

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Data Mining para análise dos resultados de Gene Expression

Luís Miguel Barroso Natividade



Mestrado Integrado em Engenharia Informática e Computação

Orientador: Rui Carlos Camacho de Sousa Ferreira da Silva

Co-orientador: Nuno Fonseca

25 de Julho de 2017

Data Mining* para análise dos resultados de *Gene Expression

Luís Miguel Barroso Natividade

Mestrado Integrado em Engenharia Informática e Computação

Aprovado em provas públicas pelo Júri:

Presidente: Jorge Barbosa

Arguente: Carlos Ferreira

Vogal: Rui Camacho

25 de Julho de 2017

Resumo

Atualmente vivemos numa era em que a tecnologia está envolvida em todas as áreas e a evolução tecnológica tem consequências diretas no estudo das diferentes áreas científicas.

Na área da biologia, a sequenciação de genomas tem sofrido enormes avanços nos últimos anos. Tornou-se mais precisa, mais rápida e menos custosa financeiramente.

Estes progressos têm como consequência uma maior utilização desta tecnologia na realização de estudos mais profundos e complexos em genómica, nomeadamente em estudos de investigação da origem genómica dos diferentes tipos de cancro.

Uma das características desta nova tecnologia de sequenciação é que requer consideráveis recursos computacionais e gera uma enorme quantidade de dados, que impossibilita a análise manual desses dados para obtenção de conclusões por parte de especialistas.

Derivado da enorme quantidade de dados gerada e da quantidade de informação disponibilizada na internet nos dias de hoje, existem já várias bases de dados acessíveis na WEB com este tipo de informação. Embora seja bastante positivo o facto de existir muita informação, em diferentes sítios WEB, torna-se trabalhoso e por vezes complexo localizar toda a informação necessária relativa a um gene. Acresce ainda a dificuldade de os identificadores das entidades biológicas, como genes, serem, frequentemente, diferentes em cada sítio WEB.

O principal objetivo desta dissertação é facilitar o trabalho dos especialistas de investigação biológica. Para isso pretendemos resolver o problema relativo à obtenção de informação relativa a genes, que estes enfrentam no seu dia a dia, assim como o problema que consiste na análise da informação obtida.

Para tal elaborámos uma plataforma WEB que permite a utilização de diferentes técnicas de data mining, técnicas de classificação e *clustering* de modo a permitir aos especialistas retirar conclusões na análise de resultados da expressão genética [Wol13]. Em complemento e de modo a simplificar o trabalho dos especialistas, a plataforma permite também a recolha de informação génica de diferentes bases de dados, sendo possível extrair essa informação para vários formatos de ficheiro, para utilização posterior. Almejando um leque de utilizadores alargado, a plataforma possui uma interface simples e intuitiva, permitindo-a ser utilizável por utilizadores sem grande experiência em informática.

A avaliação da plataforma foi feita através de uma avaliação objetiva, própria das ferramentas de data mining, e subjetiva, recorrendo a especialistas do I3S¹.

Palavras-chave: Genómica, *data mining*, classificação, *clustering*, WEB

¹<http://www.i3s.up.pt/>

Abstract

We currently live in an age when technology is involved in all areas and technological evolution has direct consequences in the study of different scientific areas.

In the area of biology, genome sequencing has undergone tremendous advances in recent years. It has become more accurate, faster and less costly financially.

These developments lead to increased use of this technology in carrying out deeper and more complex studies in genomics, in particular in research studies on the genomic origin of different types of cancer.

One of the characteristics of these new sequencing technology is that it requires considerable computational resources and generates an enormous amount of data, which makes it impossible to manually analyze these data to obtain conclusions from the experts.

Derived from the enormous amount of data generated and the amount of information available on the Internet these days, there are already several databases accessible on the WEB with this type of information. Although it is quite positive that there is a lot of information on different websites, it is arduous and complex to find all the necessary information about a gene. In addition, it gets more difficult because often each database has its own identifier for each gene.

The final objective of this dissertation is the elaboration of a platform for the use of biological research specialists, which will facilitate their work, thus allowing the development of progress in the investigation of various diseases of genomic origin, such as cancers or tumors.

In order to accomplish this we have developed a WEB Platform that allows the use of different data mining techniques, classification and *clustering* techniques in order to allow the experts to draw conclusions in the analysis of results of the genetic expression [Wol13]. In addition, and in order to simplify the work of the specialists, the platform also allows the collection of gene information from different databases, being possible to extract this information for several file formats, for later use. Targeting a wide range of users the platform has a simple and intuitive interface, allowing it to be usable by users without great experience in computing.

The evaluation of the platform was done through an objective evaluation, own of the tools of data mining, and subjective, resorting to specialists of I3S ².

Keywords: Genomics, data mining, classification, clustering, WEB

²<http://www.i3s.up.pt/>

Agradecimentos

Sendo que a conclusão da presente dissertação representa também o término de uma importante etapa da minha vida, não podia deixar de agradecer aos que de alguma forma contribuíram para o culminar deste ciclo.

Em primeiro lugar agradeço à minha família, em especial aos meus pais e ao meu irmão por todo o apoio que sempre me deram, a confiança que em mim depositaram e a compreensão que sempre tiveram, mesmo nos muitos momentos em que estive ausente.

Ao Professor Rui Camacho, meu orientador, agradeço toda a ajuda e acompanhamento prestado ao longo do desenvolvimento desta dissertação, mostrando-se sempre disponível para qualquer esclarecimento.

Aos meus amigos e colegas, que tive o prazer de conhecer neste ciclo que agora concluo, obrigado pela ajuda e companhia não só nas muitas horas de trabalho, mas também nos momentos de descontração.

À AEFEUP, que integrei nos últimos anos, e a todos os elementos com quem trabalhei e formei amizade, um sincero obrigado.

E por fim, mas não menos importante aos amigos de longa data, maioritariamente de Cabeceiras de Basto, obrigado pela companhia, pela ajuda, pela amizade.

Muito Obrigado,
Luís Natividade

*“Life is like riding a bicycle.
To keep your balance, you
must keep moving”*

Albert Einstein

Conteúdo

1	Introdução	1
1.1	Contexto	1
1.2	Motivação e objetivos	2
1.3	Projeto	2
1.3.1	Metodologia	2
1.4	Estrutura da dissertação	3
2	Conceitos Básicos, Data Mining e Tecnologias WEB	5
2.1	Estado de Arte em Biologia Molecular e Genómica	5
2.1.1	Genómica	5
2.1.2	Sequenciação	6
2.1.3	RNA-Sequencing	6
2.1.4	Expressão Génica	6
2.2	Repositórios de Dados Biológicos	7
2.2.1	ENSEMBL	8
2.2.2	GenBank	9
2.2.3	Kegg	10
2.3	APIs de Repositórios de Dados Biológicos	12
2.3.1	API Ensembl	12
2.3.2	API GenBank	14
2.3.3	API Kegg	14
2.4	Conversão de identificadores de diferentes repositórios	15
2.4.1	BioDB Hyperlink Management System	15
2.4.2	API BioDB Hyperlink Management System	16
2.5	Data Mining	16
2.5.1	Classificação	17
2.5.2	Regressão	18
2.5.3	Associação	18
2.5.4	<i>Clustering</i>	18
2.6	Classificação	19
2.6.1	Algoritmos de Classificação	19
2.6.2	Métodos de avaliação de Classificação	21
2.6.3	Métricas de Classificação	22
2.7	<i>Clustering</i>	25
2.7.1	Técnicas de <i>Clustering</i>	25
2.7.2	Algoritmos de <i>Clustering</i>	27
2.7.3	Métodos de avaliação de <i>Clustering</i>	29
2.8	Ferramentas de <i>Data Mining</i>	29

CONTEÚDO

2.8.1	RapidMiner	30
2.8.2	Weka	30
2.8.3	R	31
2.8.4	KNIME	31
2.8.5	SPSS	31
2.9	Formato <i>ARFF</i>	31
2.10	Tecnologias	32
2.10.1	Python	32
2.10.2	Django	32
2.10.3	BootStrap	34
2.10.4	JavaScript	34
2.10.5	SQLite	34
2.11	Conclusões	34
3	Plataforma WEB	35
3.1	Problema	35
3.2	Solução	36
3.3	Implementação	37
3.3.1	Arquitetura	37
3.3.2	Pesquisa de genes	38
3.3.3	Conversão de Genes	39
3.3.4	Resultados de pesquisa	39
3.3.5	Recolha de dados	39
3.3.6	Base de dados	40
3.3.7	Listagem de genes	42
3.3.8	Processamento de ficheiro <i>ARFF</i> para <i>data mining</i>	42
3.3.9	Aplicação de técnicas de <i>data mining</i>	42
3.3.10	Interface de administração	44
3.4	Funcionamento da Plataforma WEB	44
3.4.1	Menu	44
3.4.2	Casos de Uso	44
3.4.3	Pesquisa de genes	45
3.4.4	Resultados da Pesquisa e Recolha de informação	48
3.4.5	Listagem de genes	48
3.4.6	Conversão de identificadores de genes	49
3.4.7	Aplicação de técnicas de <i>data mining</i>	50
3.4.8	Resultados de aplicação de técnicas de <i>data mining</i>	53
3.5	Conclusões	54
4	Resultados e Casos de Estudo	55
4.1	Ambiente Experimental	55
4.2	Atributos estudados	55
4.3	Caso de Estudo 1	57
4.3.1	Especificação do caso	57
4.3.2	Dados analisados	57
4.3.3	Metodologia	57
4.3.4	Resultados	58
4.4	Caso de Estudo 2	62
4.4.1	Especificação do caso	62

CONTEÚDO

4.4.2	Dados analisados	62
4.4.3	Metodologia	62
4.4.4	Resultados	62
4.5	Caso de Estudo 3	64
4.5.1	Especificação do caso	64
4.5.2	Dados analisados	64
4.5.3	Metodologia	64
4.5.4	Resultados	64
4.6	Avaliação e comparação da Plataforma WEB	65
4.6.1	Procedimentos	66
4.6.2	Eficiência	67
4.7	Conclusões	68
5	Conclusões e Trabalho Futuro	69
5.1	Conclusões	69
5.2	Trabalho futuro	70
5.2.1	Adicionar novos repositórios	70
5.2.2	Adicionar outros métodos de <i>data mining</i>	70
5.2.3	Adicionar novos formatos para descarregar informação	70
A	Caso de Estudo 1	71
A.1	Dados analisados	71
A.2	Resultados	72
A.2.1	Resultados da pesquisa de informação	72
A.2.2	Resultados dos algoritmos de <i>data mining</i>	78
	Referências	99

CONTEÚDO

Lista de Figuras

2.1	Diagrama de <i>gene expression</i> (simplificado)	7
2.2	Pesquisa de um gene na interface web do projeto ENSEMBL	9
2.3	Pesquisa do gene com ID 10 na interface web do projeto GenBank	10
2.4	Pesquisa do gene A1CF na interface WEB do projeto Kegg	13
2.5	Formato de pedido à API do projeto Kegg	15
2.6	Diferentes métodos de classificação	18
2.7	Clustering	19
2.8	<i>Support Vector Machine</i> - Exemplo	20
2.9	<i>Precision vs Accuracy</i>	23
2.10	ROC-Curves	24
2.11	Exemplo do método k-means	27
2.12	Interface do WEKA	30
2.13	Interação das componentes do MVC e do utilizador	33
3.1	Arquitetura da plataforma	37
3.2	Diagrama UML da base de dados	41
3.3	Menu de navegação na Plataforma WEB, ativo na funcionalidade <i>Genes Search</i>	44
3.4	Diagrama de casos de uso para o ator Investigador	45
3.5	Diagrama de casos de uso para o ator Administrador	46
3.6	Página inicial para pesquisa de genes	46
3.7	Página para pesquisa de genes em todos os repositórios em simultâneo	47
3.8	Página com os resultados da pesquisa efetuada pelo utilizador	48
3.9	Listagem de todos os genes <i>Ensembl</i> incluídos na base de dados	49
3.10	Resultado de uma conversão	50
3.11	Página inicial relativa ao <i>data mining</i>	51
3.12	Opções disponíveis para o método <i>Expectation Maximization</i>	53
3.13	Explicação disponível para as opções do método <i>Expectation Maximization</i>	54
4.1	Resultados do algoritmo Simple K-means	59
4.2	Caracterização dos <i>clusters</i> resultantes do algoritmo Simple K-means	59
4.3	Resultados do algoritmo <i>Make Density Based Clusterer</i>	60
4.4	Caracterização dos <i>clusters</i> resultantes do algoritmo <i>Make Density Based Clusterer</i>	61
4.5	Resultados do algoritmo <i>Farthest-First</i>	63
4.6	Resultados do algoritmo <i>DBSCAN</i>	65
4.7	Procedimentos efetuados manualmente	66
4.8	Procedimentos efetuados na Plataforma WEB	67

LISTA DE FIGURAS

Lista de Tabelas

2.1	Identificador Ensembl	8
2.2	Bases de dados do projeto Kegg	11
2.3	Confusion Matrix	22
4.1	Especificações da máquina	55
4.2	Valores possíveis do atributo <i>MIM</i>	56
4.3	Descrição das opções do algoritmo <i>Simple K-means</i>	58
4.4	Descrição das opções do algoritmo <i>MDBC</i>	58
4.5	Opções escolhidas <i>Make Density Based Clusterer</i>	62
4.6	Opções escolhidas <i>DBSCAN</i>	64
4.7	Comparação entre eficiência na Plataforma WEB e manualmente	68
A.1	Simple K-means - Cluster 0 (sumarizado)	93
A.2	Simple K-means - Cluster 1 (sumarizado)	94
A.3	Simple K-means - Cluster 2 (sumarizado)	95
A.4	Make Density Based Clusterer - Cluster 0 (sumarizado)	96
A.5	Make Density Based Clusterer - Cluster 1 (sumarizado)	97
A.6	Make Density Based Clusterer - Cluster 2 (sumarizado)	98

LISTA DE TABELAS

Abreviaturas e Símbolos

API	Application Programming Interface
ARFF	Attribute-Relation File Format
AUC	Area under the curve
BIODB	Biological DataBase Network
CPU	Central Processing Unit
CSS	Cascading Style Sheets
DBSCAN	Density-based spatial clustering of applications with noise
DNA	Ácido desoxirribonucleico
DRY	Don't repeat yourself
EM	Expectation Maximization
EMBI-EBI	European Bioinformatics Institute
FF	Farthest-First
FTP	File Transfer Protocol
GUI	Graphical user interface
HTML	HyperText Markup Language
HTS	High-throughput sequencing
HTTP	Hypertext Transfer Protocol
ID	Identificador
I3S	Instituto de Investigação e Inovação em Saúde
IPATIMUP	Instituto de Patologia e Imunologia Molecular da Universidade do Porto
JSON	JavaScript Object Notation
Kegg	Kyoto Encyclopedia of Genes and Genomes
Knime	Konstanz Information Miner
MDBC	Make Density Based Clusterer
MIM	Mendelian Inheritance in Man
MVC	Model-View-Controller
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
REST	Representational State Transfer
RNA	Ácido ribonucleico
RNA-seq	RNA sequencing
SO	Sistema operativo
SVM	Support vector machine
tRNA	RNA transportador
URL	Uniform Resource Locator
WEKA	Waikato Environment for Knowledge Analysis
WTSI	Wellcome Trust Sanger Institute
XML	eXtensible Markup Language

Capítulo 1

Introdução

Atualmente a tecnologia está envolvida em praticamente todas as ações do nosso dia a dia. Todas as áreas envolvem tecnologia e a evolução desta permite avanços nas diferentes áreas.

Neste capítulo é feita uma contextualização do tema desta dissertação, uma exposição da minha motivação e dos objetivos que pretendemos alcançar com a realização da mesma. É também descrito o projeto realizado, bem como a estrutura deste documento.

1.1 Contexto

Com a evolução tecnológica que tem ocorrido nas últimas décadas também na área da biologia têm ocorrido melhorias diretamente relacionadas com esta evolução.

Dentro da biologia, a sequenciação de genomas tem sofrido enormes avanços, tornando-se mais precisa, mais rápida e menos custosa financeiramente.

Atualmente a obtenção de informação relativa a genes, proteínas ou outros produtos génicos é possível através de vários sítios WEB, que englobam dados de diferentes bases de dados, o que implica que os especialistas sejam obrigados a aceder aos diferentes sítios WEB de modo a obter o máximo informação possível relativa a um produto génico. Este é um problema que os biólogos enfrentam no seu dia a dia, agravado pelo facto de os identificadores dos genes serem, frequentemente, diferentes entre sítios WEB diferentes e do formato da informação recolhida de cada sítio WEB ser também frequentemente diferente, o que dificulta a sua agregação e análise simultânea.

Outro problema que os especialistas enfrentam no dia a dia recai sobre o facto de serem necessários consideráveis recursos computacionais para a tecnologia de sequenciação existente e desta resultam enormes quantidades de dados, tornando praticamente impossível a obtenção de conclusões manualmente, ou seja, sem o recurso a ferramentas computacionais. Estas ferramentas são muitas vezes complexas para utilizadores sem conhecimento profundo em informática, o que dificulta bastante o trabalho dos biólogos.

1.2 Motivação e objetivos

A principal motivação que tenho nesta dissertação é permitir que os especialistas da investigação biológica consigam ultrapassar alguns dos problemas que enfrentam no dia a dia, facilitando o seu trabalho e tornando-o menos custoso não só financeiramente, mas também a nível temporal.

Deste modo será possível acelerar processos na investigação biológica, podendo alcançar progressos na investigação de várias doenças relacionadas com os genes, como os vários tipos de cancro ou tumores.

Assim o principal objetivo desta dissertação é facilitar o trabalho dos especialistas de investigação biológica. Para isso pretendemos resolver o problema relativo à obtenção de informação relativa a genes, que estes enfrentam no seu dia a dia, assim como o problema que consiste na análise da informação obtida.

1.3 Projeto

Para alcançarmos os objetivos propostos concebemos uma plataforma WEB que permite procurar e recolher informação genómica de diferentes bases de dados, permitindo também aplicar diferentes técnicas de *data mining*, de modo a facilitar a obtenção de conclusões por parte dos biólogos, relativamente a um conjunto de genes.

Assim, a plataforma desenvolvida divide-se em duas partes:

- **Pesquisa e recolha de genes:** Neste módulo da plataforma é possível pesquisar e recolher informação genómica proveniente de diferentes bases de dados, relativa a um conjunto de genes. O *download* da informação pode ser feita em vários formatos, permitindo ao investigador utilizar a informação recolhida para diversos fins, não limitando o uso dessa informação na plataforma desenvolvida.
- **Análise de informação relacionada com genes:** Através do *upload* de um ficheiro no formato *arff*¹, contendo informação genómica, o utilizador pode aplicar diversas técnicas de classificação e *clustering*, podendo assim retirar conclusões relativamente ao conjunto de genes analisados. De notar que o ficheiro poderá ser obtido na plataforma desenvolvida, no entanto é também possível utilizar um ficheiro obtido de outras fontes.

1.3.1 Metodologia

O desenvolvimento deste projeto foi dividido em diferentes etapas, com o objetivo de simplificar a implementação de cada uma delas e facilitar a integração dos vários componentes.

A etapa inicial deste projeto focou na definição do problema de modo a conseguir encontrar uma solução que fosse ao encontro desse problema. Também nesta fase definimos os objetivos principais a alcançar com o desenvolvimento do projeto.

¹<http://www.cs.waikato.ac.nz/ml/weka/arff.html>

Após a fase de definição do problema e descrição da solução, estudámos as tecnologias existentes, procurando escolher as que melhor que enquadram com o projeto.

De seguida e após a escolha das tecnologias a utilizar, foi feito um estudo intensivo das diferentes áreas integradas nesta dissertação, de modo a melhor estruturar o projeto e efetuar um planeamento sólido e coerente. Assim começámos por estudar o estado da arte em biologia molecular e genómica, para melhor compreensão dos diferentes componentes dos genes e dos processos relativos a estes.

Em paralelo estudamos também os repositórios de genes existentes, de modo a escolher os que melhor se complementam para obter mais informação relativa a cada gene. Durante esta etapa pesquisamos também por ferramentas de conversão de identificadores de genes, de modo a conseguir integrar uma destas ferramentas no projeto a desenvolver.

Posteriormente foi feito um estudo intensivo aos diferentes tipos de *data mining* existentes, classificação, associação, regressão e *clustering*, de modo a perceber qual ou quais deveriam ser integrados no projeto. Após esta fase definimos quais os algoritmos mais importantes para os tipos de *data mining* escolhidos para integrar a aplicação.

Terminada a fase de estudo do estado da arte começamos a implementação definindo os requisitos da plataforma a desenvolver. Após esta fase começamos por desenvolver a pesquisa de informação relativa a genes nos repositórios selecionados.

Concluída a fase de obtenção de informação e recolha da mesma em diferentes formatos de ficheiro, começámos por implementar os algoritmos de *clustering*. Já numa fase bastante avançada do projeto implementámos os algoritmos de classificação.

Ao longo de todo o processo de implementação foram efetuados testes em cada etapa de modo a validar a integração de todos os componentes do projeto, e verificar o estado de cada fase estado da plataforma.

Após toda a implementação estar concluída foram efetuados testes, envolvendo diferentes casos de estudo de modo a permitir validar a plataforma desenvolvida.

1.4 Estrutura da dissertação

Esta dissertação é composta por cinco capítulos, começando com o capítulo atual, um capítulo introdutório ao tema que é estudado ao longo desta dissertação.

O capítulo 2, foca a explicação de alguns temas fulcrais para o bom entendimento deste projeto, contextualizando assim esta dissertação nas diversas áreas que a envolvem, com especial foco à área da genómica e à área da informática, mais especificamente o ramo do *data mining*.

É também neste capítulo que expomos algoritmos, técnicas, ferramentas e tecnologias existentes para os diferentes processos englobados neste projeto.

No capítulo 3, relativo à Plataforma desenvolvida, são referidos todos os detalhes relativos ao desenvolvimento deste projeto, apresentando detalhadamente os métodos utilizados para esse desenvolvimento.

Introdução

No capítulo 4 apresentamos os resultados deste projeto, incluindo alguns casos de estudo para avaliação do projeto.

O Capítulo 5 encerra esta dissertação descrevendo as conclusões retiradas da elaboração deste projeto, assim como qual o trabalho futuro esperado.

Capítulo 2

Conceitos Básicos, Data Mining e Tecnologias WEB

Neste capítulo é feita uma introdução aos diferentes temas relacionados com esta dissertação, de modo a facilitar a compreensão da mesma.

Começamos por apresentar uma breve explicação relativa a conceitos básicos da biologia fundamentais para o entendimento do projeto. Depois abordaremos tópicos relativos à sequenciação, expressão genética e RNA-seq.

É feita também uma revisão aos principais repositórios de genes existentes na WEB, descrevendo também as suas APIs.

Apresentamos também uma breve explicação relativa ao *data mining*, descrevendo sucintamente as principais técnicas utilizadas atualmente, dando maior ênfase à classificação e ao *clustering*, dado serem mais adequadas para a elaboração deste projeto.

Por fim serão apresentadas as tecnologias que melhor se enquadram com o desenvolvimento deste projeto, descrevendo cada uma delas.

2.1 Estado de Arte em Biologia Molecular e Genómica

2.1.1 Genómica

A genómica é uma área de conhecimento que foca no estudo do genoma de um organismo. Um genoma é o código genético que possui toda a informação hereditária de um ser. O genoma é codificado no DNA, ou em alguns vírus no RNA.

O DNA é um composto orgânico cujas moléculas contêm as instruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos e alguns vírus. O RNA é uma molécula responsável pela síntese de proteínas, pois faz a conexão entre estas e o DNA. Uma longa sequência de DNA que contém genes e outras sequências de nucleótidos é designada cromossoma.

Os genes são segmentos de DNA que contêm um código para a produção de aminoácidos, que são moléculas fundamentais para o bom funcionamento do organismo, pois são responsáveis por ajudar na formação dos tecidos corporais, enzimas, entre outros. A combinação de um número de aminoácidos forma uma proteína. No caso dos humanos vinte aminoácidos formam uma proteína.

O conjunto completo de transcritos presentes numa célula ou tecido é designado transcriptoma e é deste modo o reflexo direto da expressão dos genes.

2.1.2 Sequenciação

A compreensão do transcriptoma é essencial para interpretar os elementos funcionais do genoma, revelar os constituintes moleculares de células e tecidos nos diferentes estágios de desenvolvimento e para compreender os elementos presentes no desenvolvimento de doenças.

Obter informação de uma célula ou de um transcriptoma de um organismo é feito experiencialmente através da utilização de técnicas de sequenciação que têm evoluído ao longo dos tempos. O primeiro método utilizado foi o método *plus and minus* desenvolvido por Sanger e Coulson em 1975. Derivado da ineficácia deste método Sanger, juntamente com a sua equipa, continuou a desenvolver novos métodos e dois anos mais tarde, em 1977, apresentou um novo método, denominado Método de Sanger [FCK02]. Este método revolucionou o estudo da genómica nos anos que se seguiam, derivado da sua precisão bastante exata. No entanto este método era bastante lento e custoso, o que não permitia analisar grandes quantidades de genoma.

2.1.3 RNA-Sequencing

Nos últimos anos surgiram novas técnicas de sequenciação, bastante mais eficientes, denominadas *Next Generation Sequencing*. Estas técnicas diferem das anteriores dado que se baseiam em matrizes e combinam as técnicas desenvolvidas por Sanger para processar milhões de reações em paralelo, o que permite resultados bastante mais confiáveis, num menor período de tempo, sendo também menos custosos financeiramente [NGS].

Derivado da necessidade de melhor compreender o RNA na compreensão da expressão génica e consequência também do aumento de popularidade das técnicas NGS, surgiu o RNA-Seq.

2.1.4 Expressão Génica

Toda a informação genética de um organismo está contida no DNA do mesmo, situado no núcleo das células.

O DNA é uma sequência de bases ligadas em dupla-hélice cujas moléculas contêm as instruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos e alguns vírus. A diversidade humana é consequência das diferentes combinações possíveis entre as quatro bases existentes: Adenina (A), Citosina (C), Guanina (G) e Timina (T).

A expressão génica consiste no processo de descodificação de informação hereditária contida num gene num produto génico funcional, como proteínas ou RNA.

Este processo tem como base o código genético ou a sequência de nucleótidos de cada gene e divide-se em duas fases, a transcrição e a tradução [GENa].

A transcrição consiste na síntese de uma molécula de RNA usando como molde a sequência de uma cadeia de DNA de um gene. Vários tipos de RNA são formados nesta etapa, como o mRNA, responsável por especificar as sequências de aminoácidos que formam uma proteína, rRNA e tRNA que apenas serão importantes no processo seguinte, de tradução.

A tradução é a conversão da informação genética presente no mRNA numa sequência específica de aminoácidos. Neste processo, moléculas de tRNA reconhecem as sequências nucleótidas do mRNA e correlacionam-as com a sequência que corresponde a determinados aminoácidos.

A estrutura de um gene, de modo simplificado, baseia-se em sequências de intrões e exões. Os intrões são sequências de nucleótidos que não codificam qualquer parte da proteína, a sua principal função é separar os exões. Inicialmente são transcritos na molécula de pré-mRNA, mas são depois eliminados no processo de *splicing*¹. Os exões são também sequências de nucleótidos, mas o seu transcrito não é eliminado durante o *splicing*. Isto não acontece porque um exão pode codificar aminoácidos de uma proteína noutras moléculas de RNA, como tRNA ou rRNA.

Na figura 2.1 apresentamos um simples diagrama relativo a todo este processo que explicámos nos parágrafos anteriores [genb].

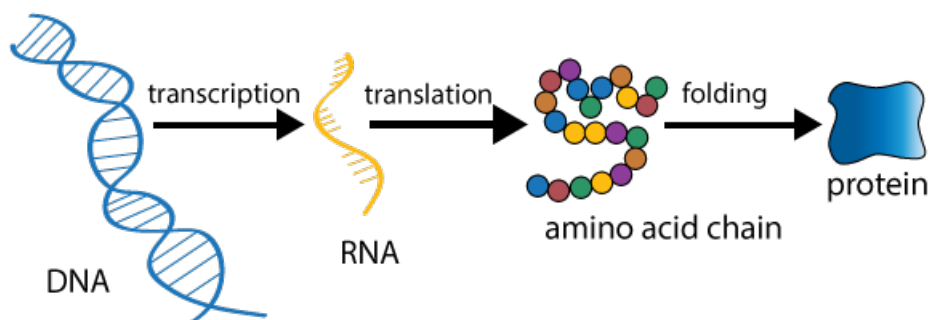


Figura 2.1: Diagrama de *gene expression* (simplificado)

2.2 Repositórios de Dados Biológicos

Embora os diferentes repositórios de dados biológicos partilhem alguma informação entre si, existe informação que apenas está disponível num repositório em específico.

De modo a obter a informação completa relativa a um gene específico, é então necessário aceder a vários repositórios disponíveis atualmente para consulta. Nesta secção apresentamos os sítios WEB que consideramos mais relevantes para os biólogos.

¹<https://www.nature.com/scitable/topicpage/rna-splicing-introns-exons-and-spliceosome-12375>

2.2.1 ENSEMBL

O *ENSEMBL*² [YAA⁺16], é um projeto conjunto do *European Bioinformatics Institute* (EMBL) e o *Wellcome Trust Sanger Institute* (WTSI), ambos localizados no *Wellcome Trust Genome Campus*, em Cambridge. Lançado em 1999, mas em constante atualização e crescimento, tem atualmente o contributo de cerca de 50 pessoas.

Este projeto destina-se maioritariamente a geneticistas e outros investigadores que estudam os genomas de espécies de vertebrados, com maior foco para a espécie humana, *Homo Sapiens*.

Embora em 1999 o genoma humano ainda não estivesse completo, era sabido, já nesta data, que a anotação de mais de 3 bilhões de pares de bases de sequência não deveria ser feita manualmente, dado que seria impensável a constante atualização a que está sujeito. Assim, o objetivo do projeto *ENSEMBL* consistia em anotar automaticamente o genoma, integrar essa anotação com outros dados biológicos disponíveis e permitir o acesso a estes dados através da *WEB*.

Ao longo dos anos os recursos deste projeto têm vindo a aumentar, englobando atualmente genómica comparativa, dados regulamentares e árvores genéticas.

Todos os dados pertencentes a este projeto estão armazenados numa base de dados MySQL, estando disponíveis gratuitamente para consulta, download ou acesso remoto, através da API disponibilizada.

Um identificador do repositório *Ensembl* consiste em cinco partes, como podemos observar na tabela 2.1.

Tabela 2.1: Identificador Ensembl

Exemplo ENSMUSG00000017167.6

Parte	1	2	3	4		5
Identificador	ENS	espécie	tipo de objeto	id	.	versão
Exemplo	ENS	MUS	G	00000017167	.	6
ENSMUSG00000017167.6						

De seguida descrevemos cada parte do identificador:

- **Parte 1:** indica que o identificador é pertence ao repositório *Ensembl*;
- **Parte 2:** código constituído por três letras que indicam a espécie. No caso de genes da espécie *hsa-homo sapiens* as três letras são omitidas.
- **Parte 3:** código constituído por uma ou duas letras que indicam o tipo de objeto:
 - **E:** Exão;
 - **FM:** Família de proteínas;

²<http://www.ensembl.org/>

- **G:** Gene;
- **GT:** Árvore de genes;
- **P:** Proteína;
- **R:** Característica regulatória;
- **T:** Transcrito;
- **Parte 4:** identificador único do objeto;
- **Parte 5:** versão do objeto.

A versão corresponde ao número de vezes que o objeto foi atualizado. É frequentemente omitida.

Na figura 2.2 podemos observar o resultado simplificado da pesquisa do gene com identificador *ENSG00000157764* na interface web do projeto *ENSEMBL*.

Figura 2.2: Pesquisa de um gene na interface web do projeto ENSEMBL

2.2.2 GenBank

O projeto *GenBank*³ [MOPT11], pertencente ao *National Center for Biotechnology Information* (NCBI) integra o *International Nucleotide Sequence Database Collaboration*, juntamente com o *DNA DataBank of Japan* (DDBJ) e o *European Nucleotide Archive* (ENA). Estas três organizações partilham informação diariamente, trabalhando várias vezes em conjunto.

Este projeto contém uma base de dados bastante abrangente de sequências de nucleótidos disponíveis publicamente para 370 000 espécies descritas. A cada dois meses é lançada uma nova atualização, mostrando assim a constante evolução deste projeto.

³<https://www.ncbi.nlm.nih.gov/genbank/>

O acesso aos dados deste projeto pode ser conseguido através da interface WEB disponibilizada, através de FTP ou através do *Entrez Programming Utilities (E-utilities)*, que funciona como uma API para as bases de dados do NCBI.

Na figura 2.3 podemos observar parte do resultado da pesquisa do gene com identificador 10 na interface web do projeto GenBank.

Gene

Gene 10[uid]

Create RSS Create alert Advanced

Full Report ▾ Send to: ▾

Showing Current Items.

NAT2 N-acetyltransferase 2 [Homo sapiens (human)]

Gene ID: 10, updated on 1-May-2017

Summary

Official Symbol NAT2 provided by [HGNC](#)

Official Full Name N-acetyltransferase 2 provided by [HGNC](#)

Primary source [HGNC:HGNC:7646](#)

See related [Ensembl:ENSG00000156006](#) [MIM:612182](#); [Vega:OTTHUMG00000130826](#)

Gene type protein coding

RefSeq status REVIEWED

Organism [Homo sapiens](#)

Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominidae; Homo

Also known as AAC2; PNAT; NAT-2

Summary This gene encodes an enzyme that functions to both activate and deactivate arylamine and hydrazine drugs and carcinogens. Polymorphisms in this gene are responsible for the N-acetylation polymorphism in which human populations segregate into rapid, intermediate, and slow acetylator phenotypes. Polymorphisms in this gene are also associated with higher incidences of cancer and drug toxicity. A second arylamine N-acetyltransferase gene (NAT1) is located near this gene (NAT2). [provided by RefSeq, Jul 2008]

Annotation Information Note: Loci in other organisms that are functionally homologous to this one are validly referred to as both NAT1 and NAT2: i.e., the functional homologs of NAT1 include mouse and rat Nat2, while the functional homologs of human NAT2 include mouse and rat Nat1. Name:sequence associations are consistent with current use in the field. [13 Feb 2013]

Orthologs [all](#)

Figura 2.3: Pesquisa do gene com ID 10 na interface web do projeto GenBank

2.2.3 Kegg

O projeto *Kyoto Encyclopedia of Genes and Genomes*⁴ [KFT⁺ 16] (*Kegg*) iniciou-se em 1995 e foi inicialmente desenvolvido com o objetivo de permitir a interpretação biológica de dados de sequência do genoma, originando o *KEGG PATHWAY*. Numa fase posterior, mas ainda inicial do projeto, este apenas englobava quatro bases de dados, *PATHWAY*, *GENES*, *COMPOUND* e *ENZYME*. Mais tarde este projeto expandiu-se, dando origem a novas bases de dados, algumas delas mais específicas que viriam a substituir as iniciais.

Atualmente é um dos projetos mais utilizados na área da bioinformática, dado que permite coletar dados relacionados com genomas, doenças, *pathways* biológicas e também substância químicas.

As várias bases de dados do projeto *Kegg* estão categorizadas, em conformidade com a informação biológica que englobam. Na tabela 2.2 podemos observar as várias bases de dados, as categorias a que pertencem, o conteúdo biológico que englobam e o identificador *Kegg* para cada base de dados.

⁴<http://www.genome.jp/kegg/>

Categoria	Base de dados	Conteúdo	ID KEGG
Sistemas	PATHWAY	Mapas para funções celulares e orgânicas	Map
	BRITE	Classificações hierárquicas de entidades biológicas	br/ko
	MODULE	Módulos ou unidades funcionais de genes	M
Genómica	ORTHOLOGY	Grupos de genes ortólogos nos genomas completos	K
	GENOME	Genomas completos	org code/ T number
	GENES	Genes e proteínas no genoma completo	org:gene
	SSDB	Similaridade sequências entre genes	
Química	COMPOUND	Compostos químicos	C
	GLYCAN	Compostos glicanos	G
	REACTION	Reacções bioquímicas	R
	RCLASS	Reacções químicas	RC
	ENZYME	Nomenclatura enzimática	EC
Saúde	DISEASE	Doenças humanas	H
	DRUG	Drogas	D
	DGROUP	Grupos de droga	DG
	ENVIRON	Drogas e substâncias relacionadas com a saúde	E

Tabela 2.2: Bases de dados do projeto Kegg

O acesso aos dados do projeto *Kegg* pode ser feito não só através da interface WEB, mas também através da API disponibilizada de forma gratuita ou através do FTP disponibilizado, sendo que este obriga a uma subscrição não gratuita.

