FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Portal Web para enriquecimento de informação Genómica e Proteómica

David Vanhuysse

Mestrado Integrado em Engenharia Informática e Computação

Orientador: Rui Camacho

Janeiro de 2017

Portal Web para enriquecimento de informação Genómica e Proteómica

David Vanhuysse

Mestrado Integrado em Engenharia Informática e Computação

---

27 de Janeiro de 2017

Resumo

Atualmente a quantidade de informação disponibilizada na Internet, no domínio da biologia molecular é enorme. As áreas da Genómica e da Proteómica não são excepção. São enumeras as Bases de Dados acessíveis na Web com informação importante para os estudos nestas áreas. O facto de existir muita informação pode ser bom pelo facto de termos imensos sítios Web onde procurar e haver com bastante frequência informação para aquilo que procuramos mas tem, como contrapartida, o facto de não ser simples localizar informação relevante e de poder haver muita informação redundante. Ascresce ainda que é frequente os identificadores das entidades biológicas (genes, proteínas, por exemplo) serem diferentes entre sítios Web diferentes.

Existem diversos sítios Web com informação relevante para os estudos em genómica e proteómica. Cada um deles tem o seu formato e a forma como se faz o acesso aos dados em cada um deles varia de sítio Web para sítio Web. Muitos destes sítios Web disponibilizam APIs (Application Programming Interface), o que permite o acesso à informação por aplicações de software, enquanto que outros guardam toda a sua informação em Bases de Dados e mostram o conteúdo apenas nas páginas HTML. Acontece que começa a ser cada vez mais importante num estudo biológico reunir informação cada vez mas abrangente e diversa. Esta necessidade obriga à consulta simultânea de cada vez mais repositórios para o mesmo estudo o que dificulta enormemente o trabalho dos biólogos. Ferramentas informáticas podem com facilidade mitigar esta dificuldade.

Acresce ainda que muitas vezes os investigadores têm que estudar uma grande quantidade de genes (resultado, por exemplo que uma sequenciação de um genoma) ou de proteínas (prever, por exemplo, nas posições da estrutura linear de aminoácidos onde irão surgir hélices – estruturas secundárias). Nestes casos que envolvem uma grande quantidade de genes ou proteínas, os métodos automáticos são extremamente valiosos. Não só pela rapidez na obtenção de resultados mas porque podem usar variados tipos de informação (recolhida na Web) para ajudar o especialista a agrupar uma grande quantidade de genes, produzindo uma descrição para cada grupo, ajudando assim a entender o fenómeno que gerou a lista de genes (ex: saída de uma análise de sequenciação (RNAseq)).

Para tal foi elaborado um Portal Web que facilita nesta tarefa de recolha da informação e ao mesmo tempo fazer este agrupamento dos genes e proteínas de forma racional. Assim sendo é feita uma recolha e conversão de referências automática gerando datasets de modo a estes depois serem analisados através de técnicas de Data Mining.

Palavras-chave: Genómica, Proteómica , Tecnologias WEB

Abstract

Currently the amount of information in the field of molecular biology is huge. The areas of Genomics and Proteomics are no exception. There are many accessible databases on the Web with important information for studies in these areas. The fact that there is too much information can be good because of huge websites terms where to look and always be information to what we seek and bad for not being simple access to information and can be a lot of redundant information, and sometimes the identifiers the items of information (genes, proteins, for example) are different for the same item in different websites.

There are several websites with information relevant to studies in genomics and proteomics. Each has its shape and the way it makes access to the data in each varies from website to website. Many of these websites offer APIs (Application Programming Interface), which allows access to information applications software, while others keep all your information in databases and show content in HTML pages. These various ways that we can refer to the information on genomics and proteomics make it difficult to access software applications, in turn hampering the work of biologists.

Furthermore, sometimes researchers have studying a large number of genes (resulting, for example a sequencing a genome) or protein (to provide, for example, oas positions of the linear structure of amino acids which will arise propellers - secondary structures). In these cases that involve a grid amount of genes or proteins, automatic methods are valuable extremanete. Not only the speed in achieving results but because they can use various types of information (collected on the Web) to help the expert to group a large number of genes or help in explaining where appear the secondary structures in proteins.

To do this will produce a Web portal that facilitates this collection task information while making this grouping of genes and proteins in a rational way, without the existence of multiple identifiers for the same item. The website will enable researchers to add new repositories containing API's or repositories that do not have any API.

Keywords: Genomics, Proteomics, Website, Web Tecnology

Agradecimentos

À minha família por todo o apoio que me deram ao longo desta dissertação.

Ao meu orientador, o professor Rui Camacho (FEUP), pela orientação que me foi dando ao longo do desenvolvimento da dissertação, por ter estado sempre disponível e por me ter esclarecido as dúvidas que me foram surgindo. Agradeço também pela preocupação com o meu trabalho e a utilidade que demonstrou que foram fundamentais para o sucesso deste projeto.

Aos meus colegas que estiveram comigo numa das etapas mais importantes e pelas sugestões que me foram dando ao longo deste período.

Conteúdo

Lista de Figuras

Lista de Tabelas

Abreviaturas e Símbolos

API	Application Programming Interface
CMD	Command Prompt
CSS	Cascading Style Sheets
DNA	DeoxyriboNucleic Acid
EUA	Estados Unidos da América
FTP	File Transfer Protocol
GUI	Graphical User Interface
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
ID	Identificador
JSON	JavaScript Object Notation
KEGG	Kyoto Encyclopedia of Genes and Genomes
NCBI	National Center for Biotechnology Information
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
RNA	Ribonucleic Acid
WWW	World Wide Web
XML	Extensible Markup Language

Capítulo 1

Introdução

Neste primeiro capítulo é descrito o contexto do trabalho de investigação, o projeto desenvolvido, a sua motivação e os objetivos que deve atingir.

1.1 Contexto

Atualmente na Biologia Molecular, o *Gene Enrichement Analysis* é uma operação utilizada frequentemente pelos biólogos. Esta operação pretende dar um significado biológico a um grupo de genes (por exemplo em experiências de *RNA-seq* e após determinação da expressão génica -*gene expression*). Nestes casos a recolha automática de informação sobre proteínas associadas, *pathways* em que participam, domínios existentes nessas proteínas, tecidos em que ocorrem, interações com outras proteínas, é extremamente valiosa para o entendimento biológico do problema em estudo.

No entanto, a quantidade de informação e o número de repositórios disponíveis na Internet, no domínio da Biologia Molecular, é enorme. Acontece também, e frequentemente, não existir uma normalização nem do formato de armazenamento da informação nem do tipo de acesso a essa informação. É habitual os identificadores das entidades biológicas (genes, proteínas, por exemplo) terem identificadores diferentes para o mesmo item em sítios *web* diferentes! Além deste problema dos identificadores, o acesso a esta informação nos diversos sítios Web é feito de forma diferente. Existem sítios *Web* que disponibilizam *APIs* (*Application Programming Interface*) de modo a permitir acesso à informação por aplicações de software enquanto que noutros sítios a informação é guardada em Bases de Dados e mostrada em páginas *Web* escritas em *HTML*, dificultando assim o acesso por aplicações de software.

1.2 Projeto

O projeto "Portal Web para Enriquecimento da informação genómica e proteómica", visa o desenvolvimento de um sítio *Web* que permita recolher e analisar informação da genómica e proteómica para problemas específicos que envolvem grandes quantidades de genes ou proteínas. O portal deve permitir a recolha de informação em diversos websites tais como: *Genbank*, *Ensembl genomes, Kegg* para os genes e *PDB* e *SwissProt* para as proteínas. Além desta recolha, e como muitas das vezes os identificadores nem sempre são os mesmos para um mesmo gene (diferenças entre os sítios *Web* dos EUA e os do Reino Unido), foi implementada uma funcionalidade de tradução. Além disto permitirá ao utilizador adicionar novos repositórios de genes ou proteínas. Um aspecto valioso deste trabalho vai ser a possibilidade de agrupar de forma automática grandes quantidades de genes (através de técnicas de *clustering conceptual*) e dar uma descrição/explicação para cada um dos grupos. Para problemas de proteómica a informação recolhida será usada em problemas de previsão das estruturas secundárias.

1.3 Motivação e Objectivos

Como já foi referido anteriormente, a quantidade de informação na área da Biologia Molecular é enorme e a forma como é disponibilizada nem sempre é feita da mesma forma.

Para superar tais problemas, foi desenvolvida uma plataforma Web que agrega informação relevante para cada tipo de estudo e ao mesmo tempo normaliza o formato de armazenamento da informação. Assim este portal *Web* irá ajudar os biólogos no seu trabalho, tendo assim a sua vida facilitada. Assim sendo, esta aplicação *Web* deve satisfazer as seguintes características:

- Desenvolvimento e manutenção incrementais, isto é, deve ser fácil adicionar um novo repositório ao repertório de repositórios conhecidos;
- Deve permitir adicionar novos repositórios que tenham *APIs*, mas também, sempre que possível, adicionar repositórios que não disponham de *APIs*.
- Deve verificar automaticamente se há alterações nos formatos ou *APIs* dos repositórios conhecidos e reportar ao administrador a ocorrência de alterações.
- Deve ser possível aos utilizadores fazer sugestões para adição de novos repositórios ou novas funcionalidades.
- Quando são estudados genomas de espécies novas ou quando há ainda pouco conhecimento sobre determinado grupo de genes, será útil também agrupar a informação de forma "racional".

Sendo essencial cumprir estas funcionalidades, surgem então diversas questões:

- "Que algoritmos de clustering devem ser utilizados de modo a agrupar os genes/proteínas para os/as quais procura informação?";
- "Perceber até que ponto existem padrões nas *APIs* dos diversos sítios *Web* e se estes podem ser tratados da mesma forma?";
- "Quais as tecnologias adequadas a este tipo de sítio Web a desenvolver?";

1.4 Estrutura da Dissertação

Para além da introdução, esta dissertação contém mais 3 capítulos. No Capítulo 2, é descrito o estado-da-arte, sendo referidos vários métodos e técnicas para agrupar os genes/proteínas. No Capítulo 3, é referida a implementação, sendo dada uma explicação detalhada do desenvolvimento do Portal Web. No Capítulo 4 estão descritos casos de estudo e apresentados os resultados. O Capítulo 5 consiste nas conclusões e o trabalho futuro.

Capítulo 2

Data Mining e Tecnologias Web para análise de Informação de Biologia Molecular

Este capítulo apresenta os conceitos de Biologia Molecular essenciais para a compreensão do trabalho realizado. Contém uma breve explicação de o que é a genómica e a proteómica, é feita uma revisão dos vários repositórios de bases de dados que estão disponíveis para genes e proteínas, apresentadas as técnicas de *Data Mining* relevantes e as suas ferramentas para fazer o tratamento dos dados. São ainda apresentadas as tecnologias Web e as bases de dados que poderão ser utilizadas para o desenvolvimento do portal web.

2.1 Conceitos de Biologia Molecular

2.1.1 Genómica

A genómica é uma área de Conhecimento dedicada ao estudo dos genomas de um organismo. Um genoma é o "índice completo" do ADN que está presente numa célula ou num organismo, isto é, contém a identificação de todos os genes conhecidos de uma espécie bem como informação relacionada com cada gene, tal como a sua localização nos cromossomas. "A Genómica envolve o estudo de processos intragenómicos tais como a epistase, a heterose e a pleiotropia assim como as interações entre locus e alelos dentro do genoma. Os campos da biologia molecular e da genética são estudos relacionados principalmente com o estudo do papel e da função dos genes, um assunto principal na pesquisa biomédica de hoje."[Genomics]

Assim sendo, a Genómica envolve o estudo de todos os genes, do ADN, do mRNA e dos mecanismos em que os genes estão envolvidos ao nível celular ou do tecido.

O termo Genómica foi inventado em 1986 por Tom Roderick, um geneticista do Laboratório de Jackson em Maine, durante uma reunião sobre o mapeamento do genoma humano.

2.1.2 Proteómica

A proteómica realiza o estudo das proteínas que estão expressas numa célula, tecido ou organismo. A proteómica pode identificar, fazer categorias e classificar as proteínas em relação às suas funções e interações estabelecidas entre elas. Hoje em dia, a proteómica está a ser aplicada na identificação de novos marcadores para o diagnóstico de doenças, identificação de novos medicamentos e determinação de proteínas envolvidas na patogénese de doenças.

O proteoma é então complementar ao genoma, uma vez que os genes podem ser transcritos em RNA (no núcleo da célula) e o RNA é traduzido em proteínas (nos ribossomas – no citoplasma).

2.2 Repositórios Web para Genómica e Proteómica

Na área da genómica e proteómica existe uma grande quantidade de sítios Web com informação sobre genes e proteínas. Nesta secção apresentamos alguns dos mais relevantes sítios para a comunidade Genómica que contém informação sobre proteínas e genes.

2.2.1 Genes

Para a recolha de informação sobre genes existem dois importantes sítios *web* que disponibilizam informação.

O *Genbank* é uma base de dados de acesso aberto. Esta base de dados é mantida pelo Centro Nacional de Informações sobre Biotecnologia (NCBI), nos Estados Unidos. O *Genbank* é a base de dados mais importante e é onde são feitas a maior parte das pesquisas em todos campos da Biologia envolvendo genes. Esta base de dados continua a crescer exponencialmente ao longo dos anos. Para podermos aceder a esta base de dados o *GenBank* dispõem de uma API, o que facilita o seu acesso.

O *Genbank* disponibiliza a informação em vários formatos- Os mais frequentes são: *XML*, asn.1 e *genbank format*. Um exemplo deste formato é apresentado na figura 1.

