- Validação dos resultados: uma vez obtidos os agrupamentos, pode-se verificar propriedades numéricas que mensurem a qualidade dos agrupamentos. Existem testes estatísticos que propõem análises destas natureza.
- Interpretação dos resultados: uma vez obtidos os agrupamentos pode-se confrontar os resultados com análises experimentais e conhecimento especialista, de modo a possibilitar conclusões acerca do objeto de pesquisa representado pelos dados;

Com base nos diferentes objetivos e passos já citados, para um mesmo conjunto de dados, vários agrupamentos podem ser obtidos. A incerteza acerca do melhor algoritmo ou melhores parâmetros para o agrupamento de um conjunto de dados é inerente à tarefa. Na maioria dos casos, a confirmação experimental ou a análise de um especialista em relação ao objeto pesquisado podem orientar a escolha de um bom agrupamento para uma finalidade específica (THEODORIDIS; KOUTROUMBAS, 2009).

Na literatura, podem ser encontradas várias classificações possíveis para os algoritmos de agrupamentos, entre elas duas características importantes do algoritmo de agrupamento que estão diretamente ligadas ao tipo de análise que se deseja realizar são: o tipo de relação de pertinência dos elementos aos grupos e a organização hierárquica ou não dos grupos. De acordo com estes critérios, os algoritmos de agrupamento podem ser divididos da seguinte forma:

- Quanto à organização:
 - Algoritmos simples ou não hierárquicos: cada elemento pertence apenas a um grupo e todos os grupos são disjuntos. Não existe a ideia de subgrupo;
 - Algoritmos hierárquicos: Os grupos são organizados em uma hierarquia, de modo que um grupo pode conter subdivisões gerando novos agrupamentos dentro do grupo. Permite análises com diferentes granularidades.
- Quanto à relação de pertinência do padrão ao grupo:

- Algoritmos Crisp: utiliza a teoria clássica dos conjuntos para relação de pertinência dos padrões a um grupo. Dessa forma, um padrão pertence ou não a um grupo, de modo que um padrão pertence a um único grupo;
- Algoritmos Fuzzy: utiliza a teoria fuzzy de conjuntos para relação de pertinência dos padrões a um grupo, de forma que um padrão possui um grau de pertinência ao grupo, geralmente no intervalo $[0, 1]$, onde o grau de pertinência 0 indica que o padrão não pertence ao grupo e o grau de pertinência 1 indica que o padrão certamente pertence ao grupo. Em algoritmos fuzzy, um padrão pode pertencer a mais de um grupo. Nestes algoritmos, a restrição $P_i \cap P_j = \emptyset$ da equação 1 não se aplica.

Definir o melhor agrupamento para um conjunto de padrões é uma atividade computacionalmente complexa diante da quantidade de combinações que cresce rapidamente em função da quantidade de padrões e pouco determinística, uma vez que diferentes parâmetros podem levar a agrupamentos distintos. Como discutido em (THEODORIDIS; KOUTROUMBAS, 2009) a incerteza é inerente à atividade de agrupamento de dados e, apesar da existência de testes estatísticos para validação dos grupos encontrados, a análise com base em conhecimento especialista e confirmação empírica de hipóteses que determinará a eficácia de um agrupamento.

Dentre os diversos algoritmos de agrupamento conhecidos, são apresentados a seguir três importantes algoritmos que possuem aplicações no reconhecimento de padrões e agrupamento de dados de biologia molecular. A escolha destes modelos foi realizada com base nas aplicações encontradas na literatura atual para análise de dados de expressão gênica e a natureza do objetivo deste trabalho, sendo escolhidos algoritmos não hierárquicos, uma vez que não serão analisados níveis de divisão entre os grupos. Também foram escolhidos algoritmos crisp, uma vez que excede o escopo a representação das incertezas inerentes à análise dos dados, onde se tem o maior potencial da lógica *fuzzy*.

3.1 K-means

O K-means, é um algoritmo de agrupamento muito utilizado em diversas áreas, inclusive em análises de dados biomoleculares, para o qual deve-se conhecer a priori a quantidade de grupos existentes nos dados.

O algoritmo funciona de forma iterativa, onde centroides, criados inicialmente de maneira aleatória são reajustados em direção ao centro dos grupos que são reorganizados a cada iteração.

Cada padrão analisado define em cada iteração como seu exemplar o centróide que minimiza a função de distância escolhida. No fim de cada iteração, os centróides são redefinidos como a média dos vetores do grupo do qual é exemplar. O algoritmo se repete até que os centróides não sejam mais reajustados ou assim que se atenda ao critério de parada estabelecido.

De forma similar a este trabalho, algumas pesquisas envolvendo análise de perfis de expressão diferencial de genes com determinada patologia utilizam o K-means para agrupamentos de dados de microarranjo. Exemplo destes trabalhos é (ZARAVINOS et al., 2011), onde são analisados perfis de expressão gênica de células de tumores e células saudáveis de pacientes com câncer de bexiga.

Além da utilização do K-means em sua forma original, alguns trabalhos propõem melhorias no algoritmo. Seguem alguns exemplos:

- Chandrasekhar et al. (2011): utiliza o K-means combinado com o algoritmo *Cluster Centre Initialization* que aprimora a escolha inicial dos centróides, ampliando a velocidade de convergência do algoritmo. A proposta do algoritmo foi validada com testes de dados de microarranjo disponíveis em bases públicas conhecidas.
- Muthukalathi et al. (2014): propõem uma estratégia conhecida por *Consensus Clustering*, que se baseia na aplicação iterativa de um algoritmo de agrupamento em reordenações dos padrões de entrada, fornecendo para parâmetros para escolha do número de grupos e permite avaliar a consistência dos grupos. A estratégia proposta por Muthukalathi et al. (2014) foi validada com a análise de dados de microarranjo de expressão gênica disponíveis em bases de dados públicas.

Considerando o objetivo de agrupamento descrito na equação 1, e os padrões de entrada a serem agrupados como vetores, pode-se propor uma formalização matemática para o K-means, que possui como entrada um conjunto X de vetores e um número de grupos m e como saída uma matriz U que representa uma partição de X , indicando a pertinência dos genes a cada grupo.

3.1.1 Formalização do Algoritmo

Sejam $X = \{x_1, x_2, \dots, x_N\}$ um conjunto de vetores de entrada e, $\beta = \{\beta_1, \beta_2, \dots, \beta_k\}$ um conjunto de centróides dispostos inicialmente de forma aleatória e; $U = u_{ij}$ uma matriz de pertinência do tipo:

$$u_{ij} = \begin{cases} 1, & \text{se } \|x_i - \beta_j\|^2 = \min_{m=1, \dots, k} \|x_i - \beta_m\|^2 \\ 0, & \text{caso contrário} \end{cases}, \quad (2)$$

o K-means tem por propósito minimizar a função:

$$J(\beta, U) = \sum_{i=1}^N \sum_{j=1}^m u_{ij} \|x_i - \beta_j\|^2. \quad (3)$$

Para minimizar a função $J(\beta, U)$, o algoritmo atualiza iterativamente a matriz U e recalcula os centroides β_j como a média dos vetores x_i , para os quais $u_{ij} = 1$, através da equação:

$$\beta_j = \frac{\sum_{i=1}^N u_{ij} x_i}{\sum_{i=1}^N u_{ij}}, j = 1, 2, \dots, k. \quad (4)$$

Decorre desta definição matemática que N é a quantidade de padrões no conjunto de entrada e k a quantidade de grupos. Apesar de ter sido utilizada a distância euclidiana como medida de similaridade entre os padrões a serem agrupados, outras distâncias podem ser utilizadas, desde que seja assegurada a convergência do algoritmo com esta nova distância.

Segue o algoritmo K-means, para um conjunto $X = \{x_1, x_2, \dots, x_N\}$, com k centroides, apresentado em (THEODORIDIS; KOUTROUMBAS, 2009) e adaptado para formulação matemática utilizada neste trabalho:

- Escolha estimativas iniciais aleatórias β_j para $j = 1, \dots, k$
- Repita
 - Para $i = 1$ até N
 - * Determine o centroide β_j mais próximo de x_i e altere a linha i matriz U de forma que $u_{ij} = 1$ e $u_{it} = 0, \forall t \neq j$
 - Fim Para
 - Para $j = 1$ até k
 - * Atualize β_j para a média dos vetores $x_i \in X$ com $u_{ij} = 1$
 - Fim Para
- Até nenhum β_j , para $j = 1, \dots, k$ ser modificado em relação a duas iterações sucessivas ou o critério de parada ser atendido.

Por ser um algoritmo iterativo, o critério de parada pode ser o número de iterações e/ou a distância dos centroides do conjunto β em relação a iterações sucessivas ou outro critério estabelecido.

3.2 Mapas auto-organizáveis

Os mapas auto-organizáveis, também conhecidos por redes SOM (do inglês - *Self Organizing Maps*) são um tipo de Rede Neural Artificial, com aprendizado não supervisionado utilizado no agrupamento de dados, proposto inicialmente por Kohonen (1982).

Assim como outros modelos de Redes Neurais, os SOMs tem inspiração no funcionamento do cérebro humano, mais especificamente do córtex cerebral, onde entradas sensoriais são mapeadas em diferentes regiões de maneira topologicamente ordenada. Dessa forma, Kohonen (1982) propôs um algoritmo capaz de mapear um conjunto de dados de dimensão arbitrária em grades de baixa dimensionalidade, onde a localização espacial do neurônio ativado está relacionada a características dos dados no espaço de entrada (HAYKIN, 1999).

O principal objetivo de um SOM é a transformação de um conjunto de dados pertencentes a um espaço de dimensão arbitrária em um mapa discreto de baixa dimensionalidade, geralmente uma ou duas dimensões, de maneira topologicamente ordenada (HAYKIN, 1999). Com base nestas características, os SOMs geralmente possuem dois tipos de aplicações: Compressão de dados (ou redução de dimensionalidade) e a disposição de dados de modo a evidenciar semelhanças entre dados agrupados.

Em relação ao seu funcionamento, os SOMs fazem parte das redes de aprendizado competitivo, ou seja, ao ser fornecida uma entrada, todos os neurônios a avaliam com uma função discriminante e aquele que obtiver a maior avaliação para a entrada tem seus pesos ajustados. Diferente de algoritmos de aprendizado competitivos, conhecidos como *winner takes all* (em uma tradução livre, "vencedor leva tudo"), em que apenas o neurônio com maior avaliação para entrada tem seus pesos ajustados, nos SOMs, os neurônios em uma vizinhança do neurônio vencedor tem seus pesos ajustados proporcionalmente à sua proximidade deste neurônio. Esta característica faz com que grupos vizinhos no mapa discreto possuam padrões semelhantes no espaço de entrada.

O agrupamento de um conjunto de dados através de um SOM perpassa pelos seguintes processos (HAYKIN, 1999):

- **Competição:** ao ser fornecida uma entrada de dados, deve ser avaliado o neurônio que possui maior valor para a função discriminante, elegendo-o como neurônio vencedor;
- **Cooperação:** a partir do neurônio vencedor deve ser definida uma região espacial na qual os neurônios serão excitados, de acordo com sua proximidade com o vencedor;

- Adaptação: os neurônios excitados devem ajustar seus pesos, em função de sua proximidade com o vencedor e do número de iterações do algoritmo.

Seguem alguns trabalhos que ilustram a aplicação de SOM na análise de dados de expressão gênica de microarranjo:

- Chavez-Alvarez et al. (2014): propõe uma metodologia para análise de perfis de expressão gênica com SOM através de um estudo de caso com agrupamento de dados de microarranjo de expressão gênica de *Saccharomyces cerevisiae* disponíveis em bases públicas.
- Monti et al. (2003): propõe um método para análise de dados de microarranjo com dados de expressão gênica baseado na utilização de um algoritmo de *Consensus Clustering*. Os estudos de caso foram realizados com dados de expressão gênica de bases pública para alguns tipos de câncer e versões do algoritmo com clusterização hierárquica e SOM.

Algumas funções são essenciais para a execução dos processos acima definidos, tanto para competição, quanto para cooperação e adaptação. Segue assim uma maior formalização matemática dos SOMs.