Na figura 2.4 podemos observar o resultado de uma pesquisa do gene *AICF*, correspondente ao identificador *Kegg* 29974 pertencente à espécie *Homo Sapiens*, na interface WEB do projeto *Kegg*.

2.3 APIs de Repositórios de Dados Biológicos

A conexão efetuada com estes repositórios é feita através das APIs disponibilizadas pelos mesmos.

De seguida demonstramos o funcionamento de cada uma dessas API.

2.3.1 API Ensembl

A API do projeto *ENSEMBL* [YBK⁺15] é desenvolvida em *PERL*⁵, disponibilizando métodos REST para possibilitar a obtenção de informação. Tal como referido na secção anterior as bases de dados deste projeto são armazenadas em MySQL.

A documentação fornecida é bastante útil, dado que engloba vários exemplos de chamadas à API em diversas linguagens.

As chamadas a esta API englobam diferentes parâmetros, dependendo do tipo de informação que se pretende obter, ou do tipo de ação que se pretende realizar. Para obtermos informação relativa a genes, é feito um pedido *HTTP POST*, contendo os identificadores pretendidos, não ultrapassando o limite imposto de 1000 caracteres por pedido. Existem vários parâmetros opcionais nestes pedidos, sendo exemplo o parâmetro *expand*, booleano, que quando ativo permite visualizar mais informação, como é o caso dos transcritos.

Um exemplo de uma chamada à API *Ensembl* é apresentado de seguida:

```
requests.post(https://rest.ensembl.org/lookup/id?expand=1, headers={"Content-Type": "application/json", "Accept": "application/json"}, data = {"ids": ["ENSG00000171428", "ENSG00000157764"]})
```

Qualquer chamada efetuada à API apenas pode ter retorno de um de três estados HTTP:

- 200: Sucesso;
- 400: Erro no pedido;
- 404: Não encontrado.

⁵<https://www.perl.org/>

KEGG **Homo sapiens (human): 29974** [Help](#)

Entry	29974 CDS T01001
Gene name	A1CF, ACF, ACF64, ACF65, APOBEC1CF, ASP
Definition	(RefSeq) APOBEC1 complementation factor
Organism	hsa Homo sapiens (human)
SSDB	Ortholog Paralog GFIT
Motif	Pfam: RRM_1 DND1_DSRM RRM_7 RRM_5 Linkain-b1 BRAP2 ERp29 Motif
Other DBs	NCBI-ProteinID: NP_001185749 NCBI-GeneID: 29974 HGNC: 24086 Ensembl: ENSG00000148584 Vega: OTTHUMG00000018240 Pharos: Q9NQ94(Tbio) UniProt: Q9NQ94 A0A024QZ19 B4E1E3
LinkDB	All DBs
Structure	PDB: 2CPD Thumbnail  Jmol
Position	10q11.23
AA seq	594 aa AA seq DB search MEAVCLGTCPEPEASMSTAIPGLKKGNNALQSIILQTLLEKENGQRKYGGPPPGWDAAPP ERGCEIFIGKLPRDLFEDELIPLCEKIGKIYEMRMMDFNNGNRGYAFVTFNSNKVEAKNA IKQLNNYEIRNGRLLGVCASVDNCRFLVGGIPKTKKREEILSEMKKVTEGVVDIVYPSA ADKTKNRGFAFVEYESHRAAMARRKLLPGRIQLWGHGIAVDWAEPEVEDEDTMSSVKI LYVRNLMSTSEEMIEKEFNKIPGAVERVKKIRDYAFVHFSNREDAVEAMKALNGKVLD GSPIEVTLAKPVDKDSYVRYTRGTGGRGTMLOGEYTYSLGQVYDPTTTYLGAPVYAPQT YAAIPSLHFPATKGHLSNRAIIRAPSVRGAAGVRGLGGRGYLAYTGLGRGYQVKGDKRED KLYDILPGMELTPMNPVTLKPQGIKLAPOILEEICQKNNWGPVYQLHSAIGQDQRQLFL YKITIPALASQNPAPIHPFTPPKLSAFVDEAKTYAAEYTLQTLGIPTDGGDGTMTAAAAA TAFPGYAVPNATAPVSAAQLKQAVTLGQDLAAYTTYEVYPTFAVTARGDGYGTF
NT seq	1785 nt NT seq atggaagcagtggtgtctgggcacatgccagagccagaagcgagcatgagcacagcaatt cctggcctcaaaaaaggaaataatgctttgcagtcctcatcttttgcaaacctttgctggaa aaggaaaaatggacaaaagaaaatattggtggccctccacctgggtgggatgctgcacccct gaaaggggctgtgaaatttttattggaaaacttccccgagaccttttgaggatgagctt ataccattatgtgaaaaaatcggtaaaatttatgaaatgagaatgatgatggattttaat ggcaacaatagaggatattgcatattgtaacattttcaataaagtggaagccaagaatgca atcaagcaacttaataattatgaaattagaattggcgccctcttaggggtttgtgccaat gtggacaactgccgattatttgttggggcatcccaaaaacaaaaagagagaagaatc ttatcggagatgaaaaaggttactgaaggtgtgtcgtatgcatcgtctaccaagcgct gcagataaaaacaaaaacaggagctttgccttcgtggagatgagagtcacgagcagct gccatggcgaggaggaaactgctaccaggaagaattcagttatggggacatggtattgca gtagactgggcagagccagaagtagaagttgatgaagatacaatgtcttcagtgaaaatc ctatatgtaagaaatcttatgctgtctacctctgaagagatgattgaaaaggaaattcaac aatatcaaacagggtgctgtggagaggggtgaagaaaattcgagactatgctttgtgcac ttcagtaaccgagaagatgcagttgaggctatgaaagctttaaatggcaagggtgctggat ggttcccccatggaagtcaacctagcaaaaacaggaggacaggttatggttaggtat accgaggcacaggtggaaggggcaccatgctgcaaggagagtatacctactctttgggc caagtttatgatcccaaccacaaccttggagctcctgtcttctatgccccccagacc tatgcagcaattcccagctcttcatttcccagccaccaaaggacatctcagcaacagagcc attatccgagccccttctgttagaggggctgcgggagtgagaggactgggcggcgtggc tatttggcatacacaggcctgggtcgaggataccaggtcaaaggagacaaaagagaagac aaactctatgacattttacctgggatggagctcaccaccaatgaatcctgtcacattaaaa

Figura 2.4: Pesquisa do gene A1CF na interface WEB do projeto Kegg

Quando uma chamada efetuada à API é efetuada com sucesso, o *output* vem no formato JSON⁶, o que facilita o tratamento dos dados.

2.3.2 API GenBank

A API disponibilizada pelo projeto *GenBank*, denominada *Entrez Programming Utilities (E-utilities)* [CLMW11], consiste no conjunto de nove ferramentas que atuam do lado do servidor, mantendo estável o sistema de consulta das bases de dados do NCBI. As *E-utilities* convertem um conjunto padrão de parâmetros de entrada nos valores necessários para acederem aos dados solicitados. Isto acontece dado que o URL possui uma sintaxe fixa.

Para a obtenção de informação relativa a genes ou proteínas, é utilizado o *ESummary*, uma das nove ferramentas pertencentes ao *E-Utils*, que requer dois parâmetros, sendo o primeiro a base de dados solicitada e o segundo um conjunto de identificadores GenBank. Este segundo parâmetro não tem limite no número de IDs, sendo que a partir dos 200 o pedido será efetuado através de um método HTTP POST.

De seguida apresentamos um exemplo de uma chamada à API *GenBank*:

```
https://ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=gene&id=1,29974
```

Tal como a API do projeto *ENSEMBL*, também retorna um de três estados HTTP:

- 200: Sucesso;
- 400: Erro no pedido;
- 404: Não encontrado.

O *output* da resposta, quando com sucesso, vem no formato JSON, facilitando o tratamento dos dados.

2.3.3 API Kegg

A API fornecida pelo projeto *Kegg* é bastante simples, contendo também uma documentação detalhada, facilitando a compreensão dos processos.

Todos os pedidos efetuados a esta API baseiam-se no formato demonstrado na figura 2.5.

Como verificámos na figura 2.5, para cada pedido escolhemos a operação pretendida e o argumento. O argumento deverá ter o formato *db:entry*, sendo *db* correspondente à base de dados pretendida e *entry* o identificador ou nome da entrada na base de dados. O argumento poderá conter várias entradas para a base de dados, sendo que cada uma delas deverá ter o formato referido anteriormente, separados pelo símbolo "+".

De seguida apresentamos um exemplo de uma chamada à API deste projeto:

⁶<http://www.json.org/>

URL form

```
http://rest.kegg.jp/<operation>/<argument>[/<argument2>][/<option>]  
<operation> = info | list | find | get | conv | link  
<argument> = <database> | <dbentries>
```

Figura 2.5: Formato de pedido à API do projeto Kegg

```
http://rest.kegg.jp/get/hsa:10458+hsa:10.
```

Sendo esta uma API bastante simples e intuitiva de usar, tem como maior problema a limitação do número de identificadores por pedido efetuado, dez. Deste modo, para uma grande quantidade de identificadores terão de ser feitos bastantes pedidos à API.

Cada pedido efetuado à API retorna a resposta em formato de texto, sendo que a resposta para cada argumento é delimitado por *tab* (\ \ \).

Esta API retorna, tal como as duas apresentadas anteriormente um de três estados HTTP:

- 200: Sucesso;
- 400: Erro no pedido;
- 404: Não encontrado.

2.4 Conversão de identificadores de diferentes repositórios

Um dos problemas com que os especialistas se confrontam no momento de obtenção de informação relativa a genes em diferentes repositórios consiste no facto de cada repositório utilizar o seu próprio identificador para cada gene. Deste modo, para permitir a pesquisa em diferentes repositórios é necessário converter esse identificador para o identificador correspondente.

Apresentamos nesta secção um sítio WEB que disponibiliza uma ferramenta que permite efetuar essa conversão.

2.4.1 BioDB Hyperlink Management System

O BioDB⁷ define um identificador para cada gene, estando conectado com os identificadores desse mesmo gene nos vários repositórios onde este se encontra. Deste modo, é possível efetuar a conversão de um identificador de uma forma simples e rápida [IN09].

Este projeto para além de estar disponível através de um sítio WEB, disponibiliza também uma API, permitindo assim que esta ferramenta seja utilizada por qualquer pessoa, ou englobada em qualquer projeto.

⁷<http://biodb.jp/>

2.4.2 API BioDB Hyperlink Management System

Para efetuar a conversão através da API basta efetuar um pedido através do URL, no formato que se segue [BIO]:

```
http://biodb.jp/convert/id\_OriginalType/id\_ResultType/[entry\_1,entry\_id2,...]
```

Sendo que os argumentos correspondem a:

- **id_OriginalType** : base de dados correspondente ao ID original submetido para conversão;
- **id_ResultType** : base de dados correspondente ao ID pretendido após a conversão;
- **[entry_1,entry_id2,...]** : IDs para conversão.

Os identificadores submetidos para conversão deverão estar separados por vírgula, sendo que o limite máximo de elementos para conversão é 100.

O resultado desta conversão é retornado em formato JSON, permitindo o tratamento dos dados de forma acessível.

2.5 Data Mining

Atualmente vivemos numa época em que uma enorme quantidade de informação é coletada diariamente [WZWD14]. A análise dessa informação é importante e é derivado disso que surge o *data mining*. Consideramos que o *data mining* surge da evolução natural do aumento de informação, em paralelo com a evolução tecnológica [HPK11].

Com o aumento da informação, e a partir do momento em que grande parte dessa informação começou a ser armazenada em sistemas informáticos, no final da década de 1980, surgiu esta nova área informática.

Data mining consiste no processo de extrair conhecimento de grandes quantidades de informação, por exemplo procurando padrões consistentes através da combinação de diferentes áreas, como estatística, inteligência artificial, aprendizagem computacional (*machine learning*) e bases de dados. Tem como objetivo transformar dados de onde não se consegue retirar conclusões em subconjuntos desses mesmos dados, que agora estruturados podem ser úteis para retirar conclusões.

Existem vários modelos relativos às diferentes fases que envolvem todo o processo de *data mining*, apresentamos de seguida as quatro fases principais, presentes na maioria desses modelos:

1. Pré-processamento
2. Data Mining
3. Validação dos resultados
4. Apresentação de conhecimento

A fase de pré-processamento engloba a limpeza, integração, seleção e transformação dos dados. A limpeza consiste em remover o ruído e os dados inconsistentes, seguindo-se pela combinação de dados de várias fontes, sendo esta a parte de integração. A seleção dos dados é a fase em que os dados relevantes são retirados da base de dados de forma a serem posteriormente analisados. Por fim acontece a transformação dos dados, de modo a estes estarem aptos para ser analisados. Existem vários métodos de transformação, sendo os mais comuns a síntese ou agregação de dados [HPK11].

A segunda etapa deste processo, *Data Mining* consiste na aplicação dos vários métodos, que iremos apresentar posteriormente nesta secção, de modo a obter conhecimento.

De modo a validar os resultados de *data mining*, nesta fase deverá acontecer a validação dos padrões reconhecidos na fase anterior.

Por fim a etapa de apresentação dos resultados, em que estes devem ser demonstrados aos utilizadores através de técnicas de representação e visualização.

Apresentamos também um modelo mais direcionado à indústria, sendo este denominado Cross Industry Standard Process for Data Mining [CRI] (CRISP-DM) que envolve as fases:

1. Entender o negócio: definir o plano para atingir objetivos, através de uma perspetiva de negócio;
2. Entender os dados: recolha dos dados e identificação de problemas;
3. Preparar os dados: construção de diferentes conjuntos de dados;
4. Modelagem: aplicação de várias técnicas de modelação de dados, sendo natural voltar à fase anterior no final desta;
5. Avaliação: verificar se o modelo atinge os objetivos do negócio;
6. Implementação: o conhecimento é adquirido e apresentado.

Existem várias técnicas de *data mining*, nesta secção apresentamos alguns dos mais utilizados atualmente, seguidos de uma breve explicação. Será dado maior foco às técnicas utilizadas para o âmbito desta dissertação.

2.5.1 Classificação

Esta técnica de *data mining* consiste em encontrar um modelo, ou função, que permita diferenciar várias classes, de modo a integrar um elemento apenas num conjunto de dados. Este modelo baseia-se na análise de conjunto de dados existentes, dados já classificados. Assim, a classificação engloba-se nos tipos de *data mining* com aprendizagem supervisionada.

Na figura 2.6 apresentamos alguns dos métodos de classificação [HPK11].

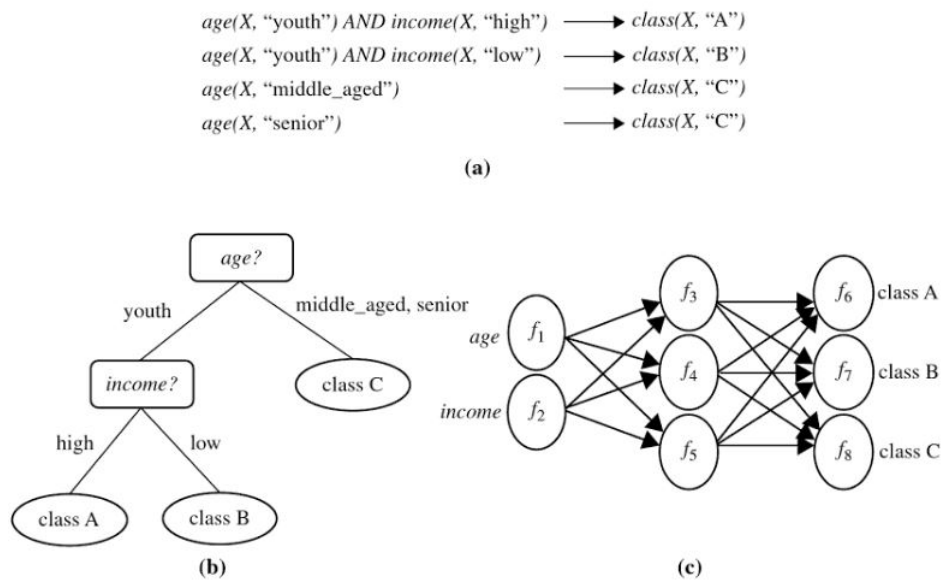


Figura 2.6: Diferentes métodos de classificação

a) *if-then*, b) árvore de decisão, c) rede neuronal

2.5.2 Regressão

A regressão consiste em analisar todos os dados fornecidos, relacionando esses mesmos dados, de modo a desenvolver uma função que permita estimar o valor de uma determinada variável, ou determinar a que conjunto de dados pertence um determinado elemento [HPK11]..

2.5.3 Associação

A associação consiste em perceber qual a probabilidade de ocorrência de um determinado elemento num determinado conjunto de dados e qual a relação desse elemento com os restantes. Consiste também na procura de elementos que implicam a presença de outros no mesmo conjunto de dados. As relações entre as ocorrências são expressas através de regras de associação, que tipicamente representam padrões existentes num conjunto de dados.

2.5.4 Clustering

Contrariamente ao que acontece com as técnicas de classificação e regressão, a técnica de *clustering*, também conhecida por agrupamento, não necessita que sejam fornecidos dados de modo a seguir um modelo de aprendizagem supervisionada [HPK11]..

Muitas vezes esses dados de aprendizagem não existem, ou fica demasiado custoso obtê-los. Esta técnica pode ser utilizada de modo a efetuar o agrupamento de vários elementos, criando um conjunto de dados específico. Esta junção de elementos num grupo baseia-se no princípio de maximizar as semelhanças dentro de um grupo e minimizar as semelhanças entre elementos de grupos distintos.

Na figura 2.7 podemos observar um exemplo bastante simples de *clustering*. Na subfigura (a) vemos todos os elementos, representados pela letra "x". Na subfigura (b) vemos os elementos já separados por *clusters*, sendo cada cluster representado por um algarismo.

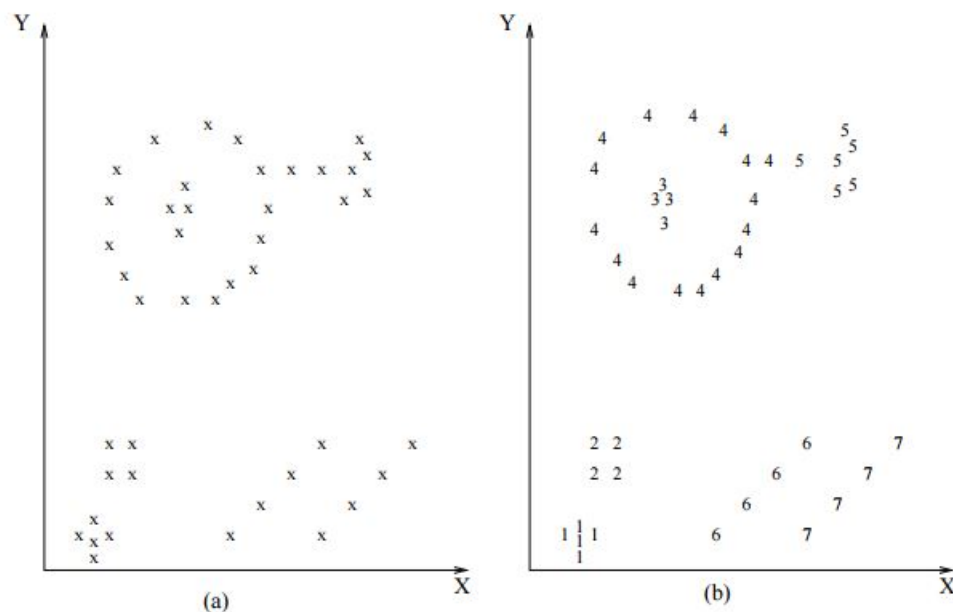


Figura 2.7: Clustering

a) Instâncias para analisar, b) Resultado final

2.6 Classificação

Após a breve explicação dada na secção 2.5.1 relativamente à classificação, e visto que esta técnica será bastante importante nesta dissertação, nesta secção apresentamos vários detalhes sobre esta técnica de *data mining*.

2.6.1 Algoritmos de Classificação

Existem vários algoritmos de classificação, maioritariamente baseados em estatística. Apresentamos de seguida alguns dos mais atuais.

2.6.1.1 C4.5

O algoritmo C4.5 foi desenvolvido por Ross Quinlan, sendo uma extensão no algoritmo ID3, baseado em árvores de decisão.

Este algoritmo constrói árvores de decisão, sendo que em cada nó da árvore escolhe o atributo que melhor particiona o conjunto de dados em subconjuntos. O atributo considerado que melhor

particiona o conjunto de dados é aquele que tem maior ganho de informação normalizado. Caso nenhum dos atributos forneça ganho de informação, este algoritmo cria um nó de decisão acima, utilizando o valor esperado.

2.6.1.2 SVM - Support Vector Machine

SVM, ou em português Máquina de Vetores de Suporte, engloba um conjunto de métodos de aprendizagem supervisionada com base em *machine learning* que analisam e reconhecem padrões. O SVM tem como entrada um conjunto de dados e para cada entrada prediz a que classe este pertence, dentro de duas classes possíveis. Deste modo concluímos que o SVM é um classificador linear binário não probabilístico [Agg15].

Assim uma SVM encontra uma linha de separação, denominada hiperplano, entre duas classes. O objetivo desta linha visa maximizar a distância entre os dois pontos mais próximos em relação a cada uma das classes.

Apresentamos na figura 2.8 uma figura que demonstra a metodologia usada por SVM.

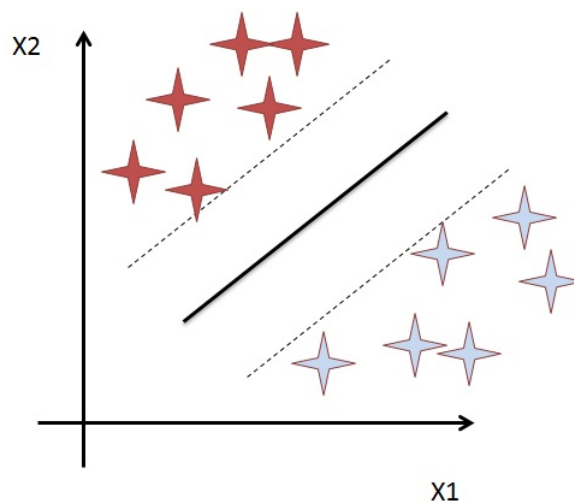


Figura 2.8: *Support Vector Machine* - Exemplo

2.6.1.3 Random Forest

Este algoritmo é baseado na combinação de várias árvores de decisão de forma a gerarem um classificador final. O nome *Forest* advém deste mesmo facto de se utilizarem várias árvores no processo. A primeira parte do nome, *random*, em português aleatório é derivado da maneira aleatória como são escolhidas as análises em cada etapa [BC12].

Tem algumas vantagens relativamente à maioria dos restantes algoritmos, de seguida apresentamos algumas:

- Muito preciso nos resultados;

- Eficiente em grandes bases de dados;
- Consegue manusear uma grande quantidade de variáveis de entrada;
- Eficaz a estimar dados em falta e mantém a precisão, mesmo quando faltam bastantes dados;
- Consegue estimar quais as variáveis mais importantes para a classificação;

2.6.1.4 Naive Bayes

Este algoritmo é baseado no teorema de Bayes, que utiliza a probabilidade para classificar os dados, supondo que existe independência entre os atributos.

É bastante utilizado atualmente para conjunto de dados bastante grandes, dada a sua simplicidade e rapidez de execução.

2.6.2 Métodos de avaliação de Classificação

Existem vários métodos de avaliação dos algoritmos de classificação. Nesta secção expomos dois dos principais tipos de avaliação de classificação. O *cross validation*, em português validação cruzada e o *bootstrap*. Após explicarmos cada um deles apresentaremos alguns métodos que se incluem nestes dois grupos principais.

2.6.2.1 Cross Validation

Este modelo de validação consiste na avaliação da capacidade de generalização de um modelo, com base num conjunto de dados. Este modelo é maioritariamente utilizado quando o objetivo é a previsão, de modo a verificar qual a precisão que o modelo terá quando aplicado a um conjunto real de dados.

O modo de atuar deste modelo consiste na divisão de um conjunto de dados em subconjuntos, utilizando alguns desses modelos como dados de treino e os restantes como dados de validação do modelo.

O modo como a divisão do conjunto inicial de dados é feita depende do método utilizado. De seguida apresentamos alguns dos métodos existentes.

- **Método *holdout***

Este método divide o conjunto inicial de dados em dois sub conjuntos mutuamente exclusivos, que não têm necessariamente a mesma dimensão. Um destes sub conjuntos é usado para treino e o outro para validação. Por norma o sub conjunto de treino tem $\frac{2}{3}$ dos dados iniciais e o conjunto de validação tem $\frac{1}{3}$.

Após esta divisão é efetuada a previsão, calculando o erro de previsão.

Este método é mais adequado para grandes quantidades de dados, dado que em conjuntos de dimensões inferiores o erro pode sofrer uma grande variação.

- **Método *k-fold***

O método *k-fold* consiste em dividir o conjunto inicial de dados em k sub conjuntos, todos com a mesma dimensão.

Depois cada sub conjunto é utilizado para treino e todos os restantes são utilizados para validação. Em cada iteração é calculado o erro, sendo que no final é calculado o erro total, de modo a obter uma medida mais confiável relativa ao modelo.

- **Método *leave-one-out***

Este método é um caso específico do método *k-fold*, sendo que o k é igual ao número total de dados (N). Deste modo são realizados N cálculos de erro, um para cada elemento.

Obviamente que este método é mais eficaz no cálculo do erro, no entanto tem um alto custo computacional, sendo apenas indicado para conjuntos de pequena dimensão.

2.6.2.2 *Bootstrap*

Contrariamente aos métodos de avaliação vistos anteriormente, os métodos baseados em *bootstrap* recorrem à reutilização dos dados para treino.

O conjunto de treino de N elementos é construído através da substituição N vezes de modo a formar um conjunto de treino com N elementos, permitindo dados repetidos.

O conjunto usado para validação é constituído pelos elementos contidos no conjunto inicial, que não estão incluídos no conjunto de treino.

De seguida apresentamos um dos métodos atualmente mais utilizados, recorrendo ao *bootstrap*:

- **0.632 *bootstrap***

Neste método em particular, cada elemento tem a probabilidade $1 - \frac{1}{N}$ de não ser escolhido para o conjunto de treino. Em contrapartida terá $(1 - \frac{1}{N})^N$ de estar incluída no conjunto de validação, para valores altos de N esta probabilidade é igual a $\frac{1}{e} = 0.368$.

Conseguimos assim concluir que o conjunto de treino contém cerca de 63.2% dos elementos e consequentemente um valor de erro estimado bastante alto.

2.6.3 Métricas de Classificação

De seguida apresentamos algumas métricas para os algoritmos de classificação, e de modo a facilitar a sua compreensão apresentamos a tabela 2.3 denominada *confusion matrix* [KP98].

Tabela 2.3: Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Apresentamos em baixo uma pequena descrição da tabela 2.3, para facilitar o seu entendimento nas fórmulas que demonstramos posteriormente.

- a é número de previsões corretas de uma instância ser negativa;
- b é o número de previsões incorretas de uma instância ser positiva;
- c é o número de previsões incorretas de uma instância ser negativa;
- d é o número de previsões corretas de uma instância ser positiva.

2.6.3.1 Accuracy

A *accuracy*, na avaliação de um algoritmo é relativa ao grau de aproximação da avaliação quando comparado com o valor real.

$$Accuracy = \frac{a + d}{a + b + c + d} \quad (2.1)$$

2.6.3.2 Precision

A *precision* refere-se à proximidade das várias avaliações efetuadas.

$$Precision = \frac{d}{b + d} \quad (2.2)$$

Quando aplicada à classificação, a *precision* refere-se ao número de verdadeiros positivos, por exemplo o número de elementos identificados corretamente a uma classe divididos pelo número total de elementos dessa mesma classe.

Na figura 2.9 podemos visualizar um exemplo de *precision* e *accuracy*, de modo a facilitar a compreensão destes dois conceitos.

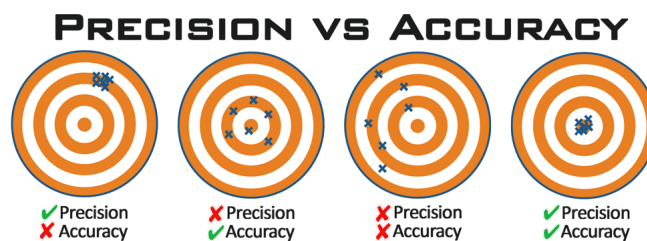


Figura 2.9: *Precision vs Accuracy*

2.6.3.3 Recall

Recall relaciona o número de elementos corretos na avaliação com o total de elementos efetivamente corretos.

Contextualizando com a classificação, *recall* pode ser definido como o número de verdadeiros positivos divididos pelo número total de elementos que efetivamente deveriam pertencer a essa classe.

$$Accuracy = \frac{d}{c + d} \quad (2.3)$$

2.6.3.4 F-Measure

F-measure utiliza o valor da *precision* e do *recall* para calcular o valor.

$$Fmeasure = \frac{precision * recall}{precision + recall} \quad (2.4)$$

2.6.3.5 AUC

Area Under Curve, relaciona os verdadeiros positivos com os falsos positivos num gráfico, de modo a perceber qual destes se encontra em maior número, obtendo conclusões através da área que fica "por baixo da curva".

Na figura 2.10 apresentamos um exemplo, em que no eixo das abcissas temos os falsos positivos e no eixo das ordenadas temos os verdadeiros positivos.

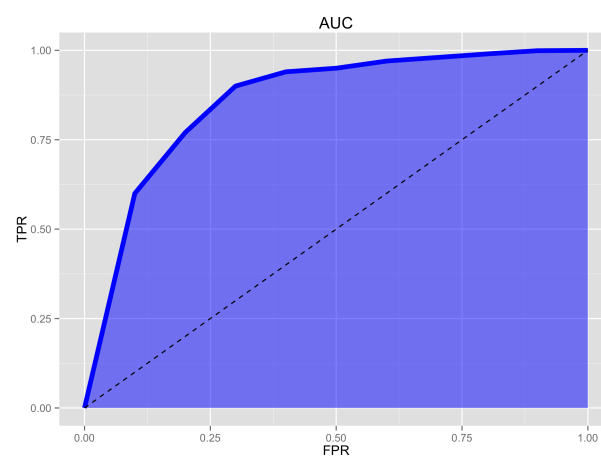


Figura 2.10: ROC-Curves

2.7 Clustering

2.7.1 Técnicas de *Clustering*

A técnica de *Clustering*, tal como explicámos na secção anterior, consiste no agrupamento de diversos elementos em vários grupos distintos, aumentando as semelhanças entre elementos do mesmo grupo e diminuindo as semelhanças entre elementos de grupos diferentes.

Esta separação é feita tendo em conta os atributos de cada elementos, usando métodos de comparação.

Visto que os elementos a analisar de modo a perceber as semelhanças e diferenças podem ser de vários tipos, não há um processo específico que possa ser utilizado em todos os casos de aplicação desta técnica de *data mining*. Desta forma existem vários métodos de *clustering* que permitem a análise de diferentes tipos de dados [JMF99].

2.7.1.1 Métodos de partição

Este método baseia-se na criação de partições de todo o conjunto de dados fornecido. Cada partição corresponde a um conjunto, também conhecido como *cluster*.

Dada uma base de dados com N elementos, este método constrói k partições, respeitando sempre a limitação $k \leq \sqrt{N}$.

Estes métodos devem satisfazer os seguintes requisitos:

1. Cada *cluster* deve conter pelo menos um objeto;
2. Cada objeto deve pertencer apenas a um *cluster*.

Este método engloba-se nos métodos de divisão, sendo que constrói as partições iniciais e iterativamente coloca os objetos noutros grupos, de modo a melhorar as partições. O objetivo depois de todas as iterações é, como foi referido anteriormente, aumentar as semelhanças entre os objetos inserido no mesmo *cluster*, aumentando também as diferenças entre objetos presentes em diferentes *clusters*.

De modo a obter o estado ótimo de um algoritmo de *cluster* seria necessário enumerar e combinar todas as partições possíveis. Dado que isto seria demasiado custoso, ou até impraticável, grande parte dos métodos de partição optam por aplicar um método heurístico para avaliar a sua qualidade. Alguns dos algoritmos mais utilizados são o *k-means* e o *k-medoids*.

O algoritmo *k-means* utiliza a média dos objetos de um determinado *cluster* para representar esse mesmo *cluster*.

O algoritmo *k-medoids* utiliza um dos objetos mais centrais do *cluster* para representar esse grupo.

2.7.1.2 Métodos hierárquicos

Os métodos hierárquicos decompõem o conjunto de dados fornecido de uma forma hierárquica, por norma representada através de árvore.

Existem dois métodos de decompor os dados, a forma aglomerativa ou divisiva.

A forma aglomerativa divide todos os objetos em *clusters* diferentes, numa fase inicial. Depois iterativamente aglomera vários objetos no mesmo grupo, até alcançar uma condição de paragem, ou caso não exista, até englobar todos os objetos no mesmo *cluster*.

A estratégia divisiva, contrariamente à aglomerativa, numa fase inicial engloba todos os objetos no mesmo *cluster*, sendo que iterativamente divide esse *cluster* em vários grupos, até atingir uma condição de paragem, ou até distribuir os objetos todos em grupos diferentes.

Este tipo de *clustering* tem uma grande vantagem relativamente a outros, um custo bastante reduzido. Este baixo custo resulta principalmente do facto de não ser possível voltar atrás, cada iteração é irreversível. Isto significa que não é possível corrigir situações que mais tarde se revelaram erradas, ou não ótimas. De modo a diminuir este tipo de decisões deve ser feito o pré-processamento, podendo este por exemplo passar por reduzir o conjunto de dados, diminuindo o conjunto de variáveis a ser considerado. Outra forma de reduzir os erros será através da utilização do algoritmo de hierarquia aglomerativa numa fase inicial e mais tarde melhorar o resultado através da utilização do algoritmo divisivo.

2.7.1.3 Métodos com base na densidade

Contrariamente aos métodos vistos anteriormente, baseados na distância entre objetos, estes métodos, tal como o nome indica, baseiam-se na densidade do *cluster*.

O objetivo destes métodos é aumentar o tamanho dos *clusters* até que o número de objetos na sua proximidade seja igual ou superior ao determinado.

Através da aplicação destes métodos, conseguimos a criação de *clusters* de forma arbitrária, sendo particularmente útil para descobrir grupos com forma irregular.

Os dois algoritmos atualmente mais utilizados, baseados em densidade são o algoritmo DBSCAN e o algoritmo OPTICS.

2.7.1.4 Métodos baseados em grelha

Os métodos baseados em grelha quantificam o espaço dos objetos num número finito de células, formando uma estrutura em grelha. Todas as operações de *clustering* são efetuadas sobre esta estrutura, revelando-se bastante mais rápidas a nível de processamento, pois este é praticamente independente do número de dados, sendo dependente do número de células e da sua dimensão.

2.7.1.5 Métodos baseados em modelos

Os métodos baseados em modelos criam um modelo matemático para cada *cluster* e encontram os objetos mais adequados a englobarem esse mesmo modelo.

Por norma estes métodos assumem que a distribuição dos objetos é feita através de probabilidades estatísticas, o que permite saber automaticamente o número de *clusters* [Mad12].