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
```

Figura 1: Formato dos Dados do Genbank, em *genbank format*

Outro repositório muito importante para a comunidade científica é o *Ensembl genomes*. Este projeto é do Instituto Europeu de Bioinformático *European Bioinformatics Institute* - EBI, e foi lançado em 2009. A maior parte dos dados do *Ensembl genomes* são armazenados em bases de dados relacionais *MySQL* e podem ser acedidos pela *API Ensembl Perl*, máquinas virtuais ou online. É possível extrair os dados em formatos *txt*, *json* ou *XML*.

2.2.2 Proteínas

Para a recolha de informação sobre proteínas, à semelhança dos genes, existem também dois grandes sítios Web que disponibilizam informação.

O PDB, *Protein Data Bank*, é uma base de dados para dados de grandes moléculas biológicas, como as proteínas. O PDB é mantido pela organização *Worldwide Protein Data Bank, wwPDB*. O PDB é um recurso fundamental nas áreas de biologia estrutural, sendo que os dados com a estrutura presente no PDB são apresentados nas maiores revistas científicas e utilizados pela maior parte dos cientistas.

Os formatos que podemos extrair a informação desta base de dados são PDB *format* ou *XML*.

O *Swiss-Prot* é uma base de dados que combina as informações extraídas da literatura científica e análise computacional. O objetivo é fornecer todas as informações relevantes sobre uma proteína em particular. Podemos ver na figura 2 um exemplo da estrutura dos dados do PDB.

Os formatos de dados do *Swiss-Prot* são em *XML*.

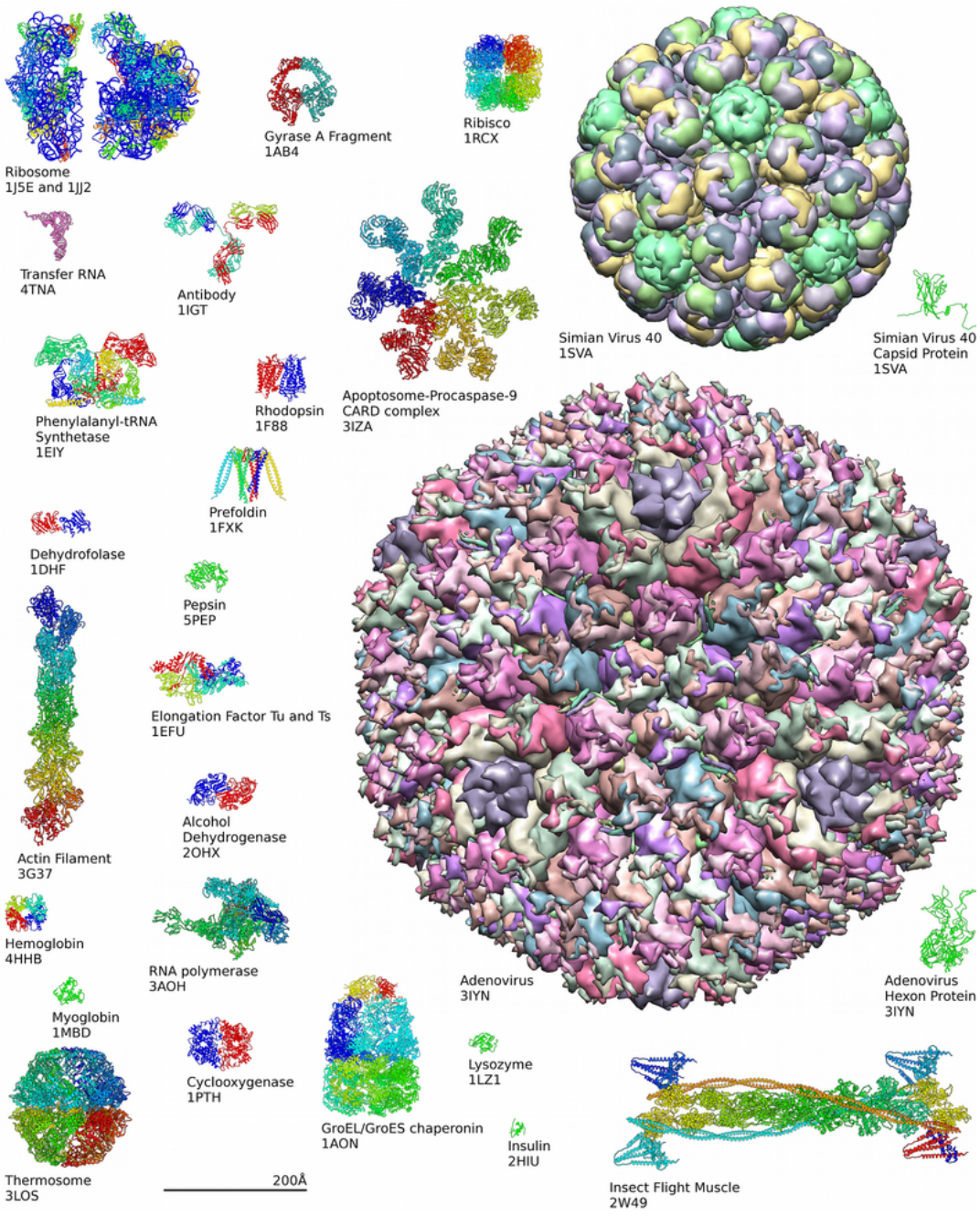


Figura 2: Exemplo de estrutura de proteínas disponíveis no PDB

2.2.3 Gene Ontology

Gene Ontology[1]é uma importante iniciativa de bioinformática para unificar a representação dos genes e atributos e relações em todas as espécies. Mais especificamente, o projeto visa manter e desenvolver os atributos dos genes, anotar os genes e fornecer ferramentas para facilitar o acesso aos dados.

2.2.4 Necessidade de Informática na Biologia, Bioinformática

O genoma humano é constituído por três biliões de pares de bases, que codificam aproximadamente 20.000 a 25.000 genes. Todavia, o genoma por si só não tem nenhuma utilidade, a menos que as localizações e relações dos genes possam ser identificadas. Uma opção para que isto aconteça é a anotação manual, através da qual uma equipa de cientistas tentam localizar genes utilizando dados experimentais juntamente com outro conhecimento recolhido em revistas científicas e bases de dados. O que torna esta tarefa bastante lenta e morosa. A alternativa a isto tudo é utilizar o computador e o seu poder de processamento identificar o padrão complexo de correspondência entre a proteína e o *DNA*, a qual é denominada de anotação automatizada.

2.2.5 Kegg

O *Kyoto Encyclopedia of Genes and Genomes* (KEGG)[2], é um projeto que conta com uma diversa coleção de bases de dados que lidam com genomas, doenças, *pathways* biológicas e substâncias químicas. Esta plataforma é muito utilizada para diversas pesquisas na área da bioinformática, incluindo análise de dados na genómica.

O projeto foi iniciado em 1995 por Minoru Kanehisa na Universidade de Kyoto, ao abrigo do programa japonês *Human Genome Program*. Prevendo-se a necessidade de existir um recurso computorizado que pudesse ser utilizado para a interpretação biológica dos dados da sequência do genoma, o projeto começou por desenvolver a base de dados do *KEGG PATHWAY*. Esta é uma base de dados com os mapas, contendo o conhecimento experimental sobre redes metabólicas e outras funções da célula e do organismo. Um exemplo pode ser visto na Figura 3. Cada um destes mapas contém uma rede de interações e reações moleculares que são utilizados para fazer a associação dos genes no genoma a, principalmente, proteínas. Assim surgiu então a chamada análise de mapeamento de via Kegg, pelo qual o conteúdo de gene no genoma é comparado com a base de dados *KEGG PATHWAY* para examinar que vias e funções associadas são prováveis de ser codificadas no genoma.

[1] http://www.geneontology.org/
[2] https://www.ncbi.nlm.nih.gov/pubmed/10592173

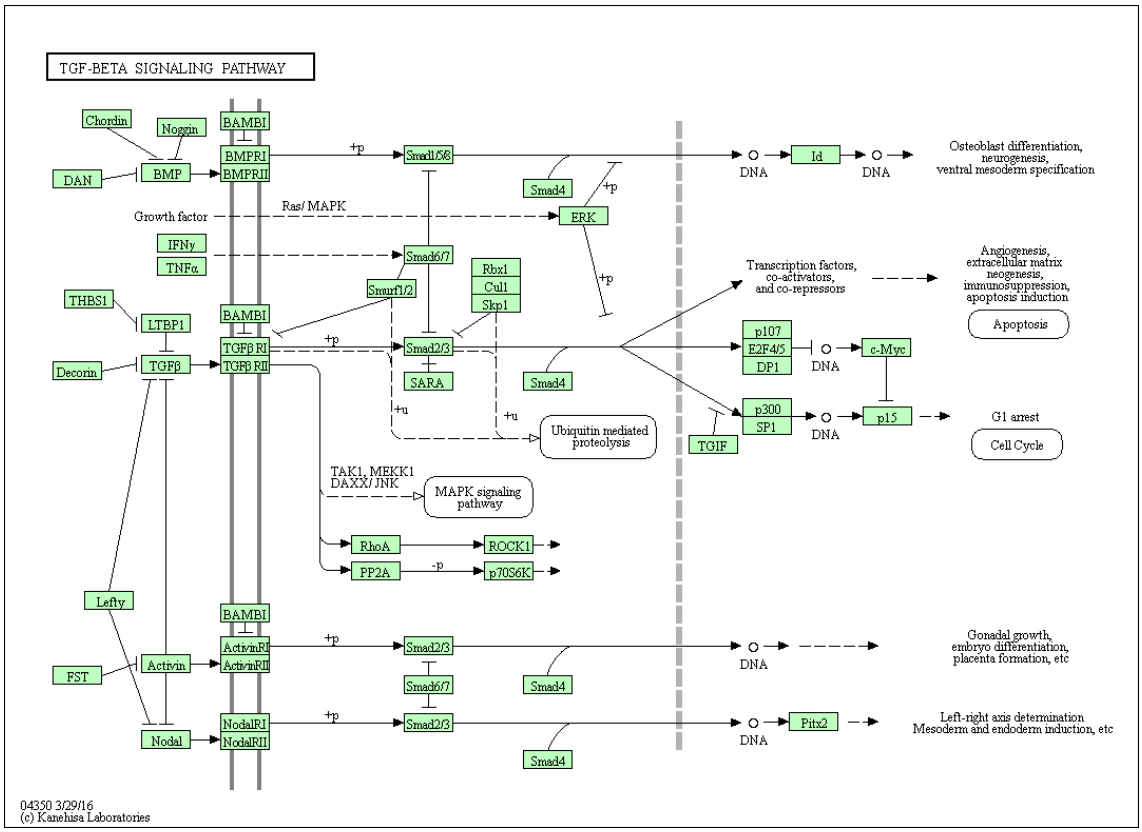


Figura 3: Exemplo de uma Pathway do Kegg

Assim sendo, o Kegg[3] é um repositório de valiosa informação do sistema biológico, em que as suas bases de dados estão categorizadas em sistemas, genómica, química e informações de saúde.

- Informações sobre sistemas:
  - PATHWAY: mapas para funções celulares e orgânicas
  - MODULE: módulos ou unidades funcionais de genes
  - BRITE: classificações hierárquicas de entidades biológicas
- Informações genómica:
  - GENOME: genomas completos
  - GENES: genes e proteínas no genoma completo
  - ORTHOLOGY: grupos de genes ortólogos nos genomas completos
- Informações químicas:
  - COMPOUND, GLYCAN: compostos químicos e glicanos
  - REACTION, RPAIR, RCLASS: reações químicas

[3] http://www.kegg.jp/kegg/rest/keggapi.html

- ENZYME: nomenclatura enzimática

- Informações de saúde:
  - DISEASE: doenças humanas
  - DRUG: fármacos
  - ENVIRON: drogas brutas e substâncias relacionadas com a saúde

2.2.6 NCBI

O *National Center for Biotechnology Information* (NCBI)[4], faz parte da Biblioteca Nacional de Medicina do Estados Unidos. Está localizada em Bethesda, Maryland e foi fundada em 1988.

O NCBI contém um série de bases de dados bastante importantes para a biotecnologia,

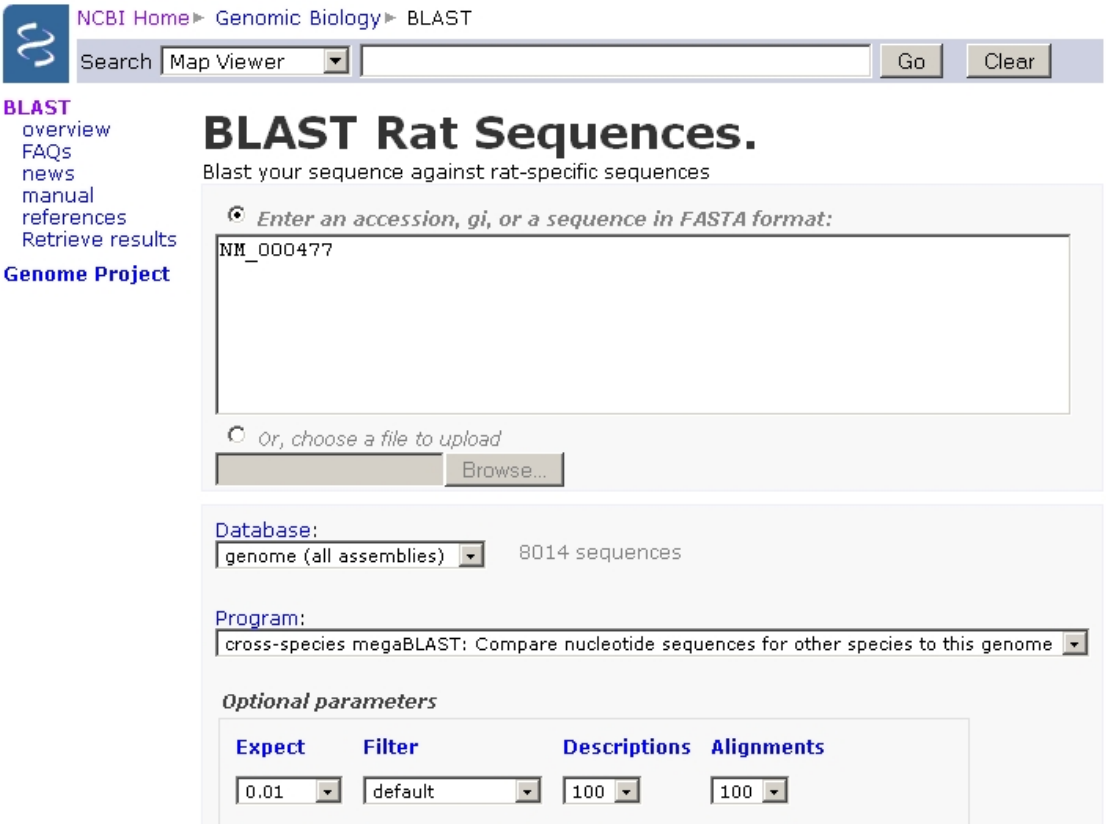


Figura 4: Exemplo de uma ferramenta do NCBI, o Blast

biomedicina e é também um recurso fundamental para a bioinformática. As suas principais bases de dados são o GenBank, focado para sequências de DNA e PubMed, uma base de dados

[4]https://www.ncbi.nlm.nih.gov/books/NBK21105/#ch1.History

bibliográfica para a literatura biomédica e outras bases de dados que incluem o NCBI Epigenomics. Estas bases de dados estão todas disponíveis através do motor de busca Entrez.

O *GenBank* está disponível desde 1992, e interliga com diversos laboratórios e outras bases de dados de sequências como o EMBL, *European Molecular Biology Laboratory* (EMBL) e o DDBJ, *DNA Data Bank of Japan*. O *GenBank* é a base de dados mais importante do NCBI e é onde são feitas a maior parte das pesquisas em todos os campos biólogos. Devido a ser uma grande base de dados, o NCBI disponibiliza ferramentas de software através de navegação WWW ou FTP. Uma destas ferramentas é o *Basic Local Alignment Search Tool* (BLAST), que é um programa de busca de similaridade de sequência DNA, fazendo comparações de uma sequência com a base de dados do *GenBank*, ver figura 4.

O *Entrez Global Query Cross-Database* é usado no NCBI para todas as principais bases de dados. O *Entrez* é um sistema de indexação e ao mesmo tempo um sistema de recuperação com dados de várias fontes para a investigação biomédica. Assim, o objetivo do Entrez é integrar os dados das diversas bases de dados e vários formatos num modelo de informação uniforme.

2.2.7 Ensembl

O *Ensembl*[5] é um projeto científico desenvolvido em conjunto entre o *European Bioinformatics Institute* e o *Wellcome Trust Sanger Institute* (WTSI), que foi lançado em 1999.

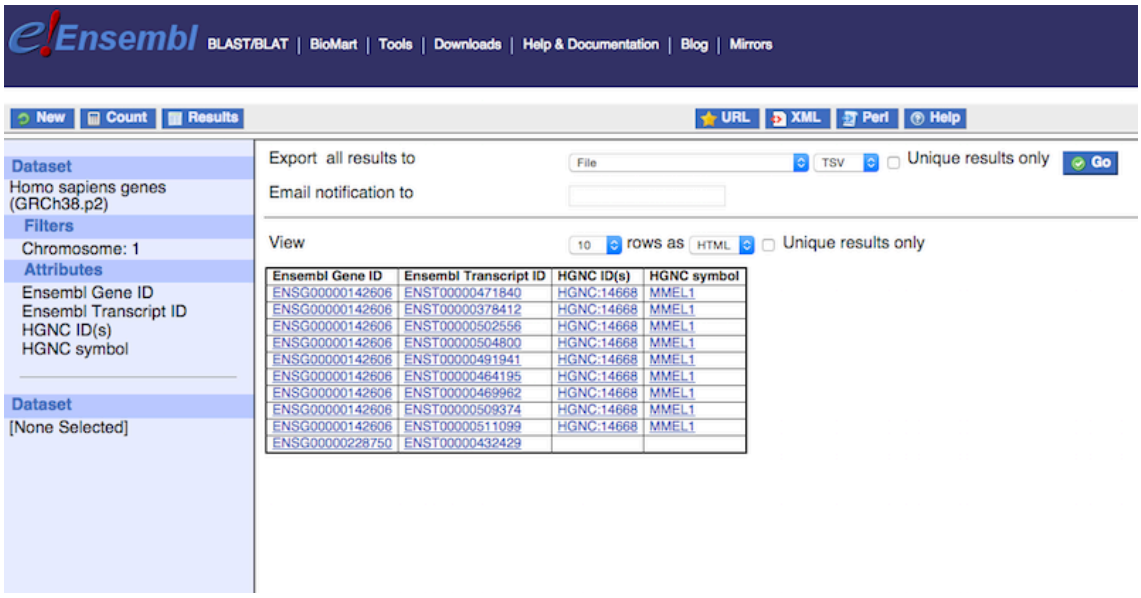


Figura 5: Exemplo de Identificadores do Ensembl

O objetivo deste projeto é ser um recurso centralizado para geneticistas, biólogos moleculares e

[5] https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq1064

outros investigadores que estudam os genomas da nossa própria espécie, *Homo Sapiens*, e de outros vertebrados e organismos-modelo. O *Ensembl* é também um dos vários navegadores do genoma para a recuperação de informações genómicas, o que o torna numa base de dados bastante similar às do NCBI.

Os dados dos genes são armazenados numa base de dados MySQL, e estão disponíveis gratuitamente para pesquisas. Todos os dados estão disponíveis para download e também para acesso remoto. Um exemplo é mostrado na Figura 5.

2.2.8 API Kegg

A API disponibilizada pelo kegg é bastante simples e intuitiva. Permite-nos facilmente listar e encontrar a informação referente a um gene que esteja disponível na base de dados do kegg.

URL form