3.2.1 Formalização do Algoritmo

De forma abstrata, um SOM pode ser visto como uma transformação não linear Φ entre um espaço contínuo de dimensão finita arbitrária $\alpha \subset R^n$ e um espaço de saída discreto A de baixa dimensionalidade, topologicamente organizado em forma de grades (HAYKIN, 1999). A figura 9 ilustra o funcionamento de uma transformação genérica $\Phi : \alpha \Rightarrow A$ implementada por um SOM.

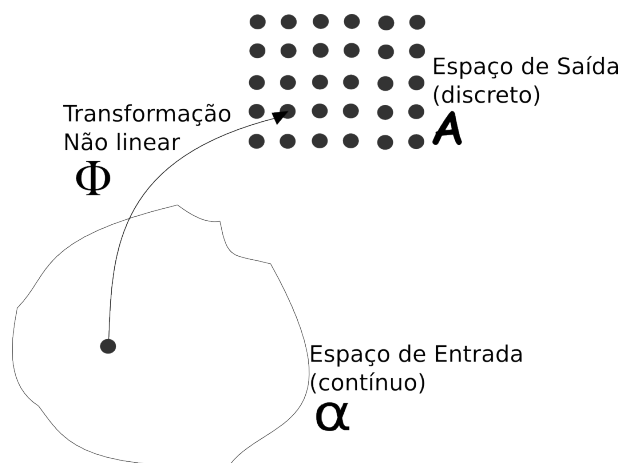


Figura 9 – Transformação não linear ϕ entre uma espaço $\alpha \subset R^n$ e um espaço discreto A implementada por um SOM. Adaptado de: (HAYKIN, 1999).

A formalização do SOM perpassa pela definição de um algumas funções e conjuntos que possibilitem a execução dos processos de competição, cooperação e adaptação. Os seguintes elementos devem ser definidos para execução do SOM para um conjunto de dados:

- $X = \{x_1, \dots, x_N\}$: um conjunto vetores de entrada de dimensão m ;
- $\omega = \{\omega_1, \dots, \omega_k\}$: um conjunto de vetores de pesos de dimensão m ;
- $d(\omega_i, \omega_j)$: uma função de distância entre o neurônio vencedor ω_i e o neurônio vizinho ω_j ;
- $i(x)$: uma função discriminante que retorna um neurônio vencedor para a entrada x_i ;
- $h_{j,i}$: uma função de vizinhança entre um neurônio j e o neurônio vencedor retornado por $i(x_i)$;
- $\eta(t)$: a taxa de aprendizado do algoritmo em função do tempo.

Para o processo de **competição**, é essencial que seja definida uma função discriminante que permita eleger o neurônio vencedor para um padrão de entrada (vetor) arbitrário. O processo de competição, pode então ser representado por uma função $i(x)$, que define o índice i , de forma que o neurônio i é o vencedor e, conseqüentemente, o vetor w_i , o vetor de pesos do neurônio vencedor. A função $i(x)$ pode ser definida da seguinte forma:

$$i(x) = \arg \min_j \|x - w_j\|^2, j = 1, 2, \dots, k. \quad (5)$$

Em redes neurais do tipo *winner takes all* usualmente vence o neurônio que possui o maior valor para a função discriminante. A utilização da menor distância euclidiana pode ser vista como critério similar, uma vez que a minimizar a distância euclidiana equivale a maximizar o produto interno entre dois vetores (HAYKIN, 1999).

Após selecionado o neurônio vencedor, para o processo de **cooperação**, deve-se definir a vizinhança na qual os neurônios em volta do vencedor serão excitados e com que proporção seus pesos serão ajustados. Pode-se considerar a função $h_{i,j}$, como uma função que define a proximidade dos padrões representados por dois neurônios i e j , para definir o quanto o ajuste de neurônio deve refletir no outro. Considerando $d_{i,j}$ a distância lateral entre os neurônios i e j , $h_{i,j}$ deve atender aos seguintes critérios (HAYKIN, 1999):

- A vizinhança deve ser simétrica em torno do ponto para o qual $d_{i,j} = 0$, possuindo maior valor para este ponto. Para o SOM, isto quer dizer que o neurônio vencedor terá seus pesos ajustados em maior proporção.

- A amplitude da vizinhança $h_{i,j}$ deve decrescer monotonicamente com o crescimento da distância lateral $d_{i,j}$. Neurônios muito distantes do vencedor não devem ter seus pesos ajustados.

Diferentes definições podem ser utilizadas para a função $h_{i,j}$, dependendo inclusive da natureza dos dados do domínio estudado. Em (HAYKIN, 1999) é apresentada como sugestão a função Gaussiana:

$$h_{j,i(x)} = e^{-\frac{d_{j,i}^2}{2\sigma^2}}, \quad (6)$$

onde a constante σ tem relação com a amplitude da parábola gerada pela função $h_{j,i(x)}$, consequentemente com a área da vizinhança ajustada em torno do neurônio vencedor. A figura 10 mostra o comportamento da função $h_{j,i(x)}$ em relação ao valor de σ .

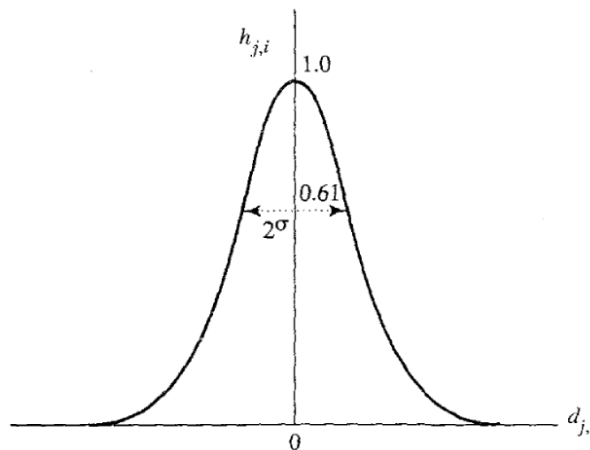


Figura 10 – Gráfico da função $h_{j,i(x)}$ em relação ao valor de σ . Adaptado de: (HAYKIN, 1999).

Haykin (1999) apresenta uma alternativa de utilização do valor de σ como uma função decrescente em relação ao número de iterações t , $\sigma(t) = \sigma_0 e^{-\frac{t}{\tau_1}}$ que faz com que a vizinhança em torno de um neurônio decresça em função do número de iterações, criando regiões especializadas para conjuntos de padrões de acordo com o avanço do processo de aprendizagem. O parâmetro σ_0 em $\sigma(t)$ é uma constante que pode ser ajustada de acordo com a aplicação e a velocidade de convergência desejada.

Por fim, é importante definir uma função a ser utilizada para o ajuste de pesos dos neurônios, possibilitando o processo de **adaptação** do algoritmo. Inicialmente, os vetores de pesos são inicializados com pequenos valores aleatórios, os dispendo aleatoriamente em relação ao espaço de entrada. O processo de adaptação desloca esses vetores em relação aos grupos nos quais atuarão como centroides, movendo-os em direção ao centro dos grupos. Assim, o ajuste dos pesos do neurônio vencedor é

definido de forma aproximá-lo do vetor de dados de entrada, considerando a diferença $(x - w_{i(x)})$ entre o vetor de pesos do neurônio vencedor $w_{i(x)}$ e o vetor de entrada de dados x .

A proporção com a qual o vetor de pesos se aproxima do vetor de entrada é conhecida parâmetro de taxa de aprendizado (η). De maneira análoga ao aprendizado humano, esta taxa deve reduzir com o aumento do número de iterações (t) do algoritmo, fazendo com que o parâmetro de taxa de aprendizado seja calculado em função do número de iterações, sendo representado pela função $\eta(t)$. Em (HAYKIN, 1999) é apresentada uma definição para a função $\eta(t)$ que possui um decaimento exponencial e pode ser ajustada através de constantes η_0 e τ . Segue a função apresentada:

$$\eta(t) = \eta_0 e^{-\frac{t}{\tau}}, t = 0, 1, 2, \dots, n \quad (7)$$

Uma vez que não só o neurônio vencedor, mas também seus vizinhos são deslocados em direção ao vetor de dados de entrada, o ajuste do peso de um neurônio qualquer para a iteração $t + 1$ do algoritmo, considerando a taxa de aprendizado e a vizinhança do neurônio é definido como (HAYKIN, 1999):

$$w_j(t + 1) = w_j(t) + \eta(t)h_{j,i(x)}(t)(x - w_j(t)). \quad (8)$$

Como pode ser observado na equação 8, a adaptação irá deslocar cada vetor de pesos em direção ao vetor de entradas fornecido, de maneira proporcional à sua relação de vizinhança com o neurônio vencedor e a taxa de aprendizado do algoritmo para a iteração em questão.

Uma vez apresentadas as funções necessárias à utilização de um SOM, segue uma representação do algoritmo:

- Escolha valores aleatórios para os vetores $w_j(0)$, com $j = 1, 2, \dots, K$, onde K é o número de neurônios do SOM;
- Repita:
 - Repita:
 - * Selecione um vetor no espaço de entrada X sem repetição;
 - * Eleja o neurônio vencedor através da equação $i(x) = \arg \min_j \|x - w_j\|^2, j = 1, 2, \dots, k$;
 - * Atualize o peso dos neurônios através da equação $w_j(t + 1) = w_j(t) + \eta(t)h_{j,i(x)}(t)(x - w_j(t))$;
 - Até todos os vetores no conjunto de entrada X serem escolhidos.

- Até que as alterações no mapa de características satisfaçam um critério de parada.

A cada iteração do algoritmo, é produzido um mapeamento entre os vetores do conjunto de padrões de entrada X e os neurônios no espaço de saída através da função $i(x)$. Este mapeamento pode ser considerado como saída do algoritmo para fins de agrupamento.

É importante ainda observar que a definição dos índices dos vetores de ω como escalares em $\omega = \{\omega_1, \dots, \omega_k\}$ e, conseqüentemente a definição da função $i(x)$ retornando um índice escalar, transformam o espaço de saída em uma grade unidimensional. A alteração destes índices para vetores de dimensões maiores e diferentes definições da função $d_{i,j}$ podem definir diferentes topologias para o espaço de saída do algoritmo.

3.3 Affinity Propagation

Alternativamente aos algoritmos citados anteriormente, o trabalho de Frey e Dueck (2007) propõe um algoritmo de agrupamento não hierárquico, no qual não é necessária a escolha a priori da quantidade de grupos. Este algoritmo é chamado de *Affinity Propagation*.

No *Affinity Propagation*, cada ponto (padrão) é visto como um nó em uma rede e todos pontos inicialmente são vistos como possíveis exemplares de um grupo. Os nós da rede trocam mensagens buscando escolher o melhor candidato a representante do grupo ao qual pertence. As trocas de mensagens entre dois pontos, baseadas em funções de similaridade, indicam a afinidade de um ponto em escolher outro como exemplar. Através das iterações do algoritmo, exemplares são definidos, formando grupos com os elementos que o elegeram como representante (FREY; DUECK, 2007).

Com as iterações do algoritmo, o número de grupos tende a crescer com o surgimento de novos exemplares. O algoritmo pode ser encerrado quando não existirem mais alterações entre os exemplares em iterações sucessivas ou quando um número de grupos preestabelecido for alcançado. Quando utilizado como critério de parada a não modificação dos exemplares, o número de grupos decorrerá da execução do algoritmo, dispensando o usuário do fornecimento deste parâmetro.

No trabalho que apresentou o algoritmo à comunidade científica, Frey e Dueck (2007) obteve bons resultados com o *Affinity Propagation* para classificação de éxons e íntrons em genes, análise de imagens, identificação de sentenças representativas em textos e eficiência de linhas aéreas.