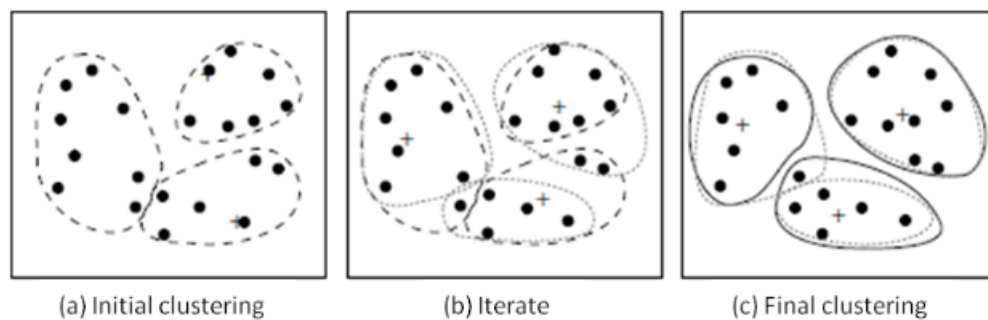


Figura 2.11: Exemplo do método k-means

O centro do cluster é apresentado pelo símbolo "+"

2.7.2 Algoritmos de *Clustering*

Existem vários algoritmos de *clustering*, sendo que a escolha do mais apropriado para o problema em causa deve ser feita baseando-se em três critérios:

1. A forma como os *clusters* são formados;
2. A estrutura dos dados;
3. A sensibilidade na utilização das técnicas de *clustering* para alterações que não afetem a estrutura dos dados.

Nesta secção apresentamos alguns dos algoritmos atualmente mais utilizados, sendo que estes serão também os algoritmos utilizados ao longo desta dissertação.

2.7.2.1 k-Means

O método *K-means* engloba-se nos métodos de partição explicados na secção 2.7.1 e utiliza a média de todos os objetos presentes num determinado *cluster* para o representar.

Este método visa particionar N dados em k *clusters*, onde cada elemento pertence ao *cluster* com a média mais próxima. O procedimento segue uma maneira simples e fácil para classificar os dados, a partir de um valor k de *clusters*.

2.7.2.2 *Farthest First*

Este método é uma extensão do método *K-means*, tem também duas fases, primeiro a escolha dos *centroids*, que corresponde ao centro de cada *cluster*, e depois a atribuição dos elementos aos *clusters*.

No entanto a escolha dos *centroids*, neste método, é feita afastando o máximo possível cada *centroid* dos outros.

2.7.2.3 Expectation–maximization

Este método, muitas vezes considerado uma extensão do algoritmo *k-means*, segue uma abordagem estatística calculando a probabilidade de cada elemento pertencer a cada um dos *clusters*.

Inicialmente, de forma aleatória, escolhe *k* objetos para representar os *centroids* dos *clusters*, depois iterativamente refina os *clusters* em dois passos:

1. **Passo E (Expectation):** associa cada objeto x_i ao *cluster* C_i através da seguinte probabilidade:

$$P(x_i \in C_k) = p(C_k/x_i) = \frac{p(C_k)p(x_i/C_k)}{p(x_i)} \quad (2.5)$$

Onde $p(x_i / C_k = N(m_k, E_k(x_i)))$ segue uma distribuição normal (Gaussiana) de probabilidade com média m_k e valor esperado E_k .

2. **Passo M (Maximization):** usa as probabilidades estimadas no passo E para re-estimar (refinar) os parâmetros do modelo:

$$m_k = \frac{1}{n} \sum_{i=1}^n \frac{x_i p(x_i \in C_k)}{\sum_j p(x_i \in C_j)} \quad (2.6)$$

2.7.2.4 DBSCAN

O método DBSCAN, abreviatura para *Density Based Spatial Clustering of Application with Noise* é um método baseado em densidade capaz de identificar *clusters* de formato arbitrário e de diferentes tamanhos, identificar e separar ruídos e detetar *clusters* naturais, sem informação prévia.

Este método tem em conta um número mínimo de elementos e o raio de vizinhança, sendo que para formar um *cluster* é necessário que exista esse número mínimo de elementos dentro do raio fornecido.

2.7.2.5 Clusterização hierárquica de ligação média

Este algoritmo é usado para calcular a distância entre *clusters* na análise de *clustering* hierárquico.

De modo a decidir quais os *clusters* que devem ser combinados ou divididos, é necessário medir a disparidade entre esses *clusters*. Neste algoritmo essa disparidade é calculada através da média da distância entre todos os objetos englobados nos dois *clusters*.

2.7.2.6 Programação Lógica Indutiva

Este algoritmo é derivado de *machine learning* que utiliza a lógica de primeira ordem para representar os dados e os modelos de modo a inferir modelos através de exemplos baseados em conhecimento prévio. Estes exemplos podem ser positivos ou negativos, sendo que os positivos são exemplos que devem ser aprendidos e os negativos são exemplos que não devem acontecer.

O conhecimento prévio são predicados que englobam toda a informação considerada útil para construir os modelos.

2.7.3 Métodos de avaliação de *Clustering*

Tal como foi dito anteriormente, o objetivo dos algoritmos de *clustering* é aumentaras semelhanças entre objetos pertencentes ao mesmo grupo, mantendo diferenças entre elementos de grupos distintos. Isto representa um critério de avaliação interno. No entanto de modo a melhorar a avaliação devemos também utilizar critérios de avaliação externos [MRS08].

Este tipo de critérios compara a estrutura de grupos descoberta com uma estrutura de grupos previamente conhecida.

2.7.3.1 Coeficiente Silhouette

O coeficiente *Silhouette* representa a similaridade dos objetos no interior dos *clusters* e a diferença entre objetos de *clusters* diferentes.

Para esta representação é utilizada a seguinte fórmula:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{if } a(i) < b(i), \\ 0 & \text{if } a(i) = b(i), \\ \frac{b(i)}{a(i)} - 1 & \text{if } a(i) > b(i) \end{cases}$$

sendo que:

- **a(i)** representa a média da disparidade entre o objeto i e todos os outros objetos do mesmo *cluster*;
- **b(i)** representa o menor valor médio de disparidade entre o objeto i e qualquer outro *cluster* onde o objeto i não esteja inserido;

Quanto maior o valor de s(i) melhor o resultado, ou seja, mais apropriado é o *cluster* para o objeto estudado, sendo que o contrário também é verdade. Assim o valor médio de s(i) de um *cluster* pode ser usado para avaliar como os dados estão agrupados. O valor médio de s(i) de toda a base de dados pode ser usado para avaliar a qualidade do *clustering* que foi feito.

Este coeficiente permite também visualizar os resultados do *clustering*, através de um gráfico que combina a *silhouette width* de todos os objetos da base de dados com a média de *silhouette width* de cada *cluster* e o coeficiente de toda a base de dados.

2.8 Ferramentas de *Data Mining*

Atualmente temos acesso a uma vasta gama de ferramentas de *data mining*, sendo que muitas delas são de utilização gratuita.

A escolha de uma ferramenta de *data mining* depende de vários fatores, sendo os mais relevantes o tipo de dados a analisar, o sistema utilizado, o orçamento existente e o conhecimento prévio na área da informática.

De seguida apresentamos algumas das ferramentas atualmente mais utilizadas.

2.8.1 RapidMiner

O *RapidMiner*⁸ é uma ferramenta gratuita para *data mining*, e é atualmente a ferramenta mais utilizada a nível mundial.

Suporta *machine learning*, *text-mining*, análise preditiva, análise de negócios e análise de dados. É também possível a produção de relatórios.

O grande número de utilizadores que esta ferramenta tem é também devida ao elevado número de modelos que suporta, dado que disponibiliza muitos operadores e muitas funções para o tratamento do dados.

A interface que apresenta ao utilizador é também uma vantagem desta ferramenta, pois é de elevada usabilidade.

2.8.2 Weka

A ferramenta *Weka* foi criada em 1993 e desenvolvida na linguagem Java.

Contém vários algoritmos de *machine learning* e disponibiliza ao utilizador funcionalidades para pré-processamento, classificação, regressão, *clustering*, regras de associação e visualização.

Esta ferramenta oferece ao utilizador uma interface gráfica, no entanto é mais utilizada através da linha de comandos dado que permite a escolha de mais parâmetros nas suas funcionalidades.

É também possível utilizar os seus algoritmos em ferramentas externas, através da API disponibilizada utilizando por exemplo a biblioteca *python-weka-wrapper*.



Figura 2.12: Interface do WEKA

⁸<https://rapidminer.com>

2.8.3 R

O R é um ambiente de programação destinado ao desenvolvimento de sistemas de apoio à decisão e análise de dados. Tem como vantagem a possibilidade de ser instalado em vários sistemas operativos e o facto de englobar diversos *packages*, abrangendo diversas técnicas de *data mining*.

O ambiente R permite a manipulação de dados e a representação gráfica de dados estatísticos, através de linha de comandos.

2.8.4 KNIME

A ferramenta *KNIME* é também gratuita e oferece ao utilizador funcionalidades que permitem a integração, processamento e análise de dados.

2.8.5 SPSS

O SPSS é uma ferramenta paga, pertencente à IBM desde 2009. Tem como principal função o apoio à tomada de decisão, incluindo *text-mining* e estatística.

Esta ferramenta integra uma interface gráfica bastante intuitiva e é uma das mais utilizadas no contexto empresarial.

2.9 Formato *ARFF*

O formato *Attribute-Relation File Format*, vulgarmente designado apenas por *arff*, é um formato desenvolvido inicialmente para a ferramenta de *data mining* *WEKA*, mas que atualmente é utilizado em várias ferramentas.

Este tipo de ficheiro divide-se em duas secções:

1. *header*, ou cabeçalho;
2. *data*, ou dados.

O cabeçalho engloba o nome da relação e a lista dos atributos, contendo também o tipo de cada atributo, apresentamos um exemplo no documento 2.9.

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%   (a) Creator: R.A. Fisher
%   (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%   (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
```

```
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Os dados contêm a lista de instâncias, com todos os atributos de cada instância. Quando um atributo não está definido numa instância é inserido o símbolo "?" nesse atributo [ARF].

No documento 2.9, que apresentamos de seguida, podemos observar um exemplo.

```
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,?,Iris-setosa
4.4,2.9,1.4,?,Iris-setosa
4.9,?,1.5,0.1,Iris-setosa
```

As linhas que começam com o símbolo % são comentários.

2.10 Tecnologias

Para o desenvolvimento da plataforma WEB são necessárias tecnologias para o *front-end*, *back-end* e armazenamento da informação.

Atualmente existem diversas tecnologias disponíveis para cada uma das etapas referidas anteriormente e a escolha deve recair sobre as mais adequadas ao projeto em questão.

De seguida apresentamos, através de uma breve explicação, as tecnologias utilizadas na elaboração desta plataforma.

2.10.1 Python

Python é uma linguagem de programação de alto nível, criada em 1991 com os objetivos de produtividade e legibilidade.

Atualmente é uma das mais usadas a nível mundial, principalmente devido ao facto de suportar múltiplos paradigmas de programação, ser orientada a objetos, funcional, processual e imperativa.

2.10.2 Django

Django é uma *full-stack* WEB *framework open-source*, desenvolvida em Python por utilizadores experientes de tecnologias para desenvolvimento de aplicações WEB.

Esta *framework* utiliza o padrão *Model-View-Controller* (MVC) e o princípio *Don't Repeat Yourself* (DRY). Este conceito de programação defende que cada porção de conhecimento em um sistema deve possuir uma representação única, de autoridade e livre de ambiguidades em todo o sistema [HT99].

Algumas das vantagens desta *framework* em complementaridade ao padrão e princípio referidos no parágrafo anterior, é a facilidade de trabalhar com bases de dados complexas, assim como a disponibilização de uma interface de administração para a base de dados bastante fácil de utilizar [FBC08].

O facto desta *framework* ser desenvolvida em Python também facilita o seu desenvolvimento, dado que esta linguagem de programação é das mais utilizadas atualmente no mundo inteiro.

2.10.2.1 Padrão de arquitetura MVC

Model-view-controller (MVC), em português modelo-visão-controlador, é um padrão de arquitetura de software que procura estruturar aplicações interativas de uma forma modular.

É composto por três componentes, *Model*, *View* e *Controller*. Na figura 2.13 apresentamos, de uma forma simplificada, o modo de interação das três componentes e do utilizador [Dea05].

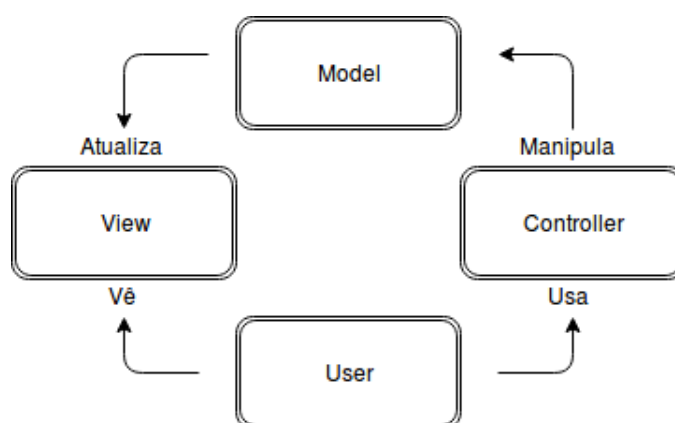


Figura 2.13: Interação das componentes do MVC e do utilizador

- **Model**

Esta componente integra o estado da aplicação e as operações que controlam esse estado, sendo responsável pela leitura e escrita de dados, assim como a sua validação.

- **View**

A componente *view* é responsável pela interação direta com o utilizador, ou seja, apresenta a informação utilizador através de uma interface GUI.

Uma *view* é atualizada ao receber uma notificação do componente *model*, podendo depois pedir a informação que necessita.

- **Controller**

A componente *controller* recebe todos os pedidos do utilizador, processa-os nos sub-controladores e posteriormente envia as transações ao *model* para serem executadas.

A cada *view* deve corresponder um *controller*.

2.10.3 Bootstrap

Bootstrap é uma *framework* de desenvolvimento de *front-end*, *open-source* criado para facilitar o desenvolvimento de aplicações WEB.

Engloba HTML e modelos de design CSS para vários componentes de interface WEB, assim como extensões JavaScript.

É atualmente a *framework* mais utilizada a nível mundial para desenvolvimento de *front-end* em aplicações WEB.

2.10.4 JavaScript

JavaScript é atualmente a principal linguagem para programação *client-side* em aplicações WEB. Já é também bastante utilizada do lado do servidor, através de ambientes como o *node.js*.

Esta linguagem foi desenvolvida para ser uma linguagem *script* com orientação a objetos.

2.10.5 SQLite

O SQL é uma linguagem de programação padrão para gerar, manipular e recuperar informação duma base de dados relacional [Bea09].

SQLite é uma biblioteca desenvolvida em C, que permite a gestão de dados de uma forma gratuita.

O uso de SQLite é recomendado para aplicações WEB de baixo ou médio tráfego. É adequado para a gestão de vários tipos de informação, incluindo ficheiros.

2.11 Conclusões

Ao longo deste capítulo contextualizámos esta dissertação com as duas áreas diretamente envolvidas no tema, a biologia, mais concretamente a genómica e a informática, focando no *data mining*.

Para isso apresentámos e descrevemos as diversas técnicas e ferramentas que existem relacionados com estes temas, dando sempre maior foco aquelas que foram úteis para o desenvolvimento desta dissertação.

Também neste capítulo descrevemos brevemente as tecnologias utilizadas ao longo do desenvolvimento deste projeto.

Capítulo 3

Plataforma WEB

Relembrando a questão inicial desta dissertação, o objetivo da implementação que descrevemos neste capítulo é facilitar o trabalho dos especialistas na área da investigação biológica.

Para tal construímos uma plataforma WEB capaz de obter informação de diferentes repositórios de informação relacionada com genes, permitindo também analisar essa informação através de diferentes técnicas de *data mining*.

Toda a plataforma foi desenvolvida com o intuito de ser utilizado por biólogos sem necessitarem de experiência na área da informática para conseguirem usufruir de todas as funcionalidades oferecidas pela aplicação.

Ao longo deste capítulo aprofundaremos o problema e a solução do problema, assim como as várias tarefas possíveis de realizar na plataforma WEB desenvolvida, descrevendo a sua implementação e o seu funcionamento.

3.1 Problema

Atualmente os especialistas da área da Biologia Molecular e da genómica enfrentam alguns desafios relacionados com a quantidade de informação disponível na WEB, as diferentes fontes de informação, a obtenção de informação repetida, os diferentes formatos na recolha da informação e consequentemente a dificuldade em retirar conclusões dessa informação.

De modo a obter toda a informação necessária relativa a um determinado gene um especialista é obrigado a executar várias etapas que se tornam morosas principalmente a nível de tempo, mas também a nível da complexidade. Isto acontece dado que existem vários sítios WEB com informação relativa a um gene em particular, sendo que cada um desses sítios WEB está ligado a uma base de dados diferente, o que implica também que grande parte deles atribuem um identificador próprio aos genes. Assim, no momento de obtenção de informação relativa a um gene, para além de obrigar o especialista a efetuar várias pesquisas em diferentes domínios, este tem também de utilizar uma ferramenta de conversão de identificadores de genes, para conseguir obter informação

relativa a um determinado gene, nas várias plataformas WEB existentes. Para além disto implicar um custo temporal bastante elevado, muitas vezes resulta na obtenção de muita informação repetida.

Após conseguir aceder a toda a informação que pretende, o especialista tem que recolher essa informação, que muitas vezes é recolhida em formatos distintos, dificultando e aumentando o seu trabalho, proceder à identificação e eliminação de informação repetida e só depois poderá trabalhar para obter conclusões relativas à informação obtida sobre um conjunto de genes.

Também esta procura de conclusões é bastante complicada para um conjunto alargado de genes, dado que atualmente existe bastante informação disponível e, derivado disso, torna-se praticamente impossível obter conclusões sem o recurso a métodos computacionais adequados.

Estes métodos adequados para análise de um conjunto alargado de dados consistem, muitas vezes, pela aplicação de técnicas de *data mining* dado que, como foi descrito na secção 2.5, estas técnicas permitem extrair conhecimento de grandes quantidades de informação, permitindo, ou facilitando a obtenção de conclusões.

O uso das ferramentas existentes para aplicação das diferentes técnicas existentes pode muitas vezes ser um obstáculo para especialistas da área da genómica, dado que por norma estes não têm conhecimento prévio em informática e é bastante frequente as ferramentas não terem uma interface amigável, pensada para todo o tipo de utilizadores, sendo que muitas delas requerem algum conhecimento em informática para uma correta utilização. Para além destas dificuldades referidas, acresce o facto das ferramentas não estarem disponíveis para todos os sistemas operativos e requererem a sua instalação nos SO suportados.

Outro problema na utilização destas ferramentas é o facto de estas requerem um grande poder computacional para conseguirem analisar um grande conjunto de dados. Esta limitação pode aumentar o desperdício de tempo por parte dos especialistas, aumentar os gastos, ou até inviabilizar o uso de certas ferramentas ou métodos, bloqueando assim a análise a alguns conjuntos de dados.

3.2 Solução

O projeto desenvolvido ao longo desta dissertação consiste numa Plataforma WEB que tenciona resolver os problemas descritos na secção anterior.

Dado um identificador de um gene, ou uma lista de identificadores de genes, através de um clique é possível efetuar a pesquisa em vários repositórios, recolhendo toda a informação fornecida por cada repositório.

Esta informação é tratada e de seguida é introduzida na base de dados, permitindo acessos mais rápidos em pesquisas posteriores. A informação é também mostrada ao utilizador, sendo possível aos utilizadores da aplicação visualizar a informação relativa aos genes, assim como efetuar o *download* dessa informação em vários formatos, facilitando assim as fases posteriores da investigação do especialista. Estes vários ficheiros possíveis de descarregar contêm informação filtrada, eliminando a informação repetida.

Após a obtenção da informação relativa a genes é possível aplicar várias técnicas de *data mining* através de algoritmos de classificação e *clustering*, de modo a analisar os genes com a finalidade de retirar conclusões, ou agrupar a informação recolhida em grupos mais dimensionados, de forma a facilitar a obtenção de conclusões, que permitam sugerir novo conhecimento que ajudem o especialista no caso de estudo.

Toda a plataforma foi construída com uma interface amigável para o utilizador, procurando desta forma que seja acessível a todo o tipo de utilizadores, com ou sem experiência informática.

3.3 Implementação

3.3.1 Arquitetura

Dado que a plataforma desenvolvida procura resolver dois problemas distintos que os investigadores enfrentam, apresentamos nas figuras 3.1 a arquitetura referente à solução de cada um desses problemas.

A primeira parte da figura é referente à recolha de informação relativa a genes enquanto a segunda parte da mesma figura representa a arquitetura relativa à aplicação de técnicas de *data mining*.

De notar que embora apresentemos a arquitetura dividida em duas partes, isso apenas acontece para facilitar a compreensão dos processos, sendo que apenas foi desenvolvida uma plataforma, que engloba os dois processos em simultâneo.

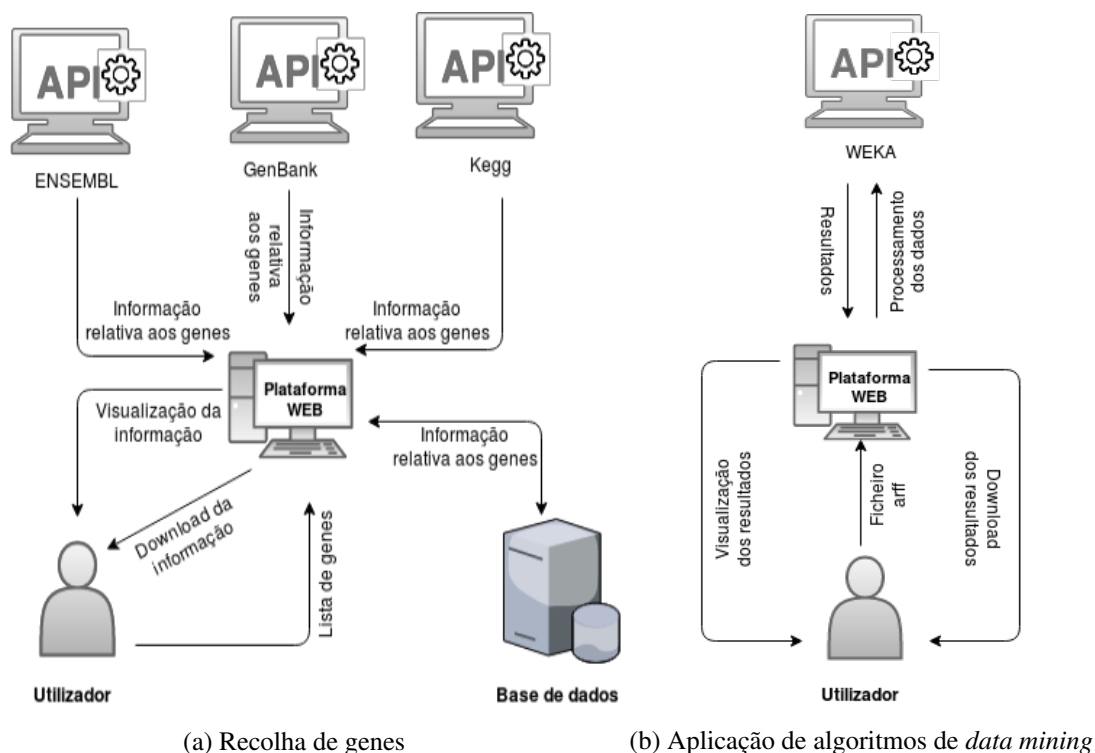


Figura 3.1: Arquitetura da plataforma

3.3.2 Pesquisa de genes

Para obter informação suficiente relativa a um determinado gene, percebemos que a pesquisa em apenas um sítio WEB não seria suficiente. Assim, após uma exaustiva pesquisa consideramos que deveríamos conectar a Plataforma WEB que desenvolvemos a três repositórios online, *Ensembl*, *GenBank* e *Kegg*. Estes três repositórios, assim como as suas APIs encontram-se descritos nas secções 2.2 e 2.3.

Os três repositórios selecionados para a Plataforma WEB são destinados à obtenção de informação relativa a produtos génicos por parte de biólogos, sendo que consideramos que se complementam relativamente à informação que possuem nas suas bases de dados. No entanto existe também bastante informação geral que encontramos nos três repositórios.

Assim, destacamos o *Ensembl* dado ser o único com bastante informação relativa aos transcritos e exões dos genes.

O *GenBank* é o repositório com maior número de genes na sua base de dados, englobando até genes já descontinuados, mas que, no entanto, poderão ainda ser úteis como termo de comparação.

O *Kegg* tem bastante informação que os restantes repositórios não têm, como é o caso dos atributos *aaseq*, *ntseq*, *module*, *pathway*, *orthology* e *structure*. Outra característica deste repositório, mais concretamente da API disponibilizada, é a capacidade de pesquisa por genes através do nome e não apenas pelo identificador, como é o caso dos restantes repositórios.

Na plataforma desenvolvida é possível pesquisar por genes em qualquer um dos repositórios referidos individualmente, ou em todos simultaneamente.

3.3.2.1 Pesquisa em um repositório

Quando o utilizador efetua a pesquisa apenas em um repositório, numa fase inicial verifica-se se cada um dos genes pesquisados já se encontram na base de dados da Plataforma WEB e, em caso afirmativo recolhe-se a informação diretamente da base de dados. Caso o gene não se encontre na base de dados da Plataforma WEB é adicionado a uma lista para que no final da verificação de todos os genes se faça o pedido à API do respetivo repositório com os identificadores necessários. O pedido é feito apenas ao repositório escolhido anteriormente, de modo a recolher a informação relativa ao genes introduzidos para pesquisa.

3.3.2.2 Pesquisa em vários repositórios

Quando a pesquisa é efetuada nos três repositórios em simultâneo, numa primeira fase efetua-se a conversão do identificador original para os identificadores dos restantes dois repositórios. Depois verifica-se quais os genes que já se encontram na base de dados da Plataforma WEB, recolhendo a informação desses e colocando numa lista os restantes. Isto é feito para todos os identificadores de cada repositório, obtidos através da conversão.

Depois de verificarmos quais os genes que não estão na base de dados da Plataforma WEB é efetuado o pedido a cada uma das três bases de dados dos repositórios de modo a recolher a informação sobre esses genes.

3.3.3 Conversão de Genes

Tal como referido anteriormente, o facto de cada repositório atribuir um identificador a cada gene dificulta o trabalho dos especialistas. Assim, na plataforma desenvolvida é também possível efetuar a conversão de genes de e para qualquer um dos três tipos.

Para efetuar esta conversão é utilizada a API da ferramenta *BioDB Hyperlink Management System*, descrita na secção 2.4.2.

Devido a limitações da API cada conversão apenas suporta um máximo de 100 identificadores, no entanto o utilizador não é confrontado com esta limitação, sendo que a Plataforma WEB efetua vários pedidos à API, se necessário, para converter todos os identificadores submetidos pelo utilizador.

Esta conversão é feita automaticamente quando o utilizador efetua uma pesquisa nos três repositórios em simultâneo, bastando para isso introduzir os identificadores de um dos três repositórios, ou o nome dos genes.

3.3.4 Resultados de pesquisa

Após efetuar a pesquisa relativa a um ou mais genes, em um ou nos três repositórios e, visto que a informação recebida das APIs dos repositórios tem diferentes formatos, toda a informação é modelada de forma a ser apresentada ao utilizador de forma legível, ocorrendo em simultâneo a inserção da informação na base de dados.

Para a formatação dos dados em XML foi utilizada a biblioteca Python *ElementTree*¹, facilitando assim o processo. Para os formatos JSON e de texto todo o processamento foi efetuado de forma manual.

3.3.5 Recolha de dados

Para além de visualizar a informação dos resultados da pesquisa, o utilizador pode também efetuar o *download* dessa informação em três formatos, texto, *arff* e *prolog*.

3.3.5.1 Ficheiro de texto

O ficheiro de texto descarregado da Plataforma WEB após a pesquisa permite ao investigador biológico guardar toda a informação num formato que permite a visualização futura sem recurso a ferramentas externas.

Cada gene é guardado com uma atributo por linha, começando pelo nome. O final do gene é indicado pelo símbolo "\\\".

No anexo A.2.1.1 mostramos um exemplo deste ficheiro.

¹<https://docs.python.org/2/library/xml.etree.elementtree.html>

3.3.5.2 Ficheiro *arff*

O ficheiro *arff* é por nós considerado o mais relevante dado que é através deste que é possível efetuar uma posterior análise à informação contida.

Este ficheiro respeita o formato convencional descrito na secção 2.9, contendo no cabeçalho o nome da relação e os atributos. Nos dados contém todos os genes pesquisados, um por linha, com os atributos separados por vírgulas.

No anexo A.2.1.2 podemos observar um exemplo deste ficheiro.

3.3.5.3 Ficheiro *prolog*

O ficheiro *prolog*, no formato *pl*, é essencialmente útil para a análise da informação através da linguagem de programação também com o nome *prolog*. Permite ao utilizador efetuar análise da informação ou visualizar essa informação através de simples programas de programação lógica.

Neste ficheiro, para além de guardar a informação total sobre os genes, guardamos também a informação completa relativa a todos os transcritos de cada gene.

Cada gene utiliza várias linhas do ficheiro, a primeira com a informação relativa ao próprio gene e as restantes com a informação relativa a todos os seus transcritos.

Na secção A.2.1.3 podemos observar um exemplo deste ficheiro.

3.3.6 Base de dados

A criação de uma base de dados torna-se essencial para tornar mais rápido o acesso à informação, evitando deste modo efetuar várias chamadas às APIs, relativamente ao mesmo gene. A base de dados torna também possível o acesso à informação de genes já pesquisados, mesmo que as APIs dos repositórios não estejam em funcionamento e ainda permite um acesso rápido a todos os genes já pesquisados através da listagens dos mesmos na Plataforma WEB.

A base de dados utilizada é SQLite, encontra-se descrita na secção 2.10.5.

Foram criadas cinco coleções de dados. Uma para cada repositório, uma para conter os identificadores dos três repositórios e uma para os transcritos dos genes.

Na figura 3.2 demonstramos os campos existentes nas coleções referidas.

Em qualquer pesquisa efetuada é inserida toda a informação relativa a esse gene na base de dados. Caso a pesquisa seja efetuada apenas em um repositório, apenas será adicionado à base de dados um elemento à tabela do repositório escolhido.

Caso a pesquisa seja efetuada nos três repositórios em simultâneo será adicionado a cada uma das tabelas dos três repositórios um elemento, mas será também adicionado um elemento à tabela *Gene*, contendo o identificador de cada um dos repositórios. Na pesquisa global, ou na pesquisa no repositório *Ensembl* é possível selecionar a opção *Include Transcripts*, que quando selecionada irá adicionar à tabela *EnsemblTranscripts* toda a informação recolhida sobre todos os transcritos encontrados relativos a cada gene pesquisado. Quando não selecionada, apenas será guardado o número de transcritos de um determinado gene. O facto desta opção ser decidida pelo utilizador

Plataforma WEB

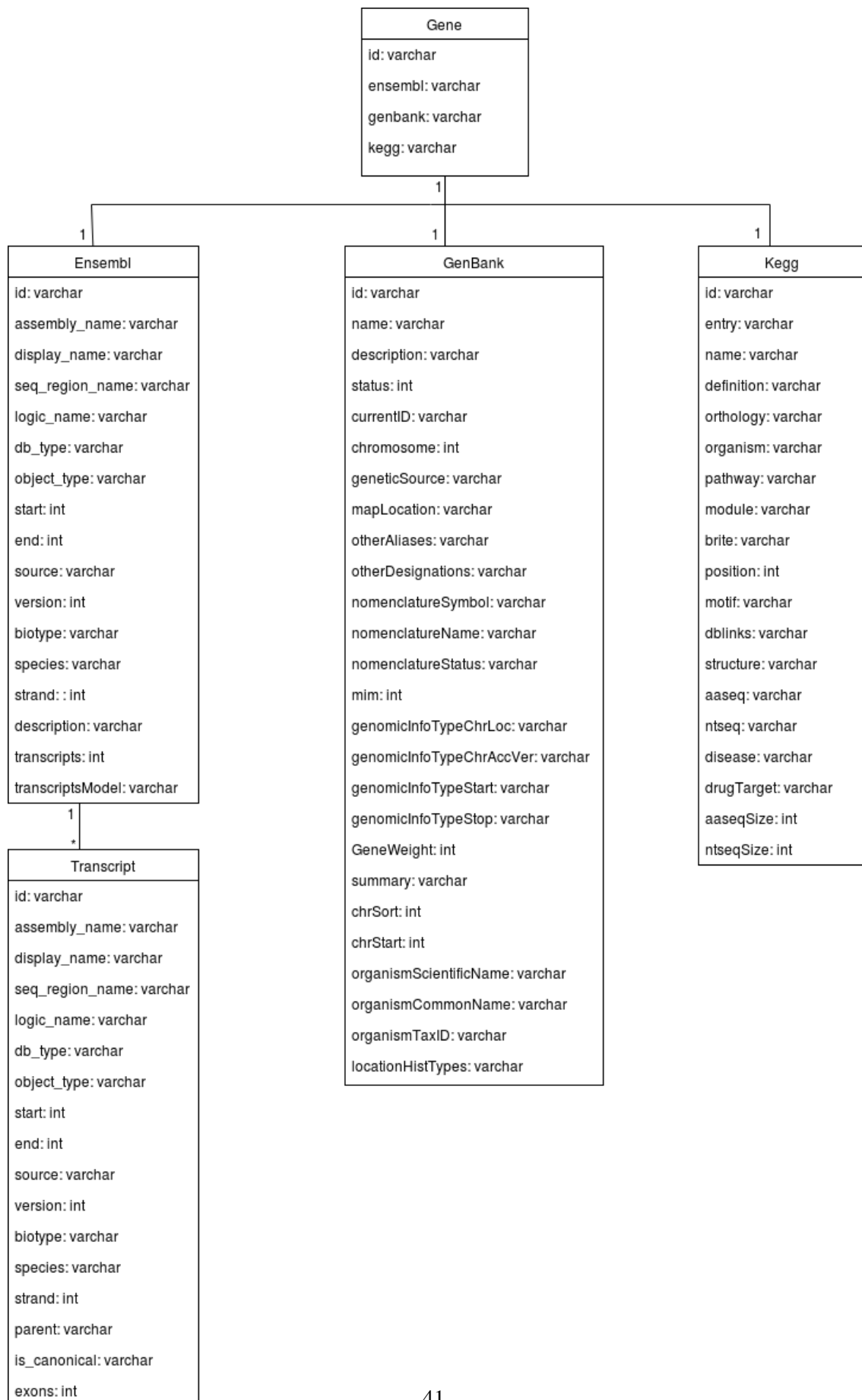


Figura 3.2: Diagrama UML da base de dados

deve-se ao facto de quando ativa aumentar bastante o tempo de pesquisa, pois cada gene poderá ter dezenas de transcritos.

3.3.7 Listagem de genes

A aplicação tem também uma página destinada à listagem dos genes presentes na sua base de dados, sendo possível listar os genes de um determinado repositório, ou todos os genes presentes na base de dados.

Para além da visualização da informação relativa aos genes, é possível seleccionar genes da listagem, podendo depois efetuar o *download* das informações detalhadas sobre os genes seleccionados em formato de texto, ou efetuar o *download* da informação em formato *arff* ou *prolog*.

3.3.8 Processamento de ficheiro *ARFF* para *data mining*

Para ser possível analisar a informação relativa a uma grande quantidade de genes, é necessário aplicar diferentes técnicas de *data mining*, adequadas à investigação ocorrente.

De forma a facilitar o trabalho do investigador, para aplicar uma técnica de *data mining* não é necessário filtrar manualmente a informação recolhida e inserida no ficheiro *arff*. Isto acontece porque a aplicação automaticamente faz a filtragem da informação contida no ficheiro *arff*, retirando valores que não são possíveis de analisar num determinado método escolhido pelo utilizador, ou transformando esses valores em atributos suportados pelo método.

Para efetuar essa filtragem é utilizada a API da ferramenta *WEKA*, que permite retirar atributos não suportados, alterar o formato de atributos que poderão ser suportados por exemplo no formato nominal mas não no formato inteiro, entre outros casos.

3.3.9 Aplicação de técnicas de *data mining*

A plataforma desenvolvida no âmbito dessa dissertação permite a aplicação de várias técnicas de *data mining* a um conjunto informação pertencente a uma grande quantidade de genes, permitindo assim a análise dessa informação tendo como objetivo final a obtenção de conhecimento, ou a distribuição dos genes em grupos mais reduzidos, aumentando as semelhanças entre genes do mesmo grupo.

Para isto decidimos implementar na Plataforma técnicas de classificação e técnicas de *clustering*, dado serem as mais adequadas ao tipo de informação que conseguimos obter em genes e às conclusões que pretendemos retirar.