```
http://rest.kegg.jp/<operation>/<argument>[/<argument2>][/<option>]

<operation> = info | list | find | get | conv | link
<argument> = <database> | <dbentries>
```

Database

```
<database> = KEGG database including KEGG organism (see Table 1)
```

Figura 6: Exemplo de um pedido à API do Kegg

Para utilizar esta REST-style KEGG API, são feitas chamadas HTTP, e pode-se escolher a operação que queremos, bem como o argumento e finalmente a base de dados.

Na API do Kegg, cada entrada da base de dados é definida por:

- db:entry, onde "db" é o nome da base de dados e "entry" é o número da entrada ou número de acessos que é atribuído na base de dados.

O output, ou seja o resultado após a chamada à API, está num formato de texto:

- Texto delimitado por tab retornado das chamadas list, find, conv and link;

- ficheiro no formato da base de dados retornado da chamada get;

- Retornada uma mensagem de texto da chamada info.

Após as chamadas efetuadas à API pode ser retornado um destes três estados HTTP:

- Código 200: A operação foi realizada com sucesso;

- Código 400: O pedido não está correto(erros de sintaxe, base de dados errada);

- Código 404: Não encontrado.

2.2.9 API NCBI

A API do NCBI-BLAST permite executar pesquisas remotamente, o que dá oportunidade à mudança da execução de pesquisas no servidor web NCBI para um provedor na nuvem, ou vice-versa, com bastante facilidade.

As chamadas à API de URL comum incluem sempre o parâmetro CMD, que pode assumir quatro argumentos diferentes. Para fazer uma submissão de uma pesquisa é utilizado o argumento PUT, GET para verificar o estado de uma submissão ou para recuperar os dados, DELETE para remover uma pesquisa e todos os seus resultados ou DisplayRIDs para listar todos os RIDs no sistema.

O output, ou seja o resultado após a chamada à API, está num formato de xml.

Após as chamadas efetuadas à API pode ser retornado um destes três estados HTTP:

- Código 200: A operação foi realizada com sucesso;

- Código 400: O pedido não está correto(erros de sintaxe, base de dados errada);

- Código 404: Não encontrado.

2.2.10 API Ensembl

O Ensembl utiliza bases de dados relacionais MySQL para armazenar a informação referente aos genes. A API do Ensembl à semelhança do kegg, é também bastante simples e intuitiva. Esta API é escrita em Perl, o que por si também disponibiliza métodos Rest para a recolha da informação.

O Output vem num formato json, o que simplifica bastante o tratamento destes dados e posteriormente a sua inserção na base de dados Mongo, que aceita ficheiros json.

Após as chamadas efetuadas à API pode ser retornado um destes três estados HTTP:

- Código 200: A operação foi realizada com sucesso;

- Código 400: O pedido não está correto(erros de sintaxe, base de dados errada);

- Código 404: Não encontrado.

2.2.11 Conversão de Id's

Um dos problemas de termos muitos websites com informação de genes é o facto de cada website utilizar os seus próprios id's para o mesmo gene. Assim sendo, se soubermos o id de um gene para um determinado website, é necessário fazer a conversão do id para ser pesquisado noutros websites. Para fazer essa conversão existem alguns websites que disponibilizam ferramentas para este fim.

2.2.11.1 DAVID Bioinformatics Resources 6.8

O *Database for Annotation, Visualization and Integrated Discovery* (DAVID), é um recurso bioinformático desenvolvido pelo *Laboratory of Immunopathogenesis and Bioinformatics*(LIB). O DAVID possui diversas ferramentas que tem como objetivo fornecer uma interpretação funcional de grandes listas de genes derivados de estudos da genómica e da

Figura 7: Exemplo de conversão de identificadores de genes no DAVID

proteómica.

De entre todas as ferramentas disponíveis do DAVID, uma delas é a ferramenta de conversão de Id's de genes, que tem disponível uma API (ver exemplo na Figura 6). O formato de entrado da lista de genes a converter é *URI Query String*. Feito então o pedido à ferramenta a resposta vem num formato JSON.

2.2.11.2 BioDB Hyperlink Management System

O BioDB define identificadores ligando-os a Id's de dados nas principais bases de dados de informação de genes e proteínas. O BioDB possui uma API, que permite fazer a submissão de um identificador no formato *URI Query String*/CRUD ou *plain text* e receber a conversão do identificador de acordo com o sistema que foi solicitado num formato JSON.

2.3 Data Mininig

Nesta secção são apresentadas as possíveis técnicas de *data mining* usadas com o objectivo de organizar a informação. *Data mining* é o processo de "extrair conhecimento a partir de grandes quantidades de dados" [DTMIN]. Este é constituído por um conjunto de técnicas e algoritmos que podem ser usados para encontrar certos padrões em grandes quantidades de dados. *Data mining* utiliza técnicas de vários campos, como a inteligência artificial, estatísticas e sistemas de bases de dados. Este objetivo é o de combinar todas estas técnicas e transformar grandes quantidades de dados em informações compreensíveis e úteis.

2.3.1 Clustering

Clustering é uma técnica de *data mining* que transforma um grupo de objetos abstratos em classes de objetos semelhantes. Para além desta formação de grupos, produz uma descrição de cada grupo que ajuda a compreender o agrupamento feito.

Uma das vantagens de utilizar *clustering*, é que este agrupamento sobre a classificação é adaptável a mudanças e ajuda a perceber quais as características que distinguem os diferentes grupos.

Esta técnica de *clustering* vai ser muito útil para fazer o tratamento dos genes, de forma a organizá-los e formar grupos de uma forma mais "racional", sendo mais simples e rápida a forma de os consultar. O *Conceptual Clustering*[6] permite não só determinar os grupos mas também associar a cada grupo uma descrição. Neste domínio da genómica a descrição de cada grupo poderá ser usada pelo especialista para compreender o agrupamento.

2.3.2 Classificação

Classificação é uma técnica de *data mining* que faz a atribuição de itens numa coleção, de modo a formar categorias ou classes. O objetivo da classificação passa por prever com bastante precisão a classe alvo para cada caso. Assim sendo, com a classificação tentamos aprender uma função que seja capaz de mapear os dados que temos em classes pré-definidas. A classificação será utilizada em problemas de previsão das posições onde ocorrem estruturas secundárias das proteínas.

[6] http://medical-dictionary.thefreedictionary.com/Conceptual+clustering

2.4 Ferramentas de Data mining

Atualmente existem várias ferramentas de *data mining* (muitas delas gratuitas), que são na maior parte das vezes personalizáveis, o que torna a sua adaptação a vários problemas fácil.

Em seguida são apresentadas algumas das ferramentas disponibilizadas de *data mining* que poderão ser utilizadas ao longo do desenvolvimento do portal *web*.

2.4.1 Weka

O Weka[7] é uma ferramenta de *data mining*, *opensource*, que implementa vários algoritmos de aprendizagem, permitindo ao utilizador aplicar facilmente esses algoritmos nas tarefas de *data mining*.

Este software foi desenvolvido na Universidade de Waikato, na Nova Zelândia em 1997, e que está ainda em desenvolvimento. O Weka suporta então várias tarefas de *data mining* comuns, como por exemplo o pré-processamento dos dados, a classificação, *clustering*, regressão e visualização de dados. As suas bibliotecas são escritas em Java e permitem uma fácil integração dos seus algoritmos de *data mining* em aplicações e código já existentes.