Além das aplicações iniciais do *Affinity Propagation* em dados de bioinformática, trabalhos recentes o utilizam na análise de perfis de expressão gênica em dados obtidos

com microarranjos de cDNA, similarmente à proposta deste trabalho. Seguem dois exemplos destas aplicações:

- Chuang et al. (2015): utilizam o *Affinity Propagation* na redução de dados para análise de expressão gênica de pacientes com leucemia linfocítica aguda, através da análise dos genes selecionados como exemplares pelo algoritmo em detrimento da análise de todas amostras do microarranjo.
- Napoleon e Baskar (2011): apresentam um comparativo de eficiência e velocidade de convergência na classificação de subtipos de câncer no qual o *Affinity Propagation* obteve bons resultados comparado a duas variações do K-means: o X-means e o K-means eficiente. Para os testes foram utilizados dados de bases públicas de expressão gênica de pacientes com variações de leucemia.

3.3.1 Formalização do Algoritmo

Considerando um conjunto de padrões de entrada $X = \{x_1, x_2, \dots, x_n\}$, as decisões acerca do algoritmo são tomadas com base em uma função ou matriz de similaridade $s(i, k)$ que associa um valor real de similaridade entre os padrões x_i e x_k , uma função ou matriz $r(i, k)$ (“responsabilidade”) que, enviada de um padrão x_i para um padrão x_k , representa a evidência acumulada de x_k ser exemplar para x_i e, uma função $a(i, k)$ (“disponibilidade”), que representa a evidência acumulada do padrão x_i ser representante do padrão x_k (FREY; DUECK, 2007).

As funções $r(i, k)$ e $a(i, k)$, podem ser definidas conforme abaixo (FREY; DUECK, 2007):

$$r(i, k) = s(i, k) - \max_{k', k' \neq k} \{a(i, k') + s(i, k')\} \quad (9)$$

$$a(i, k) = \min \left\{ 0, r(k, k) + \sum_{i', i' \ni \{i, k\}} \max\{0, r(i', k)\} \right\}, i \neq k \quad (10)$$

$$a(k, k) = \sum_{i', i' \neq k} \max\{0, r(i', k)\} \quad (11)$$

Como pode ser observado na equação 9, a função de responsabilidade $r(i, k)$ considera em sua avaliação quanto o ponto x_i é similar ao ponto x_k e reduz deste cálculo a evidência de algum outro ponto ser exemplar de x_i , considerando a similaridade e disponibilidade do candidato mais evidente a ser exemplar de x_i . Na primeira iteração, como todos os valores de $a(i, k)$ são 0, terá maior valor de responsabilidade em relação ao ponto x_i , o ponto x_k mais similar a x_i .

A equação 10 representa a disponibilidade de um ponto x_k em ser exemplar de um ponto x_i , como a tendência do ponto x_k ser um exemplar (autoresponsabilidade) acrescida da soma dos pontos que tem evidências positivas que o ponto x_k é um exemplar. As evidências negativas não são consideradas, pois os pontos x'_i para os quais $r(i', k) < 0$ tem outro ponto como exemplar. Para evitar uma predominância de valores elevados para responsabilidade, a disponibilidade é limitada superiormente por 0.

É importante observar que a equação 11 apresenta uma definição própria para a autodisponibilidade, na qual é considerada apenas a evidência positiva acumulada de outros pontos considerarem que o ponto x_k é seu exemplar, não considerando a autoresponsabilidade $r(k, k)$, caso contrário, os pontos com tendência a serem exemplares teriam $a(k, k) = 0$ por conta do efeito de um valor elevado para $r(k, k)$ na equação 10.

O algoritmo do *Affinity Propagation* pode ser resumido nos seguintes passos:

- Inicialize todas as disponibilidades $a(i, k)$ para 0
- Repita:
 - Atualize os valores de responsabilidade $r(i, k)$ com a equação 9
 - Atualize os valores de disponibilidade $a(i, k)$ com as equações 10 e 11
 - Defina para cada ponto i como exemplar o ponto k que maximiza a soma $a(i, k) + r(i, k)$, no caso de $k = i$, isto indica que o ponto é um exemplar.
- Até que a taxa de alteração de responsabilidades e disponibilidades ou o número de exemplares atendam um critério de parada.

Um ponto importante a ser observado na equação 9 é que seu resultado está fortemente relacionado à definição da função de similaridade $s(i, k)$. Esta função deve representar o quanto dois padrões i e k são similares, mas não necessariamente precisa ser uma distância. Tal característica fornece uma flexibilidade interessante ao algoritmo, possibilitando a definição de uma função de similaridade que reflita características do domínio ao qual a técnica de agrupamento está sendo aplicado. Frey e Dueck (2007) apresenta como possibilidade de função de similaridade o simétrico aditivo da distância euclidiana dos vetores, que pode ser representada por:

$$s(i, k) = -||x_i - x_k||^2. \quad (12)$$

A utilização do simétrico aditivo se justifica pelo fato da distância se comportar de maneira inversa à similaridade, ou seja, quanto menor a distância entre dois pontos, mais similares são estes pontos.

Computacionalmente, pode ser mais simples optar pela visualização das funções $a(i, k)$, $r(i, k)$ como matrizes, onde i e k são índices dos elementos da matriz. Analogamente, uma vez que as similaridades não se alteram com as iterações do algoritmo, as similaridades podem ser calculadas antes do início do algoritmo e ser fornecido como parâmetro uma matriz de similaridade ao invés da função $s(i, k)$.

A cada iteração do algoritmo é mapeado para cada vetor do conjunto X de padrões de entrada um exemplar $x_k \in X$. Para fins de agrupamento, pode ser considerado como saída do algoritmo os grupos formados pelos padrões que possuem o mesmo exemplar. Para algumas análises que tenham como propósito a redução de dados, pode ser feita a análise dos exemplares como alternativa ao estudo de todos elementos do conjunto de entrada, considerando dessa forma os exemplares como saída do algoritmo.

4 Análise dos Dados de Microarranjo

Com o objetivo de produzir análises comparáveis às apresentadas em (KANETO, 2011) os dados foram pré-processados de forma semelhante à utilizada nesse trabalho. Além disso, trabalhos similares como (FERREIRA FILHO, 2009) também utilizam metodologia similar para pré-processamento de dados de expressão gênica obtidos de análises de microarranjo.

Os genes selecionados no pré-processamento foram agrupados com os algoritmos escolhidos, buscando visualizar grupos de genes que apresentem padrões que sugiram análises de interesse para o estudo da patologia. Uma vez definidos estes grupos de interesse, os genes pertencentes a estes grupos foram analisados na base de dados do NCBI e na literatura existente em busca de informações que possam permear a discussão por hipóteses de explicações biológicas acerca dos perfis encontrados.

De modo a possibilitar uma maior compreensão acerca dos resultados obtidos através da análise de agrupamentos, antes das seções que discutem procedimentos e resultados obtidos, a seção 4.1 faz uma apresentação geral de características da Osteogênese Imperfeita importantes para as discussões que seguem.

4.1 Caracterização da Osteogênese Imperfeita

Osteogênese Imperfeita (OI) é um termo genérico para representar um conjunto de desordens genéticas que se caracterizam principalmente por deficiências na formação da matriz óssea, resultando em fragilidade óssea (WALLACE et al., 2014; VAN DIJK; SILENCE, 2014). Entre os diversos tipos da patologia é comum a produção deficiente de colágeno, o que provoca alguns sintomas sumarizados por Lindert et al. (2015) como a má formação da estrutura óssea, ossos quebradiços, esclera azul e perda de audição. De acordo com o tipo da patologia, alguns desses sintomas podem ou não se manifestar.

Com base na observação de imagens de raio-x, fichas de pacientes e análise de histórico clínico familiar, Silence et al. (1979) propuseram a classificação dos casos de OI em quatro tipos, de acordo com sua caracterização genética e seu conjunto de sintomas. Com o avanço dos estudos moleculares da OI e publicação de pesquisas envolvendo novos padrões de sintomas, surgiram propostas de novas classificações da OI, incluindo novos tipos da patologia. Em um estudo mais recente Van Dijk e Silence (2014) propuseram uma nova classificação que inclui um novo tipo de OI, a OI tipo V.

Assim, seguem os tipos de OI classificados em Van Dijk e Silence (2014) com um breve resumo de seus sintomas e características conhecidas:

- OI tipo I - OI não deformadora com esclera azul: é caracterizada pela fragilidade óssea consequente da baixa densidade do tecido. É comum uma coloração azul acinzentada na esclera e, em alguns casos, deficiências na audição no início da vida adulta;
- OI tipo II - OI letal perinatal: este é o tipo mais severo da OI, que manifesta seus sintomas durante a formação do feto, sendo detectado em fetos entre 18 e 20 semanas de gestação: má formação óssea, deformidades nos ossos longos, na estrutura da face e do crânio. Com o avanço da gestação, o feto sofre constantes fraturas nos membros, fazendo com se desenvolvam com deformidades. Mesmo quando chegam ao nascimento, relatos clínicos apontam que recém nascidos com OI tipo II apresentam dor constante. Em alguns casos, ao ser detectado este tipo de patologia em exames pré-natais, a gestação é interrompida.
- OI tipo III - OI progressivamente deformadora: comumente se manifesta desde os primeiros anos de vida, causando múltiplas fraturas que levam a deformações no esqueleto e baixa estatura. A coloração azul na esclera, mesmo quando presente em recém nascidos, não é predominante com o avançar da idade;
- OI tipo IV - OI comum variável: pacientes com este tipo de OI sofrem com fraturas recorrentes, osteoporose e variados graus de deformidade nos ossos grandes e coluna vertebral. É pouco comum deficiências na audição nestes casos e a esclera azul, quando presente em recém nascidos, tende a normalizar no decorrer da infância;
- OI tipo V - OI com calcificação de membranas interósseas: Além da fragilidade óssea média ou severa, este tipo de OI é caracterizada pela calcificação contínua de membranas interósseas, principalmente no antebraço e pernas. Pacientes com este tipo de OI geralmente apresentam maior tendência à formação de calos hiperplásticos.

Atualmente vários genes envolvidos na OI são conhecidos e catalogados. No entanto, grande parte das pesquisas atuais se concentram no estudo de alelos mutantes dos genes *COL1A1* e *COL1A2*. O direcionamento das pesquisas para mutações nestes genes pode ser explicado por estes serem responsáveis por aproximadamente 90% dos casos de OI conhecidos. No entanto existem outros genes conhecidos associados à patologia. Em (VAN DIJK; SILENCE, 2014) é mostrada uma tabela com dezessete genes conhecidos associados a cada tipo de OI.

Diante dos dados disponíveis e a diversidade dos genes e manifestações fenotípicas relacionadas com a OI, este trabalho discute somente genes possivelmente

relacionados com os tipos I e III da patologia, variantes portadas pelos pacientes dos quais as amostras foram analisadas.

4.2 Pré-processamento

Uma vez obtidos os dados resultantes da análise de microarranjo de cada amostra, estes dados foram reunidos em um único arquivo e toda análise de pré-processamento foi realizada através da biblioteca *Limma* (RITCHIE et al., 2015) do pacote *Bioconductor* do *R*, que já possui um conjunto de funções necessários ao tratamento e seleção de dados originados de análises de microarranjo por diferentes equipamentos.

Seguindo a análise ora realizada em (KANETO, 2011), com o objetivo de produzir resultados comparáveis com o uso de diferentes algoritmos de agrupamento, o pré-processamento envolveu os seguintes passos:

- a) Transformação logarítmica dos sinais (\log_2);
- b) Correção do sinal de *background* com o algoritmo *normexp* (SILVER et al., 2009);
- c) Normalização entre amostras com o método quantile (BOLSTAD et al., 2003);
- d) Retirada de redundância de diferentes sondas para o mesmo gene através da mediana;
- e) Seleção dos 100 genes com maior desvio padrão entre as amostras.

Uma vez que o objetivo do agrupamento é a análise de genes que apresentam padrões similares e não valores similares de expressão, a normalização logarítmica evita que genes sejam considerados diferentes, tendo perfis similares e valores diferentes em ordem de grandeza. Sem transformação logarítmica, dois genes que tiverem a expressão elevada e/ou reduzida nas mesmas amostras, podem ser incluídos em grupos diferentes por algoritmos baseados em distância se a expressão tiver valores muito diferentes.

A correção do sinal de *background* foi utilizada para reduzir ruídos inerentes da captura de imagens no chip de microarranjo ou do processamento computacional para extrair os níveis de expressão a partir dos *spots*.