Para a aplicação de várias técnicas existentes utilizamos a biblioteca *python-weka-wrapper*² que permite a utilização de vários algoritmos do *WEKA*, em *Python*. Esta biblioteca utiliza *javabridge*³ para iniciar, comunicar e terminar a *Java Virtual Machine*, onde todos os processos do *WEKA* se realizam.

²<http://pythonhosted.org/python-weka-wrapper/>

³<https://pypi.python.org/pypi/javabridge>

Assim, depois do utilizador escolher a técnica que pretende utilizar, introduzir ou seleccionar as opções que pretende, a plataforma aplica a técnica através da *Java Virtual Machine*, recolhe os resultados e formata-os para poderem ser apresentados ao utilizador na Plataforma WEB.

Para além de aplicar diferentes técnicas, utilizaram-se também métricas para avaliar as técnicas, sendo estes resultados também apresentados ao utilizador.

3.3.9.1 Classificação

Os algoritmos de classificação seleccionadas para a Plataforma desenvolvida foram as seguintes:

- *J48*;
- *Random Forest*;
- *Naive Bayes*.

Para cada um destes algoritmos é necessário introduzir os dados para analisar mas também dados previamente analisados, isto porque a classificação se engloba nas técnicas de *data mining* com aprendizagem supervisionada.

Assim os dados serão avaliados tendo como base os dados previamente classificados, de modo a permitir um valor mais credível nos resultados da aplicação dos algoritmos.

Existem várias opções disponíveis para cada um destes algoritmos, sendo todas elas deixadas ao critério do utilizador, através da interface da Plataforma WEB.

Os resultados mais relevantes destes algoritmos e respetiva avaliação são apresentados ao utilizador, como poderemos ver na secção 3.4.7.1. É também possível descarregar os resultados completos de cada algoritmo no formato de texto e *csv*.

3.3.9.2 Clustering

A aplicação de algoritmos de *clustering* pareceram-nos mais relevantes para estudos relacionados com informação génica e, consequentemente, decidimos integrar seis algoritmos na Plataforma WEB desenvolvida. Embora algumas pertençam ao mesmo tipo de *clustering* e algumas tenham como base o mesmo algoritmo, todos eles nos pareceram importantes para diferentes detalhes. Deste modo implementámos os seguintes algoritmos:

- Simple K-means;
- Make Density Based Clusterer;
- Hierarchical Clustering;
- Expectation Maximization
- Farthest-First;

- DBSCAN.

Para cada um destes algoritmos deixamos ao critério do utilizador todos os parâmetros de cada método, permitindo assim ao especialista customizar os algoritmos de modo a irem ao encontro das suas necessidades.

Todos estes algoritmos têm como objetivo a divisão das instâncias analisadas em diferentes *clusters*, aumentando as semelhanças entre as instâncias pertencentes ao mesmo *cluster* e aumentando as diferenças entre cada *cluster*.

Ao utilizador são apresentados os resultados considerados mais relevantes em cada algoritmo, permitindo também descarregar os resultados completos em formato de texto ou *csv*.

3.3.10 Interface de administração

Tal como referido no capítulo 2.10.2 uma das principais características que a *framework* Django possui é a sua interface de administração.

Assim um administrador da plataforma, após fazer *login* poderá apagar, acrescentar ou modificar qualquer elemento da base de dados.

Esta interface é bastante intuitiva, de simples utilização, permitindo a qualquer utilizador usufruir de todas as suas funcionalidades de uma forma simples e segura.

3.4 Funcionamento da Plataforma WEB

Nesta secção apresentamos as diferentes funcionalidades da Plataforma WEB, apresentamos alguns detalhes da mesma e descrevemos o funcionamento das ações mais relevantes.

3.4.1 Menu

Para facilitar a navegação em toda a Plataforma WEB, foi desenvolvido um menu presente no topo de todas as páginas, que permite o acesso fácil às principais funcionalidades da Plataforma desenvolvida. Neste menu conseguimos também perceber sempre em que funcionalidade nos encontramos, dado que a funcionalidade em que nos encontramos permanece ativa na navegação.

Podemos observar o menu na figura 3.3.

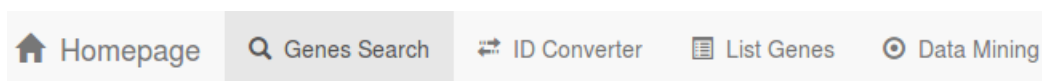


Figura 3.3: Menu de navegação na Plataforma WEB, ativo na funcionalidade *Genes Search*

3.4.2 Casos de Uso

Como foi referido na secção 3.2, a Plataforma apresenta várias funcionalidades que vão ao encontro das necessidades dos investigadores.

Assim, apresentamos na figura 3.4 as ações principais que o ator pode realizar na Plataforma desenvolvida. Decidimos atribuir o nome de "Investigador" ao ator, dado serem estes os principais utilizadores da Plataforma WEB.

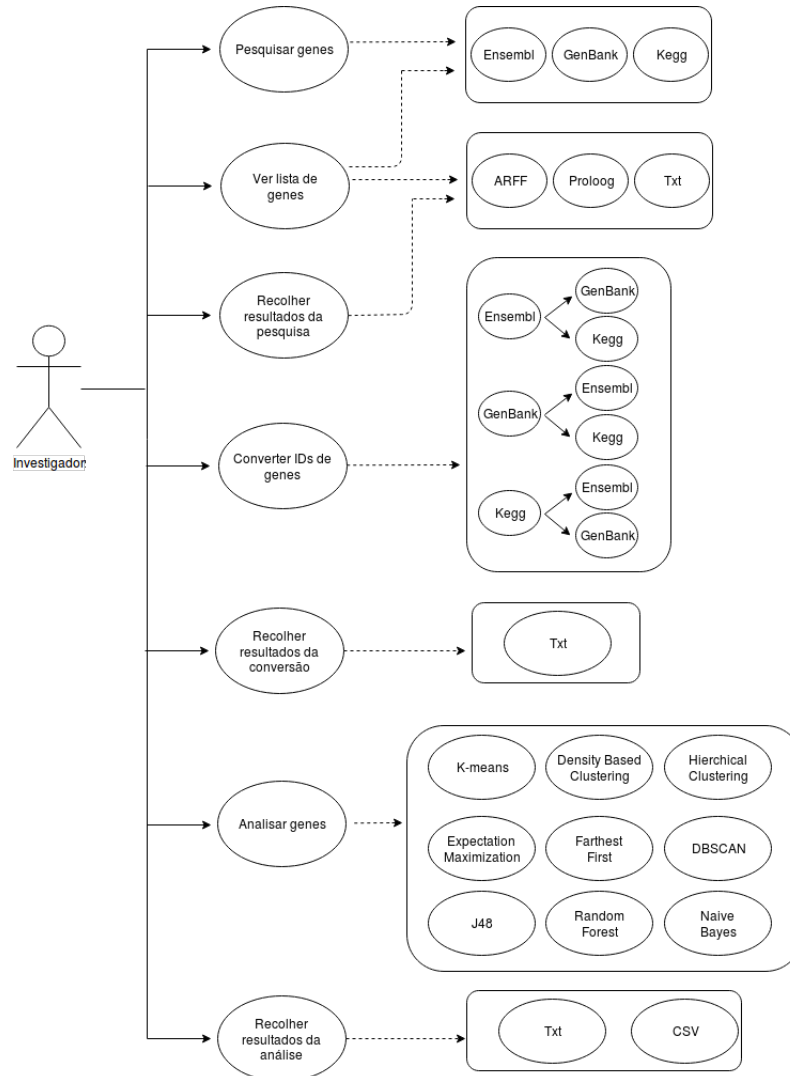


Figura 3.4: Diagrama de casos de uso para o ator Investigador

Apresentamos também na figura 3.5 as ações disponíveis para o administrador da plataforma, sendo que estas focam apenas na gestão da base de dados, permitindo adicionar, modificar ou remover qualquer elemento.

3.4.3 Pesquisa de genes

Uma das principais funcionalidades da Plataforma WEB é a pesquisa de informação relativa a genes. Na figura 3.6 apresentamos a pagina inicial para pesquisa de genes, onde podemos seleccionar o repositório onde desejamos pesquisar, ou a opção *ALL* que permite a pesquisa em todos os

Plataforma WEB

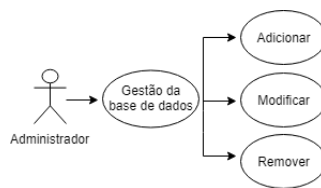


Figura 3.5: Diagrama de casos de uso para o ator Administrador

repositórios em simultâneo.

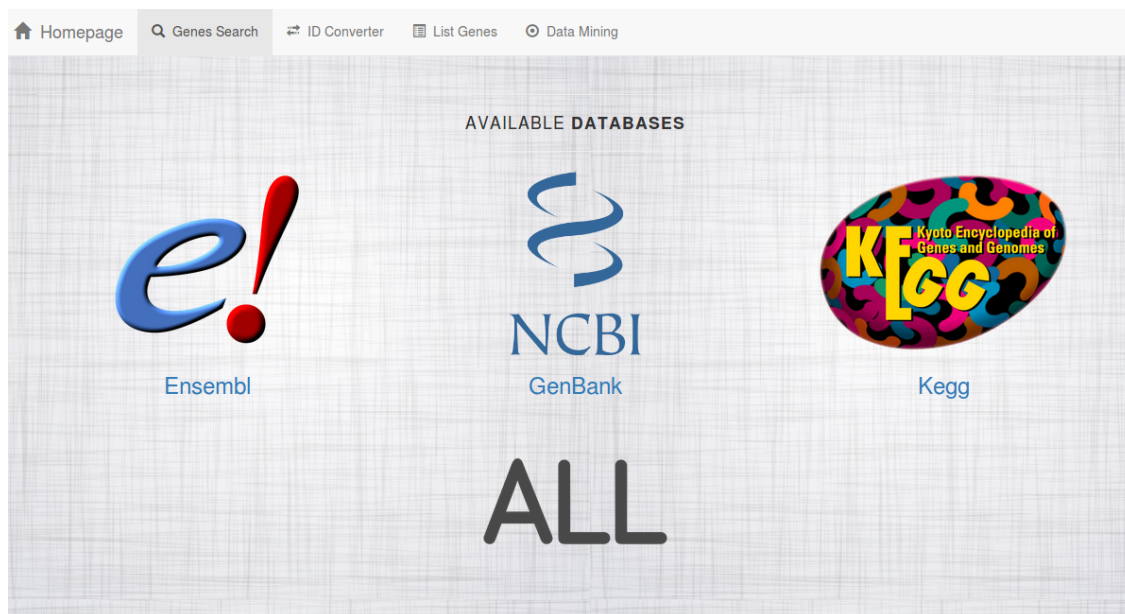


Figura 3.6: Página inicial para pesquisa de genes

3.4.3.1 Pesquisa em um repositório

Para efetuar a pesquisa de um ou vários genes, o utilizador deverá começar por escolher, na página principal de pesquisa de genes, qual o repositório em que pretende pesquisar. De seguida será direcionado para uma página específica para cada repositório, dado que cada um deles poderá ter opções de pesquisa diferentes.

Caso o repositório selecionado seja o *Ensembl* o utilizador poderá inserir uma lista de identificadores de genes *Ensembl* separados por vírgula ou poderá proceder ao *upload* de um ficheiro contendo esses identificadores também separados por vírgula. O utilizador tem ainda uma opção selecionável através de uma *checkbox* para expandir ou não a pesquisa aos transcritos dos genes pesquisados.

No caso do repositório escolhido ser o *GenBank* o utilizador apenas deverá introduzir os identificadores de genes *GenBank*, ou fazendo *upload* de um ficheiro, tal como no *Ensembl*.

O repositório *Kegg*, quando escolhido é o que oferece maior variedade nas opções. Para efetuar a pesquisa neste repositório o utilizador deve primeiro selecionar a opção *ID* ou *Name*, conforme pretenda pesquisar por identificador ou pelo nome do gene. De seguida o utilizador terá de escolher o organismo do gene em pesquisa, sendo este campo *auto-complete*, apresentando sugestões conforme o *input* do utilizador. Depois deverá introduzir os identificadores ou nomes dos genes separados por vírgula, ou através de um ficheiro, tal como acontece nos restantes dois repositórios.

3.4.3.2 Pesquisa em vários repositórios

Caso o utilizador opte por pesquisar o gene ou a lista de genes nos três repositórios suportados pela plataforma, deverá na página principal da pesquisa de genes selecionar a opção "ALL", sendo direcionado para uma página específica para o efeito.

Nesse página deverá escolher qual a base de dados do identificador do gene que pretende pesquisar, ou se pretender pesquisar através do nome deverá selecionar a opção *Kegg*, sendo que depois poderá introduzir o nome de cada gene.

Para introduzir os identificadores ou nomes dos genes que pretende pesquisar o utilizador poderá fazê-lo inserindo na plataforma os IDs ou nomes separados por vírgula, ou através do *upload* de um ficheiro com a mesma informação, também separada por vírgulas.

Neste tipo de pesquisa global a aplicação automaticamente converte o identificador de cada gene pesquisado, de forma a efetuar a pesquisa nas três bases de dados suportadas, obtendo o máximo de informação possível.

Na figura 3.7 demonstramos o cenário de pesquisa quando o utilizador seleciona a opção *ALL* e introduz genes com identificador do repositório *GenBank*.

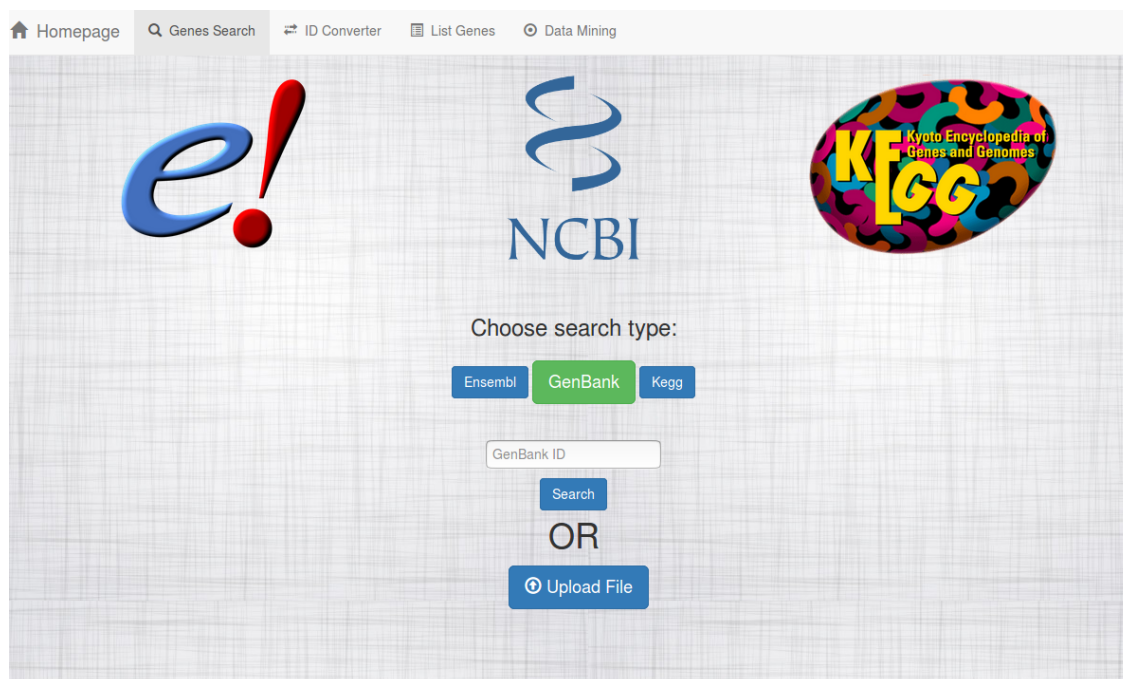


Figura 3.7: Página para pesquisa de genes em todos os repositórios em simultâneo

3.4.4 Resultados da Pesquisa e Recolha de informação

Após efetuar a pesquisa relativa aos genes pretendidos o utilizador segue automaticamente para uma página onde lhe são apresentados os resultados da sua pesquisa.

Caso a pesquisa não retorne resultados é apresentada ao utilizador uma página de erro com a informação de que não foi possível obter resultados para os IDs introduzidos.

Quando a pesquisa é efetuada com sucesso são apresentados ao utilizador todos os genes que pesquisou, assim como a informação relativa a cada um deles.

Nesta mesma página o utilizador pode proceder ao *download* da informação que lhe é apresentada, em vários formatos. Poderá descarregar a informação em formato de texto, *arff* ou *prolog*, bastando para isso clicar no botão destinado a esse efeito.

Apresentamos na figura 3.8 a página relativa aos resultados da pesquisa, englobando os botões para efetuar o download da informação.

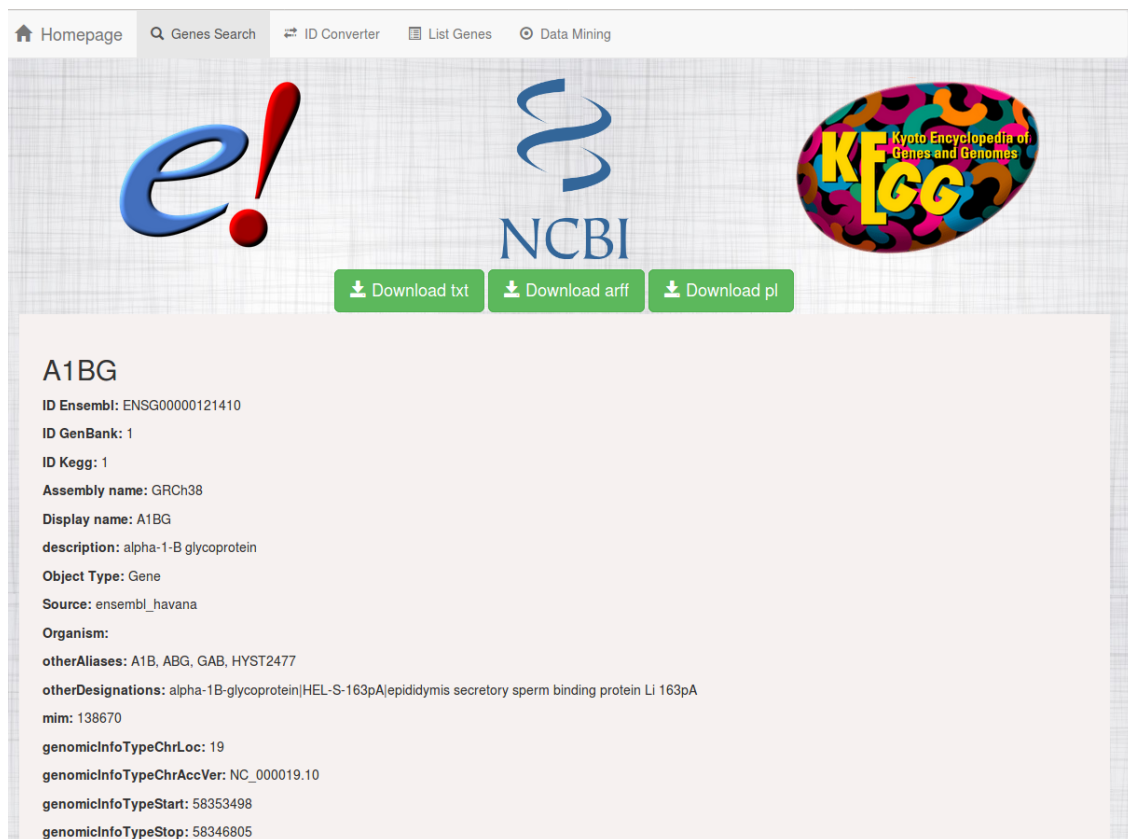


Figura 3.8: Página com os resultados da pesquisa efetuada pelo utilizador

3.4.5 Listagem de genes

É também possível ao utilizador visualizar todos os genes inseridos na base de dados. Para isso apenas terá de aceder à página *List Genes*, através do menu no topo da página, e selecionar

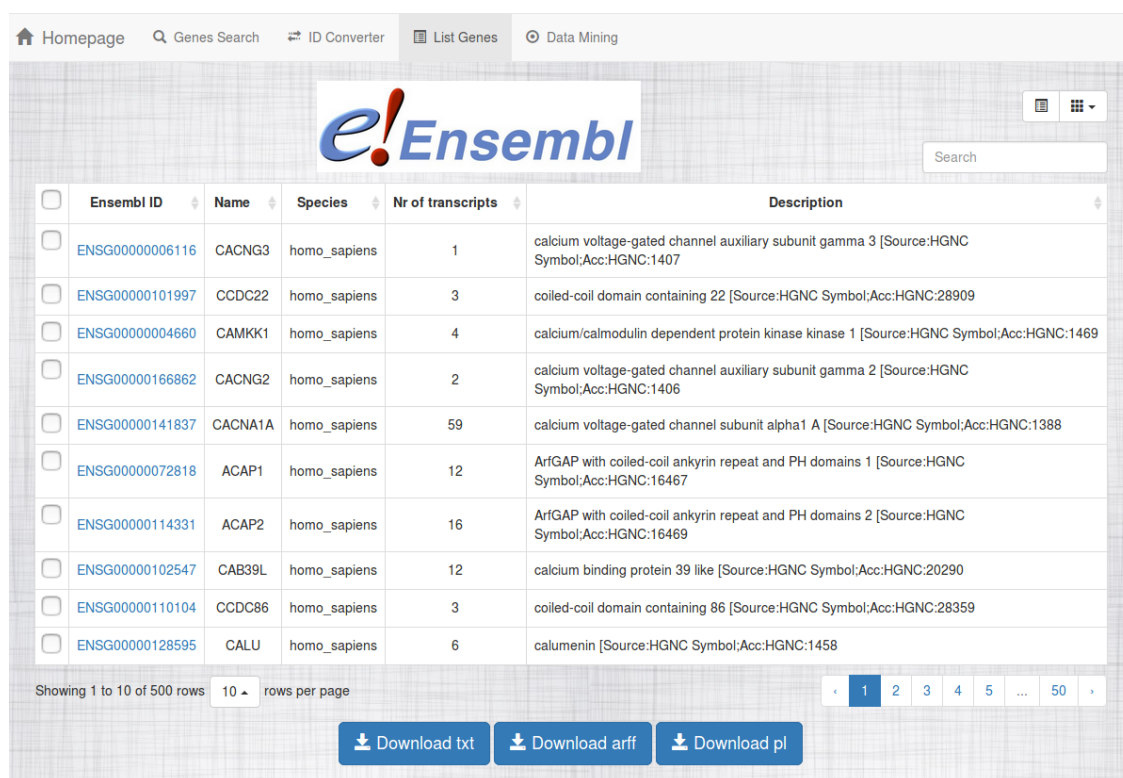
qual o repositório que pretende listar os genes, ou se pretende visualizar todos.

Nesta listagem de genes o utilizador pode observar as informações mais relevantes de cada gene, podendo também ver a informação detalhada de cada gene, bastando para isso clicar no identificador do mesmo.

Se pretender pesquisar por um gene em específico poderá fazê-lo na parte superior direita da tabela, sendo que pode pesquisar por nome, ID, espécie ou descrição.

É também possível nesta página seleccionar os genes que pretender e fazer o *download* da informação nos três formatos referidos anteriormente, texto, *arff* e *prolog*.

Na figura 3.9 demonstramos a página de listagem de genes, após seleccionar o repositório *Ensembl*.



☐	Ensembl ID	Name	Species	Nr of transcripts	Description
☐	ENSG00000006116	CACNG3	homo_sapiens	1	calcium voltage-gated channel auxiliary subunit gamma 3 [Source:HGNC Symbol;Acc:HGNC:1407]
☐	ENSG00000101997	CCDC22	homo_sapiens	3	coiled-coil domain containing 22 [Source:HGNC Symbol;Acc:HGNC:28909]
☐	ENSG00000004660	CAMKK1	homo_sapiens	4	calcium/calmodulin dependent protein kinase kinase 1 [Source:HGNC Symbol;Acc:HGNC:1469]
☐	ENSG00000166862	CACNG2	homo_sapiens	2	calcium voltage-gated channel auxiliary subunit gamma 2 [Source:HGNC Symbol;Acc:HGNC:1406]
☐	ENSG00000141837	CACNA1A	homo_sapiens	59	calcium voltage-gated channel subunit alpha1 A [Source:HGNC Symbol;Acc:HGNC:1388]
☐	ENSG00000072818	ACAP1	homo_sapiens	12	ArfGAP with coiled-coil ankyrin repeat and PH domains 1 [Source:HGNC Symbol;Acc:HGNC:16467]
☐	ENSG00000114331	ACAP2	homo_sapiens	16	ArfGAP with coiled-coil ankyrin repeat and PH domains 2 [Source:HGNC Symbol;Acc:HGNC:16469]
☐	ENSG00000102547	CAB39L	homo_sapiens	12	calcium binding protein 39 like [Source:HGNC Symbol;Acc:HGNC:20290]
☐	ENSG00000110104	CCDC86	homo_sapiens	3	coiled-coil domain containing 86 [Source:HGNC Symbol;Acc:HGNC:28359]
☐	ENSG00000128595	CALU	homo_sapiens	6	calumenin [Source:HGNC Symbol;Acc:HGNC:1458]

Showing 1 to 10 of 500 rows 10 rows per page

[Download txt](#)
[Download arff](#)
[Download pl](#)

Figura 3.9: Listagem de todos os genes *Ensembl* incluídos na base de dados

3.4.6 Conversão de identificadores de genes

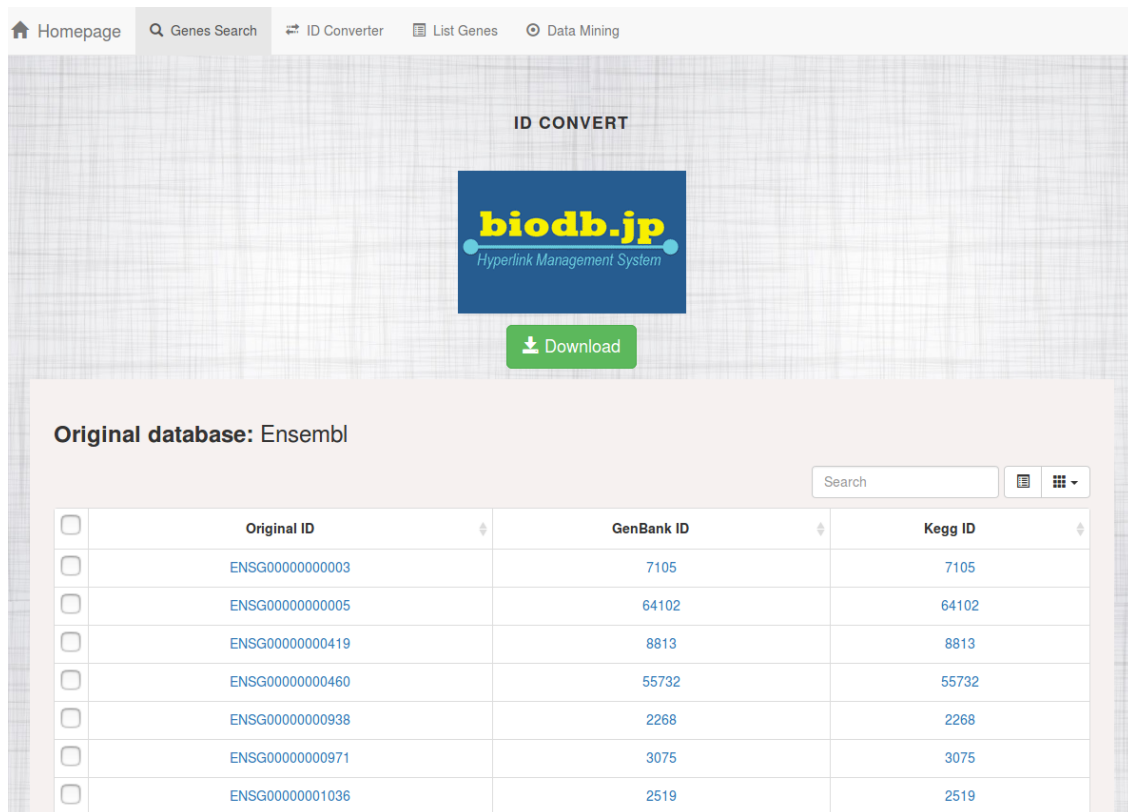
Tal como foi referido nas secções anteriores, a conversão de identificadores pode, por si só, ser bastante útil no dia a dia dos especialistas.

Assim, englobamos na plataforma desenvolvida uma página para esse efeito.

Para utilizar esta funcionalidade o utilizador apenas tem de introduzir os identificadores que pretende converter separados por vírgula e escolher a base de dados originária desses identificadores.

Serão apresentados os resultados através de uma tabela contendo o ID original e os IDs correspondentes às restantes bases de dados.

Na figura 3.10 podemos observar o resultado de uma conversão de identificadores relativos à base de dados *Ensembl*.



The screenshot shows the 'ID CONVERT' interface of the bioDB.jp Hyperlink Management System. The 'Original database' is set to 'Ensembl'. A table displays the conversion results for eight Ensembl IDs to their corresponding GenBank and Kegg IDs. A 'Download' button is visible above the table.

	Original ID	GenBank ID	Kegg ID
☐	ENSG00000000003	7105	7105
☐	ENSG00000000005	64102	64102
☐	ENSG000000000419	8813	8813
☐	ENSG000000000460	55732	55732
☐	ENSG000000000938	2268	2268
☐	ENSG000000000971	3075	3075
☐	ENSG000000001036	2519	2519

Figura 3.10: Resultado de uma conversão

Após ser efetuada a conversão é também possível descarregar os resultados no formato de texto, facilitando a utilização desses identificadores em pesquisas futuras.

3.4.7 Aplicação de técnicas de *data mining*

Para aplicar as diferentes técnicas de *data mining* suportadas pela plataforma o utilizador deverá ter um ficheiro do tipo *arff*, que pode ter sido obtido através da plataforma, ou de uma fonte externa.

O utilizador pode aceder às funcionalidades de *data mining* da plataforma desenvolvida clicando na opção *Data Mining* no menu que se encontra no topo da página.

Nesta página estão listadas as técnicas de classificação e de *clustering* disponíveis, devendo o utilizador escolher uma.

Após escolher o método que pretende aplicar o utilizador é encaminhado para uma página específica desse método onde lhe serão apresentadas opções disponíveis, seguidas de uma breve

explicação relativa a cada uma delas, para melhor entendimento do utilizador. É também nesta página que o utilizador faz o *upload* do ficheiro *arff* a ser analisado, através do botão destinado para esse efeito. Caso o algoritmo escolhido pelo utilizador seja relativo a classificação, depois de submeter o ficheiro para análise ser-lhe-á pedida a submissão de um ficheiro previamente classificado, dado que este tipo de *data mining* corresponde a uma aprendizagem supervisionada.

Na figura 3.11 apresentamos a página inicial relativa ao *data mining*, onde se encontram listados os métodos disponíveis.

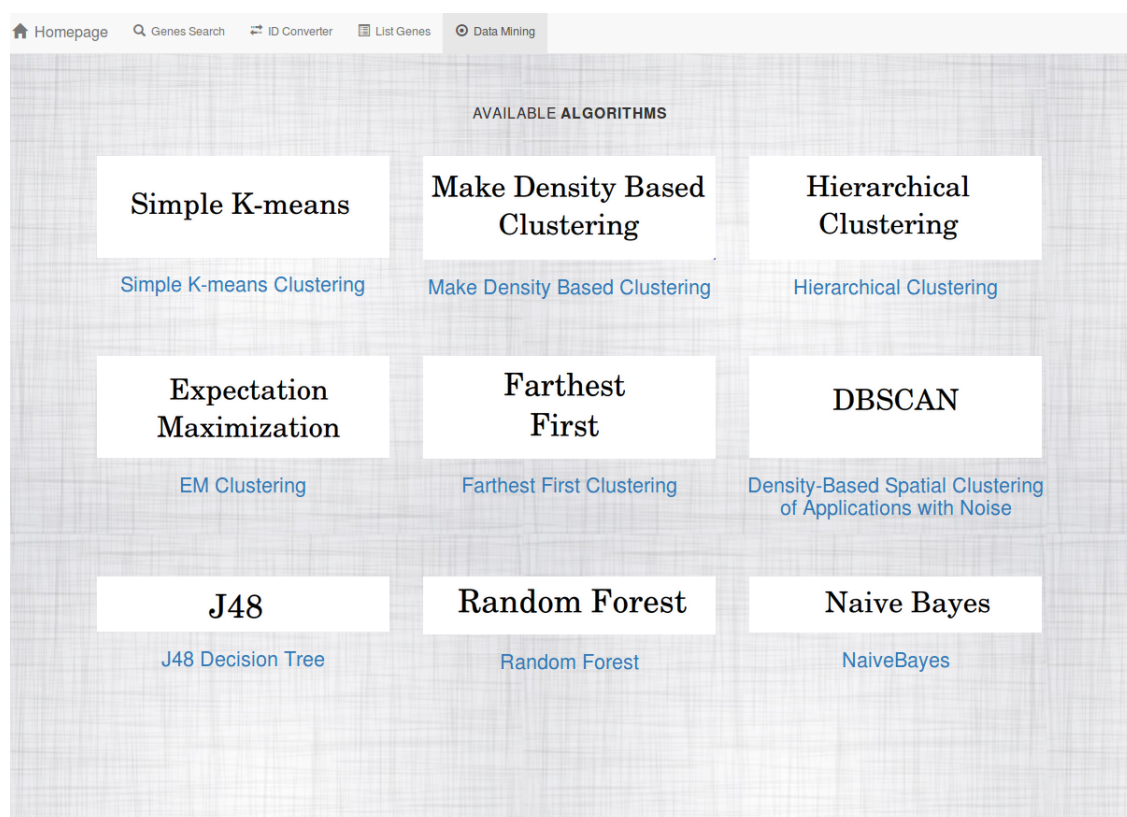


Figura 3.11: Página inicial relativa ao *data mining*

3.4.7.1 Classificação

Nesta secção apresentamos as técnicas de classificação disponíveis na plataforma, descrevendo as opções de cada uma delas.

- **J48**

Esta técnica de classificação permite a criação de modelos de decisão em árvore, utilizando o algoritmo C4.5, descrito na secção 2.6.1.1.

Tem como opção a escolha do limite para a confiança para a poda da árvore, um valor entre zero e um. O valor predefinido é 0.25. Em alternativa é possível escolher a opção para não podar a árvore.

- **Random Forest**

Esta técnica, permite também a criação de modelos de decisão em árvore, através da combinação de várias árvores de decisão de forma a gerarem um classificador final. É bastante maleável, permite que o utilizador decida várias opções, como o número máximo de iterações, a profundidade máxima da árvore, entre várias outras opções disponíveis na plataforma.

- **Naive Bayes**

Esta técnica baseada no teorema de Bayes utiliza a probabilidade para classificar dados, supondo que existe independência entre os atributos.

Permite ao utilizador escolher a estimativa de densidade *Kernel* ao invés da distribuição normal para atributos numéricos e usar discretização supervisionada, também para atributos numéricos.

3.4.7.2 *Clustering*

Nesta secção apresentamos as técnicas de *clustering* disponíveis na plataforma, descrevendo as opções de cada uma delas.

- **Simple K-means**

Esta técnica permite ao utilizador ter controlo sob praticamente todas as variáveis do algoritmo, sendo possível na Plataforma WEB escolher, por exemplo o número de *clusters* pretendidos, o método de inicialização a utilizar, o algoritmo de distância, entre outros.

- **Make Density Based Clusterer**

Este algoritmo permite ao utilizador escolher o número de *clusters* pretendido, o valor mínimo de desvio padrão para o cálculo da densidade e a opção ou não de substituir os valores em falta com a média/moda.

- **Hierarchical Clustering**

Este método aglomerativo, para além de permitir também escolher o número pretendido de *clusters*, permite também definir o algoritmo de distância entre instâncias e também entre *clusters*.

- **Expectation Maximization**

Este algoritmo não necessita de um valor predefinido para o número de *clusters*, através de *cross validation* define por si mesmo esse número. No entanto permite também que essa escolha seja feita pelo utilizador. Para além dessa opção, permite também a escolha do valor mínimo para o desvio padrão para o cálculo da densidade, o máximo de iterações, entre outros.