Além disso, o Weka também permite ser utilizado através de uma linha de comando/terminal ou através de um dos seus vários GUIs, podemos ver um exemplo de GUI do Weka na figura 8. Esta API simples e a sua arquitetura bem estruturada permite ser estendido pelos utilizadores de uma forma fácil, podendo assim ser adicionadas novas funcionalidades.

[7] http://www.cs.waikato.ac.nz/ml/weka/

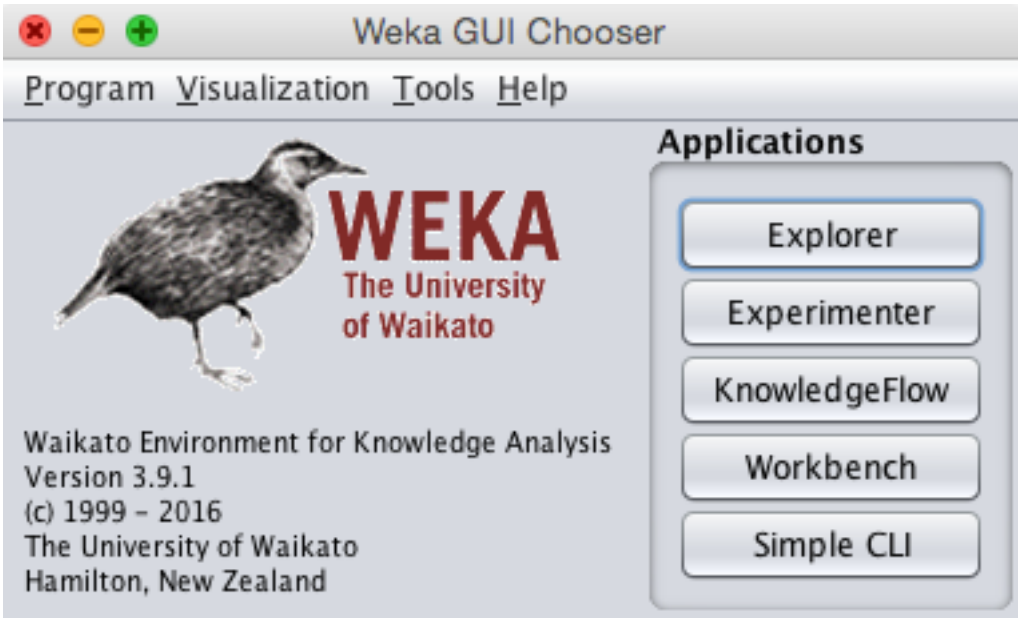


Figura 8: Interface do Weka

2.4.2 RapidMiner

O *RapidMiner* é uma ferramenta para a solução de problemas de *data mining*. É uma das ferramentas mais populares e usadas hoje em dia, as aplicações desta ferramenta abrangem vários domínios, incluindo a educação, a formação, aplicações industrias e pessoais, entre outros. Esta pode ser facilmente estendida através do uso de *plugins*, o que torna o valor desta aplicação elevada. Um dos exemplos destes *plugins* nesta área da bioinformática é o *plugin* de integração entre o *RapidMiner* e o sistema de gestão de fluxo *open-source* Taverna.

O *RapidMiner* é então uma aplicação comercial, sendo que as suas versões principais e anteriores são distribuídas sob um licença de *open-source*, oferecendo uma versão gratuita. Além disso existem múltiplas versões pagas.

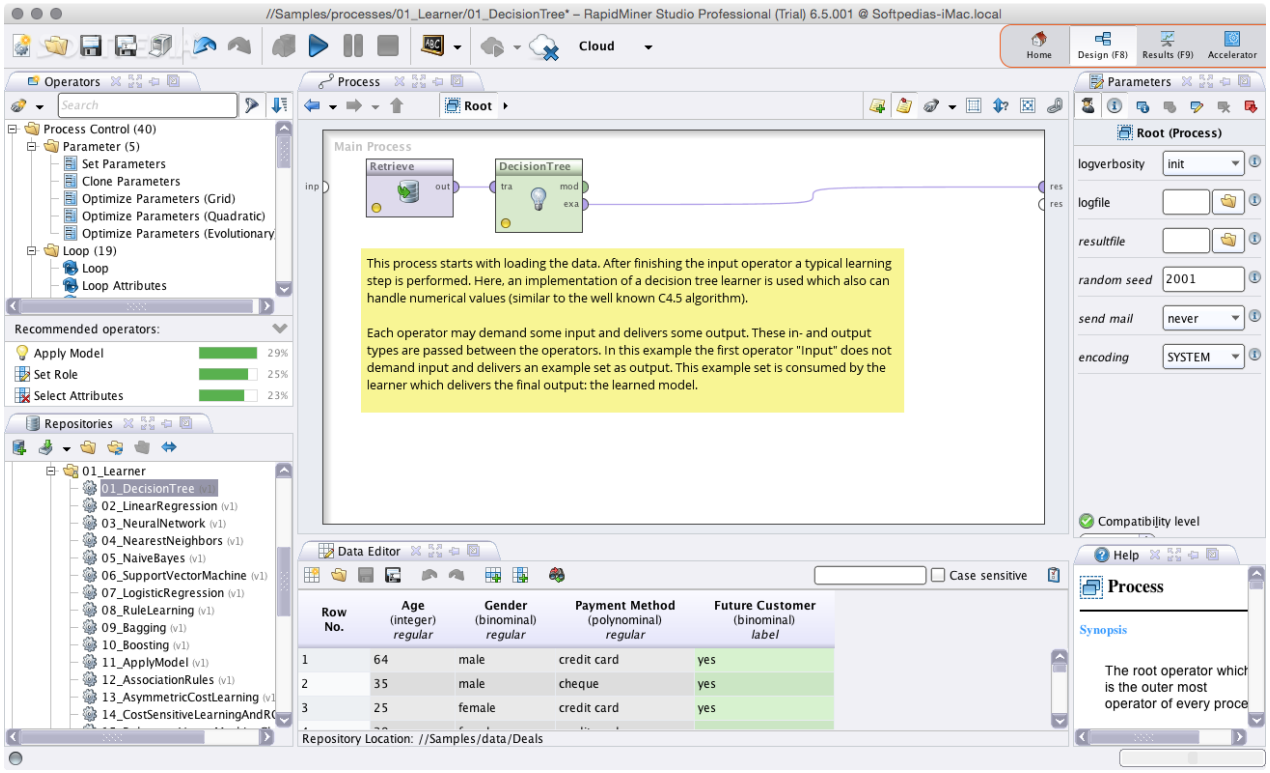


Figura 9: Interface do RapidMiner

2.4.3 R

R é uma linguagem de programação e um ambiente de software para computação e geração de gráficos estatísticos. Desenvolvido na Universidade de Auckland, na Nova Zelândia por Ross Ihaka e Robert Gentleman em 1993, continua atualmente em desenvolvimento ativo. R é Normalmente usado para estatísticas e *data mining*, seja para a análise de dados ou para o desenvolvimento de um novo software estatístico.

R está escrito numa concentração de C, Fortran e R. É possível manipular diretamente os objetos R em linguagens com C, C++, Java e Prolog. R também pode ser usado através da linha de comandos/terminal ou através dos interfaces gráficas como *Deducer*. R fornece várias técnicas de estatísticas e gráficas, incluindo modelagem linear e não-linear, testes estatísticos clássicos, classificação, *clustering*, entre outros.

R e as suas bibliotecas implementam várias técnicas estatísticas e gráficas, incluindo modelação linear e não-linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento e outros. R é facilmente extensível através de funções e extensões, e a comunidade R é notada por suas contribuições ativas em termos de pacotes. Muitas das

funções padrão de R são escritas em R, o que o torna fácil para os utilizadores seguir as escolhas algorítmicas feitas.

2.5 Tecnologias Web e Bases de Dados

Atualmente as ferramentas disponibilizadas para se construir um portal *web* têm vindo a aumentar, sendo possível fazer uma escolha de entre variadas tecnologias. Assim sendo, inicialmente será necessário escolher qual tecnologia usar para guardar os dados. Das diversas bases de dados disponíveis podem ser utilizadas *PostgreSQL* ou *MongoDB*.

PostgreSQL

PostgreSQL é um sistema de gestão de bases de dados, multiplataforma com diversos recursos: Consultas complexas, chaves estrangeiras, controle de concorrência multi-versão, suporte ao modelo híbrido objeto relacional, facilidade de acesso, indexação por texto, entre outros.

MongoDB

MongoDB é uma base de dados orientado a documentos multiplataforma livre e *open-source*. É uma base de dados *NoSQL*, e evita a estrutura das bases de dados tradicionais baseadas em tabela relacional. *MongoDB* utiliza documentos JSON com esquemas dinâmicos, tornando a sua integração em certos tipos de aplicações mais fáceis e rápidas.

Para o desenvolvimento do Website, mais propriamente do *back*-end, tecnologias como *Django* ou *AngularJs* e *NodeJS*.

Django

Django é uma *framework* grátis e *open-source*, escrito em *Python*. O objetivo desta ferramenta é facilitar a criação de websites com base de dados complexas. *Django* tem como princípio não repetir-se, pois promove a capacidade de reutilização do código, para um desenvolvimento mais rápido. Uma das vantagens de ser escrito em *Python* é que, este é usado para arquivos de configurações e modelos de dados.

AngularJS e NodeJS

AngularJS é uma *framework* em JavaScript *open-source*, mantido pela Google, que ajuda na execução de *single-page applications*.

NodeJS é um interpretador de código *JavaScript* que funciona do lado do servidor. O objetivo desta ferramenta é ajudar na criação de aplicações com grande escalabilidade, como é o caso de um servidor *web*, com a possibilidade de manipular milhares de conexões simultaneamente.

AngularJS e NodeJS estão bastantes vezes ligados um ao outro no desenvolvimento de aplicações *web*.

Em termos, da visualização do site por parte dos utilizadores, ou seja o *front*-end, temos disponível BootStrap ou Uikit.

BootStrap

BootStrap é uma *framework web* de *front-end* livre e *open-source* para a criação de websites e aplicações web. Contém HTML e modelos de design baseados em CSS para tipografia, botões, navegação e outros componentes da interface. Além disso possui extensões JavaScript opcionais.

Uikit

Uikit é uma *framework web* de *front-end* para o desenvolvimento de interfaces web rápidas. Uikit oferece uma coleção abrangente de HTML, CSS e componentes JS simples de usar, fácil de personalizar e extensível.

2.6 Conclusões e Resumos

Neste capítulo foi feita uma análise da informação sobre genes disponível na *web*. É feito um levantamento do tipo de dados que temos disponível em cada base de dados. Posteriormente verifica-se quais as ferramentas disponíveis para a análise da informação recolhida. Por fim, é feita uma exposição das tecnologias que podem ser utilizadas no desenvolvimento da aplicação.

Capítulo 3

Implementação

Neste capítulo descrevemos a implementação da solução projetada. Ao longo do capítulo são discutidos alguns detalhes da implementação bem como as tecnologias que foram utilizadas para que a solução fosse concretizada.

3.1 Visão Geral

Neste secção é apresentado uma breve descrição do problema bem como uma visão geral da solução elaborada.

3.1.1 Descrição do problema

Na *Web* temos ao nosso dispor uma enorme quantidade de informação e muitas vezes torna-se uma tarefa árdua e bastante demorada encontrar informação relevante. Este problema ocorre também nos domínios da Biologia Molecular e da Genómica e Proteómica. Além de existir muita informação, esta informação está por vezes repetida.

Além disso, existem vários websites sobre genes que não contém a informação toda sobre um gene, o que obriga em muitos casos, voltar a pesquisar esse mesmo gene noutros *websites*. Existe ainda um problema adicional de sites diferentes têm diferentes identificadores para o mesmo gene. Assim sendo, para efetuar uma pesquisa de um gene com o id de um *website* é necessário fazer a conversão do identificador para o determinado *website* a utilizar.

3.1.2 Visão geral da solução

O Portal Web descrito nesta dissertação consiste num website que visa não só tornar mais rápida a pesquisa de genes como também torná-la numa tarefa simples. Este permite então a pesquisa por genes que são introduzidos pelo utilizador.