Uma vez que os dados tiveram origem em mais de um experimento, a normalização quantílica foi utilizada para reduzir o ruído gerado pelas variações do experimento, que poderiam ser confundidas com variações nos níveis de expressão dos genes em diferentes amostras.

Como os dados originais utilizados por Kaneto (2011) possuíam amostras de mais de uma sonda para alguns genes, seguindo a estratégia adotada nesse trabalho, os vetores originados por sondas repetidas foram substituídos pela mediana dos valores das diferentes sondas para o mesmo gene.

Por fim, a seleção dos genes com maior desvio padrão entre as amostras se justifica pelo fato destes genes possuírem maior potencial de estarem envolvidos com a patologia, uma vez que apresentaram uma variância maior durante o processo de diferenciação das células para células do tecido ósseo.

4.3 Análise dos Agrupamentos

Uma vez realizado o pré-processamento dos dados, os 100 genes selecionados com maior desvio padrão entre as amostras foram agrupados utilizando os algoritmos *K-means*, Mapas auto-organizáveis e *Affinity Propagation*. Observando os resultados anteriores apresentados em (KANETO, 2011) e tendo em vista a perspectiva do surgimento de poucos grupos com informações interessantes à análise desejada, todos os algoritmos foram executados para agrupar os dados em cinco grupos.

As análises de agrupamento foram todas realizadas utilizando o ambiente de programação do R, para as quais o K-means foi executado utilizando as bibliotecas padrão do R, o SOM através do pacote *kohonen* (WEHRENS; BUYDENS, 2007) e o *Affinity Propagation* através do *apcluster* (BODENHOFER et al., 2011).

O resultado dos agrupamentos foi representado graficamente através de *heatmaps*, por esta forma permitir a observação dos grupos e perfis de expressão dos genes de cada grupo num único tipo de gráfico. As figuras 11, 12 e 13 mostram o resultado dos agrupamentos dos 100 genes selecionados pelos algoritmos *K-means*, *SOM* e *Affinity Propagation*, respectivamente.

Nos três agrupamentos, surgiu um grupo de genes que apresenta expressão **alta** nas amostras de controle e **baixa** nas amostras dos pacientes portadores da patologia e outro grupo que apresenta expressão **baixa** nas amostras de controle e **alta** nas amostras dos pacientes portadores da patologia. Com o propósito de identificar estes grupos, ora dispostos em posições diferentes nos algoritmos, e analisar seus perfis de expressão e similaridade, estes grupos foram identificados nos *heatmaps* como **Grupo A** e **Grupo B**, respectivamente.

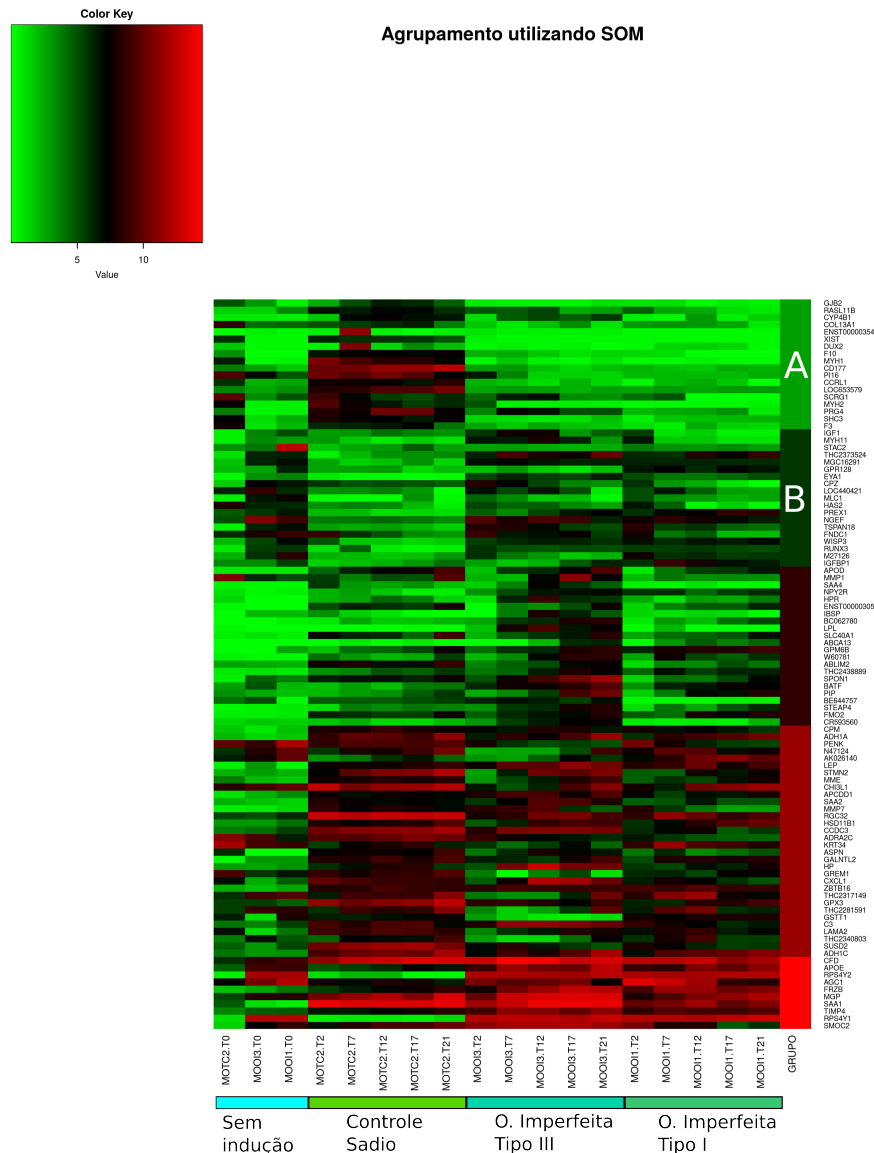


Figura 12 – Agrupamento dos 100 genes pré-selecionados utilizando o algoritmo SOM com número de grupos pré-estabelecido como cinco.

O *Affinity Propagation* revelou ainda um terceiro grupo que se mostrou interessante às análises desejadas, por possuir genes com expressão baixa nas amostras de controle e nas amostras dos pacientes com Osteogênese Imperfeita Tipo I, mas expressão alta nas amostras dos pacientes com Osteogênese Imperfeita Tipo III, que foi denominado **Grupo C** e pode contribuir na seleção de genes que sugerem um perfil de expressão específico para este tipo de OI.

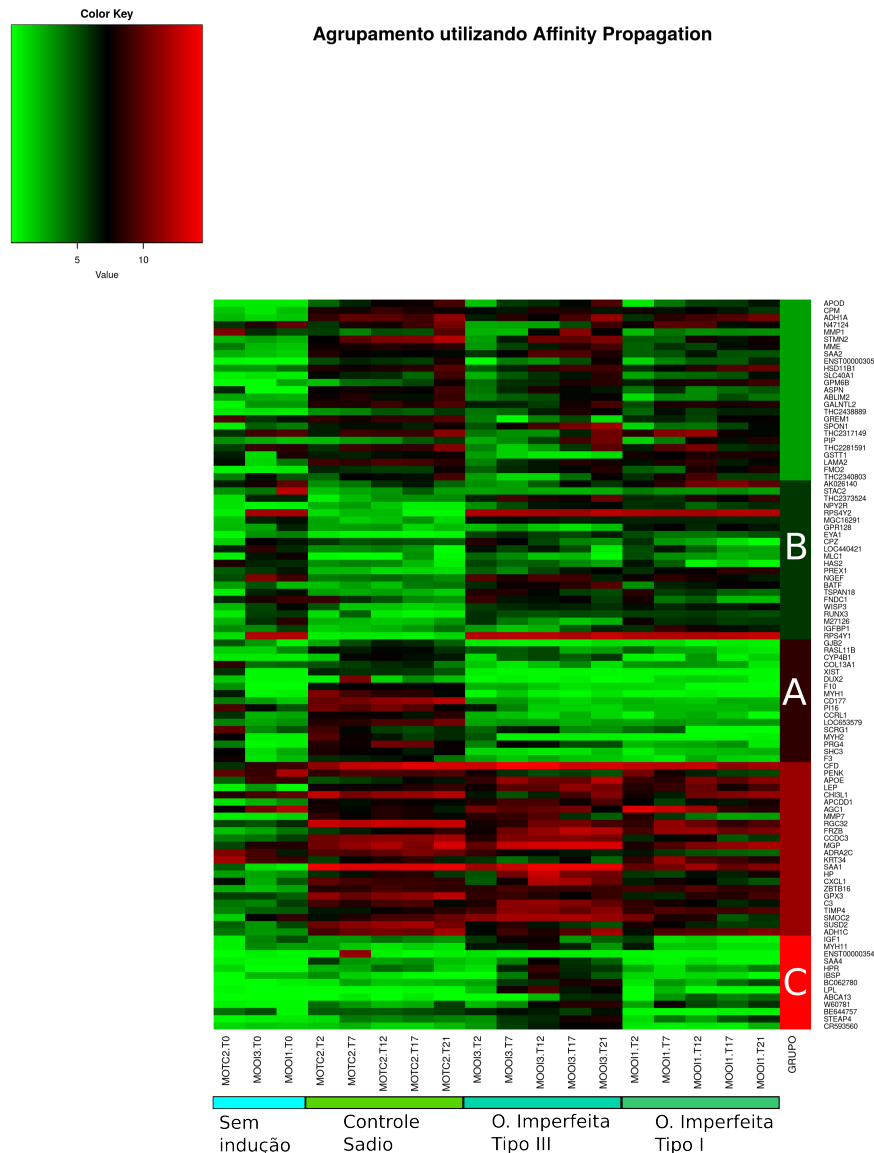


Figura 13 – Agrupamento dos 100 genes pré-selecionados utilizando o algoritmo *Affinity Propagation* com número de grupos pré-estabelecido como cinco.

4.3.1 Consistência dos Agrupamentos

Uma vez selecionados os genes presentes nos três grupos, iniciou-se a análise de similaridade dos genes nestes grupos, afim de avaliar a consistência dos agrupamentos, de acordo com a interseção dos grupos fornecidos por cada algoritmo. As figuras 14 e 15 mostram o percentual de semelhança entre os genes que pertencem a cada grupo. O percentual de genes similares entres os grupos foi considerado como o total de genes no conjunto união dos grupos, dividido pelo total de genes na interseção dos grupos,

representado pela fórmula:

$$p = \frac{\left| \bigcup_{i=1}^n G_i \right|}{\left| \bigcap_{i=1}^n G_i \right|} \quad (13)$$

onde p é o percentual de similaridade entre os grupos, n é a quantidade de algoritmos comparados e G_i o conjunto genes inserido no grupo pelo i -ésimo algoritmo.

Conforme discutido por Theodoridis e Koutroumbas (2009), determinar que um agrupamento é válido nem sempre é uma tarefa simples pois um dos principais parâmetros seria a qualidade da informação que pode se inferir a partir desta nova visualização dos dados, o que recai no não determinismo do conhecimento especialista. Com o propósito de avaliar os grupos de interesse para a análise fim deste trabalho, foram observados grupos de genes com perfis que podem estar associados à características da patologia. Na tabela 3 é possível observar com mais detalhes a interseção dos grupos gerados pelos três algoritmos.

A partir da análise da Figura 14, podemos observar que o *Grupo A* apresenta uma boa consistência, sendo em quase sua totalidade de genes idênticos nos grupos formados pelos três algoritmos.