Na figura 3.12 podemos visualizar as opções disponíveis para o método *Expectation Maximization*, seguindo-se a figura 3.13 com a explicação relativa a cada opção, assim como o seu valor predefinido. Ambas as figuras foram retiradas da Plataforma WEB desenvolvida.

Figura 3.12: Opções disponíveis para o método *Expectation Maximization*

- **Farthest-First**

Este algoritmo, apenas permite definir o número de *clusters* pretendidos pelo utilizador e o número de valores aleatórios para inicializar o método.

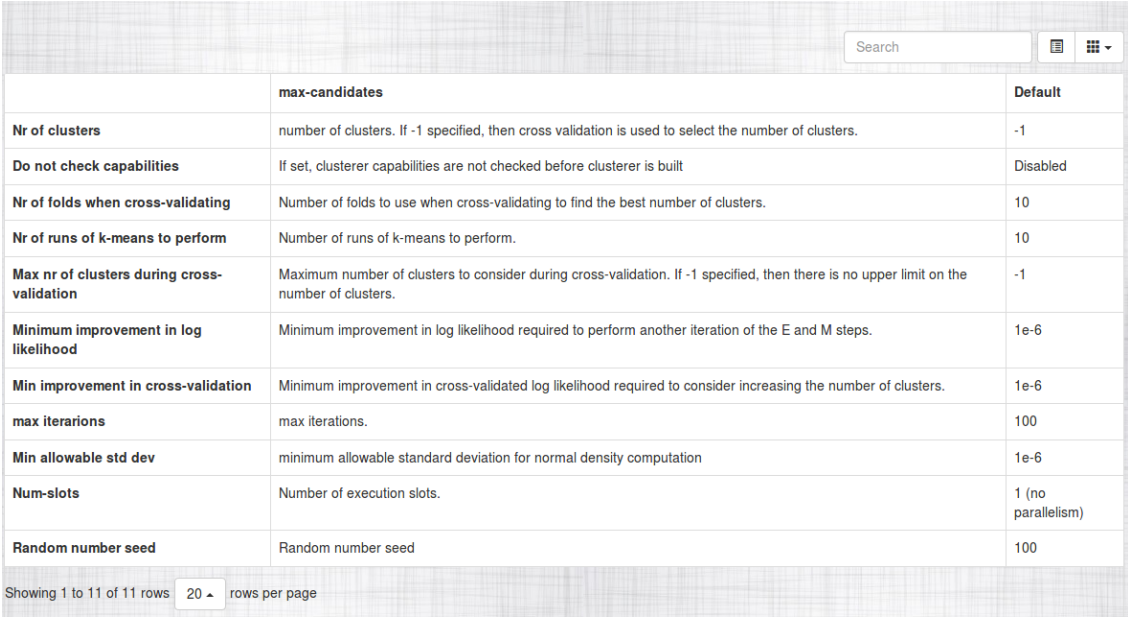
- **DBSCAN**

Este método permite escolher o algoritmo usado para calcular a distância entre instâncias, o número mínimo de elemento para formar um *cluster* e o valor máximo do raio de um *cluster*.

3.4.8 Resultados de aplicação de técnicas de *data mining*

Após aplicar uma técnica de *data mining* são exibidos ao utilizador os resultados obtidos. Estes resultados diferem entre cada um dos métodos, no entanto todos demonstram ao utilizador as conclusões possíveis de retirar após aplicar os algoritmos.

Relativamente ao algoritmos de *clustering* são apresentados os valores médios de cada atributo para todas as instâncias em conjunto e também para as instâncias de cada *cluster*, sendo assim possível comparar os valores de cada *cluster* com os valores de todos os dados. São também apresentados os valores do número de iterações, a percentagem de instâncias pertencentes a cada *cluster*, entre outros resultados considerados relevantes para o utilizador. Em complemento a esta



	max-candidates	Default
Nr of clusters	number of clusters. If -1 specified, then cross validation is used to select the number of clusters.	-1
Do not check capabilities	If set, clusterer capabilities are not checked before clusterer is built	Disabled
Nr of folds when cross-validating	Number of folds to use when cross-validating to find the best number of clusters.	10
Nr of runs of k-means to perform	Number of runs of k-means to perform.	10
Max nr of clusters during cross-validation	Maximum number of clusters to consider during cross-validation. If -1 specified, then there is no upper limit on the number of clusters.	-1
Minimum improvement in log likelihood	Minimum improvement in log likelihood required to perform another iteration of the E and M steps.	1e-6
Min improvement in cross-validation	Minimum improvement in cross-validated log likelihood required to consider increasing the number of clusters.	1e-6
max iterations	max iterations.	100
Min allowable std dev	minimum allowable standard deviation for normal density computation	1e-6
Num-slots	Number of execution slots.	1 (no parallelism)
Random number seed	Random number seed	100

Showing 1 to 11 of 11 rows 20 rows per page

Figura 3.13: Explicação disponível para as opções do método *Expectation Maximization*

informação são também apresentadas tabelas relativas a cada *cluster*, contendo todas as instâncias pertencentes, assim como a informação relativa a cada um delas.

Nos algoritmos de classificação são apresentadas as percentagens de valores corretamente classificados, os valores médios de erro, entre outros valores. Estes valores derivam do facto deste tipo de aprendizagem ser supervisionada, contendo dados previamente classificados. São também apresentados os valores de várias métricas de classificação como *precision*, *recall*, *f-measure*, *auc*, entre outros.

Em complemento aos resultados apresentados é também possível efetuar o *download* da informação apresentada e também de alguns resultados não demonstrados na Plataforma WEB, principalmente se os dados analisados forem muito extensos.

3.5 Conclusões

Neste capítulo aprofundamos o problema que nos foi proposto nesta dissertação e descrevemos pormenorizadamente a solução.

Apresentámos as várias funcionalidades disponíveis na Plataforma WEB desenvolvida, incluindo alguns detalhes que consideramos mais relevantes e descrevemos a implementação dessas funcionalidades.

Foram também apresentadas algumas figuras da plataforma de modo a demonstrar a interface amigável que possui, com o objetivo de facilitar a sua utilização e alargar o número de utilizadores.

Capítulo 4

Resultados e Casos de Estudo

Neste capítulo apresentamos os resultados do projeto que desenvolvemos, incluindo três casos de estudo que permitem a validação e verificação da solução.

Todos os dados utilizados para a validação da Plataforma WEB são reais. Foram utilizados dados provenientes de diversas fontes de modo a permitir uma validação mais profunda da plataforma desenvolvida.

4.1 Ambiente Experimental

Para efetuar os casos de estudo 4.3, 4.4 e 4.5 utilizou-se uma máquina com as especificações apresentadas na tabela 4.1.

Tabela 4.1: Especificações da máquina

Modelo	Samsung NP700Z5C-S03PT
SO	Ubuntu 16.04 LTS 64-bit
CPU	Intel Core i7-3635QM 2.40GHz
Memória	8 GB
Vel. Internet	50 Mbps

4.2 Atributos estudados

Um gene é composto por informação de vários tipos, tornando-se complicado estudar toda a informação em simultâneo. Embora o ficheiro *arff* descarregado da Plataforma WEB contenha toda a informação relativa ao gene, nestes casos de estudo não estudamos todos os atributos dos genes.

Os atributos mais relevantes para os casos de estudo que se seguem são os seguintes:

- **chromosome**

Um cromossoma corresponde a um pacote de DNA encontrado no núcleo das células. Diferentes organismos possuem um diferente número de cromossomas, sendo que nos humanos é

23. Assim este atributo, visto que todos os genes estudados pertencem à espécie *hsa - homo sapiens* varia entre 1 e 23.

- **mim**

Mendelian Inheritance in Man é uma base de dados que engloba todas as doenças humanas que tenham uma componente genética. Quando possível faz a ligação de genes a essas doenças através deste valor.

Na tabela 4.2 mostramos os possíveis valores que este atributo pode ter, assim como qual o significado desse valor.

Tabela 4.2: Valores possíveis do atributo *MIM*

Código MIM		Hereditariedade
≥ 100 000	≤ 299 000	Loci autossómicos ou fenótipos (entradas anteriores a 15 de maio de 1994)
≥ 300 000	≤ 399 000	Loci no cromossomas X ou fenótipos
≥ 400 000	≤ 499 000	Loci no cromossomas Y ou fenótipos
≥ 500 000	≤ 599 000	Loci mitocondriais ou fenótipos
≥ 600 000		Loci autossómicos ou fenótipos (entradas após 15 de maio de 1994)

- **source**

Parâmetro relativo à anotação do gene, podendo apenas ser *ensembl_havana* ou *havana*.

- **gene weight**

Este valor inteiro é calculado tendo em conta várias linhas de evidência relativas à avaliação da caracterização do gene.

Não tem valor limite.

- **aa seq size**

Este atributo é referente ao tamanho da sequência de aminoácidos do gene.

Não tem limitações de valores.

- **nt seq size**

Este atributo é referente ao tamanho da sequência de nucleótidos do gene.

Tal como o *aaseq* não tem também limitações de valores.

- **version**

Este atributo representa a versão do gene, cada vez que algum conteúdo do gene sofre modificações este atributo é incrementado¹. Deste modo é possível perceber se a informação relativa ao gene foi obtida totalmente no mesmo estudo, ou se envolveu vários estudos.

¹Nem todas as alterações obrigam a uma alteração na versão do gene. Podemos verificar o que determina essa atualização da versão neste documento http://www.ensembl.org/info/genome/stable_ids/index.html

- ***strand***

Este atributo representa a direção da cadeia a que pertence o gene.

Pode ter o valor -1 ou 1.

- ***transcripts***

Um transcrito corresponde a uma sequência de nucleótidos resultantes da transcrição de DNA para mRNA.

Um gene pode ter vários transcritos, este atributo corresponde a esse valor, e assim sendo não tem limite.

4.3 Caso de Estudo 1

4.3.1 Especificação do caso

Neste caso de estudo foi estudou-se um conjunto de 114 identificadores de genes reais.

Inicialmente efetuou-se a pesquisa dos genes na Plataforma WEB, seguindo-se de uma análise através do algoritmo *Simple K-means* e de uma análise através do algoritmo *MDBC*.

4.3.2 Dados analisados

Os dados analisados correspondem a 114 identificadores do repositório *Ensembl*, todos pertencentes à família *homeobox* e ao organismo *hsa - homo sapiens*, correspondente aos humanos. A lista dos genes pode ser consultada no anexo A.

4.3.3 Metodologia

A pesquisa de genes foi efetuada nos três repositórios, *Ensembl*, *GenBank* e *Kegg*, em simultâneo de modo a obter o máximo de informação possível.

Após a pesquisa foi efetuado o download dos ficheiros com o formato de texto, *arff* e *prolog*, sendo que o ficheiro *arff* foi posteriormente utilizado para proceder à análise dos dados.

De seguida navegou-se até à página referente ao método *Simple K-means* e definiu-se as opções que podemos observar na tabela 4.3. Nesta tabela descrevemos sucintamente também cada argumento deste método.

Após selecionar as opções pretendidas efetuou-se o *upload* do ficheiro *arff* e visualizou-se os resultados, procedendo também ao *download* dos mesmos.

Fez-se também uma análise através do algoritmo *Make Density Based Clusterer*. Para isso navegou-se até à página deste método e selecionou-se as opções visíveis na tabela 4.4. Nesta tabela apresentamos uma sucinta explicação para cada argumento deste método.

Resultados e Casos de Estudo

Tabela 4.3: Descrição das opções do algoritmo *Simple K-means*

Option	Description	Value	Default
Nr of Clusters	Number of clusters	3	2
Init method	Initialization method to use	canopy	random
Distance function	Distance function to use	ManhattanDistance	EuclideanDistance
Use Canopies	Use canopies to reduce the number of distance calculations.	Enabled	Disabled
max candidates	Maximum number of candidate canopies to retain in memory at any one time when using canopy clustering. T2 distance plus, data characteristics, will determine how many candidate canopies are formed before periodic and final pruning are performed, which might result in excess memory consumption. This setting avoids large numbers of candidate canopies consuming memory.	100	100
periodic pruning	How often to prune low density canopies when using canopy clustering.	10 000	10 000
min density	Minimum canopy density, when using canopy clustering, below which a canopy will be pruned during periodic pruning.	2	2
t2	The T2 distance to use when using canopy clustering. Values lower than 0 indicate that a heuristic based on attribute std. deviation should be used to set this.	-1.0	-1.0
t1	The T1 distance to use when using canopy clustering. A value lower than 0 is taken as a positive multiplier for T2.	-1.5	-1.5
Capabilities	Do not check capabilities	Disabled	Disabled
M	Don't replace missing values with mean/mode.	Disabled	Disabled
fast	Enables faster distance calculations, using cut-off values. Disables the calculation/output of squared errors/distances.	Disabled	Disabled
num slots	Number of execution slots.	1	1 (no parallelism)
Max iterations	Maximum number of iterations	10	10
S	Random number seed.	10	10

Tabela 4.4: Descrição das opções do algoritmo *MDBC*

Option	Description	Value	Default
Min std dev	Minimum allowable standard deviation for normal density computation	0.000001	1e-6
Nr of clusters	Number of clusters	3	2
S	Random number seed	10	10
Replace missing values	Replace missing values with mean/mode	Disabled	Disabled

4.3.4 Resultados

Apresentamos agora os resultados gerados pela Plataforma WEB.

O documento *arff* obtido da Plataforma, resultado da pesquisa dos genes pode ser consultado no anexo A.

Na figuras 4.1 e 4.2 mostramos os resultados da aplicação do algoritmo *Simple K-means*.

Os resultados complementares, como a visualização de todos os elementos de cada *cluster*, assim como os seus atributos podem ser consultados no anexo A.

Após visualizarmos todos os resultados podemos retirar algumas conclusões dos *clusters* resultantes do algoritmo *Simple K-means* aplicado.

O **cluster 0**, que contém 55 instâncias, 48% do total tem as seguintes características:

- atributo *strand* = 1 em todas as instâncias;

Resultados e Casos de Estudo

Number of iterations: 4
Instances: 114
Within cluster sum of squared errors: -1
Initial starting points (random):

- **Cluster 0:** 0 , 0 , 11.152851 , 508920.245262 , 2104.232143 , 318.714286 , 1185.678571 , 12.055556 , 70318629.660714 , 7.678571 , 70346845.821429 , 0.714286 , 3.964286 , {56} <0 , 1 , 2>
- **Cluster 1:** 0 , 0 , 9.288141 , 529323.941043 , 2187.962963 , 344.666667 , 1111.62963 , 10.482853 , 66027407.703704 , 10.851852 , 66061219.962963 , -1 , 5.592593 , {27} <0 , 1 , 2>
- **Cluster 2:** 0 , 0 , 11.325688 , 198197.5 , 3829.571429 , 299.214286 , 900.642857 , 51.791005 , 67228380.214286 , 7.214286 , 67239223.285714 , -1 , 3.214286 , {14} <0 , 1 , 2>

Clustered Instances

- **Cluster 0:** 55 - 48 %
- **Cluster 1:** 36 - 32 %
- **Cluster 2:** 23 - 20 %

Figura 4.1: Resultados do algoritmo Simple K-means

Final Cluster Centroids		Cluster #		
Attribute	Full Data	0	1	2
status	0	0	0	0
currentID	0	0	0	0
chromosome	10	8	10	12
mim	600036.5	600647	604578	142970
geneWeight	1913	1805	1574.5	3077
aaseqSize	315	291	346	269
ntseqSize	990	987	1050	810
seq_region_name	10	8	11.5	12
start	58232439	76681219	56998874	48621159
version	8	9	8	7
end	58250303.5	76867285	57018500	48626356
strand	1	1	-1	-1
transcript	2	2	3	2

Figura 4.2: Caracterização dos *clusters* resultantes do algoritmo Simple K-means

Resultados e Casos de Estudo

- atributo *source* = *ensembl_havana* 84 % das instâncias;

O **cluster 1**, que contém 36 instâncias, 32% do total tem as seguintes características:

- atributo *strand* = **-1** em todas as instâncias;
- atributo *mim* $\geq$ **600 000** em 81 % das instâncias, sendo que 17 % das instâncias não estão classificadas neste atributo;

O **cluster 2**, que contém 23 instâncias, 20% do total tem as seguintes características:

- atributo *source* = *ensembl_havana* em todas as instâncias;
- atributo $100\,000 \leq \textit{mim} \leq 200\,000$ em 91% das instâncias;
- atributo *version* $\leq$ **20** em 91% das instâncias.

Nas figuras 4.3 e 4.4 mostramos os resultados do algoritmo *Make Density Based Clusterer*.

```
Number of iterations: 11
Instances: 114
Within cluster sum of squared errors: 51.875635143086996
Initial starting points (random):
  • Cluster 0: 0 , 0 , 1 , 606066 , 1259 , 397 , 1194 , 1 , 197911902 , 15 , 197935478 , 1 , 8
  • Cluster 1: 0 , 0 , 19 , 438617.081633 , 943 , 375 , 1128 , 19 , 47403124 , 17 , 47419523 , -1 , 14
  • Cluster 2: 0 , 0 , 10 , 607158 , 1913 , 258 , 777 , 13 , 83907301 , 2 , 83907781 , -1 , 1
Clustered Instances
  • Cluster 0: 50 - 44 %
  • Cluster 1: 24 - 21 %
  • Cluster 2: 40 - 35 %
Log likelihood: -66.75763
```

Figura 4.3: Resultados do algoritmo *Make Density Based Clusterer*

Resultados e Casos de Estudo

Attribute	Full Data	Cluster #		
		0	1	2
status	0	0	0	0
currentID	0	0	0	0
chromosome	9.7798	8.5289	11.2936	10.6286
mim	438617.0816	476792.1077	161877.12	577389.8711
geneWeight	2504.3246	2179.3333	2861.76	2750.4286
aaseqSize	379.9386	367.2778	320.76	441.7429
ntseqSize	1271.7719	1339.7593	965.28	1385.8
seq_region_name	33.037	46.7058	14.0844	25.4857
start	79174515.5351	96344273.3519	55246955.44	69775146.4
version	8.7632	8.6852	7.8	9.5714
end	79202730.0088	96375857.8889	55267785.28	69803436.0857
strand	0.1053	1	-0.28	-1
transcript	4.614	4.3519	3.2	6.0286

Figura 4.4: Caracterização dos *clusters* resultantes do algoritmo *Make Density Based Clusterer*

Os restantes resultados, como a visualização de todos os elementos de cada *cluster*, ou a média e desvio padrão de cada atributo em cada *cluster*, podem ser consultados no anexo A.

Após visualizarmos todos os resultados podemos retirar algumas conclusões dos *clusters* resultantes do algoritmo aplicado.

O **cluster 0**, que contém 50 instâncias, 44% do total tem as seguintes características:

- atributo ***strand* = 1** em 90 % das instâncias;
- atributo ***version* ≤ 10** em 72 % das instâncias;
- atributo ***gene_weight* ≤ 2 000** em 76 % das instâncias;
- atributo ***transcript* ≤ 4** em 78 % das instâncias.

O **cluster 1**, que contém 24 instâncias, 21% do total tem as seguintes características:

- atributo **$100\ 000 \leq \text{mim} \leq 300\ 000$** em 96 % das instâncias;
- atributo ***transcript* ≤ 4** em 83 % das instâncias;
- atributo **$2\ 000 \leq \text{gene_weight} \leq 7\ 000$** em 71 % das instâncias;
- atributo ***version* ≤ 10** em todas as instâncias;

O **cluster 2**, que contém 40 instâncias, 35% do total tem as seguintes características:

- atributo ***strand* = -1** em 80 % das instâncias;

- atributo *mim* $\geq 600\,000$ em 88 % das instâncias;
- atributo *version* ≥ 7 em 68 % das instâncias;

4.4 Caso de Estudo 2

4.4.1 Especificação do caso

Neste caso de estudo foi estudou-se um conjunto de 8548 genes reais.

Inicialmente efetuou-se a pesquisa de genes na Plataforma WEB, seguindo-se de uma análise através do algoritmo *Farthest-First*.

4.4.2 Dados analisados

Os dados analisados correspondem a 8548 nomes de genes reais, todos pertencentes ao organismo *hsa - homo sapiens*, correspondente aos humanos.

4.4.3 Metodologia

A pesquisa foi efetuada nos três repositórios em simultâneo de modo a obter o máximo de informação possível.

Após a pesquisa foi efetuado o download do ficheiro com o formato *arff* para posteriormente proceder à análise dos dados.

De seguida navegou-se até à página referente ao método de *clustering Farthest-First* e definiu-se as seguintes opções:

Tabela 4.5: Opções escolhidas *Make Density Based Clusterer*

Option	Value
Nr of Clusters	5
Random nr seed	10 (default)

4.4.4 Resultados

Na figura 4.5 mostramos os resultados do algoritmo *Farthest-First*.

Após visualizarmos todos os resultados podemos retirar algumas conclusões dos *clusters* resultantes do algoritmo aplicado.

O **cluster 0**, que contém 4516 instâncias, 53% do total tem as seguintes características:

- atributo *strand* = -1 em 83 % das instâncias;

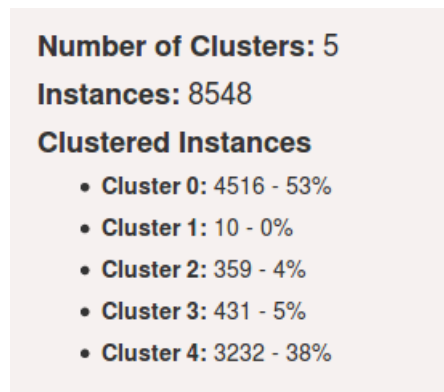


Figura 4.5: Resultados do algoritmo *Farthest-First*

- atributo *source* = *ensembl_havana* em 99 % das instâncias;
- atributo *geneWeight* $\leq 10\ 000$ em 91 % das instâncias.

O **cluster 1**, apenas contém 10 instâncias, correspondente a menos de 1% do total e sendo assim é considerado irrelevante para o estudo.

O **cluster 2**, que contém 359 instâncias, 4% do total tem as seguintes características:

- atributo *strand* = 1 em todas as instâncias;
- atributo $100\ 000 \leq \textit{mim} \leq 300\ 000$ em 93 % das instâncias;
- atributo *version* ≥ 10 em 86 % das instâncias.

O **cluster 3**, que contém 431 instâncias, 5% do total tem as seguintes características:

- atributo *strand* = -1 em 97 % das instâncias;
- atributo *seq_region_name* ≤ 10 em 90 % das instâncias;
- atributo *version* ≥ 10 em 78 % das instâncias.

O **cluster 4**, que contém 3232 instâncias, 38% do total tem as seguintes características:

- atributo *strand* = 1 em 98 % das instâncias;
- atributo *version* ≥ 10 em 72 % das instâncias;
- atributo *transcript* ≤ 10 em 71 % das instâncias.

4.5 Caso de Estudo 3

4.5.1 Especificação do caso

Neste caso de estudo analisou-se um conjunto de 948 genes reais.

Inicialmente efetuou-se a pesquisa dos genes na Plataforma WEB, seguindo-se de uma análise através do algoritmo DBSCAN.

4.5.2 Dados analisados

Os dados analisados correspondem a 948 genes mitocondriais.

4.5.3 Metodologia

Foi efetuada a pesquisa dos genes nos três repositórios em simultâneo, *Ensembl*, *GenBank* e *Kegg*, através do *upload* de um ficheiro com os nomes de todos os genes.

De seguida descarregou-se da Plataforma WEB o ficheiro *arff* contendo a informação relativa aos genes, no formato apropriado para serem aplicadas técnicas de *data mining* também na plataforma desenvolvida.

Navegou-se até à página referente ao método *Density-based spatial clustering of applications with noise (DBSCAN)* e escolheu-se as opções visíveis na tabela 4.6.

Tabela 4.6: Opções escolhidas *DBSCAN*

Option	Value
Epsilon	0.45
MinPoints	6 (default)
Distance type	EuclideanDataObject (default)

4.5.4 Resultados

Na figura 4.6 mostramos o sumário dos resultados da aplicação do algoritmo *DBSCAN*.

Para além deste sumário dos resultados, na Plataforma WEB é possível observar a informação relativa a cada instância, seguida do *cluster* atribuído. A Plataforma permite também visualizar as instâncias divididas pelos *clusters*.

Após analisar essa informação é possível retirar conclusões relativas a cada *cluster*.

O **cluster 0** contém 401 instâncias, 43 % do total, e apresenta as seguintes características:

- atributo **strand = 1** em todas as instâncias;
- atributo **mim** $\geq$ **600 000** em 88 % das instâncias, sendo que as restantes instâncias não têm este atributo definido.

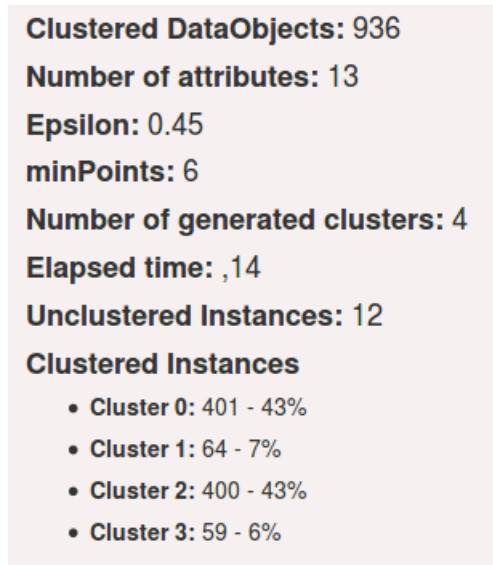


Figura 4.6: Resultados do algoritmo *DBSCAN*

O **cluster 1** contém 64 instâncias, 7 % do total, e apresenta as seguintes características:

- atributo **strand = 1** em todas as instâncias;
- atributo $100\ 000 \leq \text{mim} < 300\ 000$ em 78 % das instâncias;
- atributo $300\ 000 \leq \text{mim} < 400\ 000$ em 22 % das instâncias.

O **cluster 2** abrange 400 instâncias, 43 % do total, e apresenta as seguintes características:

- atributo **strand = -1** em todas as instâncias;
- atributo $\text{mim} \geq 600\ 000$ em 87 % das instâncias, sendo que as restantes instâncias não têm este atributo definido.

O **cluster 3** integra 59 instâncias, 6 % do total, e apresenta as seguintes características:

- atributo **strand = -1** em todas as instâncias;
- atributo $100\ 000 \leq \text{mim} < 300\ 000$ em 75 % das instâncias;
- atributo $300\ 000 \leq \text{mim} < 400\ 000$ em 25 % das instâncias.

4.6 Avaliação e comparação da Plataforma WEB

Após o desenvolvimento e utilização da Plataforma WEB desenvolvida, é necessário proceder à avaliação da sua metodologia e resultados, comparando também com as metodologias e resultados existentes em outras ferramentas existentes.

4.6.1 Procedimentos

Nesta secção demonstramos quais os procedimentos necessários de efetuar para obter informação relativa a um conjunto de genes nos três repositórios, *Ensembl*, *GenBank* e *Kegg* e aplicar algoritmos de *data mining* a essa informação.

4.6.1.1 Manualmente

No diagrama 4.7 mostramos todos os passos para alcançar o objetivo final, de uma forma manual.

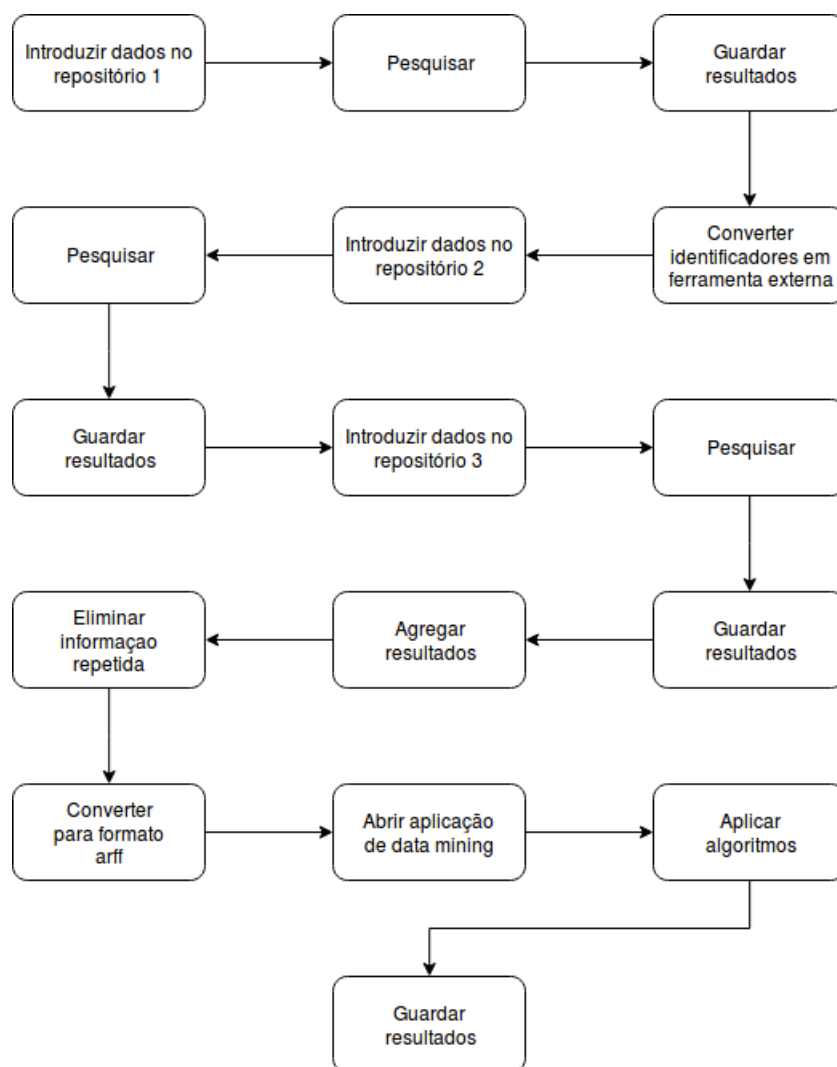


Figura 4.7: Procedimentos efetuados manualmente

Como podemos verificar são várias as etapas que teremos que efetuar, algumas delas com alguma complexidade, requerendo algum conhecimento informático.

Para obter a informação completa relativa aos genes é necessário efetuar a conversão dos identificadores, introduzir os respetivos identificadores em cada repositório, guardar os resultados obtidos em cada base de dados e no final agregar todos os resultados, filtrando a informação repetida. A etapa seguinte é a conversão dos resultados para o formato *arff* que pode ser feita manualmente ou através de ferramentas externas. Após este longo processo é ainda necessário instalar e abrir a aplicação de *data mining* e aplicar os algoritmos desejados. Nesta etapa poderá ser necessário filtrar o ficheiro *arff* para cada algoritmo, eliminando por exemplo atributos *string*, no entanto algumas ferramentas de *data mining* já o fazem automaticamente.

4.6.1.2 Na Plataforma WEB

No diagrama 4.8 mostramos todas as etapas para atingir o objetivo final, utilizando a Plataforma WEB desenvolvida.

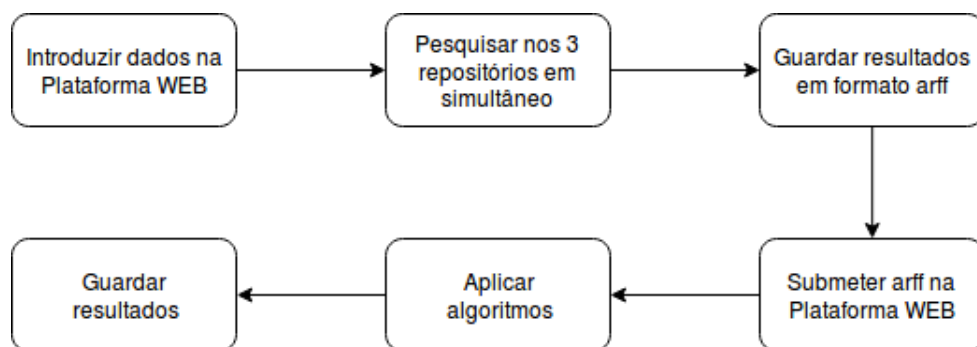


Figura 4.8: Procedimentos efetuados na Plataforma WEB

Facilmente identificamos a simplicidade ao efetuar este processo na Plataforma WEB.

É apenas necessário introduzir os identificadores na Plataforma WEB, poderá ser através de um ficheiro contendo os identificadores, ou introduzindo-os manualmente na plataforma, pesquisar, sendo que a plataforma pesquisa em simultâneo nos três repositórios, sem necessidade de efetuar conversão de identificadores manualmente. De seguida podemos fazer o *download* da informação diretamente num ficheiro *arff*, já com a informação filtrada, sendo que o passo seguinte consiste em submeter esse ficheiro na plataforma, escolher o algoritmo desejado e aplicá-lo. De seguida serão apresentados os resultados, sendo possíveis transferi-los.

4.6.2 Eficiência

Um dos principais objetivos do desenvolvimento desta Plataforma WEB passa obrigatoriamente por ser eficiente. Para considerarmos a plataforma eficiente esta tem que realizar as tarefas propostas num tempo inferior ao tempo de realização das mesmas propostas manualmente.

Assim utilizaremos o caso de estudo 4.3, constituído por 114 genes para pesquisar nos três repositórios, *Ensembl*, *GenBank* e *Kegg*. De seguida serão aplicados os algoritmos *Simple K-means* e *Make Density Based Clusterer*.

A informação existente relativa a cada gene para efetuar a pesquisa é o seu identificador na base de dados *Ensembl* e deste modo teremos que proceder à conversão para as restantes bases de dados.

Ao efetuar este procedimento manualmente demoramos cerca de 36 segundos para cada base de dados, totalizando 72 segundos, ou seja 1 minuto e 12 segundos.

Para efetuar manualmente a pesquisa de cada gene no repositório *Ensembl* e exportar o resultado demora cerca de 48 segundos, no repositório *GenBank* demora 34 segundos e no repositório *Kegg* cerca de 81 segundos, ou seja, 1 minuto e 21 segundos. Deste modo, para cada gene são necessários 187 segundos, 3 minutos e 7 segundos.

Para obter informação dos 114 genes manualmente precisaríamos de cerca de 21 318 segundos, equivalente a quase 6 horas. Juntando a este tempo a agregação dos resultados, a filtragem destes e a aplicação dos algoritmos, o processo demoraria pelo menos 8 horas a ser realizado.

Na Plataforma WEB todo o processo demorou 2 minutos e 11 segundos. Este resultado deve-se principalmente ao facto do utilizador apenas ter de inserir os dados uma vez, sendo responsabilidade da Plataforma WEB a conversão e pesquisa nos três repositórios, através das APIs disponibilizadas, que permitem uma pesquisa de centenas de genes em simultâneo. Também o processo automatizado de agregação e filtragem dos resultados, disponibilizando o conteúdo diretamente no formato *arff* contribui para o valor temporal ser tão reduzido. Por fim, o facto do utilizador poder facilmente introduzir as opções desejadas em cada algoritmo, através de uma interface desenvolvida especificamente para facilitar e acelerar este processo permite realizar todo o processo com um valor temporal bastante reduzido.

Tabela 4.7: Comparação entre eficiência na Plataforma WEB e manualmente

Conjunto de dados	Origem Identificador	Pesquisa e análise na Plataforma WEB	Pesquisa e análise manualmente
114	Ensembl	2 min 11 segundos	≥ 8 horas

4.7 Conclusões

Após a realização dos casos de estudos 4.3, 4.4 e 4.5 e após efetuar a comparação entre os processos realizados manualmente ou através da Plataforma WEB, verificámos que a simplicidade de processos e o tempo de execução dos mesmos é bastante inferior na plataforma desenvolvida do que manualmente.

Para além da simplicidade e rapidez dos processos, percebemos também que a utilização da plataforma desenvolvida não implica conhecimento prévio em informática, não sendo necessário por exemplo instalar aplicações, saber como funcionam as diferentes ferramentas que permitem aplicar técnicas de *data mining*, ou até utilizar a linha de comandos, como acontece em algumas ferramentas de *data mining*.

Capítulo 5

Conclusões e Trabalho Futuro

Neste capítulo final são apresentadas as conclusões da dissertação, é efetuada uma comparação entre os objetivos propostos e os objetivos alcançados e apresentamos também as previsões para o trabalho futuro para este projeto.