Dada uma lista de genes, a aplicação *web* verifica os identificadores dos genes introduzidos e faz a pesquisa deles pelos websites. A informação é recolhida das bases de dados externas e em seguida introduzida na base de dados da aplicação. Ao mesmo tempo que é introduzida na base de dados, a informação também está disponível para visualização pelos utilizadores. Além disso são gerados *datasets* para depois serem analisados atrvés de técnicas de *data mining* de modo a gerar resultados. A figura 10 mostra-nos uma visão geral do website tendo as bases de dados externas de onde são recolhidas informações sobre genes, o portal web e a sua base de dados.

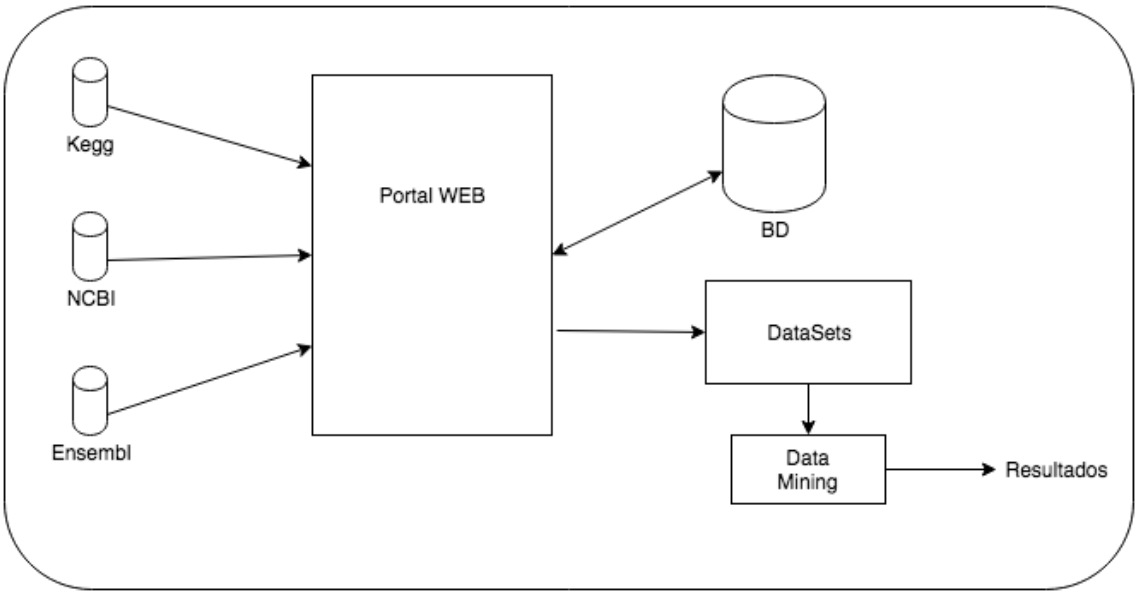


Figura 10: Visão geral do website

3.2 Pesquisa e Análise de Genes

Antes de começar a elaborar a solução, foi necessário fazer uma pesquisa exaustiva sobre a informação referente a genes disponibilizada na *Web*.

Para um dado gene, existem diversos websites que disponibilizam informação. Um dos grandes problemas são os id's que diferem de website para website. De entre os *websites* consultados, os mais relevantes para o trabalho foram o *Ensembl, Kegg e NCBI*.

O KEGG é uma base de dados com informação sobre *pathways* biológicas, doenças e substâncias químicas. Costuma ser utilizado para pesquisas na educação e em bioinformática, incluindo análise de dados em genómica e proteómica. Para tal, a informação referente a cada gene que podemos extrair desta base de dados está relacionada com a caracterização do Kegg em si. Sendo que está então disponível informação como: *pathways*, *orthology*, *module*, *structure*, *aaseq* e *ntseq*.

O *Ensembl* é outra base de dados que tem como principal objetivo ser um recurso centralizados para biólogos moleculares e outros investigadores que estudam os genomas da nossa própria espécie e de outros vertebrados e organismos. Assim sendo, o *Ensembl* é uma importante base de dados de genomas para a recuperação de informações genómicas. Por ser uma base de dados que contém apenas a informação de espécies como o *Homo Sapiens*, e outros organismos modelo, a informação que podemos extrair desta base de dados é mais extensa e detalhada. Para além dos dados que conseguimos extrair do *kegg,* nesta base de dados do *ensembl* é possível obter outro tipo informação para cada gene, como é o caso dos *Transcripts* e dos *Exon*.

O *NCBI, National Center for Biotechnology Information,* inclui uma série de bases de dados, não só relevantes para a biomedicina e biotecnologia como também é um recurso fundamental para diversas ferramentas e serviços da bioinformática. As principais bases de dados são o *GenBank*, uma base de dados bibliográfica para a literatura biomédica, para sequências de *DNA* e *PubMed*; e a base de dados do *NCBI Epigenomics*. Estas bases de dados estão acessíveis online através do seu motor de busca *Entrez*. Assim sendo, a informação que pode ser extraída desta base de dados está ligada com as sequências de *DNA*.

3.2.1 Conversão de id's de genes

Um dos grandes problemas para fazer a recolha de informação sobre um determinado gene, é que nem todos os websites utilizam o mesmo id para um determinado gene. Isto faz com que uma pesquisa completa sobre um gene se torne numa tarefa difícil e demorada. Pois é necessário pesquisar manualmente o gene num website, e para pesquisar esse mesmo gene noutro website diferente é necessário fazer a conversão desse id para o respectivo *website* desejado.

Para tal foi utilizado uma ferramenta de conversão, o biodb.jp, sendo enviada uma lista de genes a converter num formato URI através de um pedido HTTP do tipo *get* enviado para o servidor, que por sua vez envia outro pedido HTTP do tipo get para a API do biodb.jp, sendo devolvido uma lista com os id's dos genes convertidos.

3.3 Recolha de informação sobre Genes

Após a análise feita a cada website sobre a informação dos genes, é necessário fazer uma recolha destes dados.

Para a recolha de informação, existem API's que facilitam e tornam mais rápido o acesso a esta informação. Através da API disponibilizada por cada website.

Estas API's servem então como uma camada intermediária entre os esquemas de base de dados subjacentes e programas de aplicações mais específicas. Além disso as API's visam encapsular o *layout* da base de dados fornecendo um acesso de alto nível a tabelas de dados e isola as aplicações de alterações de *layout* de dados.

3.4 Criação e Armazenamento na Base de Dados

3.4.1 Criação da base de dados

De forma a tornar mais rápido o acesso à informação sobre cada gene, é necessário armazena-la numa base de dados local, evitando assim que sejam feitas diversas chamadas às APIs externas nas próximas pesquisas pelo mesmo gene. Além disto, caso as bases de dados destes grandes websites estejam em baixo, o portal web não fica dependente destes para mostrar a informação pretendida pelo utilizador.

Assim sendo, foi criada uma base de dados em mongodb. A opção pela base de dados em Mongo deveu-se ao facto de esta facilitar no acesso à informação, ou seja, como não existem transações ou joins torna a sua consulta simples e mais rápida que as bases de dados baseadas em sql. Além disso Mongo é apropriado para armazenar enormes quantidade de dados.

Foram então criadas 4 coleções na base de dados. A primeira visa armazenar os id's de um gene para cada base de dados. Tem um id geral para o gene e depois contem os id's desse gene para o ensembl, kegg e ncbi.

Na segunda coleção é armazenada toda a informação dos genes que vêm da base de dados do Kegg. A terceira armazena dados do NCBI e a quarta os dados do Ensembl. A figura 11 ilustra a base de dados criada através de um diagrama UML.

3.4.2 Armazenamento na Base de Dados

Os dados vão sendo armazenados na base de dados de acordo com as pesquisas dos utilizadores. Inicialmente a base de dados não contém informação sobre nenhum gene. Quando é feita uma pesquisa no Portal WEB, o utilizador dá uma lista de id's de genes de qualquer website. Ao iniciar para fazer a pesquisa é feita a conversão do id do gene para os outros websites. Após termos os id's do gene para os diferentes *websites*, são feitas chamadas à API de cada *website* com o respetivo id, de forma a obter toda a informação disponível para esse determinado gene. É então imediatamente armazenado os id's de cada website para cada gene na primeira coleção. Em seguida são armazenados os dados de cada gene nas outras três

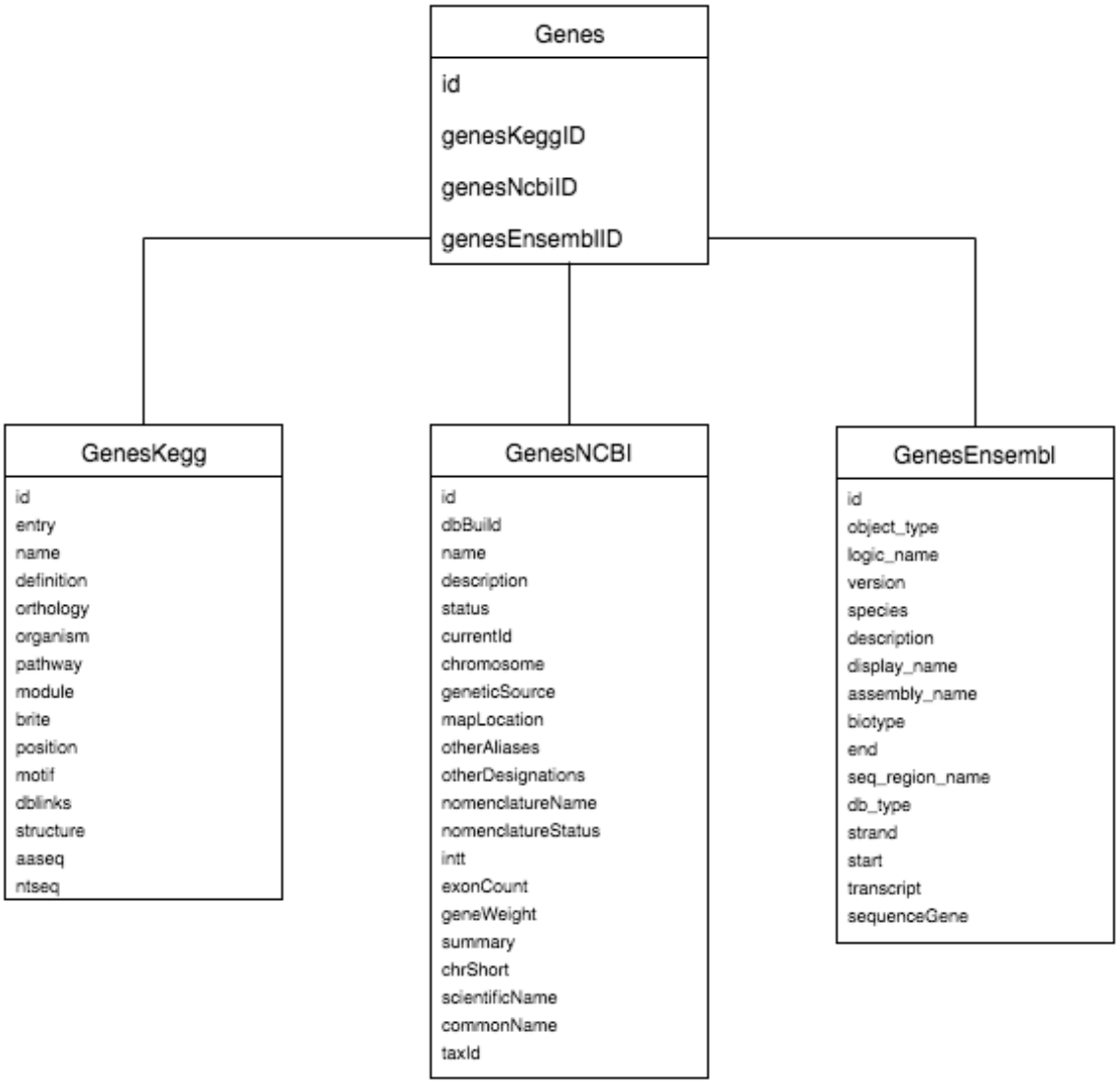


Figura 11: Diagrama UML da base de dados

coleções respetivamente.

3.5 Desenvolvimento do Portal WEB

O desenvolvimento do Portal WEB visa várias etapas até à sua conclusão. As etapas vão sendo descritas nas secções que se seguem.

3.5.1 Aplicação WEB

O Portal WEB é uma aplicação *web* que disponibiliza ao utilizador uma ferramenta de pesquisa de informação de genes. A aplicação *web* inclui dois módulos fundamentais: o *Front-End* e o *Back-End*, que são fundamentais numa aplicação *web* e que são descritos nas próximas secções.

3.5.1.1 Front-End

Numa aplicação *web*, o *front-end* consiste na parte da aplicação que interage com o utilizador e faz a ligação entre o utilizador e a parte do servidor. Para o desenvolvimento do *front-end* deste Portal Web utilizei as tecnologias HTML5, CSS3, Javascript, jQuery(libraria de javascript) e a *framework* Bootstrap.