Similaridade de genes do Grupo A em diferentes algoritmos

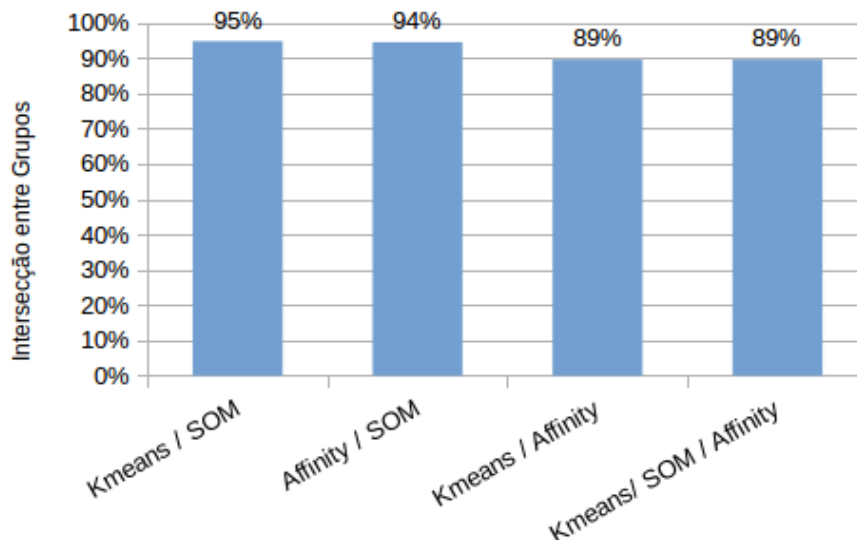


Figura 14 – Percentual de similaridade entre os genes inseridos no grupo A nos diferentes algoritmos.

Tabela 3 – Genes Pertencentes a Grupos A e B por Algoritmo.

Grupo A				Grupo B			
Gene	<i>Kmeans</i>	<i>SOM</i>	<i>AP*</i>	Gene	<i>Kmeans</i>	<i>SOM</i>	<i>AP*</i>
GJB2	X	X	X	IGF1	X	X	
RASL11B	X	X	X	MYH11	X	X	
CYP4B1	X	X	X	AK026140			X
COL13A1	X	X	X	SAA4	X		
ENST00000354854	X	X		STAC2	X	X	X
XIST	X	X	X	THC2373524	X	X	X
DUX2	X	X	X	NPY2R	X		X
F10	X	X	X	RPS4Y2			X
MYH1	X	X	X	MGC16291	X	X	X
CD177	X	X	X	GPR128	X	X	X
PI16	X	X	X	HPR	X		
CCRL1	X	X	X	EYA1	X	X	X
LOC653579	X	X	X	CPZ	X	X	X
SCRG1	X	X	X	LOC440421	X	X	X
MYH2	X	X	X	IBSP	X		
BE644757	X			MLC1	X	X	X
PRG4	X	X	X	BC062780	X		
SHC3	X	X	X	HAS2	X	X	X
F3	X	X	X	PREX1	X	X	X
				NGEF		X	X
				LPL	X		
				ABCA13	X		
				W60781	X		
				BATF	X		X
				TSPAN18	X	X	X
				FNDC1	X	X	X
				WISP3	X	X	X
				RUNX3	X	X	X
				M27126	X	X	X
				STEAP4	X		
				IGFBP1	X	X	X
				CR593560	X		
				RPS4Y1			X
*Affinity Propagation							

Já observando a Figura 15, tem-se a primeira impressão que o *Grupo B* apresenta menor semelhança entre os agrupamentos. No entanto, se melhor observado nas tabelas 3 e 4, pode-se observar que o *Kmeans* agrupou um conjunto maior de genes no **Grupo B** que os demais algoritmos. O *Affinity Propagation* encontrou padrões específicos e dividiu estes genes entre os grupos **B** e **C**. O *SOM* conseguiu identificar semelhança em grande parte dos genes presentes no **Grupo B** do *Affinity Propagation*, mas não revelou padrões nos demais genes, que formariam um grupo semelhante ao **Grupo C** do *Affinity Propagation*.

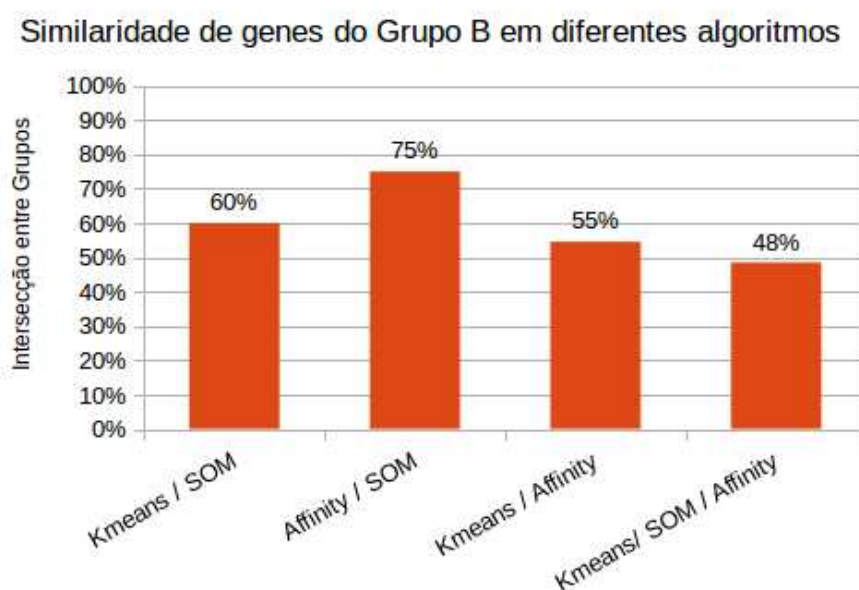


Figura 15 – Percentual de similaridade entre os genes inseridos no grupo **B** nos diferentes algoritmos.

Uma vez que os grupos encontrados mostraram-se semelhantes entre os grupos e o agrupamento fornecido pelo *Affinity Propagation* possibilitou uma melhor visualização para analisar relações entre os perfis de expressão gênica e processos/fenótipos relacionados com a patologia estudada, a organização dos grupos resultante deste algoritmo foi adotada para as próximas análises realizadas no trabalho.

Outros trabalhos que apresentam comparativos de desempenho de diferentes algoritmos de agrupamento, também apresentam resultados nos quais o *Affinity Propagation* realizou classificações mais adequadas ao problema estudado, como em (DUECK; FREY, 2007), que comparou o desempenho dos algoritmos K-centers e *Affinity Propagation* para classificação de imagens de faces de humanos em relação ao percentual de erros e o *Affinity Propagation* apresentou o melhor desempenho e, (RANGREJ et al., 2011) que comparou resultados de classificações de textos utilizando K-means, *Singular Value Decomposition* (SVD) e *Affinity Propagation*, onde o *Affinity Propagation* apresentou

percentual de erros de classificação expressivamente menores.

Tabela 4 – Genes do Grupo C no *Affinity Propagation* e intersecção com o Grupo B do K-means.

Sigla do Gene	Grupo B do K-means
IGF1	Sim
MYH11	Sim
ENST00000354854	Não
SAA4	Sim
HPR	Sim
IBSP	Sim
BC062780	Não
LPL	Sim
ABCA13	Sim
W60781	Sim
BE644757	Não
STEAP4	Sim
CR593560	Sim

Apesar destes resultados, em alguns trabalhos, como em (VLASBLOM; WODAK, 2009) que comparou a utilização do *Affinity Propagation* e o *Markov Clustering Algorithm* (MCL) para particionamento de grafos de interação de proteínas e o MCL apresentou melhor convergência e tolerância a ruídos, outros algoritmos podem apresentar melhores resultados que o *Affinity Propagation* de acordo com a natureza dos dados e a análise desejada.

4.3.2 Análise dos Perfis Encontrados

Uma vez selecionados os grupos de interesse para análise, foi utilizada a base de dados de genes do NCBI para busca de anotações e publicações relacionadas aos genes agrupados de forma a subsidiar a formulação de hipóteses acerca do significado biológico destes perfis de expressão e sua relação com a Osteogênese Imperfeita.

Cabe ressaltar que diante da complexidade dos processos biológicos que sucedem a expressão gênica, como a regulação pós-transcricional, as interações entre proteínas para diferentes funções e outras análises que remetem à biologia de sistemas, o fato de não encontrar relações nas bases de dados existentes entre os genes estudados e a patologia não garante que não estejam relacionados com a mesma. De modo análogo, uma vez que análise de microarranjo é um registro estático do processo dinâmico de

expressão gênica, a expressão diferencial destes genes, se relacionados à patologia, deve ser avaliada em novos experimentos laboratoriais.

Os genes do grupo **A**, conforme já discutido, possuem expressão relativa maior nas amostras de controle que nas amostras dos pacientes com OI. Pode-se esperar que alguns destes genes estejam relacionados à produção de produtos gênicos envolvidos na formação do tecido ósseo, resultando em falhas no processo de osteogênese. Como pode ser observado na Figura 16, que mostra a expressão média dos genes do grupo em função do tempo, esses genes tem sua expressão relativa aumentada nas amostras de controle sadio em níveis mais elevados que nas amostras de pacientes com a patologia.

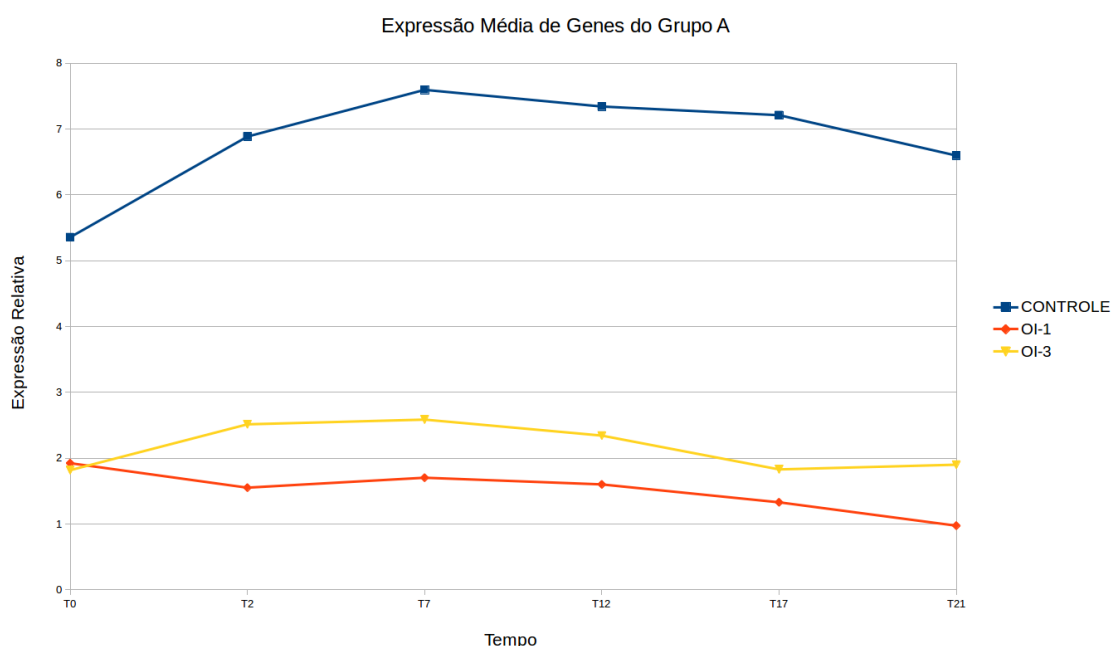


Figura 16 – Expressão relativa dos genes do Grupo A em função do tempo.

Na Figura 17 é possível observar a expressão do gene CCRL1, gene tomado como exemplar do **Grupo A** pelo *Affinity Propagation*, onde pode-se observar que após a indução da diferenciação das células, este gene mantém sua expressão relativa aumentada durante todo processo de análise para a amostra de controle e mantém sua expressão relativa baixa nas amostras com a patologia. É importante observar que a seleção do gene CCRL1 como exemplar pelo *Affinity Propagation* não permite inferências diretas sobre seu grau de relação com a patologia. O exemplar de um grupo no *Affinity Propagation* pode apenas ser considerado como o gene que melhor representa o padrão numérico existente nos elementos do grupo, de acordo com o critério (função) de similaridade adotado.