5.1 Conclusões

Este projeto tinha como principal objetivo facilitar o trabalho dos especialistas na investigação biológica.

Para isso comprometemo-nos a desenvolver uma plataforma que resolvesse dois problemas que os investigadores biológicos enfrentam no seu dia a dia, a recolha de informação relativa a produtos génicos e a análise dessa informação.

Toda a plataforma foi pensada e desenvolvida de forma a ser utilizada por especialistas biológicos com pouco conhecimento informático, sendo também um dos principais objetivos deste projeto a interface amigável da plataforma desenvolvida, quer a nível da sua utilização quer a nível da gestão da sua base de dados.

Através da utilização da plataforma desenvolvida no âmbito desta dissertação o especialista de investigação biológica ultrapassa tarefas como a conversão dos identificadores dos genes, a pesquisa em vários repositórios individualmente, a agregação dos resultados das pesquisas nos diferentes repositórios, a conversão dos resultados para os formatos funcionais nas ferramentas de *data mining*, a filtragem da informação repetida, a instalação e aprendizagem de utilização de ferramentas de *data mining*, entre outros obstáculos com que se confronta no seu dia a dia.

Findado o desenvolvimento deste projeto consideramos que os principais objetivos foram alcançados. A Plataforma WEB torna a pesquisa de genes uma tarefa mais rápida e menos complexa, permitindo a pesquisa em três repositórios em simultâneo, *Ensembl*, *GenBank* e *Kegg*. Permite também a recolha da informação pesquisada em vários formatos, para diferentes fins. Após essa recolha de informação permite também de forma acessível aplicar algoritmos de classificação e *clustering* a essa mesma informação, facilitando a obtenção de conclusões por parte dos especialistas.

Após a realização de testes e vários casos de estudo, é notória a diferença de eficiência e simplicidade de processos na pesquisa de genes e na análise dos mesmos, permitindo reduções de tempo na ordem dos 200 %, sem implicar algum tipo de custo financeiro.

5.2 Trabalho futuro

Apesar de considerarmos que os objetivos propostos foram atingidos, ao longo do desenvolvimento da Plataforma WEB surgiram aspetos que consideramos que podem e devem ser melhorados no futuro.

5.2.1 Adicionar novos repositórios

Atualmente a Plataforma WEB permite a pesquisa de genes apenas em três repositórios de genes, *Ensembl*, *GenBank* e *Kegg*. No futuro esperamos que esta pesquisa seja alargada a mais repositórios, procurando sempre obter o máximo de informação possível relativa a cada gene.

5.2.2 Adicionar outros métodos de *data mining*

A plataforma desenvolvida integra nove algoritmos referentes a dois tipos de *data mining*. Três algoritmos de classificação e seis algoritmos de *clustering*. Espera-se que no futuro sejam adicionados mais métodos, permitindo uma escolha mais alargada por parte do especialista, facilitando o seu trabalho e permitindo-o obter mais e melhores conclusões relativas a um conjunto de genes.

5.2.3 Adicionar novos formatos para descarregar informação

Após a pesquisa de genes, a recolha de informação é atualmente suportada em três formatos distintos, texto, *arff* e *prolog*. Consideramos que esta recolha deve ser expandida a outros formatos.

Também a recolha de resultados após a aplicação de algoritmos de *data mining*, neste momento apenas em formato de texto e CSV pode e deve ser alargada a outros formatos.

Anexo A

Caso de Estudo 1

Neste anexo apresentamos os dados analisados no caso de estudo 4.3, assim como os resultados completos, possíveis de descarregar na Plataforma WEB.

A.1 Dados analisados

O documento A.1 contém os identificadores para pesquisa na Plataforma WEB, sendo através do *upload* deste documento que se inicia a pesquisa. Estes identificadores são referentes ao repositório *Ensembl*.

```
ENSG00000227059,ENSG00000178928,ENSG00000230465,ENSG00000215037,ENSG00000253569,
  ENSG00000259021,ENSG00000265293,ENSG00000179528,ENSG00000150051,ENSG00000257364
  ,ENSG00000271672,ENSG00000151650,ENSG00000237872,ENSG00000259849,
  ENSG00000101544,ENSG00000128713,ENSG00000179277,ENSG00000236380,ENSG00000254105
  ,ENSG00000264743,ENSG00000123407,ENSG00000180806,ENSG00000119614,
  ENSG00000131721,ENSG00000180438,ENSG00000091656,ENSG00000147421,ENSG00000162624
  ,ENSG00000160199,ENSG00000105419,ENSG00000170689,ENSG00000188816,
  ENSG00000177426,ENSG00000165495,ENSG00000106038,ENSG00000107807,ENSG00000136367
  ,ENSG00000170178,ENSG00000215612,ENSG00000205857,ENSG00000186103,
  ENSG00000213921,ENSG00000174306,ENSG00000143995,ENSG00000130675,ENSG00000143355
  ,ENSG00000169554,ENSG00000120068,ENSG00000121454,ENSG00000135638,
  ENSG00000165259,ENSG00000177045,ENSG00000180613,ENSG00000197757,ENSG00000235602
  ,ENSG00000120093,ENSG00000037965,ENSG00000169840,ENSG00000172789,
  ENSG00000164900,ENSG00000167157,ENSG00000188620,ENSG00000101883,ENSG00000106006
  ,ENSG00000006377,ENSG00000128645,ENSG00000170166,ENSG00000164438,
  ENSG00000165606,ENSG00000198353,ENSG00000236375,ENSG00000106852,ENSG00000134138
  ,ENSG00000119608,ENSG00000108511,ENSG00000128652,ENSG00000175879,
  ENSG00000134438,ENSG00000163666,ENSG00000197576,ENSG00000168779,ENSG00000171476
  ,ENSG00000105997,ENSG00000185610,ENSG00000120075,ENSG00000106031,
  ENSG00000106410,ENSG00000043039,ENSG00000123364,ENSG00000123576,ENSG00000128709
  ,ENSG00000175329,ENSG00000159387,ENSG00000215271,ENSG00000280635,
  ENSG00000115297,ENSG00000115507,ENSG00000173976,ENSG00000136630,ENSG00000170549
  ,ENSG00000170561,ENSG00000188909,ENSG00000153779,ENSG00000156150,
```

```
ENSG00000280774,ENSG00000116132,ENSG00000118707,ENSG00000144355,ENSG00000165588
,ENSG00000122592,ENSG00000128710,ENSG00000173917,ENSG00000165462,
ENSG00000179981
```

Listing A.1: Documento contendo os identificadores enviados para a Plataforma WEB

A.2 Resultados

Nesta secção encontram-se os ficheiros descarregados da Plataforma WEB.

Na secção A.2.1 demonstrámos os ficheiros *txt*, *arff* e *prolog* resultantes da pesquisa de informação relativa aos genes introduzidos. O ficheiro *arff* foi utilizado posteriormente para aplicar as diferentes técnicas de *data mining*.

Nas secção A.2.2 apresentamos os ficheiros descarregados da Plataforma WEB após a aplicação das técnicas de *data mining*. Estes ficheiros encontram-se sumarizados, não apresentando os dados relativos a todos os genes, apenas devido à extensão desses documentos. Dado que estes ficheiros englobam, entre outra informação, os dados relativos a cada *cluster* no formato de tabela, mas que derivado da extensão de cada linha ficam desformatados ao englobar neste documento, apresentamos também, em complemento a estes ficheiros, as tabelas relativas a cada *cluster*, de modo a facilitar a leitura dessa informação.

A.2.1 Resultados da pesquisa de informação

A.2.1.1 Formato de texto

O documento demonstrado encontra-se reduzido, derivado da sua longa extensão, englobando apenas a informação relativa ao gene *NOBOX*. O documento completo engloba a informação relativa aos restantes 113 genes pesquisados, no mesmo formato do gene apresentado. É possível identificar o fim da informação relativa a um gene através de uma linha com os símbolos `\\`.

```
NOBOX
Ensembl ID: ENSG00000106410
GenBank ID: 135935
Kegg ID: 135935
Assembly Name: GRCh38
Display Name: NOBOX
Organism: hsa Homo sapiens (human)
Description: NOBOX oogenesis homeobox
OtherAliases: OG-2, OG2, OG2X, POF5, TCAG_12042
OtherDesignations: homeobox protein NOBOX|newborn ovary homeobox-encoding
Status: 0
CurrentID: 0
Chromosome: 7
Mim: 610934
GenomicInfoTypeChrLoc: 7
```

Caso de Estudio 1

GenomicInfoTypeChrAccVer: NC_000007.14
GenomicInfoTypeStart: 144410437
GenomicInfoTypeStop: 144396804
GenomicInfoTypeExonCount: 10
GeneWeight: 2908
Summary: This homeobox gene encodes a transcription factor that is thought to play a role in oogenesis. In mice, it is essential for folliculogenesis and regulation of oocyte-specific genes. Defects in this gene result in premature ovarian failure type 5.[provided by RefSeq, May 2011]
ChrSort: 7
ChrStart: 144396804
Position: 7q35
Motif: Pfam: Homeobox Eaf7
DBLinks: NCBI-ProteinID: NP_001073882
NCBI-GeneID: 135935
OMIM: 610934
HGNC: 22448
Ensembl: ENSG00000106410
Vega: OTTHUMG00000158051
Pharos: O60393 (Tbio)
UniProt: O60393
Aaseq Size: 691
Aaseq: MALLLTLTSPDLEGTWDTRDKDGFKAQEGPPLAVPEFPVCGLYRIYGVCGSFSSFFIIRC
SLCALETLKSPQHDPLEIPEQSLKLIPLVSGKRELTRGQKAGEKPLAAGPGEEELLRGS
PHAQDTQSEELPPSCTISGEKKPPAVSGEATGADAGRLCPPRSRAPHKDRTLARSRPQT
QGEDCSLPVGEVKIGKRSYSPAPGKQKKPNAMGLAPTSSPGAPNSARATHNPVPCGSGRG
PCHLANLLSTLAQSNQNRDQKGPPEVTCQIRKKTRTLYRSDQLEELEKIFQEDHYPDS
KRREIAQTVGVTPQRIIMVKGAGSLVAGWSGGGPTIETLELQSERSAVAVWVFQNRRAKWR
KMEKLNKESKDNPAAPGPASSQCSSAAEILPAVPMPEKPDFFPQESPLDTFPEPPMLLT
SDQTLAPTQPSGAQRVVTPLFSPPPVRADLPFPLGPVHTPQLMPLLMVDVAGSDSSHK
DGPCGSGWTSITLPPPCSYLEELPQDYQQSNQPGPFQFSQAPQPLFQSPQPKLPYLPT
FFFSMPSSLTLPPEDSLFMFPCGPGSGGTSQGYCPGASSGQILMQPPAGNIGTASWSDPC
LPELFPFPGPFQALGHPPGGDGYFPDLFPTPCPQALGRQPSSALSWMPEGARPGTGPLL
SKAKEEPPAASLDQPSALEEARGDDKNSHVP
Ntseq Size: 2076
Ntseq: atggctctccttttgacactaacatcaccagacctggagggtacctgggacaccagagac
aaggatggcttcaaagcccaggagggcgccctggctgtacctgaatttcctgtgtgt
ggactgtaccggatctacggagtctgtggctctttcagctccttcttcatcatccggtgc
agcctttgtgctctggagaccctcaaatacaccacacatgatcccttagagatacctgaa
cagtcctcaaactcataccctgggtgtctgggaaaagggaactcacaaggggccagaaa
gctggagagaagccctggctgcaggacccggggaggaggaactgctccggggctcagcc
cctcatgctcaggacactcagagtgggaactgccaccctcctgcaccatctcaggagag
aagaagccgccagcagtctctggagaagccacggggctgatgctgggagactgtgccc
ccccccgctccagggtccccacaaagacagaactctagcccgctccaggccccagact
cagggggaagattgttcctcccagtgaggaggtgaagataggaaagaggtcctattct
ccagccccgggaagcagaaaaagcctaatagccatgggtctggcccaacatcatctccg
ggtgccctaactcagccgtgccacacacacacccagtgccctgtgggtcaggccgggg
cctgccacctggccaatctcctcagtacattggcgagagcaacaaaacagagaccac
aagcagggggcccccggaagtgcctgccaaattaggaaaaagacacgaaccctataccgc

Caso de Estudo 1

```
t cagatcagctggaggagctagagaagatattccaagaagaccactatcctgacagtgat
aaacgccgagagattgccagacggtgggggtgacccccagcgcatcatggtaaagggg
gccggctcactggtggcaggggtggagtggcggagggcccaccattgaaacactcgaattg
cagagtgagcgtcagcggtagcctgggtgtggttccagaatcgccgggccaagtggcga
aaaatggagaaaactgaatgggaaagaaagcaaggacaatcctgcagcccctggccctgcc
agcagtcaatgcagctctgcagctgagatcctacctgctgtgcccatggagccaaagcct
gaccctttccctcaggagtcctcttgatacctttccagagcccccatgctgctgact
tctgaccagactttggccccacccaaccagtgaggggtgctcagaggggtggtgaccccc
ccactcttcagccccccacctgtgcgaagggccgatcttccctttcccccctggccctgtc
cacaccccccaactgatgccactgctgatggatgttgctggcagtgacagcagccacaag
gacggccccctgtgggtcctgggggacaagcatcacctgccacccccctgttcataattg
gaggagctggagccccaggattaccaacagagcaaccagccaggacccttccagttctcc
caggctccacagcccccgcttttccagtcctcctcagcccaagttgccctacctccccact
ttcccttctccatgccagttcactgacgcttccaccgccgaagactctctctttatg
tttcctgtggccccagcgggggacatcgagggtattgccaggtgcctcctcagga
cagatcctgatgaaccacctgctgggaatataggtacagcctcctggagtgaccctgt
ttgccagagctgcccttccctgggtccgttctgccacaagctctggggcatccccagga
ggggatggctactttcctgatctatttccaactccctgccccaggtctgggcaggcag
ccttcgtcagctctctcatggatgctgaaggggccagaccagggaactgggcccttactc
agcaaggcaaaaggaaccaccagctgcttccctggatcagccctcagcactggaggag
gccagaggggatgacaagaatagccatgtcccctag
```

Disease: DISEASE H00627 Premature ovarian failure

Seq region name: 7

Object type: Gene

Start: 144397240

Source: ensembl_havana

Version: 14

BioType: protein_coding

End: 144410227

Nr of transcripts: 2

\\

A.2.1.2 Formato ARFF

Apresentamos nesta secção o ficheiro *arff* descarregado da Plataforma WEB. Derivado da extensão do mesmo, apresentamos uma versão sumarizada, apenas contendo a informação relativa ao gene *ANHX*. O ficheiro completo engloba a informação dos restantes 113 genes, no mesmo formato apresentado, sendo que cada linha corresponde a um gene. É também este o documento enviado para a Plataforma WEB para se efetuar as análises de *data mining*.

```
@RELATION Ensembl_GenBank_Keggfilename
```

```
@ATTRIBUTE IDensembl string
```

```
@ATTRIBUTE IDgenbank string
```

```
@ATTRIBUTE IDkegg string
```

```
@ATTRIBUTE name string
```

Caso de Estudio 1

```

@ATTRIBUTE status numeric
@ATTRIBUTE currentID numeric
@ATTRIBUTE chromosome numeric
@ATTRIBUTE otherAliases string
@ATTRIBUTE otherDesignations string
@ATTRIBUTE mim numeric
@ATTRIBUTE genomicInfoTypeChrLoc string
@ATTRIBUTE genomicInfoTypeChrAccVer string
@ATTRIBUTE genomicInfoTypeStart string
@ATTRIBUTE genomicInfoTypeStop string
@ATTRIBUTE genomicInfoTypeExonCount string
@ATTRIBUTE geneWeight numeric
@ATTRIBUTE summary string
@ATTRIBUTE organism string
@ATTRIBUTE orthology string
@ATTRIBUTE pathway string
@ATTRIBUTE module string
@ATTRIBUTE brite string
@ATTRIBUTE position string
@ATTRIBUTE motif string
@ATTRIBUTE dblinks string
@ATTRIBUTE structure string
@ATTRIBUTE aaseqSize numeric
@ATTRIBUTE aaseq string
@ATTRIBUTE ntseqSize numeric
@ATTRIBUTE ntseq string
@ATTRIBUTE disease string
@ATTRIBUTE drugTarget string
@ATTRIBUTE assembly_name string
@ATTRIBUTE display_name string
@ATTRIBUTE seq_region_name numeric
@ATTRIBUTE object_type string
@ATTRIBUTE start numeric
@ATTRIBUTE source string
@ATTRIBUTE version numeric
@ATTRIBUTE bio_type string
@ATTRIBUTE end numeric
@ATTRIBUTE strand numeric
@ATTRIBUTE transcript numeric
@DATA
ENSG00000227059,647589,647589,ANH,0,0,12,?,anomaloushomeoboxprotein|hCG2007354
,?,12,NC_000012.12,133236219,133218311,12,430,?,hsaHomoSapiens (human)
,?,?,?,12q24.33,Pfam:HomeoboxHomeobox_KNSIX1_SD,NCBI-ProteinID:
NP_001177983NCBI-GeneID:647589HGNC:40024Ensembl:ENSG00000227059Vega:
OTTHUMG00000167949Pharos:E9PGG2 (Tdark)UniProt:E9PGG2,?,379,
MQSFLTLKEHEDTCAPPAELVTLAGRLCRDFQDDLAQLQPLVLTALDSQLRLHLLDNADVALACARVLDQQ
EQQQAACRLLEGCGVPGGSQELVQLWNDIHYRLVMRRLGVAALTPVQKFRCKRKNPPPSLCPEGLKSRNFP
REVREKLHNFAVGVNTNPSKAERENLAETSLTPEQVYNWFANYRRRQRALPQHMKPAQQATAEDPGARERG
PDLLQPSGNPRVDSGFVDRPQWSEEREKGGPPQSPQTTQGPWEPLALAPDFPADETIVSKPLDVSGHPQSVQL

```

Caso de Estudo 1

```
EEGLGTSSGRTELRVGSFLVTQPPLQAPFILTQSPPELAPAPSAFFPGPVSAMELSQALPSSQVQCSDSQAS
GDAFWGARMLLLEFSGSSSLG, 1140 atgcagagcttcctgactctgctgaaggagcatgaggacacctgtgc
acccccggcgagctggtgaccttgcgggcagactgtgccgggacttccaggatgaccttgcccaactgca
gcctttggtcacagccattctggacagccagctccgcctgcatctcctggacaacgcagatgtggccctggc
gtgcgccctgtcctggaccagcaggagcagcagcaggcggttgcgcctcctggaggggtgccaggtgcc
gggaggcagccaggagtttagtgagctctggaacgacatccactaccgtctggtcatgaggaggctggcggt
ggctgcgtcaccgccgtgcagaagttccgctgcaggaagaggaacccccgccccctccctctgccaga
ggggtgaagagccggaacttccccagagaggttcgtgagaagctgcacaatttcgctgtgggggtgaacac
caaccccgcaaggctgagaggagaacttggcattggagacgagcttgaccttgagcaggtgtacaactg
gtttgccaattaccggcgccgcaaaagagcccttccccagcacatgaagccagccagcaggccacagctga
agaccctggtgcgaggagaggggtcctgacctcctgcagccctcaggcaacccccgtgtgactctgggtt
tgtggacaggcctcagtgtgcaggaacgtgaggaaaaggcctccacagtccccacagaccaccaagg
accatgggagccactggccttagccccggactttccgcagatgagacagtctcaaagccactggatgtcag
cggccatccccagagcgtgcaattggaggagggtctgggcacaagcagtgagcgacagagctacgggtggg
cagcttcctggtgacacagccccactgcaagctcctgaattcatcctcaccagagccctccagagctggc
cccagccccatctgccttccccggcctgtgtctgccatggagctgagccaggccctgccctccagccaggt
gcagtgctctgatagccaggcctctggtgatgccttctggggagccaggatgctccttgagtttccaggag
cagcctgggctga, ?, ?, GRCh38, ANHX, 12, Gene, 133218312, ensembl\_havana, 6,
protein\_coding, 133236095, -1, 2
```

A.2.1.3 Formato *Prolog*

Apresentamos também o ficheiro descarregado da Plataforma WEB no formato *prolog*, também este sumariado. A informação demonstrada é apenas relativa ao gene *NOBOX*, sendo que o ficheiro original contém informação relativa aos 114 genes.

A primeira linha contém informação do gene e nas linhas seguintes a informação dos seus transcritos, um por linha. Este é o formato para todo o documento.

```
NOBOX, ENSG00000106410, 135935, 135935, 0, 0, 7, OG-20G20G2XPOF5TCAG_12042,
homeoboxproteinNOBOX|newbornovaryhomeobox-encoding, 610934, 7, NC_000007
.14, 144410437, 144396804, 10, 2908,
Thishomeoboxgeneencodesatranscriptionfactorthatisthoughttoplayaroleinoogenesis.
Inmiceitisessentialforfolliculogenesisandregulationofoocyte-specificgenes.
Defectsinthisgeneresultinprematureovarianfailuretype5.[providedbyRefSeqMay2011
], hsaHomo sapiens (human), ?, ?, ?, 7q35, Pfam: HomeoboxEaf7, NCBI-GeneID: 135935 NCBI-
ProteinID: NP_001073882 OMIM: 610934 HGNC: 22448 Ensembl: ENSG00000106410 Vega:
OTTHUMG00000158051 Pharos: O60393 (Tbio) UniProt: O60393, ?, 691,
MALLLTLTSPDLEGTWDRDKDGFKAQEGPPLAVPEFPVCGLYRIYGVCGSFSSFFIIRC
SLCAETLTKSPQHDPLEIPEQSLKLIPLVSGKRELTRGQKAGEKPLAAGPGEEELLRGS
PHAQDTQSEELPSPCTISGEKKPPAVSGEATGADAGRLCPPRSRAPHKDRTLARSRPQT
QGEDCSLPVGEVKIGKRSYPAPGKQKKPNAMGLAPTSSPGAPNSARATHNPVPCGSGRG
PCHLANLLSTLAQSNQNRDHKQGPPEVTCQIRKTRTLYRSQLEELEKIFQEDHYPDS
KRREIAQTVGVTPQRIMVKAGSLVAGWSGGPTIETLELQSERSAVAWVWFQNRRAKWR
KMEKLNKESKDNPAAPGPASSQCSAAEILPAVPMEPKPDFFPQESPLDTFPEPMLLT
SDQTLAPTQPSGAQRVVTPLFSPPPVRRADLPFLGPVHTPQLMPLMDVAGSDSSHK
DGPCGSWGTSITLPPPCSYLEELEPQDYQQSNQPGPFQFSQAPQPPLFQSPQPKLPYLPT
```

Caso de Estudio 1

FFFSMPSSLTLPPEDSLFMFPCGPGSGGTSQGYCPGASSGQILMQPPAGNIGTASWSDPC
LPELPFPGPFCPQALGHPPGGDGYFDFLFTPCPQALGRQPSSALSWMPEGARPGTGPLL
SKAKEPPAASLDQPSALEEARGDDKNSHVP, 2076,

atggctctccttttgacactaacatcaccagacctggagggtacctgggacaccagagac
aaggatggcttcaaagcccaggaggggcccctggctgtacctgaatttcctgtgtgt
ggactgtaccgatctacggagtctgtggctctttcagctccttctcatcatccggtgc
agcctttgtgctctggagaccctcaaataccccaacatgatcccttagagataacctgaa
cagtcctcaaactcataccctgggtgtctgggaaaagggaactcacaaggggcccagaaa
gctggagagaagcccctggctgcaggacccggggaggaggaactgctccggggctcagcc
cctcatgctcaggacactcagagtgaggaactgccaccctcctgcacatctcaggagag
aagaagccgcccagcagctctctggagaagccacggggctgatgctgggagactgtgcccg
ccccccgctccagggtccccacaaagacagaactctagcccgctccaggccccagact
cagggggaagattgttccctcccagtgaggaggtgaagataggaagaggtcctattct
ccagccccgggaagcagaaaaagcctaataccatgggtctggccccaacatcatctccg
ggtgccctaaactcagcccggtccacacacaacccagtgccctgtgggtcaggccggggg
ccttgccacctggccaatctcctcagtaattggcgagagcaacaaaaacagagaccac
aagcagggggccccgggaagtgcctgccaataaggaaaaagacgaaccctataccgc
tcagatcagctggaggagctagagaagataattccaagaagaccactatcctgacagtgat
aaacgcgagagattgccagacggtgggggtgacccccagcgcatcatggtaaagggg
gccggctcactggtggcagggtggagtggcgaggggccaccattgaaacactcgaattg
cagagtgaagcgtcagcggtagcctgggtgtggttcagaatcgccgggccaagtggcga
aaaatggagaaactgaatgggaaagaaagcaaggacaatcctgcagcccctggccctgcc
agcagtcaatgcagctctgcagctgagatcctacctgtgtgcccatggagccaaagcct
gaccctttccctcaggagtcccctctggatacctttccagagccccccatgctgctgact
tctgaccagactttggccccacccaacccagtgagggtgctcagagggtggtgaccccc
ccactcttcagccccccacctgtgcgaagggccgatcttctttcccccttggccctgtc
cacaccccccaactgatgccactgctgatggatggtgctggcagtgacagcagccacaag
gacggcccctgtgggtcctgggggacaagcatcacctgccacccccctgttcataatttg
gaggagctggagccccaggattaccaacagagcaaccagccaggacccttcagttctcc
cagggtccacagcccccgcttttccagtcacctcagcccaagttgcctacctccccact
ttcccccttctccatgccagttcactgacgcttcaccgcccgaagactctctctttatg
tttccctgtgccccagcgggggacatcgagggtattgccaggtgectcctcagga
cagatcctgatgcaaccacctgctgggaatataggtacagcctcctggagtacccctgt
ttgccagagctgcccttccctggtccgttctgccacaagctctggggcatccccagga
ggggatggctactttcctgatctatttccaactcctgccccaggtctgtggcaggcag
ccttcgtcagctctctcatggatgcctgaaggggcccagaccagggactgggcccttactc
agcaaggcaaaagaggaaccacagctgcttccctggatcagccctcagcactggaggag
gccagaggggatgacaagaatagccatgtcccttag, DISEASEH00627Prematureovarianfailure, ?,
GRCh38, NOBOX, 7, Gene, 144397240, ensembl_havana, 14, protein_coding
, 144410227, -1, 2

```
Transcript_Ensembl, ENST00000467773, gene_id: ENSG00000106410, logic_name:  
ensembl_havana_transcript, object_type: Transcript, source: ?, parent:  
ENSG00000106410, seq_region_name 7, db_type: core, is_canonical: 1, strand: -1, version  
: 1, species: homo_sapiens, assembly_name: GRCh38, display_name: NOBOX-201, end  
: 144410227, biotype: protein_coding, start: 144397240, nr_of_exons 10  
Transcript_Ensembl, ENST00000483238, gene_id: ENSG00000106410, logic_name: havana,  
object_type: Transcript, source: ?, parent: ENSG00000106410, seq_region_name 7, db_type  
: core, is_canonical: 0, strand: -1, version: 5, species: homo_sapiens, assembly_name:
```

Caso de Estudo 1

```
GRCh38,display_name:NOBOX-202,end:144410227,biotype:protein_coding,start  
:144397240,nr_of_exons10\\
```

A.2.2 Resultados dos algoritmos de *data mining*

A.2.2.1 Resultados do algoritmo Simple K-means

Nesta secção apresentamos os resultados completos relativos à aplicação do algoritmo *Simple K-means*, englobando uma tabela para cada *cluster*, contendo a informação do mesmo.

```
kMeans
```

```
=====
```

```
Number of iterations: 4
```

```
Sum of within cluster distances: 148.67261622812168
```

```
Initial starting points (canopy):
```

```
T2 radius: 1,134
```

```
T1 radius: 1,418
```

```
Cluster 0:
```

```
0,0,11.152851,508920.245262,2104.232143,318.714286,1185.678571,12.055556,  
70318629.660714,7.678571,70346845.821429,0.714286,3.964286,{56} <0,1,2>
```

```
Cluster 1: 0,0,9.288141,529323.941043,2187.962963,344.666667,1111.62963,10.482853,  
66027407.703704,10.851852,66061219.962963,-1,5.592593,{27} <0,1,2>
```

```
Cluster 2: 0,0,11.325688,198197.5,3829.571429,299.214286,900.642857,51.791005,  
67228380.214286,7.214286,67239223.285714,-1,3.214286,{14} <0,1,2>
```

```
Reduced number of distance calculations by using canopies.
```

```
Missing values globally replaced with mean/mode
```

```
Final cluster centroids:
```

Attribute	Cluster#			
	Full Data	0	1	2
	(114.0)	(55.0)	(36.0)	(23.0)
=====				
status	0	0	0	0
currentID	0	0	0	0
chromosome	10	8	10	12
mim	600036.5	600647	604578	142970
geneWeight	1913	1805	1574.5	3077
aaseqSize	315	291	346	269
ntseqSize	990	987	1050	810
seq_region_name	10	8	11.5	12

Caso de Estudio 1

start	58232439	76681219	56998874	48621159
version	8	9	8	7
end	58250303.5	76867285	57018500	48626356
strand	1	1	-1	-1
transcript	2	2	3	2

Clustered Instances

0	55 (48%)
1	36 (32%)
2	23 (20%)

Cluster0

clusterNr	EnsemblID	GenBankID	KeggID	Display_Name	Assembly_Name	Source	status	currentID	chromosome	mim	geneWeight	aaseqSize	ntseqSize	seq_region_name	start	version	end	strand	transcript
0	ENSG00000179277	4213	4213	MEIS3P1	GRCh38	havana	0	0	17	?	307	0	2872	17	15786618	9	15787575	1	1
0	ENSG00000128709	3235	3235	HOXD9	GRCh38	ensembl_havana	0	0	2	142982	3609	352	1059	2	176122720	12	176124937	1	1
0	ENSG00000143355	56956	56956	LHX9	GRCh38	ensembl_havana	0	0	1	606066	1259	397	1194	1	197911902	15	197935478	1	8
0	ENSG00000169840	219409	219409	GSX1	GRCh38	ensembl_havana	0	0	13	616542	884	264	795	13	27792643	4	27794768	1	1
0	ENSG00000147421	79618	79618	HMBX1	GRCh38	ensembl_havana	0	0	8	?	2401	420	1263	8	28890394	17	29064764	1	16
0	ENSG00000165495	63876	63876	PKNOX2	GRCh38	ensembl_havana	0	0	11	613066	1782	472	1419	11	125164687	15	125433389	1	14
0	ENSG00000186103	503582	503582	ARGFX	GRCh38	ensembl_havana	0	0	3	611164	790	315	948	3	121570704	3	121586634	1	1
0	ENSG00000164438	30012	30012	TLX3	GRCh38	ensembl_havana	0	0	5	604640	4534	291	876	5	171309284	5	171312134	1	1
0	ENSG00000175879	3234	3234	HOXD8	GRCh38	ensembl_havana	0	0	2	142985	1838	290	873	2	176129694	8	176132695	1	4
0	ENSG00000118707	60436	60436	TGIF2	GRCh38	ensembl_havana	0	0	20	607294	3091	237	714	20	36573488	9	36593950	1	8
0	ENSG00000237872	645682	645682	POU5F1P4	GRCh38	havana	0	0	1	?	774	0	1083	1	155433178	4	155434262	1	1
0	ENSG00000115297	3196	3196	TLX2	GRCh38	ensembl_havana	0	0	2	604240	1937	284	855	2	74513463	10	74517147	1	4
0	ENSG00000259849	139538	139538	VENTXP1	GRCh38	havana	0	0	?	?	395	0	2716	?	26558337	1	26561052	1	1
0	ENSG00000213921	342900	342900	LEUTX	GRCh38	ensembl_havana	0	0	19	?	472	168	507	19	39776595	7	39786167	1	2
0	ENSG00000006377	1750	1750	DLX6	GRCh38	ensembl_havana	0	0	7	600030	1974	293	882	7	97005548	10	97011039	1	3

Caso de Estudio 1

```

0, ENSG00000101544, 22850, 22850, ADNP2, GRCh38, ensembl_havana, 0, 0, 18, 617422,
    1008, 1131, 3396, 18, 80109031, 8, 80147523, 1, 5,
0, ENSG00000236380, 391518, 391518, VENTXP7, GRCh38, havana, 0, 0, 3, ?, 373, 0,
    983, 3, 21405737, 5, 21406508, 1, 1,
0, ENSG00000043039, 8538, 8538, BARX2, GRCh38, ensembl_havana, 0, 0, 11, 604823,
    1531, 279, 840, 11, 129375940, 6, 129452279, 1, 2,
0, ENSG00000257364, 27287, 27287, VENTX, GRCh38, havana, 0, 0, 10, 607158, 1913,
    258, 777, 12, 74292324, 1, 74293096, 1, 1,
0, ENSG00000271672, 503637, 503637, DUXAP8, GRCh38, havana, 0, 0, 22, ?, 461, 0,
    2107, 22, 15826566, 1, 15827187, 1, 1,
0, ENSG00000177426, 7050, 7050, TGIF1, GRCh38, ensembl_havana, 0, 0, 18, 602630,
    9347, 401, 1206, 18, 3411608, 20, 3459978, 1, 23,
0, ENSG00000131721, 84528, 84528, RHOXF2, GRCh38, ensembl_havana, 0, 0, ?, 300447,
    2150, 288, 867, ?, 120158561, 5, 120165630, 1, 1,
0, ENSG00000159387, 79190, 79190, IRX6, GRCh38, ensembl_havana, 0, 0, 16, 606196,
    798, 446, 1341, 16, 55323760, 7, 55330760, 1, 2,
0, ENSG00000143995, 4211, 4211, MEIS1, GRCh38, ensembl_havana, 0, 0, 2, 601739,
    9661, 390, 1173, 2, 66433452, 19, 66573869, 1, 17,
0, ENSG00000179981, 10194, 10194, TSHZ1, GRCh38, ensembl_havana, 0, 0, 18, 614427,
    1443, 1077, 3234, 18, 75210755, 9, 75289950, 1, 5,
0, ENSG00000153779, 60436, 60436, TGIF2, GRCh38, ensembl_havana, 0, 0, 20, 607294,
    3091, 237, 714, ?, 89921882, 10, 89922883, 1, 2,
0, ENSG00000188620, 340784, 340784, HMX3, GRCh38, ensembl_havana, 0, 0, 10, 613380,
    851, 357, 1074, 10, 123135962, 9, 123137741, 1, 1,
0, ENSG00000264743, 503645, 503645, DPRXP4, GRCh38, havana, 0, 0, 17, ?, 235, 0,
    670, 17, 30975387, 2, 30975954, 1, 1,
0, ENSG00000170166, 3233, 3233, HOXD4, GRCh38, ensembl_havana, 0, 0, 2, 142981,
    2473, 255, 768, 2, 176151222, 5, 176153226, 1, 1,
0, ENSG00000107807, 3195, 3195, TLX1, GRCh38, ensembl_havana, 0, 0, 10, 186770,
    5483, 330, 993, 10, 101130505, 12, 101137789, 1, 5,
0, ENSG00000136630, 3142, 3142, HLX, GRCh38, ensembl_havana, 0, 0, 1, 142995, 4937,
    488, 1467, 1, 220879400, 12, 220885059, 1, 3,
0, ENSG00000135638, 2016, 2016, EMX1, GRCh38, ensembl_havana, 0, 0, 2, 600034,
    1405, 290, 873, 2, 72916260, 13, 72936071, 1, 6,
0, ENSG00000180613, 170825, 170825, GSX2, GRCh38, ensembl_havana, 0, 0, 4, 616253,
    1302, 304, 915, 4, 54099523, 10, 54102505, 1, 5,
0, ENSG00000116132, 5396, 5396, PRRX1, GRCh38, ensembl_havana, 0, 0, 1, 167420,
    3435, 245, 738, 1, 170662728, 11, 170739419, 1, 8,
0, ENSG00000121454, 89884, 89884, LHX4, GRCh38, ensembl_havana, 0, 0, 1, 602146,
    3702, 390, 1173, 1, 180230286, 5, 180278982, 1, 3,
0, ENSG00000106038, 2128, 2128, EVX1, GRCh38, ensembl_havana, 0, 0, 7, 142996,
    1332, 407, 1224, 7, 27242700, 12, 27250493, 1, 4,
0, ENSG00000144355, 1745, 1745, DLX1, GRCh38, ensembl_havana, 0, 0, 2, 600029,
    1745, 255, 768, 2, 172084740, 14, 172089677, 1, 7,
0, ENSG00000188816, 3167, 3167, HMX2, GRCh38, ensembl_havana, 0, 0, 10, 600647,
    1100, 273, 822, 10, 123148122, 3, 123150672, 1, 1,
0, ENSG00000162624, 431707, 431707, LHX8, GRCh38, ensembl_havana, 0, 0, 1, 604425,
    2603, 356, 1071, 1, 75128434, 14, 75161533, 1, 3,