A interface tem um design simples, com poucas cores e texto o que facilita a utilização do *website*.

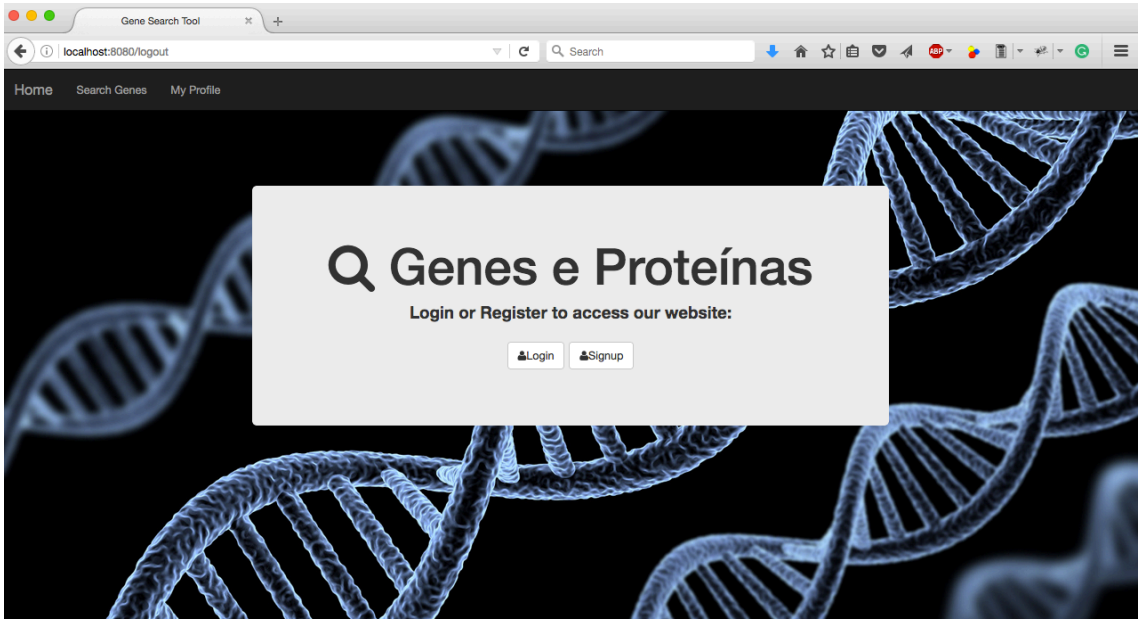


Figura 12: Página principal do Portal Web

3.5.1.2 Back-End

Em certas as aplicações web é necessário fazer uma ligação entre o *front-end* e a base de dados, e o back-end consiste então em fazer essa ligação. Para o desenvolvimento da aplicação foi utilizada a tecnologia *NodeJS*. Foi escolhida esta tecnologia por ser uma ferramenta inovadora e por permitir uma integração com a base de dados *mongoDB*.

Também no *back-end* encontra-se o servidor que se encarrega de fazer as chamadas à API do Kegg, Ensembl e NCBI. As chamadas às API's são feitas por um pedido HTTP do tipo get ao servidor, que por sua vez faz outro pedido HTTP do tipo get às API's do Kegg, Ensembl e NCBI. Os resultados deste pedido nem sempre vem num formato igual. No caso do Ensembl a informação recolhida vem num formato JSON, pelo que não é preciso tratar a informação. Já as chamadas à API do Kegg os resultados vem num formato de texto, pelo que o servidor se encarrega de converter a informação recebida para um formato JSON. As chamadas à API do NCBI os resultados estão num formato xml, pelo que também é feita a conversão para um formato JSON. Esta informação depois de tratada é enviada pelo servidor para a base de dados do Portal Web de modo a ser armazenada. Além disto o servidor também se encarrega de utilizar o programa weka através da linha de comandos para correr algoritmos de clustering.

3.5.2 Registo

Para a utilização do Portal WEB é necessário efetuar um registo no website, de modo a que seja permitida a utilização de todas as funções e ferramentas presentes no Portal WEB.

Para um utilizador efetuar um registo é necessário que insira algumas informações pessoais, como o nome, instituição e um email válido. Além destes dados o utilizador deve escolher uma password. Para proteger o utilizador as passwords nunca são armazenadas numa base dados. O utilizador ao se registar, a *password* é encriptada e é gerado um código que, este sim é guardado na base de dados.

3.5.3 Login

Para utilizar qualquer funcionalidade é necessário que o utilizador esteja autenticado. Para tal, o utilizador depois de já ter feito o seu registo apenas têm de introduzir o email e a password, que a *password* é encriptada e gerado o código para comparar com o código armazenado na base de dados, de modo a proteger as informações dos utilizadores.

3.5.4 Menu

Para facilitar a navegação no website foi feito um barra superior que está presente em todas as páginas. Esta barra permite um fácil acesso a todas as funcionalidades do Portal WEB.

3.5.5 Pesquisa de Genes

Para facilitar a pesquisa de genes na página de pesquisa estão disponíveis várias opções para o utilizador selecionar de acordo com a pesquisa que pretende fazer. O utilizador carrega para o website uma lista com os genes que pretende, e depois seleciona as opções.

Está disponível uma pesquisa de genes por id e por base de dados que pretende.

3.5.5.1 Pesquisa de genes de uma base de dados

Ao clicar para fazer uma pesquisa de genes numa base de dados, o Portal Web verifica em primeiro lugar se a informação existe na base de dados, caso não exista na base de dados, a aplicação faz um pedido HTTP do tipo *get* ao servidor que por sua vez faz outro pedido HTTP do tipo *get* à API da base de dados selecionada. Este processo é feito para cada id de gene introduzido, ou seja, se forem introduzidos cinquenta id's de genes são feitos cinquenta pedidos ao servidor e este faz outros 50 pedidos. Estes pedidos retornam informação com dados dos genes sendo feita uma conversão desta informação para um formato JSON de modo a estar pronto a ser inserido na base de dados MongoDB criada anteriormente.

3.5.5.2 Pesquisa de genes em várias bases de dados

Ao clicar para fazer uma pesquisa de genes em várias bases de dados, o Portal Web verifica em primeiro lugar se a informação existe na base de dados, caso não exista a aplicação começa por fazer um pedido HTTP do tipo *get* ao servidor que por sua vez faz outro pedido HTTP do tipo *get* à API do *biodb.jp*, de modo a converter os id's dos genes introduzidos para as outras bases de dados. Esta informação é recebida num formato de texto. Depois de ter os id's de todos os genes das diferentes bases de dados são feitos três pedidos HTTP do tipo get para cada gene: um pedido à API do kegg, outro à do Ensembl e um último à do NCBI. Efetuados todos estes pedidos o servidor faz a conversão da informação recebida para o formato JSON, no caso do pedido à API do Ensembl não é preciso converter a informação para um formato JSON pois esta já vem nesse formato. Após estas conversões, o servidor encarrega-se de armazenar a informação recolhida na base de dados.

3.5.6 Download de Datasets

Para fazer a recolha de informação dos genes pesquisados o servidor vai à base de dados local buscar a informação sobre cada gene, prepara a informação para esta ficar disponível para download num formato arff.

3.5.7 Pesquisa de Proteínas

As pesquisas de proteínas são feitas através da página do website de pesquisar proteínas. Esta pesquisa é feita no PDB, e é mostrada uma lista ao utilizador com uma breve descrição de proteína. Para realizar esta pesquisa o utilizador tem de introduzir os id´s das proteínas que quer procurar. O Portal Web efetua um pedido HTTP do tipo get ao servidor com os id's das proteínas, e o servidor por sua vez efetua outro pedido HTTP do tipo *get* à API do PDB. A informação é retornada em formato XML, pelo que o servidor depois se encarrega de tratar os dados para o utilizador fazer download da informação. Além disso o utilizador pode a qualquer momento ver detalhes sobre uma proteína ao clicar no id da proteína, sendo mostrado ao utilizador informação detalhada da proteína em questão.

3.6 Funcionamento do Portal WEB

3.6.1 Casos de Uso

Um diagrama de casos de uso descreve o que o sistema faz do ponto de vista do utilizador, ou seja, é onde estão presentes as principais funcionalidades do sistema e como é feita a interação dessas mesmas funcionalidades com o utilizador do sistema. Assim, o diagrama de casos de uso é composto por atores, casos de uso e relacionamentos(associações entre atores e casos de uso; generalizações entre os atores; generalizações, extends e includes entre os casos de uso) entre estes elementos. [Rib16]

Como mostra a figura 13, o visitante pode visualizar a página principal do website e registar-se. Já o utilizador registado pode autenticar-se, visualizar a sua página de perfil, fazer uma pesquisa de genes ou proteínas, o que inclui selecionar as diversas bases de dados e as opções de pesquisa, fazer download da informação e utilizar algoritmos de *clustering* e terminar sessão.

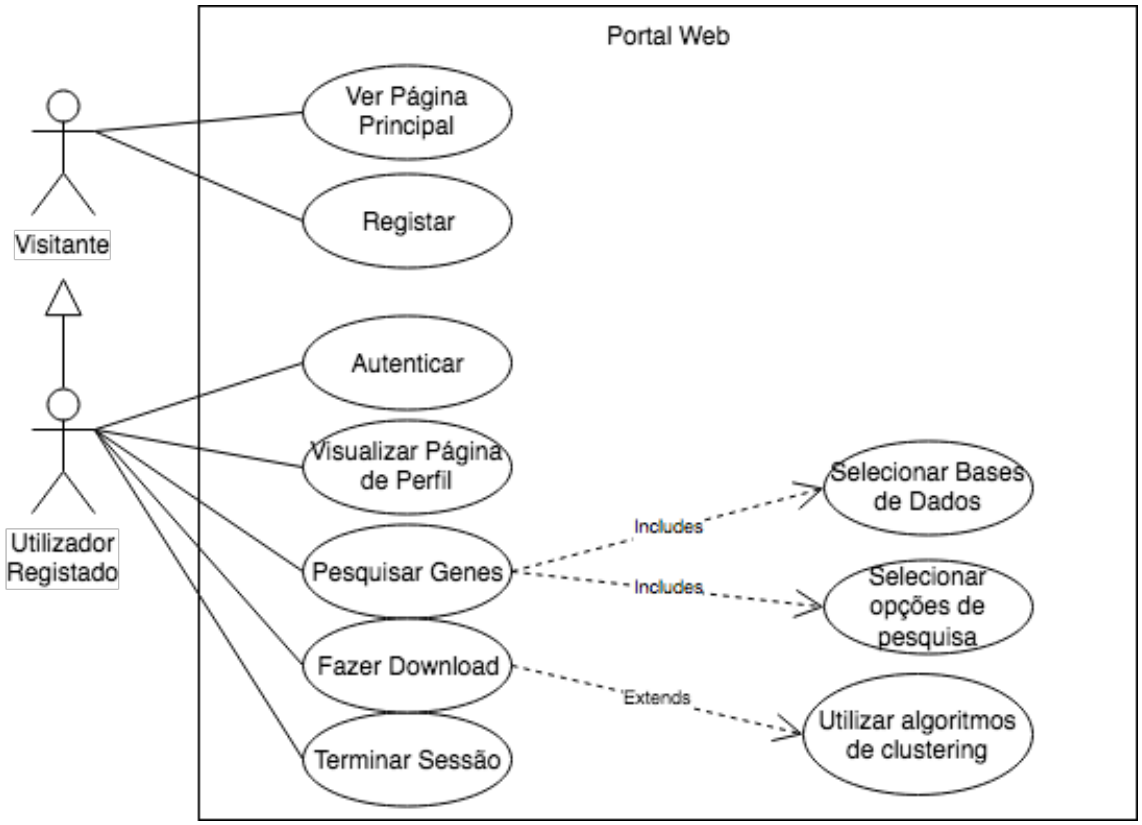


Figura 13: Diagrama de casos de uso

3.6.2 Pesquisa de genes

Para fazer uma pesquisa de genes no Portal Web é necessário estar autenticado. Após estar autenticado no sistema, para a pesquisa é necessário o utilizador introduzir os id's dos genes a pesquisar. Após a introdução dos genes o utilizador escolhe as bases de dados em que quer procurar, podemos ver um exemplo na figura 14. Após essa introdução da lista de genes e escolha das bases de dados e outras opções de pesquisa existem vários cenários:

- O utilizador seleciona apenas uma base de dados. Caso o utilizador escolha apenas a base de dados cujos id's foram introduzidos, o portal web faz a pesquisa dos genes, e mostra uma lista com a descrição para cada gene ao utilizador. Ao mesmo tempo que é mostrada a informação para o utilizador, esta também é inserida na base de dados. Além disso, caso o utilizador queira ver mais detalhes sobre um gene, pode clicar em cima do id que é apresentada informação mais detalhada do gene.

- O Utilizador seleciona várias bases de dados. Neste caso o portal web inicialmente começa por fazer automaticamente a conversão dos id's dos genes para as outras bases de dados

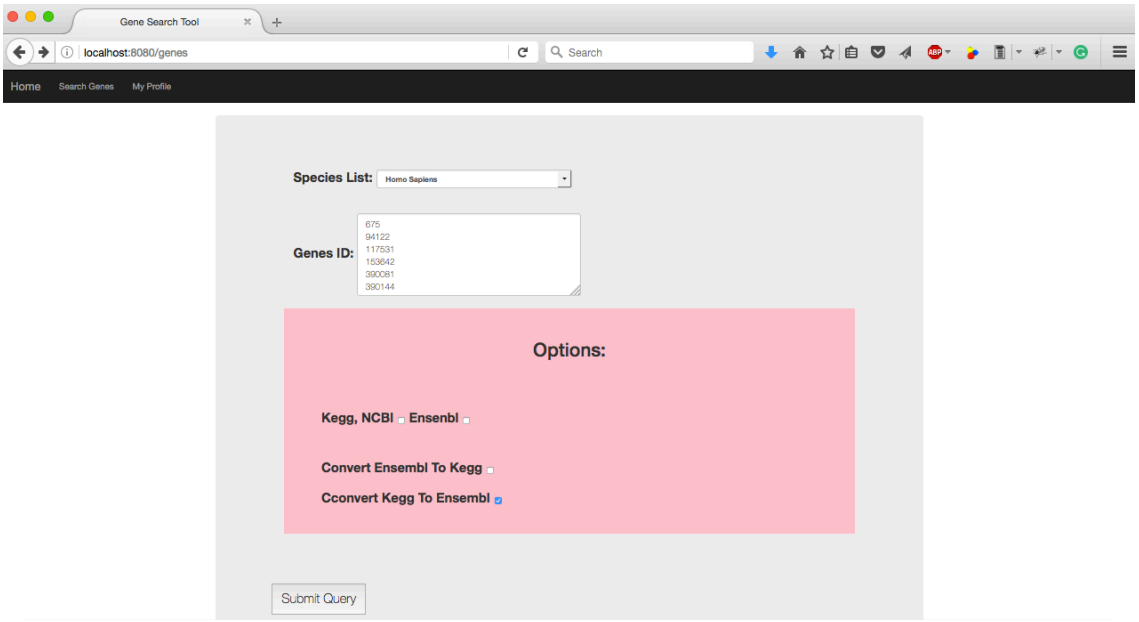


Figura 14: Página de pesquisa de genes do Portal Web

selecionadas. Feita esta conversão, é feita a pesquisa dos genes nas respetivas bases de dados com os id's que foram convertidos anteriormente. Em seguida é mostrada ao utilizador uma lista com os diferentes genes e os diferentes id's de cada base de dados para cada gene, figura 15, ao mesmo tempo que esta informação é armazenada na base de dados. É dada também a opção ao utilizador de abrir a página individual do gene, que contém informação mais detalhada sobre cada gene, figura 16.

Toda esta informação sobre os genes está preparada para a qualquer momento o utilizador fazer download.

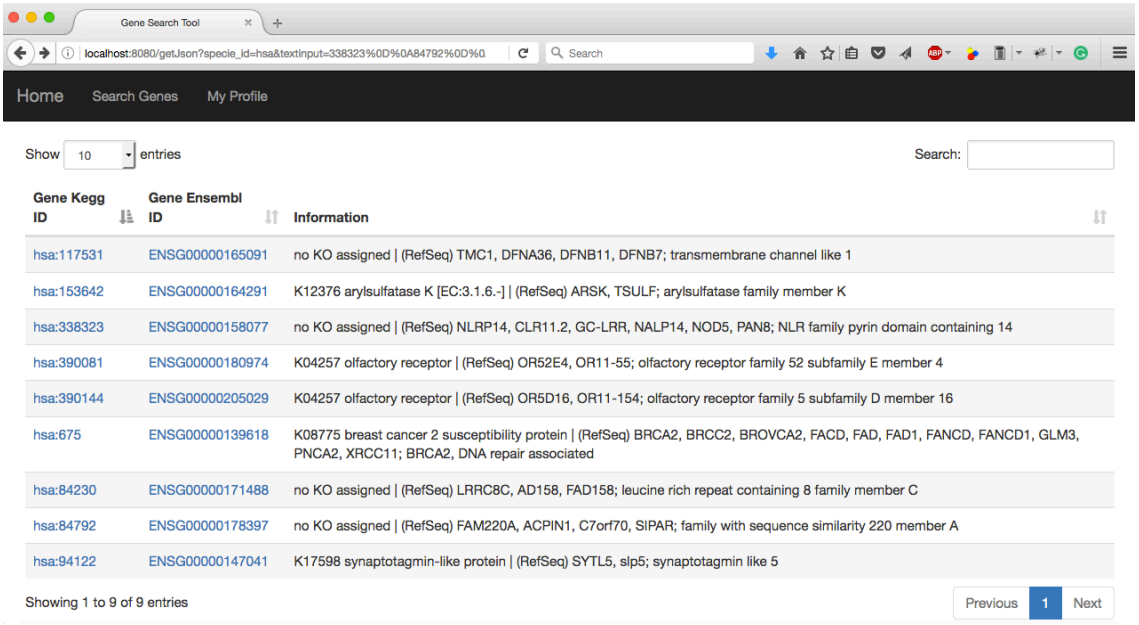


Figura 15: Página com informação dos genes pesquisados do Portal Web

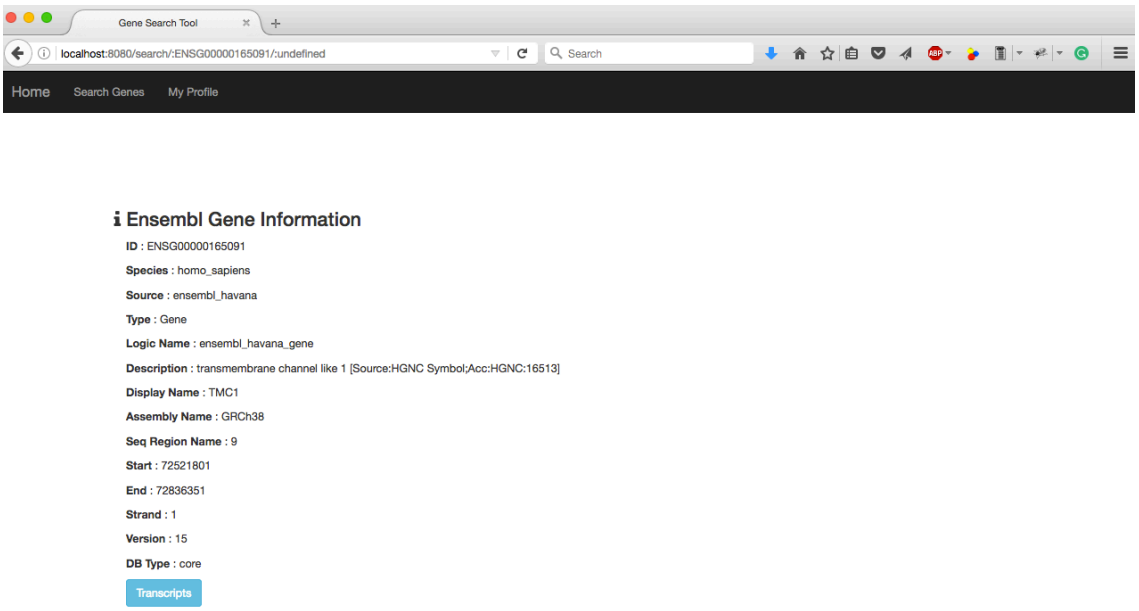


Figura 16: Página com informação detalhada de um gene do Portal Web

3.6.3 Algoritmos do Weka

O Portal Web tem disponível uma integração com a ferramenta weka através de chamadas pela linha de comandos. É possível o utilizador introduzir um ficheiro arff gerado anteriormente na pesquisa de genes e submete-lo para fazer análises de *clustering*. Estão disponíveis duas análises de *clustering*, o *Make Density Based Clustery* e o *Simple K-Means*.

3.6.3.1 Make Density Based Clusterer (MDBC)

Para utilizar este algoritmo, o Portal Web faz uma chamada ao programa weka através da linha de comandos e passa como argumentos o nome do algoritmo *Make Density Based Clusterer* e os parâmetros. São retornados *clusters* com densidade e distribuição que são apresentados ao utilizador no Portal Web.

3.6.3.2 Simple K-Means

Neste algoritmo é feito uma chamada ao programa weka através da linha comandos, com os parâmetros K(número de *clusters* que são pretendidos) desde k=2 até k= 10% dos dados no documento arff. São guardados os valores dos erros para cada k. Verifica-se qual k tem o menor erro e volta a ser feita uma chamada ao programa weka para o melhor k, ou seja, o k com menor erro. Em seguida é guardada toda a informação proveniente do weka e mostrada ao utilizador.

3.7 Conclusões e Resumos

Neste capítulo foram apresentadas as etapas do desenvolvimento do Portal Web, tendo sido explicados os métodos utilizados e que levaram a aplicação ao seu estado final. Inicialmente foi feita uma pesquisa e analise dos genes. É explicada como se procedeu à recolha desta informação sobre genes. Em seguida é explicada como foi criada a base de dados e como vão sendo os dados armazenados. Finalmente é descrito o desenvolvimento do Portal Web e dos algoritmos de *clustering* para a análise de informação dos genes.

Capítulo 4

Resultados

Neste capítulo são apresentados dois casos de estudo utilizados como prova do conceito desenvolvido. É feita uma caracterização do ambiente experimental, dos conjuntos de dados utilizados e os resultados obtidos.

4.1 Especificação dos Casos de estudo

Os dados utilizados neste estudo foram obtidos a partir do Portal Web desenvolvido durante esta dissertação.

Em seguida é feita uma explicação detalhada, ilustrada com algumas imagens, desde o momento em que o utilizador se depara com o Portal Web até chegar aos resultados pretendidos.

4.1.1 Registo/Login

Ao entrar no Portal Web o utilizador é levado para a página principal do *website*, onde tem a opção de se registar ou fazer login no *website*. Para se registar basta introduzir todos os dados que são pedidos no formulário de registo, como é mostrado na figura 17. Feito o registo, nas próximas visitas do utilizador ao Portal Web, apenas necessita de efetuar o login, introduzindo apenas o email e a *password*, como é ilustrado na figura 18.

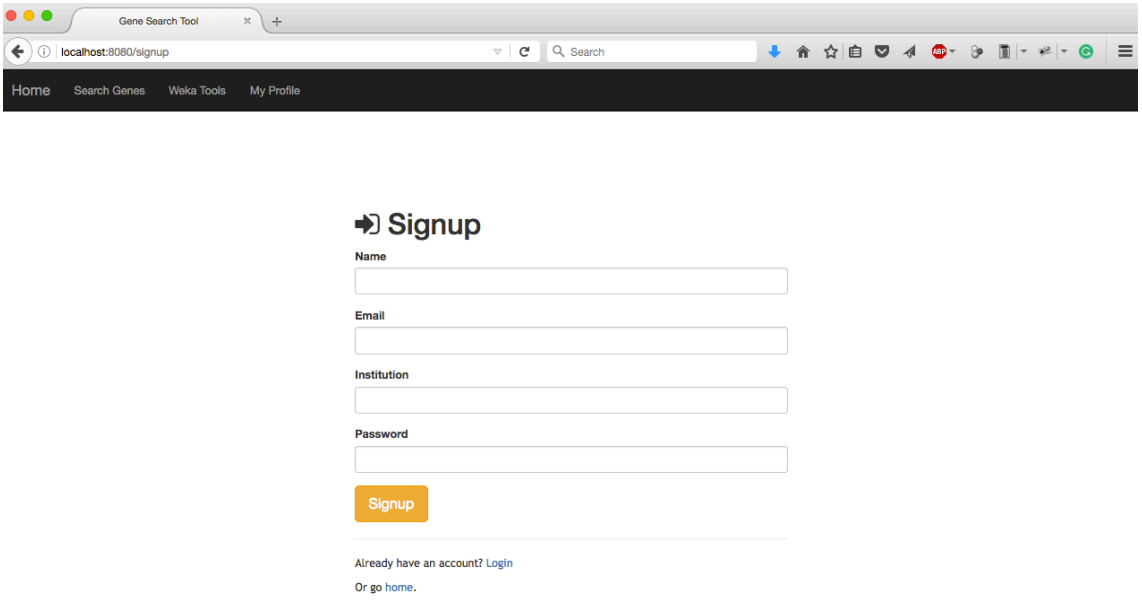


Figura 17: Página de registo do Portal Web

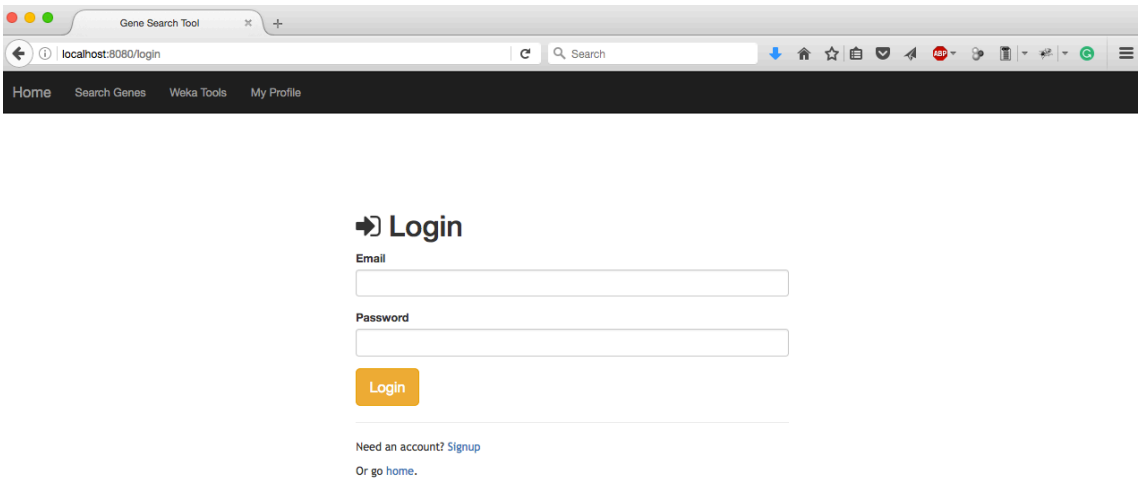


Figura 18: Página de login do Portal Web

4.1.2 Efetuar uma pesquisa de genes

Para efetuar qualquer pesquisa de genes é necessário o utilizador introduzir os identificadores dos genes. Os id's devem estar todos separados por parágrafo, ou seja, um id por linha como ilustra a figura 19. Feita esta introdução dos id's é necessário o utilizador selecionar as opções de pesquisa e escolher as bases de dados em que pretende efetuar a pesquisa e escolher qual conversão de identificadores precisa de fazer.

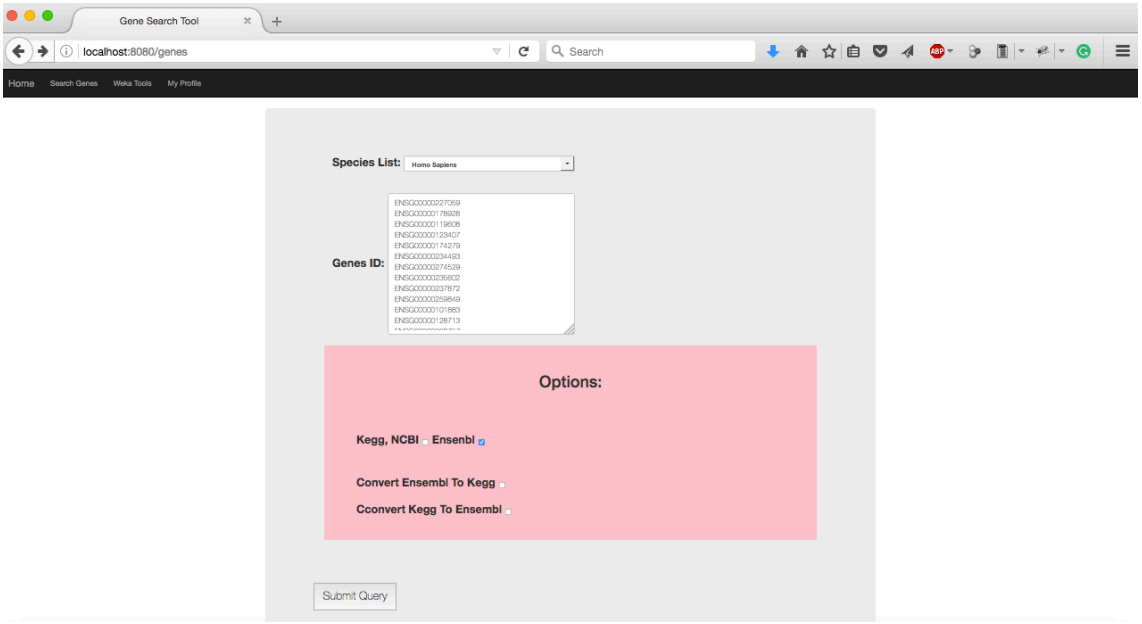


Figura 19: Página de pesquisa de genes do Portal Web

4.1.3 Exportar dados para formato arff

Após a conclusão da pesquisa, o Portal Web mostra ao utilizador uma lista com os id's dos genes e uma breve descrição. Nesta página, é possível o utilizador fazer download da informação completa de cada gene para um ficheiro compatível com o weka, um ficheiro arff, através do botão verde ilustrado nas figuras 20 e 21. Na figura 20 temos o exemplo de uma pesquisa de genes efetuada em apenas uma base de dados, o Ensembl. Já na figura 21 é mostrada uma pesquisa efetuada nas três bases de dados disponíveis no Portal Web, o Kegg, o

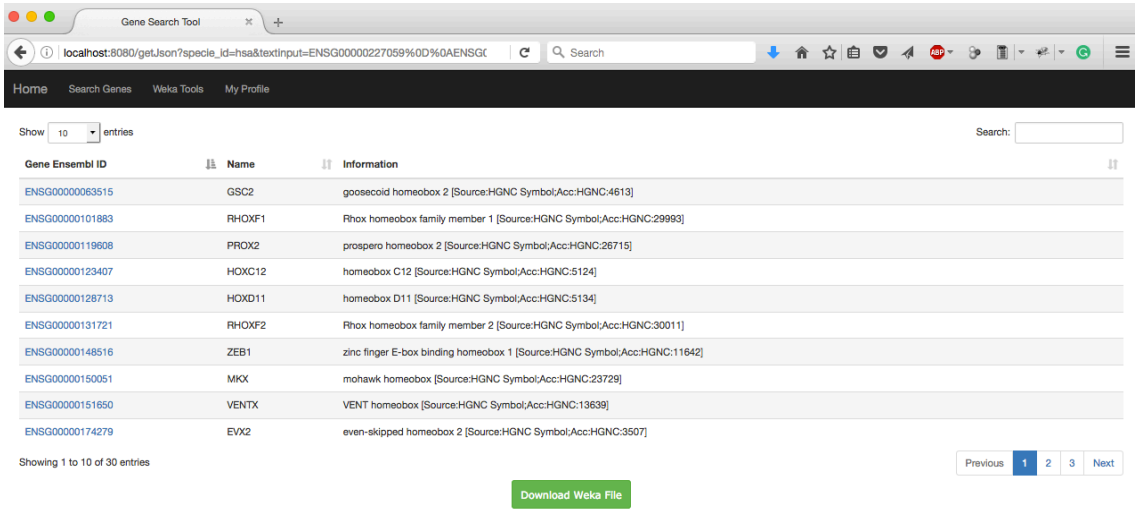


Figura 20: Página com informação dos genes pesquisados de apenas uma base de dados

Ensembl e o NCBI. Apenas foi necessário introduzir os id's de uma base de dados, encarregando-se o Portal Web de os converter para as outras bases de dados.

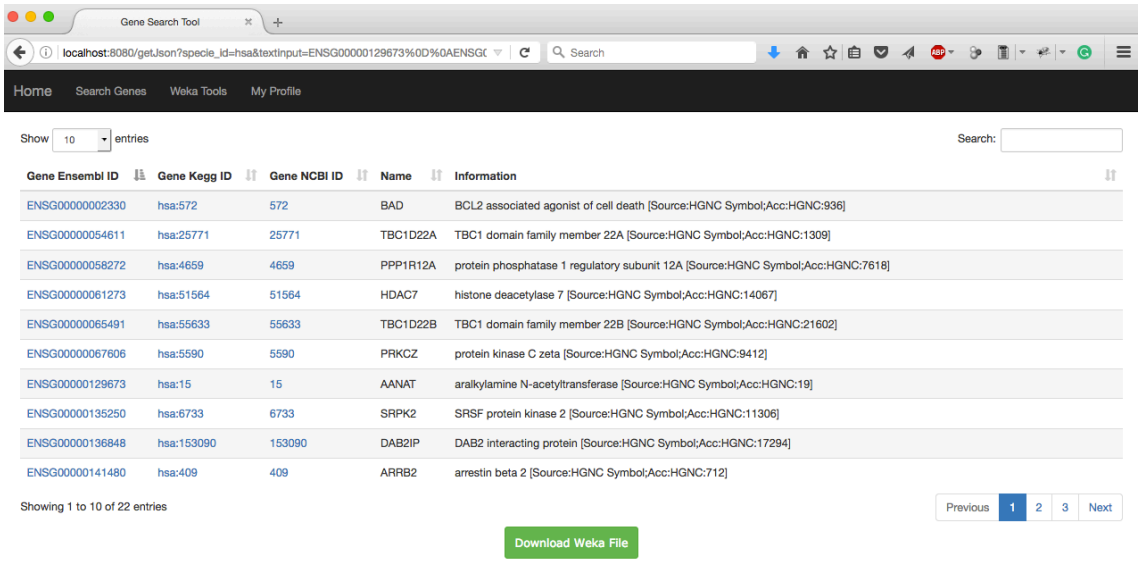


Figura 21: Página com informação dos genes pesquisados de três bases de dados: Ensembl, Kegg e NCBI

4.1.4 Utilização de algoritmos de clustering do weka

Finalmente para a utilização de algoritmos de *clustering* do weka, o utilizador tem de carregar um ficheiro no formato arff para o Portal Web. Seleciona os algoritmos que pretende, e são então apresentados numa tabela os resultados depois da realização dos algoritmos weka.

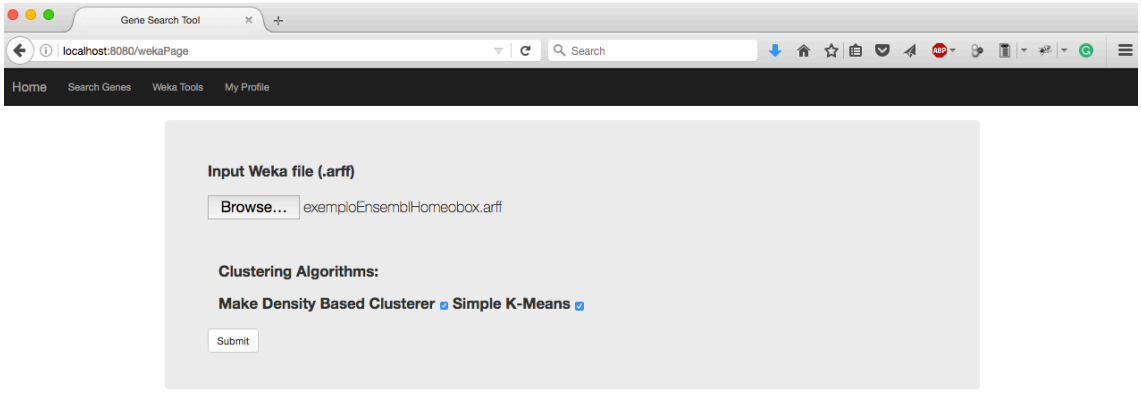


Figura 22: Página de efetuar algoritmos de clustering

4.2 Casos de Estudo

4.2.1 Ambiente Experimental para as experiências

O ambiente utilizado para efetuar os testes foram realizados na mesma máquina. A tabela 1 mostra algumas das especificações da máquina. O desempenho da máquina não foi uma grande preocupação, a preocupação está na conexão à internet, pois por vezes, o tamanho dos dados genes pode ser muito grande, podendo demorar algumas horas para conexões demasiados lentas.

Especificações	**Máquina: MacBook Pro (13-inch, Early 2011)**
Sistema Operativo	OS X Yosemite Versão 10.10.5 (14F1713)
CPU	2.7GHz Intel Core i7 (2 núcleos)
Memória	4GB 1333MHz DDR3
Conexão à Internet	100 Mbps

Tabela 1: Especificações da máquina utilizada para a realização dos casos de estudo

4.2.2 Caso de estudo 1

O caso de estudo 1 contém analises dos dados na base de dados do Ensembl utilizando algoritmos de clustering Simple K-Means e MDBC.

4.2.2.1 Conjunto dos dados analisados