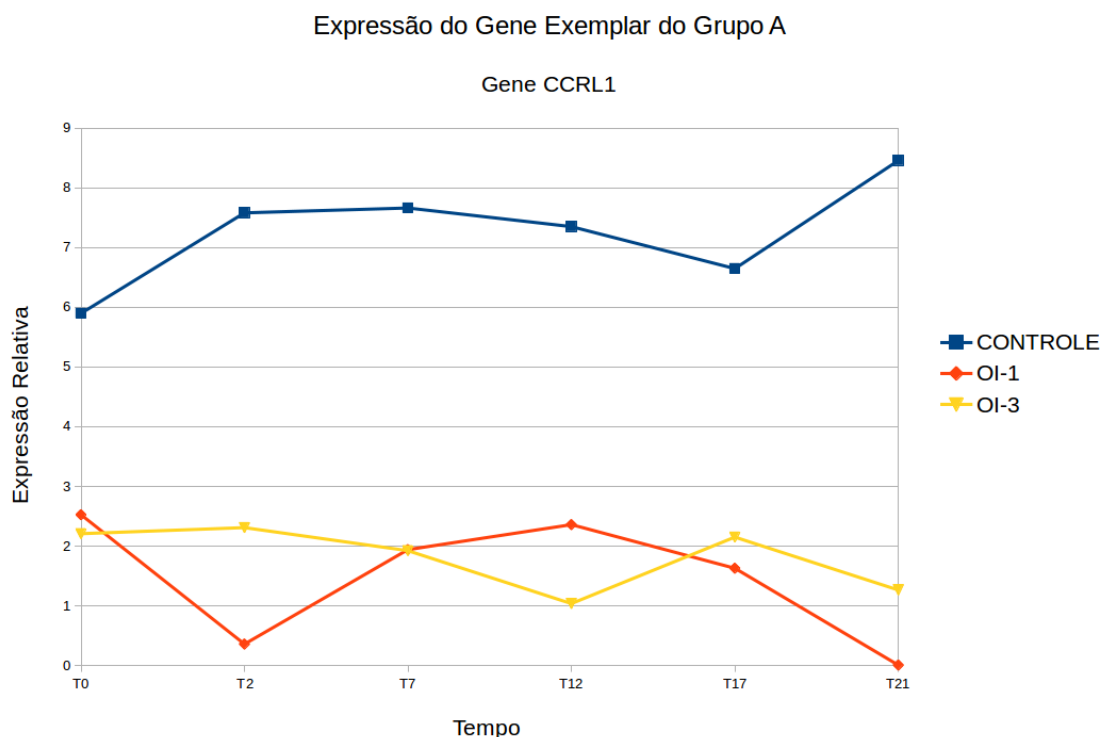


Figura 17 – Expressão relativa do gene CCRL1, exemplar do Grupo A selecionado pelo *Affinity Propagation*.

Entre os genes encontrados no **Grupo A**, está o gene PRG4, que gera uma proteína que tem sua função associada a absorção elástica e a dissipação de energia, na cartilagem articular. A redução na expressão deste gene pode estar relacionada com a fragilidade característica de vários tipos de OI.

Os genes do **Grupo B**, por outro lado, possuem expressão relativa maior nas amostras dos pacientes com a patologia que nas amostras de controle. No entanto, como pode-se observar nos heatmaps das figuras 11, 12 e 13 e nos níveis de expressão mostrados na Figura 18, os genes do *Grupo B* já possuíam uma expressão relativa maior entre as amostras com a patologia antes mesmo da indução da diferenciação das células e tem sua expressão reduzida em função do tempo. Este comportamento torna menos provável a relação dos genes deste grupo com o processo de osteogênese, uma vez que a indução da diferenciação não provocou comportamentos distintos entre as amostras de controle sadio e as amostras com a patologia. O padrão de níveis de expressão dos genes do **Grupo B** pode ser melhor observado através do gene MGC16291, exemplar do grupo, mostrado na Figura 19.

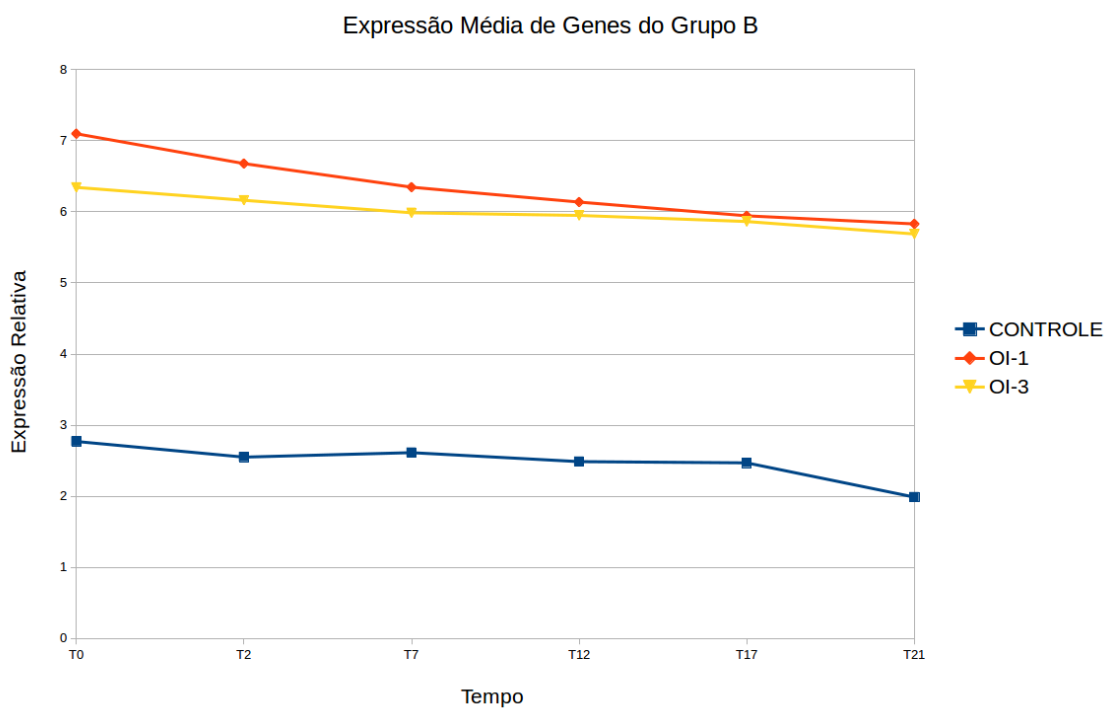


Figura 18 – Expressão relativa dos genes do Grupo B em função do tempo.

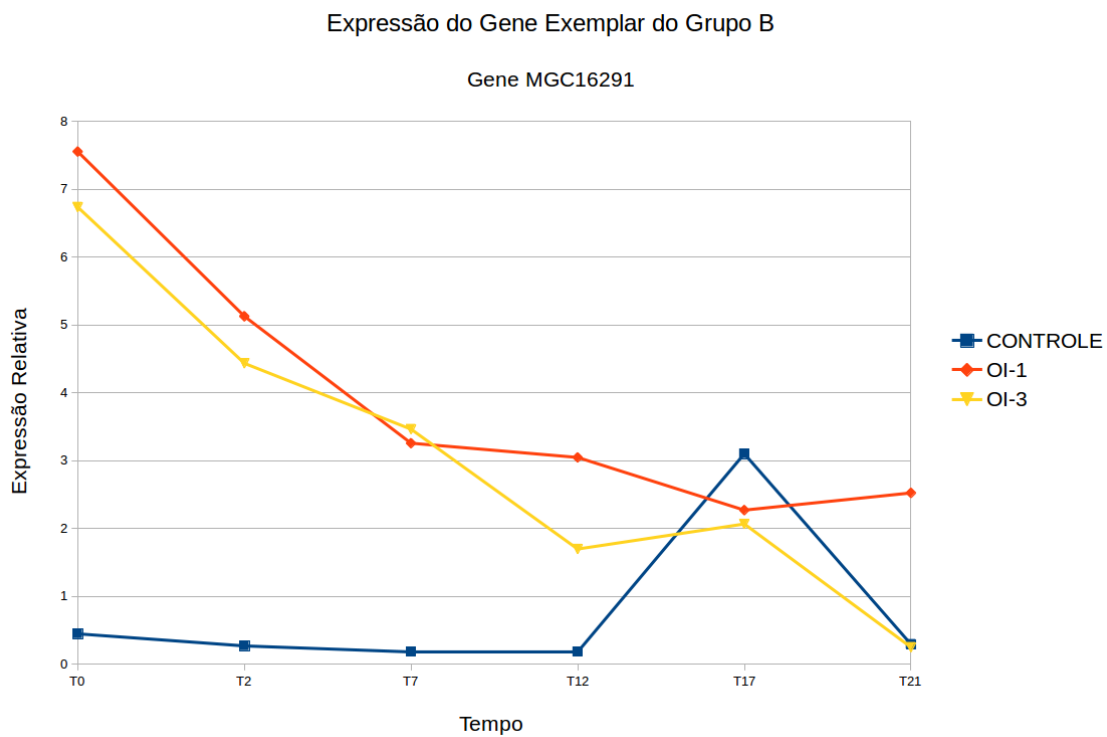


Figura 19 – Expressão relativa do gene MGC16291, exemplar do Grupo B selecionado pelo *Affinity Propagation*.

Não foram encontrados no **Grupo B** associação direta entre a função dos genes estudados e características da OI com buscas na bases de genes do NCBI. Este fato por si só não garante que tal relação não existe, podendo ser realizadas novas análises de bioinformática e/ou laboratoriais para confirmar esta hipótese.

Já os genes do **Grupo C**, tem seu perfil modificado após a indução da diferenciação das células apenas nas amostras de pacientes portadores da Osteogênese Imperfeita Tipo III, tendo seus níveis de expressão relativa aumentados nestas amostras. Este comportamento leva à hipótese que alguns destes genes podem estar relacionados com a patologia, uma vez que expressão de genes que não deviam ser expressos nestas células podem gerar novas proteínas, alterando a composição das células do tecido ósseo dos pacientes com esta variação da patologia em relação a pessoas sem a patologia, ou gerar produtos gênicos que podem atuar como reguladores da expressão gênica, comprometendo a produção de proteínas importantes para a osteogênese. O perfil de expressão dos genes do **Grupo C** pode ser observado na Figura 20 e o padrão do grupo pode ser observado com mais clareza através do gene CR593560, exemplar do grupo, mostrado na Figura 21.

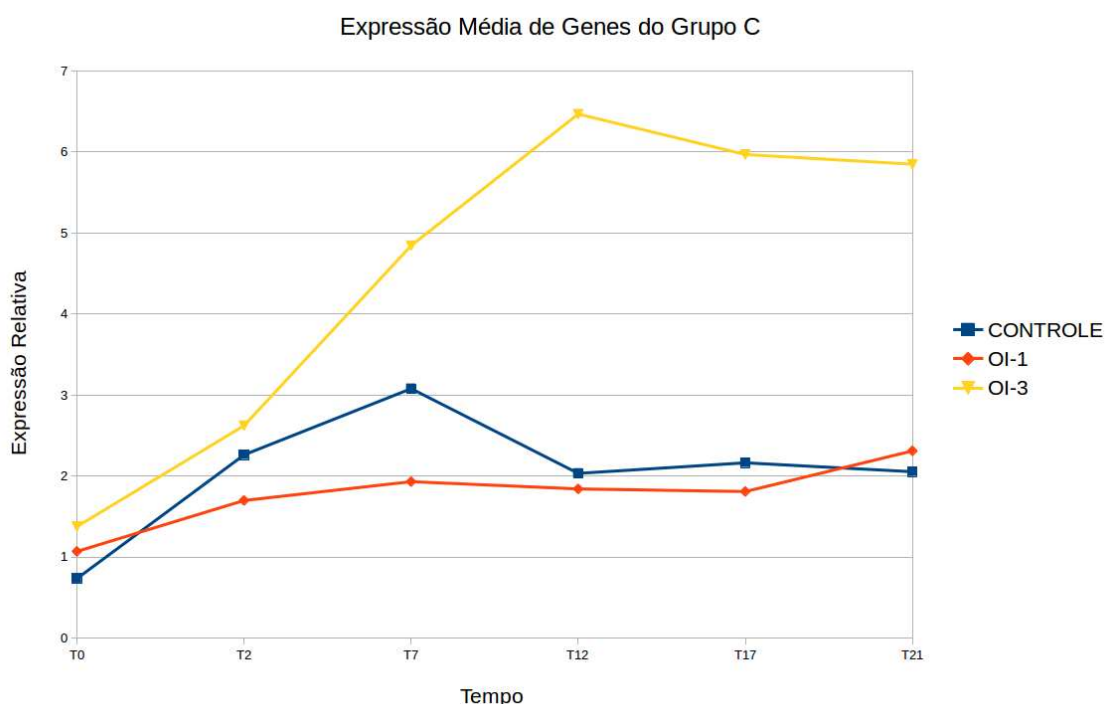


Figura 20 – Expressão relativa dos genes do Grupo C em função do tempo.