```

Caso de Estudo 1

```

0, ENSG00000115507, 5013, 5013, OTX1, GRCh38, ensembl_havana, 0, 0, 2, 600036,
    3222, 354, 1065, 2, 63050057, 9, 63057836, 1, 6,
0, ENSG00000280774, 342900, 342900, LEUTX, GRCh38, ensembl_havana, 0, 0, 19, ?,
    472, 168, 507, 2021, 39776595, 2, 39786167, 1, 2,
0, ENSG00000151650, 27287, 27287, VENTX, GRCh38, ensembl_havana, 0, 0, 10, 607158,
    1913, 258, 777, 10, 133237404, 7, 133241929, 1, 1,
0, ENSG00000128645, 3231, 3231, HOXD1, GRCh38, ensembl_havana, 0, 0, 2, 142987,
    1982, 328, 987, 2, 176188579, 14, 176190907, 1, 1,
0, ENSG00000170549, 79192, 79192, IRX1, GRCh38, ensembl_havana, 0, 0, 5, 606197,
    1805, 480, 1443, 5, 3596054, 3, 3601403, 1, 1,
0, ENSG00000180438, 348825, 348825, TPRXL, GRCh38, havana, 0, 0, 3, 611167, 686, 0,
    2276, 3, 13937273, 15, 14082811, 1, 14,
0, ENSG00000160199, 5316, 5316, PKNOX1, GRCh38, ensembl_havana, 0, 0, 21, 602100,
    3777, 436, 1311, 21, 42974510, 14, 43033931, 1, 11,
0, ENSG00000254105, 27287, 27287, VENTX, GRCh38, havana, 0, 0, 10, 607158, 1913,
    258, 777, 8, 73651289, 1, 73651602, 1, 1,
0, ENSG00000170178, 3231, 3231, HOXD1, GRCh38, ensembl_havana, 0, 0, 2, 142987,
    1982, 328, 987, 2, 176099730, 6, 176101193, 1, 2,
0, ENSG00000167157, 51450, 51450, PRRX2, GRCh38, ensembl_havana, 0, 0, 9, 604675,
    1503, 253, 762, 9, 129665641, 10, 129722674, 1, 1,
0, ENSG00000175329, 91464, 91464, ISX, GRCh38, ensembl_havana, 0, 0, 22, 612019,
    981, 245, 738, 22, 35066136, 12, 35087387, 1, 2,
0, ENSG00000205857, 360030, 360030, NANOGNB, GRCh38, ensembl_havana, 0, 0, 12, ?,
    612, 188, 567, 12, 7765216, 2, 7774121, 1, 2,
0, ENSG00000128652, 3232, 3232, HOXD3, GRCh38, ensembl_havana, 0, 0, 2, 142980,
    4374, 432, 1299, 2, 176136612, 11, 176173102, 1, 4,
0, ENSG00000091656, 79776, 79776, ZFX4, GRCh38, ensembl_havana, 0, 0, 8, 606940,
    1404, 3616, 10851, 8, 76681219, 15, 76867285, 1, 12,
0, ENSG00000128713, 3231, 3231, HOXD1, GRCh38, ensembl_havana, 0, 0, 2, 142987,
    1982, 328, 987, 2, 176104216, 13, 176109754, 1, 2,
0, ENSG00000128710, 3231, 3231, HOXD1, GRCh38, ensembl_havana, 0, 0, 2, 142987,
    1982, 328, 987, 2, 176108790, 5, 176119942, 1, 3,

```

Cluster1

```

clusterNr, EnsemblID, GenBankID, KeggID, Display_Name, Assembly_Name, Source,status
    , currentID, chromosome, mim, geneWeight, aaseqSize, ntseqSize, seq_region_name
    , start, version, end, strand, transcript,
1, ENSG00000106410, 135935, 135935, NOBOX, GRCh38, ensembl_havana, 0, 0, 7, 610934,
    2908, 691, 2076, 7, 144397240, 14, 144410227, -1, 2,
1, ENSG00000235602, 642559, 642559, POU5F1P3, GRCh38, havana, 0, 0, 12, ?, 777, 0,
    1084, 12, 8133772, 5, 8134849, -1, 1,
1, ENSG00000280635, 5626, 5626, PROP1, GRCh38, havana, 0, 0, 5, 601538, 9847, 226,
    681, 525, 177996182, 1, 177996994, -1, 1,
1, ENSG00000230465, 27287, 27287, VENTX, GRCh38, havana, 0, 0, 10, 607158, 1913,
    258, 777, 3, 26346681, 1, 26347638, -1, 1,
1, ENSG00000165259, 139324, 139324, HDX, GRCh38, ensembl_havana, 0, 0, ?, 300994,
    689, 690, 2073, ?, 84317874, 13, 84502479, -1, 6,
1, ENSG00000259021, 284355, 284355, TPRX1, GRCh38, havana, 0, 0, 19, 611166, 457,
    411, 1236, 10, 79499666, 2, 79499984, -1, 1,

```

Caso de Estudio 1

```
1, ENSG00000174306, 23051, 23051, ZHX3, GRCh38, ensembl_havana, 0, 0, 20, 609598,
    1374, 956, 2871, 20, 41178448, 21, 41317672, -1, 21,
1, ENSG00000105419, 56917, 56917, MEIS3, GRCh38, ensembl_havana, 0, 0, 19, ?, 943,
    375, 1128, 19, 47403124, 17, 47419523, -1, 14,
1, ENSG00000156150, 257, 257, ALX3, GRCh38, ensembl_havana, 0, 0, 1, 606014, 1693,
    343, 1032, 1, 110059994, 6, 110070700, -1, 1,
1, ENSG00000215271, 57594, 57594, HOMEZ, GRCh38, ensembl_havana, 0, 0, 14, 608119,
    1290, 550, 1653, 14, 23272422, 7, 23299447, -1, 4,
1, ENSG00000134438, 30062, 30062, RAX, GRCh38, ensembl_havana, 0, 0, 18, 601881,
    1978, 346, 1041, 18, 59267035, 9, 59274086, -1, 4,
1, ENSG00000165588, 5015, 5015, OTX2, GRCh38, ensembl_havana, 0, 0, 14, 600037,
    10053, 289, 870, 14, 56799905, 16, 56810479, -1, 7,
1, ENSG00000236375, 100009667, 100009667, POU5F1P5, GRCh38, havana, 0, 0, 10, ?,
    405, 0, 937, 10, 68010205, 3, 68010862, -1, 1,
1, ENSG00000171476, 84525, 84525, HOPX, GRCh38, ensembl_havana, 0, 0, 4, 607275,
    5424, 73, 222, 4, 56647988, 21, 56681899, -1, 15,
1, ENSG00000265293, 503582, 503582, ARGFX, GRCh38, havana, 0, 0, 3, 611164, 790,
    315, 948, 17, 32150433, 2, 32151504, -1, 1,
1, ENSG00000150051, 283078, 283078, MKX, GRCh38, ensembl_havana, 0, 0, 10, 601332,
    1640, 352, 1059, 10, 27672875, 13, 27746060, -1, 4,
1, ENSG00000185610, 440097, 440097, DBX2, GRCh38, ensembl_havana, 0, 0, 12, ?, 435,
    339, 1020, 12, 45014672, 6, 45051099, -1, 1,
1, ENSG00000170561, 153572, 153572, IRX2, GRCh38, ensembl_havana, 0, 0, 5, 606198,
    1509, 471, 1416, 5, 2745845, 12, 2751662, -1, 2,
1, ENSG00000173976, 30062, 30062, RAX, GRCh38, ensembl_havana, 0, 0, 18, 601881,
    1978, 346, 1041, 19, 3769089, 15, 3772221, -1, 2,
1, ENSG00000119608, 283571, 283571, PROX2, GRCh38, ensembl_havana, 0, 0, 14,
    615094, 662, 592, 1779, 14, 74852871, 12, 74871940, -1, 3,
1, ENSG00000168779, 6474, 6474, SHOX2, GRCh38, ensembl_havana, 0, 0, 3, 602504,
    4303, 331, 996, 3, 158095954, 19, 158106503, -1, 6,
1, ENSG00000163666, 8820, 8820, HESX1, GRCh38, ensembl_havana, 0, 0, 3, 601802,
    4951, 185, 558, 3, 57197843, 8, 57226521, -1, 3,
1, ENSG00000179528, 85474, 85474, LBX2, GRCh38, ensembl_havana, 0, 0, 2, 607164,
    679, 198, 597, 2, 74497517, 15, 74503316, -1, 6,
1, ENSG00000164900, 2636, 2636, GBX1, GRCh38, ensembl_havana, 0, 0, 7, 603354, 959,
    363, 1092, 7, 151148589, 4, 151174745, -1, 2,
1, ENSG00000165606, 644168, 644168, DRGX, GRCh38, ensembl_havana, 0, 0, 10, 606701,
    901, 263, 792, 10, 49364181, 8, 49396016, -1, 2,
1, ENSG00000227059, 647589, 647589, ANHX, GRCh38, ensembl_havana, 0, 0, 12, ?, 430,
    379, 1140, 12, 133218312, 6, 133236095, -1, 2,
1, ENSG00000134138, 4212, 4212, MEIS2, GRCh38, ensembl_havana, 0, 0, 15, 601740,
    3685, 477, 1434, 15, 36889204, 19, 37101299, -1, 27,
1, ENSG00000177045, 147912, 147912, SIX5, GRCh38, ensembl_havana, 0, 0, 19, 600963,
    2469, 739, 2220, 19, 45764785, 7, 45769226, -1, 4,
1, ENSG00000178928, 284355, 284355, TPRX1, GRCh38, ensembl_havana, 0, 0, 19,
    611166, 457, 411, 1236, 19, 47801243, 8, 47819051, -1, 3,
1, ENSG00000253569, 27287, 27287, VENTX, GRCh38, havana, 0, 0, 10, 607158, 1913,
    258, 777, 8, 33722305, 1, 33723079, -1, 1,
```

Caso de Estudio 1

```

1, ENSG00000169554, 9839, 9839, ZEB2, GRCh38, ensembl_havana, 0, 0, 2, 605802,
    19995, 1214, 3645, 2, 144364364, 18, 144524583, -1, 47,
1, ENSG00000215037, 27287, 27287, VENTX, GRCh38, havana, 0, 0, 10, 607158, 1913,
    258, 777, 13, 83907301, 2, 83907781, -1, 1,
1, ENSG00000165462, 401, 401, PHOX2A, GRCh38, ensembl_havana, 0, 0, 11, 602753,
    4205, 284, 855, 11, 72239077, 5, 72245664, -1, 3,
1, ENSG00000106852, 26468, 26468, LHX6, GRCh38, ensembl_havana, 0, 0, 9, 608215,
    1913, 363, 1092, 9, 122202577, 15, 122229626, -1, 12,
1, ENSG00000188909, 390259, 390259, BSX, GRCh38, ensembl_havana, 0, 0, 11, 611074,
    923, 233, 702, 11, 122977570, 4, 122981720, -1, 1,
1, ENSG00000136367, 85446, 85446, ZFXH2, GRCh38, ensembl_havana, 0, 0, 14, ?, 493,
    2572, 7719, 14, 23520855, 13, 23556192, -1, 5,

```

Cluster2

```

clusterNr, EnsemblID, GenBankID, KeggID, Display_Name, Assembly_Name, Source,status
    , currentID, chromosome, mim, geneWeight, aaseqSize, ntseqSize, seq_region_name
    , start, version, end, strand, transcript,
2, ENSG00000106006, 3203, 3203, HOXA6, GRCh38, ensembl_havana, 0, 0, 7, 142951,
    1275, 233, 702, 7, 27145396, 6, 27150603, -1, 2,
2, ENSG00000120075, 3215, 3215, HOXB5, GRCh38, ensembl_havana, 0, 0, 17, 142960,
    4370, 269, 810, 17, 48591257, 5, 48593961, -1, 1,
2, ENSG00000197576, 3201, 3201, HOXA4, GRCh38, ensembl_havana, 0, 0, 7, 142953,
    3249, 320, 963, 7, 27128507, 13, 27130799, -1, 4,
2, ENSG00000120068, 3218, 3218, HOXB8, GRCh38, ensembl_havana, 0, 0, 17, 142963,
    1731, 243, 732, 17, 48611377, 6, 48614939, -1, 3,
2, ENSG00000170689, 3219, 3219, HOXB9, GRCh38, ensembl_havana, 0, 0, 17, 142964,
    5337, 250, 753, 17, 48621159, 9, 48626356, -1, 1,
2, ENSG00000173917, 3212, 3212, HOXB2, GRCh38, ensembl_havana, 0, 0, 17, 142967,
    3077, 356, 1071, 17, 48540894, 10, 48544989, -1, 3,
2, ENSG00000105997, 3200, 3200, HOXA3, GRCh38, ensembl_havana, 0, 0, 7, 142954,
    1626, 443, 1332, 7, 27106184, 22, 27152581, -1, 7,
2, ENSG00000197757, 3223, 3223, HOXC6, GRCh38, ensembl_havana, 0, 0, 12, 142972,
    3245, 235, 708, 12, 53990624, 7, 54030823, 1, 4,
2, ENSG00000123364, 3229, 3229, HOXC13, GRCh38, ensembl_havana, 0, 0, 12, 142976,
    3334, 330, 993, 12, 53938765, 4, 53946544, 1, 1,
2, ENSG00000120093, 3213, 3213, HOXB3, GRCh38, ensembl_havana, 0, 0, 17, 142966,
    3097, 431, 1296, 17, 48548870, 11, 48604912, -1, 13,
2, ENSG00000123407, 3228, 3228, HOXC12, GRCh38, ensembl_havana, 0, 0, 12, 142975,
    861, 282, 849, 12, 53954834, 3, 53958956, 1, 1,
2, ENSG00000198353, 3221, 3221, HOXC4, GRCh38, ensembl_havana, 0, 0, 12, 142974,
    2748, 264, 795, 12, 54016931, 7, 54056030, 1, 3,
2, ENSG00000101883, 158800, 158800, RHOF1, GRCh38, ensembl_havana, 0, 0, ?,
    300446, 1367, 184, 555, ?, 120109053, 4, 120115937, -1, 1,
2, ENSG00000215612, 3166, 3166, HMX1, GRCh38, ensembl_havana, 0, 0, 4, 142992,
    1219, 348, 1047, 4, 8846076, 7, 8871817, -1, 2,
2, ENSG00000123576, 80712, 80712, ESX1, GRCh38, ensembl_havana, 0, 0, ?, 300154,
    2200, 406, 1221, ?, 104250038, 5, 104254933, -1, 1,
2, ENSG00000106031, 3209, 3209, HOXA13, GRCh38, ensembl_havana, 0, 0, 7, 142959,
    6866, 388, 1167, 7, 27193503, 7, 27200106, -1, 2,

```

Caso de Estudo 1

```
2, ENSG00000130675, 3110, 3110, MNX1, GRCh38, ensembl_havana, 0, 0, 7, 142994,
5766, 401, 1206, 7, 156994051, 14, 157010651, -1, 8,
2, ENSG00000172789, 3222, 3222, HOXC5, GRCh38, ensembl_havana, 0, 0, 12, 142973,
1418, 222, 669, 12, 54032853, 3, 54035358, 1, 1,
2, ENSG00000119614, 338917, 338917, VSX2, GRCh38, ensembl_havana, 0, 0, 14, 142993,
3058, 361, 1086, 14, 74239472, 2, 74262738, 1, 1,
2, ENSG00000122592, 3204, 3204, HOXA7, GRCh38, ensembl_havana, 0, 0, 7, 142950,
3901, 230, 693, 7, 27153716, 7, 27157936, -1, 3,
2, ENSG00000180806, 3225, 3225, HOXC9, GRCh38, ensembl_havana, 0, 0, 12, 142971,
2568, 260, 783, 12, 53994895, 4, 54003337, 1, 3,
2, ENSG00000037965, 3224, 3224, HOXC8, GRCh38, ensembl_havana, 0, 0, 12, 142970,
3649, 242, 729, 12, 54009106, 5, 54012362, 1, 1,
2, ENSG00000108511, 3216, 3216, HOXB6, GRCh38, ensembl_havana, 0, 0, 17, 142961,
3561, 224, 675, 17, 48595751, 9, 48604992, -1, 4,
```

Nas tabelas A.1, A.2 e A.3 apresentamos, de uma forma mais legível a informação referente a cada *cluster* resultante. Esta informação está contida no documento A.2.2.1, sendo estas tabelas apenas para facilitar a leitura da informação.

A.2.2.2 Resultados do algoritmo Make Density Based Clusterer

Nesta secção apresentamos os resultados relativos à aplicação do algoritmo *make density based clusterer*, englobando também tabelas referentes aos *clusters* resultantes.

```
MakeDensityBasedClusterer:
```

```
Wrapped clusterer:
```

```
kMeans
```

```
=====
```

```
Number of iterations: 11
```

```
Within cluster sum of squared errors: 51.875635143086996
```

```
Initial starting points (random):
```

```
Cluster 0: 0,0,1,606066,1259,397,1194,1,197911902,15,197935478,1,8
```

```
Cluster 1: 0,0,19,438617.081633,943,375,1128,19,47403124,17,47419523,-1,14
```

```
Cluster 2: 0,0,10,607158,1913,258,777,13,83907301,2,83907781,-1,1
```

```
Missing values globally replaced with mean/mode
```

```
Final cluster centroids:
```

	Cluster#			
Attribute	Full Data	0	1	2
	(114.0)	(54.0)	(25.0)	(35.0)
=====				
status	0	0	0	0

Caso de Estudo 1

currentID	0	0	0	0
chromosome	9.7798	8.5289	11.2936	10.6286
mim	438617.0816	476792.1077	161877.12	577389.8711
geneWeight	2504.3246	2179.3333	2861.76	2750.4286
aaseqSize	379.9386	367.2778	320.76	441.7429
ntseqSize	1271.7719	1339.7593	965.28	1385.8
seq_region_name	33.037	46.7058	14.0844	25.4857
start	79174515.5351	96344273.3519	55246955.44	69775146.4
version	8.7632	8.6852	7.8	9.5714
end	79202730.0088	96375857.8889	55267785.28	69803436.0857
strand	0.1053	1	-0.28	-1
transcript	4.614	4.3519	3.2	6.0286

Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.4701

Attribute: status

Normal Distribution. Mean = 0 StdDev = 0

Attribute: currentID

Normal Distribution. Mean = 0 StdDev = 0

Attribute: chromosome

Normal Distribution. Mean = 8.5289 StdDev = 6.891

Attribute: mim

Normal Distribution. Mean = 476792.1077 StdDev = 181068.2467

Attribute: geneWeight

Normal Distribution. Mean = 2179.3333 StdDev = 1893.1074

Attribute: aaseqSize

Normal Distribution. Mean = 367.2778 StdDev = 490.1419

Attribute: ntseqSize

Normal Distribution. Mean = 1339.7593 StdDev = 1450.2666

Attribute: seq_region_name

Normal Distribution. Mean = 46.7058 StdDev = 271.3296

Attribute: start

Normal Distribution. Mean = 96344273.3519 StdDev = 61724206.7051

Attribute: version

Normal Distribution. Mean = 8.6852 StdDev = 4.8945

Attribute: end

Normal Distribution. Mean = 96375857.8889 StdDev = 61719051.1127

Attribute: strand

Normal Distribution. Mean = 1 StdDev = 0.9988

Attribute: transcript

Normal Distribution. Mean = 4.3519 StdDev = 4.8502

Cluster: 1 Prior probability: 0.2222

Attribute: status

Caso de Estudio 1

```
Normal Distribution. Mean = 0 StdDev = 0
Attribute: currentID
Normal Distribution. Mean = 0 StdDev = 0
Attribute: chromosome
Normal Distribution. Mean = 11.2936 StdDev = 3.9765
Attribute: mim
Normal Distribution. Mean = 161877.12 StdDev = 51201.6052
Attribute: geneWeight
Normal Distribution. Mean = 2861.76 StdDev = 1544.8115
Attribute: aaseqSize
Normal Distribution. Mean = 320.76 StdDev = 106.2545
Attribute: ntseqSize
Normal Distribution. Mean = 965.28 StdDev = 318.7636
Attribute: seq_region_name
Normal Distribution. Mean = 14.0844 StdDev = 8.0301
Attribute: start
Normal Distribution. Mean = 55246955.44 StdDev = 31830679.5727
Attribute: version
Normal Distribution. Mean = 7.8 StdDev = 4.4091
Attribute: end
Normal Distribution. Mean = 55267785.28 StdDev = 31836077.4108
Attribute: strand
Normal Distribution. Mean = -0.28 StdDev = 0.96
Attribute: transcript
Normal Distribution. Mean = 3.2 StdDev = 2.7713

Cluster: 2 Prior probability: 0.3077

Attribute: status
Normal Distribution. Mean = 0 StdDev = 0
Attribute: currentID
Normal Distribution. Mean = 0 StdDev = 0
Attribute: chromosome
Normal Distribution. Mean = 10.6286 StdDev = 5.6116
Attribute: mim
Normal Distribution. Mean = 577389.8711 StdDev = 63222.5576
Attribute: geneWeight
Normal Distribution. Mean = 2750.4286 StdDev = 3749.1512
Attribute: aaseqSize
Normal Distribution. Mean = 441.7429 StdDev = 434.464
Attribute: ntseqSize
Normal Distribution. Mean = 1385.8 StdDev = 1265.4002
Attribute: seq_region_name
Normal Distribution. Mean = 25.4857 StdDev = 85.8438
Attribute: start
Normal Distribution. Mean = 69775146.4 StdDev = 46985740.4219
Attribute: version
Normal Distribution. Mean = 9.5714 StdDev = 6.2073
Attribute: end
```

Caso de Estudo 1

Normal Distribution. Mean = 69803436.0857 StdDev = 46984717.9395

Attribute: strand

Normal Distribution. Mean = -1 StdDev = 0.9988

Attribute: transcript

Normal Distribution. Mean = 6.0286 StdDev = 9.1853

Clustered Instances

```
0      50 ( 44%)
1      24 ( 21%)
2      40 ( 35%)
```

Log likelihood: -66.75763

Cluster0

```
clusterNr, EnsemblID, GenBankID, KeggID, Display_Name, Assembly_Name, Source,status
, currentID, chromosome, mim, geneWeight, aaseqSize, ntseqSize, seq_region_name
, start, version, end, strand, transcript,
0, ENSG00000280635, 5626, 5626, PROP1, GRCh38, havana, 0, 0, 5, 601538, 9847, 226,
681, 525, 177996182, 1, 177996994, -1, 1,
0, ENSG00000165259, 139324, 139324, HDX, GRCh38, ensembl_havana, 0, 0, ?, 300994,
689, 690, 2073, ?, 84317874, 13, 84502479, -1, 6,
0, ENSG00000179277, 4213, 4213, MEIS3P1, GRCh38, havana, 0, 0, 17, ?, 307, 0, 2872,
17, 15786618, 9, 15787575, 1, 1,
0, ENSG00000128709, 3235, 3235, HOXD9, GRCh38, ensembl_havana, 0, 0, 2, 142982,
3609, 352, 1059, 2, 176122720, 12, 176124937, 1, 1,
0, ENSG00000143355, 56956, 56956, LHX9, GRCh38, ensembl_havana, 0, 0, 1, 606066,
1259, 397, 1194, 1, 197911902, 15, 197935478, 1, 8,
0, ENSG00000186103, 503582, 503582, ARGFX, GRCh38, ensembl_havana, 0, 0, 3, 611164,
790, 315, 948, 3, 121570704, 3, 121586634, 1, 1,
0, ENSG00000164438, 30012, 30012, TLX3, GRCh38, ensembl_havana, 0, 0, 5, 604640,
4534, 291, 876, 5, 171309284, 5, 171312134, 1, 1,
0, ENSG00000175879, 3234, 3234, HOXD8, GRCh38, ensembl_havana, 0, 0, 2, 142985,
1838, 290, 873, 2, 176129694, 8, 176132695, 1, 4,
0, ENSG00000237872, 645682, 645682, POU5F1P4, GRCh38, havana, 0, 0, 1, ?, 774, 0,
1083, 1, 155433178, 4, 155434262, 1, 1,
0, ENSG00000115297, 3196, 3196, TLX2, GRCh38, ensembl_havana, 0, 0, 2, 604240,
1937, 284, 855, 2, 74513463, 10, 74517147, 1, 4,
0, ENSG00000259849, 139538, 139538, VENTXP1, GRCh38, havana, 0, 0, ?, ?, 395, 0,
2716, ?, 26558337, 1, 26561052, 1, 1,
0, ENSG00000213921, 342900, 342900, LEUTX, GRCh38, ensembl_havana, 0, 0, 19, ?,
472, 168, 507, 19, 39776595, 7, 39786167, 1, 2,
0, ENSG00000006377, 1750, 1750, DLX6, GRCh38, ensembl_havana, 0, 0, 7, 600030,
1974, 293, 882, 7, 97005548, 10, 97011039, 1, 3,
0, ENSG00000101544, 22850, 22850, ADNP2, GRCh38, ensembl_havana, 0, 0, 18, 617422,
1008, 1131, 3396, 18, 80109031, 8, 80147523, 1, 5,
0, ENSG00000236380, 391518, 391518, VENTXP7, GRCh38, havana, 0, 0, 3, ?, 373, 0,
983, 3, 21405737, 5, 21406508, 1, 1,
```

Caso de Estudio 1

```

0, ENSG00000043039, 8538, 8538, BARX2, GRCh38, ensembl_havana, 0, 0, 11, 604823,
    1531, 279, 840, 11, 129375940, 6, 129452279, 1, 2,
0, ENSG00000257364, 27287, 27287, VENTX, GRCh38, havana, 0, 0, 10, 607158, 1913,
    258, 777, 12, 74292324, 1, 74293096, 1, 1,
0, ENSG00000271672, 503637, 503637, DUXAP8, GRCh38, havana, 0, 0, 22, ?, 461, 0,
    2107, 22, 15826566, 1, 15827187, 1, 1,
0, ENSG00000131721, 84528, 84528, RHOF2, GRCh38, ensembl_havana, 0, 0, ?, 300447,
    2150, 288, 867, ?, 120158561, 5, 120165630, 1, 1,
0, ENSG00000159387, 79190, 79190, IRX6, GRCh38, ensembl_havana, 0, 0, 16, 606196,
    798, 446, 1341, 16, 55323760, 7, 55330760, 1, 2,
0, ENSG00000179981, 10194, 10194, TSHZ1, GRCh38, ensembl_havana, 0, 0, 18, 614427,
    1443, 1077, 3234, 18, 75210755, 9, 75289950, 1, 5,
0, ENSG00000153779, 60436, 60436, TGIF2, GRCh38, ensembl_havana, 0, 0, 20, 607294,
    3091, 237, 714, ?, 89921882, 10, 89922883, 1, 2,
0, ENSG00000188620, 340784, 340784, HMX3, GRCh38, ensembl_havana, 0, 0, 10, 613380,
    851, 357, 1074, 10, 123135962, 9, 123137741, 1, 1,
0, ENSG00000264743, 503645, 503645, DPRXP4, GRCh38, havana, 0, 0, 17, ?, 235, 0,
    670, 17, 30975387, 2, 30975954, 1, 1,
0, ENSG00000170166, 3233, 3233, HOXD4, GRCh38, ensembl_havana, 0, 0, 2, 142981,
    2473, 255, 768, 2, 176151222, 5, 176153226, 1, 1,
0, ENSG00000101883, 158800, 158800, RHOF1, GRCh38, ensembl_havana, 0, 0, ?,
    300446, 1367, 184, 555, ?, 120109053, 4, 120115937, -1, 1,
0, ENSG00000136630, 3142, 3142, HLX, GRCh38, ensembl_havana, 0, 0, 1, 142995, 4937,
    488, 1467, 1, 220879400, 12, 220885059, 1, 3,
0, ENSG00000135638, 2016, 2016, EMX1, GRCh38, ensembl_havana, 0, 0, 2, 600034,
    1405, 290, 873, 2, 72916260, 13, 72936071, 1, 6,
0, ENSG00000180613, 170825, 170825, GSX2, GRCh38, ensembl_havana, 0, 0, 4, 616253,
    1302, 304, 915, 4, 54099523, 10, 54102505, 1, 5,
0, ENSG00000116132, 5396, 5396, PRRX1, GRCh38, ensembl_havana, 0, 0, 1, 167420,
    3435, 245, 738, 1, 170662728, 11, 170739419, 1, 8,
0, ENSG00000121454, 89884, 89884, LHX4, GRCh38, ensembl_havana, 0, 0, 1, 602146,
    3702, 390, 1173, 1, 180230286, 5, 180278982, 1, 3,
0, ENSG00000144355, 1745, 1745, DLX1, GRCh38, ensembl_havana, 0, 0, 2, 600029,
    1745, 255, 768, 2, 172084740, 14, 172089677, 1, 7,
0, ENSG00000188816, 3167, 3167, HMX2, GRCh38, ensembl_havana, 0, 0, 10, 600647,
    1100, 273, 822, 10, 123148122, 3, 123150672, 1, 1,
0, ENSG00000227059, 647589, 647589, ANHX, GRCh38, ensembl_havana, 0, 0, 12, ?, 430,
    379, 1140, 12, 133218312, 6, 133236095, -1, 2,
0, ENSG00000162624, 431707, 431707, LHX8, GRCh38, ensembl_havana, 0, 0, 1, 604425,
    2603, 356, 1071, 1, 75128434, 14, 75161533, 1, 3,
0, ENSG00000115507, 5013, 5013, OTX1, GRCh38, ensembl_havana, 0, 0, 2, 600036,
    3222, 354, 1065, 2, 63050057, 9, 63057836, 1, 6,
0, ENSG00000280774, 342900, 342900, LEUTX, GRCh38, ensembl_havana, 0, 0, 19, ?,
    472, 168, 507, 2021, 39776595, 2, 39786167, 1, 2,
0, ENSG00000151650, 27287, 27287, VENTX, GRCh38, ensembl_havana, 0, 0, 10, 607158,
    1913, 258, 777, 10, 133237404, 7, 133241929, 1, 1,
0, ENSG00000128645, 3231, 3231, HOXD1, GRCh38, ensembl_havana, 0, 0, 2, 142987,
    1982, 328, 987, 2, 176188579, 14, 176190907, 1, 1,

```

Caso de Estudio 1

```

0, ENSG00000170549, 79192, 79192, IRX1, GRCh38, ensembl_havana, 0, 0, 5, 606197,
  1805, 480, 1443, 5, 3596054, 3, 3601403, 1, 1,
0, ENSG00000254105, 27287, 27287, VENTX, GRCh38, havana, 0, 0, 10, 607158, 1913,
  258, 777, 8, 73651289, 1, 73651602, 1, 1,
0, ENSG00000170178, 3231, 3231, HOXD1, GRCh38, ensembl_havana, 0, 0, 2, 142987,
  1982, 328, 987, 2, 176099730, 6, 176101193, 1, 2,
0, ENSG00000167157, 51450, 51450, PRRX2, GRCh38, ensembl_havana, 0, 0, 9, 604675,
  1503, 253, 762, 9, 129665641, 10, 129722674, 1, 1,
0, ENSG00000175329, 91464, 91464, ISX, GRCh38, ensembl_havana, 0, 0, 22, 612019,
  981, 245, 738, 22, 35066136, 12, 35087387, 1, 2,
0, ENSG00000136367, 85446, 85446, ZFHX2, GRCh38, ensembl_havana, 0, 0, 14, ?, 493,
  2572, 7719, 14, 23520855, 13, 23556192, -1, 5,
0, ENSG00000205857, 360030, 360030, NANOGNB, GRCh38, ensembl_havana, 0, 0, 12, ?,
  612, 188, 567, 12, 7765216, 2, 7774121, 1, 2,
0, ENSG00000128652, 3232, 3232, HOXD3, GRCh38, ensembl_havana, 0, 0, 2, 142980,
  4374, 432, 1299, 2, 176136612, 11, 176173102, 1, 4,
0, ENSG00000091656, 79776, 79776, ZFHX4, GRCh38, ensembl_havana, 0, 0, 8, 606940,
  1404, 3616, 10851, 8, 76681219, 15, 76867285, 1, 12,
0, ENSG00000128713, 3231, 3231, HOXD1, GRCh38, ensembl_havana, 0, 0, 2, 142987,
  1982, 328, 987, 2, 176104216, 13, 176109754, 1, 2,
0, ENSG00000128710, 3231, 3231, HOXD1, GRCh38, ensembl_havana, 0, 0, 2, 142987,
  1982, 328, 987, 2, 176108790, 5, 176119942, 1, 3,

```

Cluster1

```

clusterNr, EnsemblID, GenBankID, KeggID, Display_Name, Assembly_Name, Source, status
, currentID, chromosome, mim, geneWeight, aaseqSize, ntseqSize, seq_region_name
, start, version, end, strand, transcript,
1, ENSG00000106006, 3203, 3203, HOXA6, GRCh38, ensembl_havana, 0, 0, 7, 142951,
  1275, 233, 702, 7, 27145396, 6, 27150603, -1, 2,
1, ENSG00000120075, 3215, 3215, HOXB5, GRCh38, ensembl_havana, 0, 0, 17, 142960,
  4370, 269, 810, 17, 48591257, 5, 48593961, -1, 1,
1, ENSG00000197576, 3201, 3201, HOXA4, GRCh38, ensembl_havana, 0, 0, 7, 142953,
  3249, 320, 963, 7, 27128507, 13, 27130799, -1, 4,
1, ENSG00000120068, 3218, 3218, HOXB8, GRCh38, ensembl_havana, 0, 0, 17, 142963,
  1731, 243, 732, 17, 48611377, 6, 48614939, -1, 3,
1, ENSG00000170689, 3219, 3219, HOXB9, GRCh38, ensembl_havana, 0, 0, 17, 142964,
  5337, 250, 753, 17, 48621159, 9, 48626356, -1, 1,
1, ENSG00000173917, 3212, 3212, HOXB2, GRCh38, ensembl_havana, 0, 0, 17, 142967,
  3077, 356, 1071, 17, 48540894, 10, 48544989, -1, 3,
1, ENSG00000105997, 3200, 3200, HOXA3, GRCh38, ensembl_havana, 0, 0, 7, 142954,
  1626, 443, 1332, 7, 27106184, 22, 27152581, -1, 7,
1, ENSG00000197757, 3223, 3223, HOXC6, GRCh38, ensembl_havana, 0, 0, 12, 142972,
  3245, 235, 708, 12, 53990624, 7, 54030823, 1, 4,
1, ENSG00000123364, 3229, 3229, HOXC13, GRCh38, ensembl_havana, 0, 0, 12, 142976,
  3334, 330, 993, 12, 53938765, 4, 53946544, 1, 1,
1, ENSG00000120093, 3213, 3213, HOXB3, GRCh38, ensembl_havana, 0, 0, 17, 142966,
  3097, 431, 1296, 17, 48548870, 11, 48604912, -1, 13,
1, ENSG00000123407, 3228, 3228, HOXC12, GRCh38, ensembl_havana, 0, 0, 12, 142975,
  861, 282, 849, 12, 53954834, 3, 53958956, 1, 1,