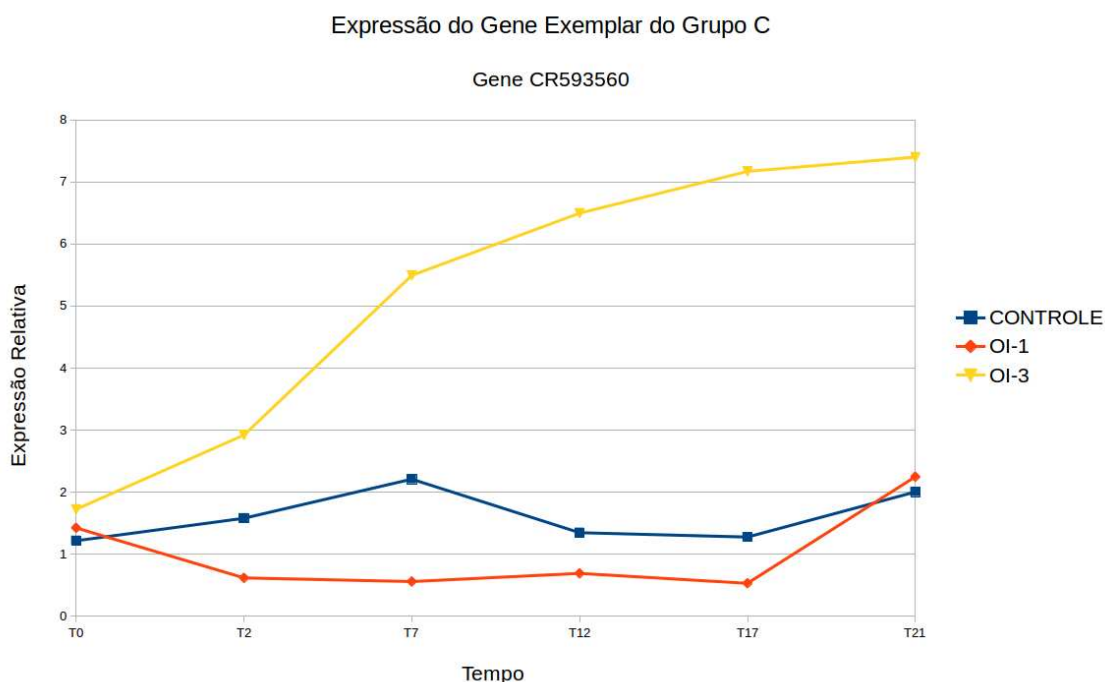


Figura 21 – Expressão relativa do gene CR593560, exemplar do Grupo C selecionado pelo *Affinity Propagation*.

Entre os genes presentes no **Grupo C** está o gene IBSP que gera a maior proteína estrutural da matriz óssea. No entanto esta é uma das proteínas não colagenosas e sua produção em excesso pode estar relacionada a uma das características principais da OI que é a produção reduzida ou defeituosa de colágeno. Além destes genes, foram encontrados os genes W60781 e STEAP4, para os quais existem estudos que apontam que as proteínas geradas por estes genes estão relacionados à formação de adipócitos, reforçando a hipótese apresentada por Kaneto (2011) de um desvio de osteogênese para adipogênese nos portadores de OI Tipo III.

Nenhum dos genes classificados nos grupos A,B e C fazem parte da lista de genes com relações conhecidas com a OI apresentados por Van Dijk e Sillence (2014) na caracterização molecular dos cinco da patologia. No entanto, uma vez que os genes encontrados na literatura por Van Dijk e Sillence (2014) estão relacionados com a patologia através de algumas mutações nestes genes, eles podem levar a produção defeituosa de proteínas, mesmo não interferindo diretamente nos perfis de expressão gênica.

5 Considerações Finais

Neste trabalho foram analisados dados de expressão gênica de células tronco mesenquimais da medula óssea de pacientes portadores de Osteogênese Imperfeita (tipos I e III) utilizando os algoritmos de agrupamento K-means, Mapas auto-organizáveis e *Affinity Propagation* para evidenciar perfis de expressão diferencial que sugiram grupos de genes associados à patologia.

A análise dos perfis de expressão gênica por diferentes algoritmos permitiu observar grupos de genes com perfil interessante para compreensão da patologia, reforçando a semelhança numérica dos padrões encontrados, quando os mesmos genes são agrupados de forma semelhante através de diferentes algoritmos.

Nos três agrupamentos, surgiu um grupo de genes que apresenta expressão **alta** nas amostras de controle e **baixa** nas amostras dos pacientes portadores da patologia, sugerindo a produção reduzida de um conjunto de produtos gênicos e, outro grupo que apresenta expressão **baixa** nas amostras de controle e **alta** nas amostras dos pacientes portadores da patologia, o que pode alterar a composição do tecido resultante por possuir proporções elevadas de outras proteínas.

Especificamente para o conjunto de amostras analisadas no estudo de caso deste trabalho, o *Affinity Propagation* foi capaz de evidenciar os perfis de expressão mais específicos encontrando um grupo de genes que teve sua expressão relativa aumentada apenas nos pacientes portadores da OI Tipo III, podendo fornecer informações mais específicas sobre este tipo da patologia.

A análise de função dos genes envolvidos na patologia mostrou entre os grupos de interesse poucos genes envolvidos com a composição da matriz óssea como o *PRG4* e o *IBSP* e genes que reforçam uma hipótese de desvio da osteogênese para adipogênese como o *W60781* e o *STEAP4*. Apesar do pequeno número de genes encontrados não refletir diretamente a complexidade genética da patologia, este fato pode ser explicado por grande parte dos genes envolvidos com a patologia apresentarem mutações que levam à produção de proteínas defeituosas ou impedem a tradução completa das proteínas através de inserção prematura de códons de terminação de cadeia, características que podem não interferir diretamente na expressão gênica.

Além disso, a compreensão destes perfis de expressão característicos da patologia podem auxiliar no desenvolvimento de marcadores moleculares para diferentes tipos da patologia, bem como seus estágios, podendo apoiar diagnósticos e prognósticos com exames pouco invasivos.

Os resultados das buscas de funções conhecidas na base de genes do NCBI e lite-

ratura disponível mostrou-se pouco eficiente para descoberta de relações entre os genes agrupados uma vez que muitos genes mostrados ainda possuem pouca informação sobre sua função. É interessante complementar informações sobre estes genes com o estudo de mecanismos de regulação dos mesmos, microRNAs e RNAs de interferência, e funções associadas às proteínas geradas.

É importante ainda ressaltar que a partir do estudo detalhado destes genes pode ser necessário validar os níveis de expressão encontrados nos microarranjos através de técnicas mais específicas como o PCR de tempo real, além do estudo destes perfis para um conjunto maior de pacientes.

Como trabalhos futuros em relação à Osteogênese Imperfeita, ainda cabem outras análises de expressão gênica similares, principalmente que permitam comparação dos níveis de expressão gênica durante a osteogênese para os cinco tipos da patologia catalogados, que forneça parâmetros para uma caracterização molecular mais específica destas variações. O estudo de expressão gênica direcionado para o desenvolvimento de marcadores moleculares para os diferentes tipos da doença pode contribuir significativamente com diagnósticos e prognósticos.

Em relação aos algoritmos estudados, principalmente o *Affinity Propagation* que é relativamente recente, é interessante o desenvolvimento de outros trabalhos de análise de expressão gênica diferencial em condições de alterações fisiológicas (relacionadas ou não a patologias) com a combinação de diferentes algoritmos e o desenvolvimento de parâmetros e modelos matemáticos que direcionem a escolha adequada do algoritmo em função da natureza e organização dos dados.

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.
- ANDERSON, M. W.; SCHRIJVER, I. Next generation DNA sequencing and the future of genomic medicine. **Genes**, v. 1, n. 1, p. 38–69, 2010. ISSN 20734425.
- BALDI, P.; BRUNAK, S. **Bioinformatics: the machine learning approach**. 2. ed. Massachusetts: The MIT Press, 2001. ISBN 026202506X.
- BARTON, G. J. Creation and Analysis of Protein Multiple Sequence Alignments. In: BAXEVANIS, A. D.; OUELLETTE, B. F. F. (Ed.). **Bioinformatics: A practical guide to the analysis of genes an proteins**. 2. ed. New York: Wiley Interscience, 2001. cap. 9, p. 215–232.
- BERG, J.; TYMOCZKO, J.; STRYER, L. **Biochemistry**. 5. ed. [S.l.]: W. H. Freeman and Company, 2002. 1050 p. ISBN 0-7167-3051-0.
- BODENHOFER, U.; KOTHMEIER, A.; HOCHREITER, S. Apcluster: An R package for affinity propagation clustering. **Bioinformatics**, v. 27, n. 17, p. 2463–2464, 2011. ISSN 13674803.
- BOLSTAD, B.; IRIZARRY, R.; ASTRAND, M.; SPEED, T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. **Bioinformatics**, v. 19, n. 2, p. 185–193, 2003. ISSN 1367-4803.
- CHANDRASEKHAR, T.; THANGAVEL, K.; ELAYARAJA, E. Effective Clustering Algorithms for Gene Expression Data. **International Journal of Computer Applications**, v. 32, n. 4, p. 25–29, 2011.
- CHAVEZ-ALVAREZ, R.; CHAVOYA, A.; MENDEZ-VAZQUEZ, A. Discovery of Possible Gene Relationships through the Application of Self-Organizing Maps to DNA Microarray Databases. **PLoS ONE**, v. 9, n. 4, p. e93233, 2014. ISSN 1932-6203.
- CHUANG, C.-c.; LI, Y.-c.; JENG, J.-t.; CHANG, C.-k.; WANG, Z.-q. Feature Genes Selection of Adult ALL Microarray Data with Affinity Propagation Clustering. **International Conference on Consumer Electronics**, p. 230–231, 2015.
- CONSORTIUM, I. H. G. Initial sequencing and analysis of the human genome. **Nature**, v. 420, n. February, p. 520–562, 2002.
- CORTESI, L.; RAZZABONI, E.; TOSS, A.; De Matteis, E.; MARCHI, I.; MEDICI, V.; TAZZIOLI, G.; ANDREOTTI, A.; De Santis, G.; PIGNATTI, M.; FEDERICO, M. A rapid genetic counselling and testing in newly diagnosed breast cancer is associated with high rate of risk-reducing mastectomy in BRCA1/2-positive Italian women. **Annals of oncology : official journal of the European Society for Medical Oncology / ESMO**, v. 25, n. 1, p. 57–63, jan. 2014. ISSN 1569-8041.

COVELL, D. G.; WALLQVIST, A.; RABOW, A. a.; THANKI, N. Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data. **Molecular cancer therapeutics**, v. 2, n. 3, p. 317–332, 2003. ISSN 1535-7163.

DESRIAC, N.; POSTOLLEC, F.; COROLLER, L.; SOHIER, D.; ABEE, T.; BESTEN, H. M. W. den. Prediction of *Bacillus weihenstephanensis* acid resistance: the use of gene expression patterns to select potential biomarkers. **International journal of food microbiology**, Elsevier B.V., v. 167, n. 1, p. 80–6, out. 2013. ISSN 1879-3460.

DI STEFANO, V.; ZACCAGNINI, G.; CAPOGROSSI, M. C.; MARTELLI, F. microRNAs as peripheral blood biomarkers of cardiovascular disease. **Vascular pharmacology**, Elsevier Inc., v. 55, n. 4, p. 111–8, out. 2011. ISSN 1879-3649.