```

Caso de Estudio 1

```

1, ENSG00000198353, 3221, 3221, HOXC4, GRCh38, ensembl_havana, 0, 0, 12, 142974,
    2748, 264, 795, 12, 54016931, 7, 54056030, 1, 3,
1, ENSG00000107807, 3195, 3195, TLX1, GRCh38, ensembl_havana, 0, 0, 10, 186770,
    5483, 330, 993, 10, 101130505, 12, 101137789, 1, 5,
1, ENSG00000215612, 3166, 3166, HMX1, GRCh38, ensembl_havana, 0, 0, 4, 142992,
    1219, 348, 1047, 4, 8846076, 7, 8871817, -1, 2,
1, ENSG00000106038, 2128, 2128, EVX1, GRCh38, ensembl_havana, 0, 0, 7, 142996,
    1332, 407, 1224, 7, 27242700, 12, 27250493, 1, 4,
1, ENSG00000123576, 80712, 80712, ESX1, GRCh38, ensembl_havana, 0, 0, ?, 300154,
    2200, 406, 1221, ?, 104250038, 5, 104254933, -1, 1,
1, ENSG00000106031, 3209, 3209, HOXA13, GRCh38, ensembl_havana, 0, 0, 7, 142959,
    6866, 388, 1167, 7, 27193503, 7, 27200106, -1, 2,
1, ENSG00000130675, 3110, 3110, MNX1, GRCh38, ensembl_havana, 0, 0, 7, 142994,
    5766, 401, 1206, 7, 156994051, 14, 157010651, -1, 8,
1, ENSG00000172789, 3222, 3222, HOXC5, GRCh38, ensembl_havana, 0, 0, 12, 142973,
    1418, 222, 669, 12, 54032853, 3, 54035358, 1, 1,
1, ENSG00000119614, 338917, 338917, VSX2, GRCh38, ensembl_havana, 0, 0, 14, 142993,
    3058, 361, 1086, 14, 74239472, 2, 74262738, 1, 1,
1, ENSG00000122592, 3204, 3204, HOXA7, GRCh38, ensembl_havana, 0, 0, 7, 142950,
    3901, 230, 693, 7, 27153716, 7, 27157936, -1, 3,
1, ENSG00000180806, 3225, 3225, HOXC9, GRCh38, ensembl_havana, 0, 0, 12, 142971,
    2568, 260, 783, 12, 53994895, 4, 54003337, 1, 3,
1, ENSG00000037965, 3224, 3224, HOXC8, GRCh38, ensembl_havana, 0, 0, 12, 142970,
    3649, 242, 729, 12, 54009106, 5, 54012362, 1, 1,
1, ENSG00000108511, 3216, 3216, HOXB6, GRCh38, ensembl_havana, 0, 0, 17, 142961,
    3561, 224, 675, 17, 48595751, 9, 48604992, -1, 4,

```

Cluster2

```

clusterNr, EnsemblID, GenBankID, KeggID, Display_Name, Assembly_Name, Source,status
    , currentID, chromosome, mim, geneWeight, aaseqSize, ntseqSize, seq_region_name
    , start, version, end, strand, transcript,
2, ENSG00000106410, 135935, 135935, NOBOX, GRCh38, ensembl_havana, 0, 0, 7, 610934,
    2908, 691, 2076, 7, 144397240, 14, 144410227, -1, 2,
2, ENSG00000235602, 642559, 642559, POU5F1P3, GRCh38, havana, 0, 0, 12, ?, 777, 0,
    1084, 12, 8133772, 5, 8134849, -1, 1,
2, ENSG00000230465, 27287, 27287, VENTX, GRCh38, havana, 0, 0, 10, 607158, 1913,
    258, 777, 3, 26346681, 1, 26347638, -1, 1,
2, ENSG00000259021, 284355, 284355, TPRX1, GRCh38, havana, 0, 0, 19, 611166, 457,
    411, 1236, 10, 79499666, 2, 79499984, -1, 1,
2, ENSG00000169840, 219409, 219409, GSX1, GRCh38, ensembl_havana, 0, 0, 13, 616542,
    884, 264, 795, 13, 27792643, 4, 27794768, 1, 1,
2, ENSG00000147421, 79618, 79618, HMBOX1, GRCh38, ensembl_havana, 0, 0, 8, ?, 2401,
    420, 1263, 8, 28890394, 17, 29064764, 1, 16,
2, ENSG00000174306, 23051, 23051, ZHX3, GRCh38, ensembl_havana, 0, 0, 20, 609598,
    1374, 956, 2871, 20, 41178448, 21, 41317672, -1, 21,
2, ENSG00000105419, 56917, 56917, MEIS3, GRCh38, ensembl_havana, 0, 0, 19, ?, 943,
    375, 1128, 19, 47403124, 17, 47419523, -1, 14,
2, ENSG00000165495, 63876, 63876, PKNOX2, GRCh38, ensembl_havana, 0, 0, 11, 613066,
    1782, 472, 1419, 11, 125164687, 15, 125433389, 1, 14,

```

Caso de Estudio 1

```
2, ENSG00000156150, 257, 257, ALX3, GRCh38, ensembl_havana, 0, 0, 1, 606014, 1693,
    343, 1032, 1, 110059994, 6, 110070700, -1, 1,
2, ENSG00000215271, 57594, 57594, HOMEZ, GRCh38, ensembl_havana, 0, 0, 14, 608119,
    1290, 550, 1653, 14, 23272422, 7, 23299447, -1, 4,
2, ENSG00000118707, 60436, 60436, TGIF2, GRCh38, ensembl_havana, 0, 0, 20, 607294,
    3091, 237, 714, 20, 36573488, 9, 36593950, 1, 8,
2, ENSG00000134438, 30062, 30062, RAX, GRCh38, ensembl_havana, 0, 0, 18, 601881,
    1978, 346, 1041, 18, 59267035, 9, 59274086, -1, 4,
2, ENSG00000165588, 5015, 5015, OTX2, GRCh38, ensembl_havana, 0, 0, 14, 600037,
    10053, 289, 870, 14, 56799905, 16, 56810479, -1, 7,
2, ENSG00000236375, 100009667, 100009667, POU5F1P5, GRCh38, havana, 0, 0, 10, ?,
    405, 0, 937, 10, 68010205, 3, 68010862, -1, 1,
2, ENSG00000177426, 7050, 7050, TGIF1, GRCh38, ensembl_havana, 0, 0, 18, 602630,
    9347, 401, 1206, 18, 3411608, 20, 3459978, 1, 23,
2, ENSG00000171476, 84525, 84525, HOPX, GRCh38, ensembl_havana, 0, 0, 4, 607275,
    5424, 73, 222, 4, 56647988, 21, 56681899, -1, 15,
2, ENSG00000143995, 4211, 4211, MEIS1, GRCh38, ensembl_havana, 0, 0, 2, 601739,
    9661, 390, 1173, 2, 66433452, 19, 66573869, 1, 17,
2, ENSG00000265293, 503582, 503582, ARGFX, GRCh38, havana, 0, 0, 3, 611164, 790,
    315, 948, 17, 32150433, 2, 32151504, -1, 1,
2, ENSG00000150051, 283078, 283078, MKX, GRCh38, ensembl_havana, 0, 0, 10, 601332,
    1640, 352, 1059, 10, 27672875, 13, 27746060, -1, 4,
2, ENSG00000185610, 440097, 440097, DBX2, GRCh38, ensembl_havana, 0, 0, 12, ?, 435,
    339, 1020, 12, 45014672, 6, 45051099, -1, 1,
2, ENSG00000170561, 153572, 153572, IRX2, GRCh38, ensembl_havana, 0, 0, 5, 606198,
    1509, 471, 1416, 5, 2745845, 12, 2751662, -1, 2,
2, ENSG00000173976, 30062, 30062, RAX, GRCh38, ensembl_havana, 0, 0, 18, 601881,
    1978, 346, 1041, 19, 3769089, 15, 3772221, -1, 2,
2, ENSG00000119608, 283571, 283571, PROX2, GRCh38, ensembl_havana, 0, 0, 14,
    615094, 662, 592, 1779, 14, 74852871, 12, 74871940, -1, 3,
2, ENSG00000168779, 6474, 6474, SHOX2, GRCh38, ensembl_havana, 0, 0, 3, 602504,
    4303, 331, 996, 3, 158095954, 19, 158106503, -1, 6,
2, ENSG00000163666, 8820, 8820, HESX1, GRCh38, ensembl_havana, 0, 0, 3, 601802,
    4951, 185, 558, 3, 57197843, 8, 57226521, -1, 3,
2, ENSG00000179528, 85474, 85474, LBX2, GRCh38, ensembl_havana, 0, 0, 2, 607164,
    679, 198, 597, 2, 74497517, 15, 74503316, -1, 6,
2, ENSG00000164900, 2636, 2636, GBX1, GRCh38, ensembl_havana, 0, 0, 7, 603354, 959,
    363, 1092, 7, 151148589, 4, 151174745, -1, 2,
2, ENSG00000165606, 644168, 644168, DRGX, GRCh38, ensembl_havana, 0, 0, 10, 606701,
    901, 263, 792, 10, 49364181, 8, 49396016, -1, 2,
2, ENSG00000134138, 4212, 4212, MEIS2, GRCh38, ensembl_havana, 0, 0, 15, 601740,
    3685, 477, 1434, 15, 36889204, 19, 37101299, -1, 27,
2, ENSG00000177045, 147912, 147912, SIX5, GRCh38, ensembl_havana, 0, 0, 19, 600963,
    2469, 739, 2220, 19, 45764785, 7, 45769226, -1, 4,
2, ENSG00000178928, 284355, 284355, TPRX1, GRCh38, ensembl_havana, 0, 0, 19,
    611166, 457, 411, 1236, 19, 47801243, 8, 47819051, -1, 3,
2, ENSG00000253569, 27287, 27287, VENTX, GRCh38, havana, 0, 0, 10, 607158, 1913,
    258, 777, 8, 33722305, 1, 33723079, -1, 1,
```

Caso de Estudo 1

```
2, ENSG00000169554, 9839, 9839, ZEB2, GRCh38, ensembl_havana, 0, 0, 2, 605802,
  19995, 1214, 3645, 2, 144364364, 18, 144524583, -1, 47,
2, ENSG00000215037, 27287, 27287, VENTX, GRCh38, havana, 0, 0, 10, 607158, 1913,
  258, 777, 13, 83907301, 2, 83907781, -1, 1,
2, ENSG00000165462, 401, 401, PHOX2A, GRCh38, ensembl_havana, 0, 0, 11, 602753,
  4205, 284, 855, 11, 72239077, 5, 72245664, -1, 3,
2, ENSG00000106852, 26468, 26468, LHX6, GRCh38, ensembl_havana, 0, 0, 9, 608215,
  1913, 363, 1092, 9, 122202577, 15, 122229626, -1, 12,
2, ENSG00000180438, 348825, 348825, TPRXL, GRCh38, havana, 0, 0, 3, 611167, 686, 0,
  2276, 3, 13937273, 15, 14082811, 1, 14,
2, ENSG00000160199, 5316, 5316, PKNOX1, GRCh38, ensembl_havana, 0, 0, 21, 602100,
  3777, 436, 1311, 21, 42974510, 14, 43033931, 1, 11,
2, ENSG00000188909, 390259, 390259, BSX, GRCh38, ensembl_havana, 0, 0, 11, 611074,
  923, 233, 702, 11, 122977570, 4, 122981720, -1, 1,
```

Nas tabelas A.4, A.5 e A.6 apresentamos, de uma forma mais legível a informação referente a cada *cluster* resultante. Esta informação está contida no documento A.2.2.2, sendo estas tabelas apenas para facilitar a leitura da informação.

Caso de Estudo 1

Tabela A.1: Simple K-means - Cluster 0 (sumarizado)

Name	Source	chromosome	mim	gene Weight	aaseq Size	ntseq Size	seq region name	version	strand	transcript
MEIS3P1	havana	17	?	307	0	2872	17	9	1	1
HOXD9	ensembl_havana	2	142982	3609	352	1059	2	12	1	1
LHX9	ensembl_havana	1	606066	1259	397	1194	1	15	1	8
GSX1	ensembl_havana	13	616542	884	264	795	13	4	1	1
HMBOX1	ensembl_havana	8	?	2401	420	1263	8	17	1	16
PKNOX2	ensembl_havana	11	613066	1782	472	1419	11	15	1	14
ARGFX	ensembl_havana	3	611164	790	315	948	3	3	1	1
TLX3	ensembl_havana	5	604640	4534	291	876	5	5	1	1
HOXD8	ensembl_havana	2	142985	1838	290	873	2	8	1	4
TGIF2	ensembl_havana	20	607294	3091	237	714	20	9	1	8
POU5F1P4	havana	1	?	774	0	1083	1	4	1	1
TLX2	ensembl_havana	2	604240	1937	284	855	2	10	1	4
VENTXP1	havana	?	?	395	0	2716	?	1	1	1
LEUTX	ensembl_havana	19	?	472	168	507	19	7	1	2
DLX6	ensembl_havana	7	600030	1974	293	882	7	10	1	3
ADNP2	ensembl_havana	18	617422	1008	1131	3396	18	8	1	5
VENTXP7	havana	3	?	373	0	983	3	5	1	1
BARX2	ensembl_havana	11	604823	1531	279	840	11	6	1	2
VENTX	havana	10	607158	1913	258	777	12	1	1	1
DUXAP8	havana	22	?	461	0	2107	22	1	1	1
TGIF1	ensembl_havana	18	602630	9347	401	1206	18	20	1	23
RHOXF2	ensembl_havana	?	300447	2150	288	867	?	5	1	1
IRX6	ensembl_havana	16	606196	798	446	1341	16	7	1	2
MEIS1	ensembl_havana	2	601739	9661	390	1173	2	19	1	17
TSHZ1	ensembl_havana	18	614427	1443	1077	3234	18	9	1	5
TGIF2	ensembl_havana	20	607294	3091	237	714	?	10	1	2
HMX3	ensembl_havana	10	613380	851	357	1074	10	9	1	1
DPRXP4	havana	17	?	235	0	670	17	2	1	1
HOXD4	ensembl_havana	2	142981	2473	255	768	2	5	1	1
TLX1	ensembl_havana	10	186770	5483	330	993	10	12	1	5
HLX	ensembl_havana	1	142995	4937	488	1467	1	12	1	3
EMX1	ensembl_havana	2	600034	1405	290	873	2	13	1	6
GSX2	ensembl_havana	4	616253	1302	304	915	4	10	1	5
PRRX1	ensembl_havana	1	167420	3435	245	738	1	11	1	8
LHX4	ensembl_havana	1	602146	3702	390	1173	1	5	1	3
EVX1	ensembl_havana	7	142996	1332	407	1224	7	12	1	4
DLX1	ensembl_havana	2	600029	1745	255	768	2	14	1	7
HMX2	ensembl_havana	10	600647	1100	273	822	10	3	1	1
LHX8	ensembl_havana	1	604425	2603	356	1071	1	14	1	3
OTX1	ensembl_havana	2	600036	3222	354	1065	2	9	1	6
LEUTX	ensembl_havana	19	?	472	168	507	2021	2	1	2
VENTX	ensembl_havana	10	607158	1913	258	777	10	7	1	1
HOXD1	ensembl_havana	2	142987	1982	328	987	2	14	1	1
IRX1	ensembl_havana	5	606197	1805	480	1443	5	3	1	1
TPRXL	havana	3	611167	686	0	2276	3	15	1	14
PKNOX1	ensembl_havana	21	602100	3777	436	1311	21	14	1	11
VENTX	havana	10	607158	1913	258	777	8	1	1	1
HOXD1	ensembl_havana	2	142987	1982	328	987	2	6	1	2
PRRX2	ensembl_havana	9	604675	1503	253	762	9	10	1	1
ISX	ensembl_havana	22	612019	981	245	738	22	12	1	2
NANOGNB	ensembl_havana	12	?	612	188	567	12	2	1	2
HOXD3	ensembl_havana	2	142980	4374	432	1299	2	11	1	4
ZFHX4	ensembl_havana	8	606940	1404	3616	10851	8	15	1	12
HOXD1	ensembl_havana	2	142987	1982	328	987	2	13	1	2
HOXD1	ensembl_havana	2	142987	1982	328	987	2	5	1	3

Caso de Estudo 1

Tabela A.2: Simple K-means - Cluster 1 (sumarizado)

Name	Source	chromosome	mim	gene Weight	aaseq Size	ntseq Size	seq region name	version	strand	transcript
NOBOX	ensembl_havana	7	610934	2908	691	2076	7	14	-1	2
POU5F1P3	havana	12	?	777	0	1084	12	5	-1	1
PROP1	havana	5	601538	9847	226	681	525	1	-1	1
VENTX	havana	10	607158	1913	258	777	3	1	-1	1
HDX	ensembl_havana	?	300994	689	690	2073	?	13	-1	6
TPRX1	havana	19	611166	457	411	1236	10	2	-1	1
ZHX3	ensembl_havana	20	609598	1374	956	2871	20	21	-1	21
MEIS3	ensembl_havana	19	?	943	375	1128	19	17	-1	14
ALX3	ensembl_havana	1	606014	1693	343	1032	1	6	-1	1
HOMEZ	ensembl_havana	14	608119	1290	550	1653	14	7	-1	4
RAX	ensembl_havana	18	601881	1978	346	1041	18	9	-1	4
OTX2	ensembl_havana	14	600037	10053	289	870	14	16	-1	7
POU5F1P5	havana	10	?	405	0	937	10	3	-1	1
HOPX	ensembl_havana	4	607275	5424	73	222	4	21	-1	15
ARGFX	havana	3	611164	790	315	948	17	2	-1	1
MKX	ensembl_havana	10	601332	1640	352	1059	10	13	-1	4
DBX2	ensembl_havana	12	?	435	339	1020	12	6	-1	1
IRX2	ensembl_havana	5	606198	1509	471	1416	5	12	-1	2
RAX	ensembl_havana	18	601881	1978	346	1041	19	15	-1	2
PROX2	ensembl_havana	14	615094	662	592	1779	14	12	-1	3
SHOX2	ensembl_havana	3	602504	4303	331	996	3	19	-1	6
HESX1	ensembl_havana	3	601802	4951	185	558	3	8	-1	3
LBX2	ensembl_havana	2	607164	679	198	597	2	15	-1	6
GBX1	ensembl_havana	7	603354	959	363	1092	7	4	-1	2
DRGX	ensembl_havana	10	606701	901	263	792	10	8	-1	2
ANHX	ensembl_havana	12	?	430	379	1140	12	6	-1	2
MEIS2	ensembl_havana	15	601740	3685	477	1434	15	19	-1	27
SIX5	ensembl_havana	19	600963	2469	739	2220	19	7	-1	4
TPRX1	ensembl_havana	19	611166	457	411	1236	19	8	-1	3
VENTX	havana	10	607158	1913	258	777	8	1	-1	1
ZEB2	ensembl_havana	2	605802	19995	1214	3645	2	18	-1	47
VENTX	havana	10	607158	1913	258	777	13	2	-1	1
PHOX2A	ensembl_havana	11	602753	4205	284	855	11	5	-1	3
LHX6	ensembl_havana	9	608215	1913	363	1092	9	15	-1	12
BSX	ensembl_havana	11	611074	923	233	702	11	4	-1	1
ZFHX2	ensembl_havana	14	?	493	2572	7719	14	13	-1	5

Caso de Estudo 1

Tabela A.3: Simple K-means - Cluster 2 (sumarizado)

Name	Source	chromosome	mim	gene Weight	aaseq Size	ntseq Size	seq region name	version	strand	transcript
NOBOX	ensembl_havana	7	610934	2908	691	2076	7	14	-1	2
POU5F1P3	havana	12	?	777	0	1084	12	5	-1	1
PROP1	havana	5	601538	9847	226	681	525	1	-1	1
VENTX	havana	10	607158	1913	258	777	3	1	-1	1
HDX	ensembl_havana	?	300994	689	690	2073	?	13	-1	6
TPRX1	havana	19	611166	457	411	1236	10	2	-1	1
ZHX3	ensembl_havana	20	609598	1374	956	2871	20	21	-1	21
MEIS3	ensembl_havana	19	?	943	375	1128	19	17	-1	14
ALX3	ensembl_havana	1	606014	1693	343	1032	1	6	-1	1
HOMEZ	ensembl_havana	14	608119	1290	550	1653	14	7	-1	4
RAX	ensembl_havana	18	601881	1978	346	1041	18	9	-1	4
OTX2	ensembl_havana	14	600037	10053	289	870	14	16	-1	7
POU5F1P5	havana	10	?	405	0	937	10	3	-1	1
HOPX	ensembl_havana	4	607275	5424	73	222	4	21	-1	15
ARGFX	havana	3	611164	790	315	948	17	2	-1	1
MKX	ensembl_havana	10	601332	1640	352	1059	10	13	-1	4
DBX2	ensembl_havana	12	?	435	339	1020	12	6	-1	1
IRX2	ensembl_havana	5	606198	1509	471	1416	5	12	-1	2
RAX	ensembl_havana	18	601881	1978	346	1041	19	15	-1	2
PROX2	ensembl_havana	14	615094	662	592	1779	14	12	-1	3
SHOX2	ensembl_havana	3	602504	4303	331	996	3	19	-1	6
HESX1	ensembl_havana	3	601802	4951	185	558	3	8	-1	3
LBX2	ensembl_havana	2	607164	679	198	597	2	15	-1	6
GBX1	ensembl_havana	7	603354	959	363	1092	7	4	-1	2
DRGX	ensembl_havana	10	606701	901	263	792	10	8	-1	2
ANHX	ensembl_havana	12	?	430	379	1140	12	6	-1	2
MEIS2	ensembl_havana	15	601740	3685	477	1434	15	19	-1	27
SIX5	ensembl_havana	19	600963	2469	739	2220	19	7	-1	4
TPRX1	ensembl_havana	19	611166	457	411	1236	19	8	-1	3
VENTX	havana	10	607158	1913	258	777	8	1	-1	1
ZEB2	ensembl_havana	2	605802	19995	1214	3645	2	18	-1	47
VENTX	havana	10	607158	1913	258	777	13	2	-1	1
PHOX2A	ensembl_havana	11	602753	4205	284	855	11	5	-1	3
LHX6	ensembl_havana	9	608215	1913	363	1092	9	15	-1	12
BSX	ensembl_havana	11	611074	923	233	702	11	4	-1	1
ZFHX2	ensembl_havana	14	?	493	2572	7719	14	13	-1	5

Caso de Estudo 1

Tabela A.4: Make Density Based Clusterer - Cluster 0 (sumarizado)

Name	Source	chromosome	mim	gene Weight	aaseq Size	ntseq Size	seq region name	version	strand	transcript
PROP1	havana	5	601538	9847	226	681	525	1	-1	1
HDX	ensembl_havana	?	300994	689	690	2073	?	13	-1	6
MEIS3P1	havana	17	?	307	0	2872	17	9	1	1
HOXD9	ensembl_havana	2	142982	3609	352	1059	2	12	1	1
LHX9	ensembl_havana	1	606066	1259	397	1194	1	15	1	8
ARGFX	ensembl_havana	3	611164	790	315	948	3	3	1	1
TLX3	ensembl_havana	5	604640	4534	291	876	5	5	1	1
HOXD8	ensembl_havana	2	142985	1838	290	873	2	8	1	4
POU5F1P4	havana	1	?	774	0	1083	1	4	1	1
TLX2	ensembl_havana	2	604240	1937	284	855	2	10	1	4
VENTXP1	havana	?	?	395	0	2716	?	1	1	1
LEUTX	ensembl_havana	19	?	472	168	507	19	7	1	2
DLX6	ensembl_havana	7	600030	1974	293	882	7	10	1	3
ADNP2	ensembl_havana	18	617422	1008	1131	3396	18	8	1	5
VENTXP7	havana	3	?	373	0	983	3	5	1	1
BARX2	ensembl_havana	11	604823	1531	279	840	11	6	1	2
VENTX	havana	10	607158	1913	258	777	12	1	1	1
DUXAP8	havana	22	?	461	0	2107	22	1	1	1
RHOXF2	ensembl_havana	?	300447	2150	288	867	?	5	1	1
IRX6	ensembl_havana	16	606196	798	446	1341	16	7	1	2
TSHZ1	ensembl_havana	18	614427	1443	1077	3234	18	9	1	5
TGIF2	ensembl_havana	20	607294	3091	237	714	?	10	1	2
HMX3	ensembl_havana	10	613380	851	357	1074	10	9	1	1
DPRXP4	havana	17	?	235	0	670	17	2	1	1
HOXD4	ensembl_havana	2	142981	2473	255	768	2	5	1	1
RHOXF1	ensembl_havana	?	300446	1367	184	555	?	4	-1	1
HLX	ensembl_havana	1	142995	4937	488	1467	1	12	1	3
EMX1	ensembl_havana	2	600034	1405	290	873	2	13	1	6
GSX2	ensembl_havana	4	616253	1302	304	915	4	10	1	5
PRRX1	ensembl_havana	1	167420	3435	245	738	1	11	1	8
LHX4	ensembl_havana	1	602146	3702	390	1173	1	5	1	3
DLX1	ensembl_havana	2	600029	1745	255	768	2	14	1	7
HMX2	ensembl_havana	10	600647	1100	273	822	10	3	1	1
ANHX	ensembl_havana	12	?	430	379	1140	12	6	-1	2
LHX8	ensembl_havana	1	604425	2603	356	1071	1	14	1	3
OTX1	ensembl_havana	2	600036	3222	354	1065	2	9	1	6
LEUTX	ensembl_havana	19	?	472	168	507	2021	2	1	2
VENTX	ensembl_havana	10	607158	1913	258	777	10	7	1	1
HOXD1	ensembl_havana	2	142987	1982	328	987	2	14	1	1
IRX1	ensembl_havana	5	606197	1805	480	1443	5	3	1	1
VENTX	havana	10	607158	1913	258	777	8	1	1	1
HOXD1	ensembl_havana	2	142987	1982	328	987	2	6	1	2
PRRX2	ensembl_havana	9	604675	1503	253	762	9	10	1	1
ISX	ensembl_havana	22	612019	981	245	738	22	12	1	2
ZFHX2	ensembl_havana	14	?	493	2572	7719	14	13	-1	5
NANOGNB	ensembl_havana	12	?	612	188	567	12	2	1	2
HOXD3	ensembl_havana	2	142980	4374	432	1299	2	11	1	4
ZFHX4	ensembl_havana	8	606940	1404	3616	10851	8	15	1	12
HOXD1	ensembl_havana	2	142987	1982	328	987	2	13	1	2
HOXD1	ensembl_havana	2	142987	1982	328	987	2	5	1	3

Caso de Estudo 1

Tabela A.5: Make Density Based Clusterer - Cluster 1 (sumarizado)

Name	Source	chromosome	mim	gene Weight	aaseq Size	ntseq Size	seq region name	version	strand	transcript
HOXA6	ensembl_havana	7	142951	1275	233	702	7	6	-1	2
HOXB5	ensembl_havana	17	142960	4370	269	810	17	5	-1	1
HOXA4	ensembl_havana	7	142953	3249	320	963	7	13	-1	4
HOXB8	ensembl_havana	17	142963	1731	243	732	17	6	-1	3
HOXB9	ensembl_havana	17	142964	5337	250	753	17	9	-1	1
HOXB2	ensembl_havana	17	142967	3077	356	1071	17	10	-1	3
HOXA3	ensembl_havana	7	142954	1626	443	1332	7	22	-1	7
HOXC6	ensembl_havana	12	142972	3245	235	708	12	7	1	4
HOXC13	ensembl_havana	12	142976	3334	330	993	12	4	1	1
HOXB3	ensembl_havana	17	142966	3097	431	1296	17	11	-1	13
HOXC12	ensembl_havana	12	142975	861	282	849	12	3	1	1
HOXC4	ensembl_havana	12	142974	2748	264	795	12	7	1	3
TLX1	ensembl_havana	10	186770	5483	330	993	10	12	1	5
HMX1	ensembl_havana	4	142992	1219	348	1047	4	7	-1	2
EVX1	ensembl_havana	7	142996	1332	407	1224	7	12	1	4
ESX1	ensembl_havana	?	300154	2200	406	1221	?	5	-1	1
HOXA13	ensembl_havana	7	142959	6866	388	1167	7	7	-1	2
MNX1	ensembl_havana	7	142994	5766	401	1206	7	14	-1	8
HOXC5	ensembl_havana	12	142973	1418	222	669	12	3	1	1
VSX2	ensembl_havana	14	142993	3058	361	1086	14	2	1	1
HOXA7	ensembl_havana	7	142950	3901	230	693	7	7	-1	3
HOXC9	ensembl_havana	12	142971	2568	260	783	12	4	1	3
HOXC8	ensembl_havana	12	142970	3649	242	729	12	5	1	1
HOXB6	ensembl_havana	17	142961	3561	224	675	17	9	-1	4

Caso de Estudo 1

Tabela A.6: Make Density Based Clusterer - Cluster 2 (sumarizado)

Name	Source	chromosome	mim	gene Weight	aaseq Size	ntseq Size	seq region name	version	strand	transcript
NOBOX	ensembl_havana	7	610934	2908	691	2076	7	14	-1	2
POU5F1P3	havana	12	?	777	0	1084	12	5	-1	1
VENTX	havana	10	607158	1913	258	777	3	1	-1	1
TPRX1	havana	19	611166	457	411	1236	10	2	-1	1
GSX1	ensembl_havana	13	616542	884	264	795	13	4	1	1
HMBOX1	ensembl_havana	8	?	2401	420	1263	8	17	1	16
ZHX3	ensembl_havana	20	609598	1374	956	2871	20	21	-1	21
MEIS3	ensembl_havana	19	?	943	375	1128	19	17	-1	14
PKNOX2	ensembl_havana	11	613066	1782	472	1419	11	15	1	14
ALX3	ensembl_havana	1	606014	1693	343	1032	1	6	-1	1
HOMEZ	ensembl_havana	14	608119	1290	550	1653	14	7	-1	4
TGIF2	ensembl_havana	20	607294	3091	237	714	20	9	1	8
RAX	ensembl_havana	18	601881	1978	346	1041	18	9	-1	4
OTX2	ensembl_havana	14	600037	10053	289	870	14	16	-1	7
POU5F1P5	havana	10	?	405	0	937	10	3	-1	1
TGIF1	ensembl_havana	18	602630	9347	401	1206	18	20	1	23
HOPX	ensembl_havana	4	607275	5424	73	222	4	21	-1	15
MEIS1	ensembl_havana	2	601739	9661	390	1173	2	19	1	17
ARGFX	havana	3	611164	790	315	948	17	2	-1	1
MKX	ensembl_havana	10	601332	1640	352	1059	10	13	-1	4
DBX2	ensembl_havana	12	?	435	339	1020	12	6	-1	1
IRX2	ensembl_havana	5	606198	1509	471	1416	5	12	-1	2
RAX	ensembl_havana	18	601881	1978	346	1041	19	15	-1	2
PROX2	ensembl_havana	14	615094	662	592	1779	14	12	-1	3
SHOX2	ensembl_havana	3	602504	4303	331	996	3	19	-1	6
HESX1	ensembl_havana	3	601802	4951	185	558	3	8	-1	3
LBX2	ensembl_havana	2	607164	679	198	597	2	15	-1	6
GBX1	ensembl_havana	7	603354	959	363	1092	7	4	-1	2
DRGX	ensembl_havana	10	606701	901	263	792	10	8	-1	2
MEIS2	ensembl_havana	15	601740	3685	477	1434	15	19	-1	27
SIX5	ensembl_havana	19	600963	2469	739	2220	19	7	-1	4
TPRX1	ensembl_havana	19	611166	457	411	1236	19	8	-1	3
VENTX	havana	10	607158	1913	258	777	8	1	-1	1
ZEB2	ensembl_havana	2	605802	19995	1214	3645	2	18	-1	47
VENTX	havana	10	607158	1913	258	777	13	2	-1	1
PHOX2A	ensembl_havana	11	602753	4205	284	855	11	5	-1	3
LHX6	ensembl_havana	9	608215	1913	363	1092	9	15	-1	12
TPRXL	havana	3	611167	686	0	2276	3	15	1	14
PKNOX1	ensembl_havana	21	602100	3777	436	1311	21	14	1	11
BSX	ensembl_havana	11	611074	923	233	702	11	4	-1	1

Referências

- [Agg15] Charu C. Aggarwal. *Data classification : algorithms and applications*. 2015.
- [ARF] Arff. <http://www.cs.waikato.ac.nz/ml/weka/arff.html>. Acedido em: 2017-04-13.
- [BC12] Leo Breiman e Adele Cutler. Breiman and Cutler's random forests for classification and regression. *Package 'randomForest'*, page 29, 2012. URL: <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>, arXiv:1609-3631, doi:10.5244/C.22.54.
- [Bea09] Alan Beaulieu. *Learning SQL*. 2009. URL: <http://books.google.com/books?id=1PgCCVryjOQC>, arXiv:arXiv:1011.1669v3, doi:10.1017/CBO9781107415324.004.
- [BIO] Biodb api. http://biodb.jp/help/ws_en.html. Acedido em: 2017-04-21.
- [CLMW11] Peter Cooper, Melissa Landrum, Ilene Mizrahi e Jane Weisemann. Entrez Sequences Quick Start. *Ncbi*, (Md):1–9, 2011.
- [CRI] Crisp-dm. <http://www.sv-europe.com/crisp-dm-methodology/>. Acedido em: 2017-07-20.
- [Dea05] John Deacon. Model-view-controller (mvc) architecture. *Computer Systems Development*, pages 1–6, 2005.
- [FBC08] Jeff Forcier, Paul Bissex e Wesley Chun. *Python Web Development with Django*. Addison-Wesley Professional, 1 edition, 2008.
- [FCK02] Lilian T. C. França, Emanuel Carrilho e Tarso B. L. Kist. A review of DNA sequencing techniques. *Quarterly Reviews of Biophysics*, 35(02), 2002. doi:10.1017/S0033583502003797.
- [GENa] Gene expression and regulation. university of leicester. <http://www2.le.ac.uk/departments/genetics/vgec/schoolscolleges/topics/geneexpression-regulation>. Acedido em: 2016-12-04.
- [genb] genee. Acedido em: 2017-01-28. URL: <http://biosocialmethods.isr.umich.edu/epigenetics-tutorial/epigenetics-tutorial-gene-expression-from-dna-to-protein/>.
- [HPK11] Jiawei Han, Jian Pei e Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

REFERÊNCIAS

- [HT99] Andrew Hunt e David Thomas. *The Pragmatic Programmer*. 1999. doi:10.1093/carcin/bgs054.
- [IN09] Tadashi Imanishi e Hajime Nakaoka. Hyperlink management system and ID converter system: Enabling maintenance-free hyperlinks among major biological databases. *Nucleic Acids Research*, 37(SUPPL. 2), 2009. doi:10.1093/nar/gkp355.
- [JMF99] A. K. Jain, M. N. Murty e P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999. URL: <http://portal.acm.org/citation.cfm?doid=331499.331504>, arXiv:arXiv:1101.1881v2, doi:10.1145/331499.331504.
- [KFT⁺16] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato e Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(November 2016):1–15, 2016. arXiv:1611.06654, doi:10.1093/nar/gkw1002.
- [KP98] Ron Kohavi e Foster Provost. Glossary of Terms. *Machine Learning*, 30(2-3):271–274, 1998. doi:10.1023/A:1017181826899.
- [Mad12] T. Soni Madhulatha. AN OVERVIEW ON CLUSTERING METHODS. *IOSR Journal of Engineering*, 02(04):719–725, 2012. arXiv:1205.1117, doi:10.9790/3021-0204719725.
- [MOPT11] Donna Maglott, Jim Ostell, Kim D. Pruitt e Tatiana Tatusova. Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Research*, 39(SUPPL. 1), 2011. doi:10.1093/nar/gkq1237.
- [MRS08] Christopher D Manning, Prabhakar Raghavan e Hinrich Schütze. Introduction to Information Retrieval. 2008, 1(c):496, 2008. arXiv:0521865719 9780521865715, doi:10.1109/LPT.2009.2020494.
- [NGS] Next generation sequencing. <http://www.atdbio.com/content/58/Next-generation-sequencing>. Acedido em: 2016-12-17.
- [Wol13] Jochen B W Wolf. Principles of transcriptome analysis and gene expression quantification: An RNA-seq tutorial. *Molecular Ecology Resources*, 13(4):559–572, 2013. arXiv:arXiv:1301.5277v2, doi:10.1111/1755-0998.12109.
- [WZWD14] Xindong Wu, Xingquan Zhu, Gong Qing Wu e Wei Ding. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):97–107, 2014. doi:10.1109/TKDE.2013.109.
- [YAA⁺16] Andrew Yates, Wasiu Akanni, Amode et al. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, 2016. doi:10.1093/nar/gkv1157.
- [YBK⁺15] Andrew Yates, Kathryn Beal, Keenan et al. The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics*, 31(1):143–145, 2015. doi:10.1093/bioinformatics/btu613.