DOUGHERTY, E. R. The fundamental role of pattern recognition for gene-expression/microarray data in bioinformatics. **Pattern Recognition**, v. 38, n. 12, p. 2226–2228, dez. 2005. ISSN 00313203.

DUECK, D.; FREY, B. J. Non-metric affinity propagation for unsupervised image categorization. **Iccv**, 2007.

FERREIRA FILHO, D. **Estudo de expressão gênica em citros utilizando modelos lineares**. Tese (Mestrado em Ciências) — Universidade de São Paulo, 2009.

FREY, B. J.; DUECK, D. Clustering by Passing Messages Between Data Points. **Science**, v. 315, n. February, p. 972–976, 2007. ISSN 0036-8075.

GRIFFITH, F. The Significance of Pneumococcal Types. **Journ. of Hyg.**, v. 27, n. 2, p. 113–159, 1928. ISSN 0022-1724.

GRIFFITHS, A. J. F.; WESSLER, S. R.; LEWONTIN, R. C.; CARROLL, S. B. **Introdução à Genética**. 9. ed. Rio de Janeiro: Guanabara Koogan, 2008. ISBN 9788527714976.

HAYKIN, S. **Neural Networks: A comprehensive foundation**. 2. ed. New Jersey: Prentice-Hall, 1999.

HERSHEY, A. D.; CHASE, M. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. **The Journal of General Physiology**, p. 39–56, 1952.

HU, J.; HE, X. Enhanced quantile normalization of microarray data to reduce loss of information in gene expression profiles. **Biometrics**, v. 63, n. 1, p. 50–59, 2007. ISSN 0006341X.

KANETO, C. M. **Análise da Expressão Gênica durante a diferenciação osteogênica de células mesenquimais estromais de medula óssea de pacientes portadores de Osteogênese Imperfeita**. 1–106 p. Tese (Tese de Doutorado) — Faculdade de Medicina de Ribeirão Preto/USP, 2011.

KASABOV, N. Global, local and personalised modeling and pattern discovery in bioinformatics: An integrated approach. **Pattern Recognition Letters**, v. 28, n. 6, p. 673–685, abr. 2007. ISSN 01678655.

KOHONEN, T. Self-organized formation of topologically correct feature maps. **Biological Cybernetics**, v. 43, p. 59–69, 1982. ISSN 0340-1200.

KULTERER, B.; FRIEDL, G.; JANDROSITZ, A.; SANCHEZ-CABO, F.; PROKESCH, A.; PAAR, C.; SCHEIDELER, M.; WINDHAGER, R.; PREISEGGER, K.-h.; TRAJANOSKI, Z. Gene expression profiling of human mesenchymal stem cells derived from bone marrow during expansion and osteoblast differentiation. **BMC Genomics**, v. 15, p. 1–15, 2007.

LEHNINGER, A.; NELSON, D. L.; COX, M. M. **Principles of biochemistry**. 5. ed. New York: W.H. Freeman and Company, 2005. 1100 p. ISBN 9780716771081.

LI, C.; PEI, F.; ZHU, X.; DUAN, D. D.; ZENG, C. Circulating microRNAs as novel and sensitive biomarkers of acute myocardial Infarction. **Clinical biochemistry**, The Canadian Society of Clinical Chemists, v. 45, n. 10-11, p. 727–32, jul. 2012. ISSN 1873-2933.

LIEW, A. W.-C.; YAN, H.; YANG, M. Pattern recognition techniques for the emerging field of bioinformatics: A review. **Pattern Recognition**, v. 38, n. 11, p. 2055–2073, nov. 2005. ISSN 00313203.

LIN, T.-C.; LIU, R.-S.; CHEN, C.-Y.; CHAO, Y.-T.; CHEN, S.-Y. Pattern classification in DNA microarray data of multiple tumor types. **Pattern Recognition**, v. 39, n. 12, p. 2426–2438, dez. 2006. ISSN 00313203.

LINDERT, U.; KRAENZLIN, M.; CAMPOS-XAVIER, A. B.; BAUMGARTNER, M. R.; BONAFÉ, L.; GIUNTA, C.; ROHRBACH, M. Urinary pyridinoline cross-links as biomarkers of osteogenesis imperfecta. **Orphanet Journal of Rare Diseases**, Orphanet Journal of Rare Diseases, v. 10, n. 1, p. 104, 2015. ISSN 1750-1172.

LODISH, H.; Arnold Berk; MATSUDAIRA, P.; KEISER, C. A.; KRIEGER, M.; SCOTT, M. P.; ZIPURSKY, L.; James Darnell. **Molecular Cell Biology**. 5. ed. [S.l.]: W. H. Freeman, 2003. ISBN 0716788756.

MENDELL, J. T.; OLSON, E. N. MicroRNAs in stress signaling and human disease. **Cell**, Elsevier Inc., v. 148, n. 6, p. 1172–87, mar. 2012. ISSN 1097-4172.

METZKER, M. L. Sequencing technologies - the next generation. **Nature reviews. Genetics**, Nature Publishing Group, v. 11, n. 1, p. 31–46, 2010. ISSN 1471-0056. Disponível em: <<http://dx.doi.org/10.1038/nrg2626>>.

MONTI, S.; TAMAYO, P.; MESIROV, J.; GOLUB, T.; SEBASTIANI, P.; KOHANE, I. S.; RAMONI, M. F. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. **Machine Learning**, v. 52, n. i, p. 91–118, 2003. ISSN 0885-6125, 1573-0565.

MOUNT, D. W. **Bioinformatics: sequence and genome analysis**. 2. ed. Tucson: University of Arizona, 2004.

MUTHUKALATHI, S.; RAMANUJAM, R.; THALAMUTHU, A. Consensus Clustering for Microarray Gene Expression Data. **Bonfring International Journal of Data Mining**, v. 4, n. 4, p. 26–33, 2014. ISSN 2250107X.

NAPOLEON, D.; BASKAR, G. An Efficient K-Means with Microarray Gene Expression Using Affinity Propagation for Cancer Dataset. **International Journal of Advanced Research in Computer Science**, v. 2, n. 3, p. 172–176, 2011.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of molecular biology**, v. 48, n. 3, p. 443–453, 1970. ISSN 00222836.

Qizheng Sheng, Yves Moreau, Frank De Smet, Kathleen Marchal, B. D. M. Advances in Cluster Analysis of Microarray Data. In: AZUAJE, F.; DOPAZO, J. (Ed.). **Data analysis and visualization in genomics and proteomics**. West Sussex: Wiley, 2005. cap. 10, p. 153–173. ISBN 0470094397.

QUACKENBUSH, J. COMPUTATIONAL ANALYSIS OF MICROARRAY DATA. **Nature reviews. Genetics**, v. 2, n. June, p. 418–427, 2001.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2015. Disponível em: <<https://www.R-project.org/>>.

RANGREJ, A.; KULKARNI, S.; TENDULKAR, A. V. Comparative Study of Clustering Techniques for Short Text Documents. **Proceedings of the 20th international conference companion on World wide web**, p. 111–112, 2011.

REID, G.; KIRSCHNER, M. B.; ZANDWIJK, N. van. Circulating microRNAs: Association with disease and potential use as biomarkers. **Critical reviews in oncology/hematology**, Elsevier Ireland Ltd, v. 80, n. 2, p. 193–208, nov. 2011. ISSN 1879-0461.

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, 2015. ISSN 0305-1048.

RITCHIE, M. E.; SILVER, J.; OSHLACK, A.; HOLMES, M.; DIYAGAMA, D.; HOLLOWAY, A.; SMYTH, G. K. A comparison of background correction methods for two-colour microarrays. **Bioinformatics**, v. 23, n. 20, p. 2700–2707, 2007. ISSN 1367-4803.

RUSPINI, E. H. A new approach to clustering. **Information and Control**, v. 15, p. 22–32, 1969. ISSN 00199958.

SAMISH, I.; BOURNE, P. E.; NAJMANOVICH, R. J. Achievements and challenges in structural bioinformatics and computational biophysics. **Bioinformatics (Oxford, England)**, v. 31, n. 1, p. 146–50, 2015. ISSN 1367-4811.

SANGER, F.; NICKLEN, S. DNA sequencing with chain-terminating. **PNAS**, v. 74, p. 5463–5467, 1977. ISSN 07407378.

SCHULER, G. D. Sequence Alignment and Database Searching. In: **Bioinformatics: A practical guide to the analysis of genes and proteins**. [S.l.: s.n.], 2001. cap. 8, p. 187–214.

SILLENCE, D. O.; SENN, A.; DANKS, D. M. Genetic Heterogeneity in osteogenesis imperfecta. **Journal of Medical Genetics**, v. 16, p. 101–116, 1979.

SILVER, J. D.; RITCHIE, M. E.; SMYTH, G. K. Microarray background correction: Maximum likelihood estimation for the normal-exponential convolution. **Biostatistics**, v. 10, n. 2, p. 352–363, 2009. ISSN 14654644.

SLAVKOV, I.; DZEROSKI, S.; PETERLIN, B.; LOVRECIĆ, L. Analysis of Huntington's Disease Gene Expression Profiles with Predictive Clustering Trees. **Informatica Medica Slovenica**, v. 11, p. 1–7, 2006.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. 4. ed. San Diego: Elsevier, 2009.

VAN DIJK, F.; SILLENCE, D. Osteogenesis imperfecta: Clinical diagnosis, nomenclature and severity assessment. **American Journal of Medical Genetics Part A**, v. 164, n. 6, p. 1470–1481, 2014. ISSN 15524825.

VAN 'T VEER, L. J.; DAI, H.; VIJVER, M. J. Van de; HE, Y. D.; HART, A. a. M.; MAO, M.; PETERSE, H. L.; KOOY, K. van der; MARTON, M. J.; WITTEVEEN, A. T.; SCHREIBER, G. J.; KERKHOVEN, R. M.; ROBERTS, C.; LINSLEY, P. S.; BERNARDS, R.; FRIEND, S. H. Gene expression profiling predicts clinical outcome of breast cancer. **Nature**, v. 415, n. 6871, p. 530–536, 2002. ISSN 00280836.

VENTER, J. C. et. al. The Sequence of the Human Genome. **Science**, v. 291, n. 5507, p. 1304–1351, 2001. ISSN 0036-8075.

VLASBLOM, J.; WODAK, S. J. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. **BMC bioinformatics**, v. 10, p. 99, 2009. ISSN 1471-2105.

WALLACE, D. J.; CHAU, F. Y.; SANTIAGO-TURLA, C.; HAUSER, M.; CHALLA, P.; LEE, P. P.; HERNDON, L. W.; ALLINGHAM, R. R. Osteogenesis imperfecta and primary open angle glaucoma: genotypic analysis of a new phenotypic association. **Molecular vision**, v. 20, n. August, p. 1174–81, 2014. ISSN 1090-0535.

WATSON, J. D.; CRICK, F. H. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. **Nature**, v. 171, p. 737–738, 1953. ISSN 00348376.

WEHRENS, R.; BUYDENS, L. M. C. Self- and Super-organizing Maps in R: The kohonen Package. **Journal of Statistical Software**, v. 21, n. 5, p. 1–19, 2007. ISSN 1548-7660.

WITTMANN, J.; JÄCK, H.-M. Serum microRNAs as powerful cancer biomarkers. **Biochimica et biophysica acta**, Elsevier B.V., v. 1806, n. 2, p. 200–7, dez. 2010. ISSN 0006-3002.

ZARAVINOS, A.; LAMBROU, G. I.; BOULALAS, I.; DELAKAS, D.; SPANDIDOS, D. a. Identification of common differentially expressed genes in urinary bladder cancer. **PloS one**, v. 6, n. 4, p. e18135, 2011. ISSN 1932-6203.

ZHAO, Y.-Y.; LIN, R.-C. UPLC-MS(E) application in disease biomarker discovery: the discoveries in proteomics to metabolomics. **Chemico-biological interactions**, Elsevier Ireland Ltd, v. 215, p. 7–16, maio 2014. ISSN 1872-7786.