O conjunto dos dados de análise deste caso de estudo são constituídos por 30 identificadores de genes do *Homeobox*, uma sequência de DNA. Os genes reais utilizados para os testes estão listados na Tabela 2. É de referir que o Portal Web só necessita de receber os identificadores dos genes, Nomes de genes não são aceites como dados de entrada.

Gene Name	Ensembl Gene ID
ANHX	ENSG00000227059
DPRXP3	ENSG00000282308
DUX4L31	ENSG00000231411
DUX4L51	ENSG00000250482
DUX4L52	ENSG00000258336
DUXAP11	ENSG00000270222
DUXAP3	ENSG00000270552
DUXAP8	ENSG00000206195
DUXB	ENSG00000282757
EVX2	ENSG00000174279
GSC2	ENSG00000063515
HOXC12	ENSG00000123407
HOXD11	ENSG00000128713
MKX	ENSG00000150051
NANOGP10	ENSG00000231750
NANOGP2	ENSG00000228670
NANOGP5	ENSG00000231697
NANOGP6	ENSG00000227351
NANOGP9	ENSG00000231809
POU5F1P3	ENSG00000235602
POU5F1P4	ENSG00000237872
PROX2	ENSG00000119608
RHOXF1	ENSG00000101883
RHOXF1P1	ENSG00000234493
RHOXF2	ENSG00000131721
SEBOX	ENSG00000274529
TPRX1	ENSG00000178928
VENTX	ENSG00000151650
VENTXP1	ENSG00000259849
ZEB1	ENSG00000148516

Tabela 2: Genes da família Homeobox utilizados para a realização do caso de estudo

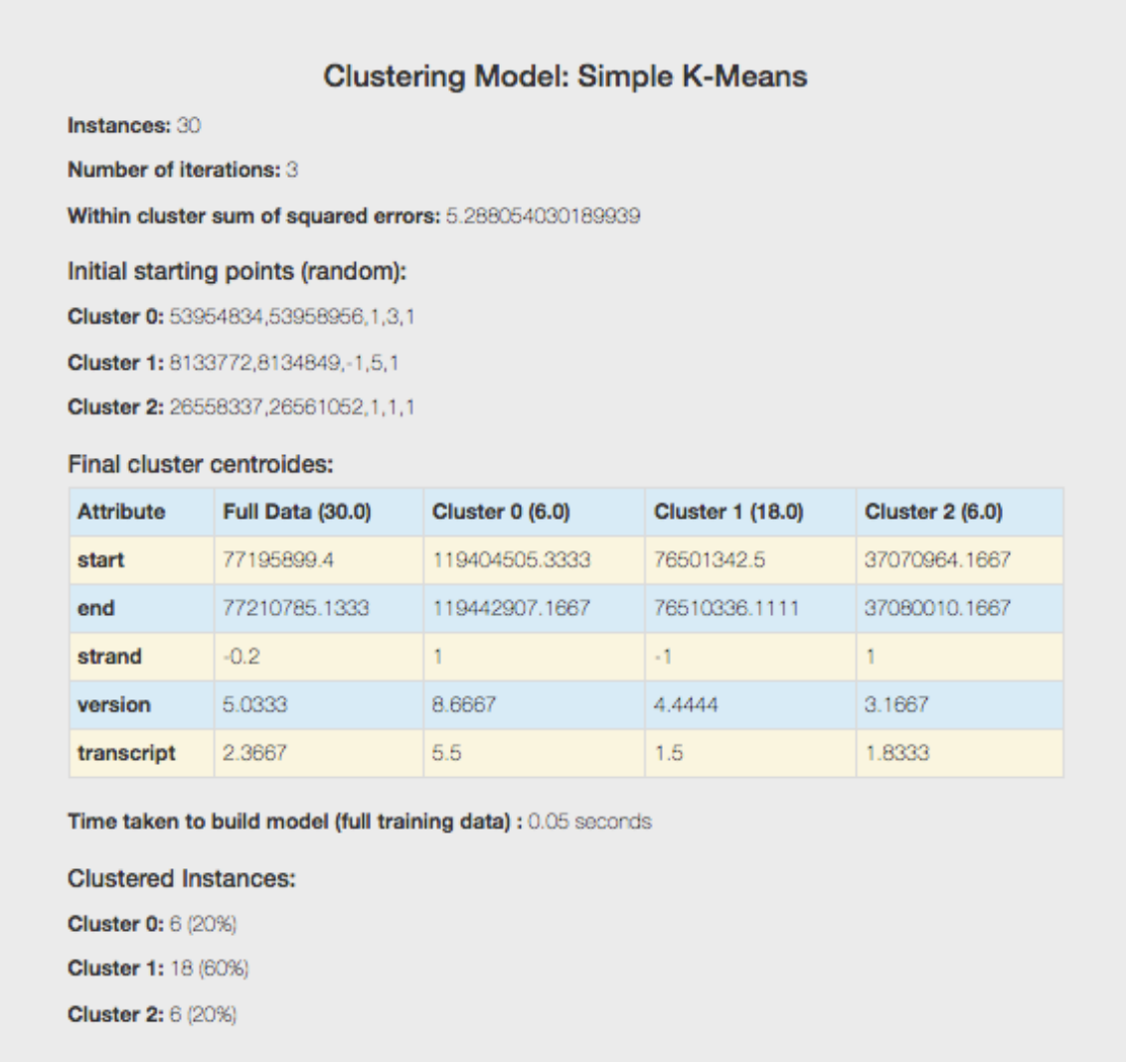
4.2.2.2 Metodologia

Para assegurar máxima eficiência e de modo a assegurar que o tempo de execução dependa apenas do hardware da máquina, várias restrições foram aplicadas. Em primeiro lugar, foi assegurado que a máquina estava a correr o seu sistema operativo com as mínimas aplicações requeridas. Em seguida a conexão de rede foi verificada, de modo a ter certeza que a velocidade de conexão à internet era a referida na tabela 1.

A experiência foi então executada com genes de apenas uma base de dados, o Ensembl. A experiência foi realizada até se obter os resultados dos algoritmos de *clustering Simple K-means* e *Make Density Based Clusterer*. Esta experiência utilizou através do Portal Web, dados de genes do Ensembl e a utilização da ferramenta weka.

4.2.2.3 Resultados

Os resultados gerados através do Portal Web pela ferramenta do weka são apresentados nas figuras 23 e 24. Estes resultados são estatísticas que servem para dividir as instâncias por *clusters*.



Clustering Model: Simple K-Means

Instances: 30

Number of iterations: 3

Within cluster sum of squared errors: 5.288054030189939

Initial starting points (random):

Cluster 0: 53954834,53958956,1,3,1

Cluster 1: 8133772,8134849,-1,5,1

Cluster 2: 26558337,26561052,1,1,1

Final cluster centroides:

Attribute	Full Data (30.0)	Cluster 0 (6.0)	Cluster 1 (18.0)	Cluster 2 (6.0)
start	77195899.4	119404505.3333	76501342.5	37070964.1667
end	77210785.1333	119442907.1667	76510336.1111	37080010.1667
strand	-0.2	1	-1	1
version	5.0333	8.6667	4.4444	3.1667
transcript	2.3667	5.5	1.5	1.8333

Time taken to build model (full training data) : 0.05 seconds

Clustered Instances:

Cluster 0: 6 (20%)

Cluster 1: 18 (60%)

Cluster 2: 6 (20%)

Figura 23: Output do algoritmo de clustering Simple K-Means

Clustering Model: Make Density Based Clusterer

Wrapped clusterer: K-Means

Instances: 30

Number of iterations: 2

Within cluster sum of squared errors: 6.470762134857711

Initial starting points (random):

Cluster 0: 53954834,53958956,1,3,1

Cluster 1: 8133772,8134849,-1,5,1

Final cluster centroides:

Attribute	Full Data (30.0)	Cluster 0 (12.0)	Cluster 1 (18.0)
start	77195899.4	78237734.75	76501342.5
end	77210785.1333	78261458.6667	76510336.1111
strand	-0.2	1	-1
version	5.0333	5.9167	4.4444
transcript	2.3667	3.6667	1.5

Fitted estimators (with ML estimates of variance):

Cluster: 0 **Prior probability:** 0.4063

Attribute: start Normal Distribution. Mean = 78237734.75 StdDev = 53670469.6184

Attribute: end Normal Distribution. Mean = 78261458.6667 StdDev = 53652071.1057

Attribute: strand Normal Distribution. Mean = 1 StdDev = 0.9965

Attribute: version Normal Distribution. Mean = 5.9167 StdDev = 5.6636

Attribute: transcript Normal Distribution. Mean = 3.6667 StdDev = 6.587

Cluster: 1**Prior probability:** 0.5938

Attribute: start Normal Distribution. Mean = 76501342.5 StdDev = 57365372.628

Attribute: end Normal Distribution. Mean = 76510336.1111 StdDev = 57363227.0698

Attribute: strand Normal Distribution. Mean = -1 StdDev = 0.9965

Attribute: version Normal Distribution. Mean = 4.4444 StdDev = 3.3536

Attribute: transcript Normal Distribution. Mean = 1.5 StdDev = 0.6872

Time taken to build model (full training data): 0.13 seconds

Clustered Instances:

Cluster 0: 3(10%)

Cluster 1: 27(90%)

Log likelihood: -44.50168

Figura 24: Output do algoritmo de clustering MDBC

Com base no algoritmo de *clustering Simple K-Means*, é possível agruparmos os dados de cada gene, ou seja, cada instância em diferentes *clusters*. As tabelas 3,4 e 5 mostram que a ferramenta encontrou três *clusters* de genes dentro do conjunto de dados. Com este agrupamento de instâncias em *clusters*, é possível fazermos uma caracterização de cada *cluster*.

Neste caso de estudo, o menor erro encontrado foi para um número de *clusters* igual a três.

Observe-se que o cluster 0 é caracterizado por ter o atributo strand=1 em todas as suas instâncias. Por outro lado, o cluster 1 é caracterizado por ter o atributo strand=-1 em todas as instâncias e pelos genes possuírem apenas 1 transcript em 61,6% das instâncias. Já o cluster 2, tem o atributo strand=1 e os genes possuírem 1 trancript em 83,3% das instâncias.

Cluster 0:

Instância	ID	species	source	type	display_name	assembly_name	seq_region_name	start	end	strand	version	db_type	transcript
9	ENSG00000237872	homo_sapiens	havana	Gene	POU5F1P4	GRCh38	1	155433178	155434262	1	4	core	1
12	ENSG00000128713	homo_sapiens	ensembl_havana	Gene	HOXD11	GRCh38	2	176104216	176109754	1	12	core	3
20	ENSG00000148516	homo_sapiens	ensembl_havana	Gene	ZEB1	GRCh38	10	31318495	31529814	1	21	core	25
22	ENSG00000131721	homo_sapiens	ensembl_havana	Gene	RHOXF2	GRCh38	0	120158561	120165630	1	5	core	1
23	ENSG00000151650	homo_sapiens	ensembl_havana	Gene	VENTX	GRCh38	10	133237404	133241929	1	7	core	1
27	ENSG00000231697	homo_sapiens	havana	Gene	NANOGP5	GRCh38	9	100175178	100176054	1	3	core	2

Tabela 3: Cluster 0 gerado pelo algoritmo Simple-Kmeans

Cluster 1:

Instância	ID	species	source	type	display_name	assembly_name	seq_region_name	start	end	strand	version	db_type	transcript
1	ENSG00000227059	homo_sapiens	ensembl_havana	Gene	ANHX	GRCh38	12	133218312	133236095	-1	6	core	2
2	ENSG00000178928	homo_sapiens	ensembl_havana	Gene	TPRX1	GRCh38	19	47801243	47819051	-1	8	core	3
3	ENSG00000119608	homo_sapiens	ensembl_havana	Gene	PROX2	GRCh38	14	74852871	74871940	-1	12	core	3
5	ENSG00000174279	homo_sapiens	ensembl_havana	Gene	EVX2	GRCh38	2	176077472	176083913	-1	4	core	1
6	ENSG00000234493	homo_sapiens	havana	Gene	RHOXF1P1	GRCh38	0	120010718	120015544	-1	2	core	2
7	ENSG00000274529	homo_sapiens	ensembl_havana	Gene	SEBOX	GRCh38	17	28364268	28365244	-1	5	core	2
8	ENSG00000235602	homo_sapiens	havana	Gene	POU5F1P3	GRCh38	12	8133772	8134849	-1	5	core	1
11	ENSG00000101883	homo_sapiens	ensembl_havana	Gene	RHOXF1	GRCh38	0	120109053	120115937	-1	4	core	1
13	ENSG00000282757	homo_sapiens	havana	Gene	DUXB	GRCh38	16	75694434	75700152	-1	2	core	2
14	ENSG00000231411	homo_sapiens	havana	Gene	DUX4L31	GRCh38	0	10171590	10172725	-1	1	core	1
17	ENSG00000258336	homo_sapiens	havana	Gene	DUX4L52	GRCh38	12	61600067	61600843	-1	2	core	1
19	ENSG00000150051	homo_sapiens	ensembl_havana	Gene	MKX	GRCh38	10	27672875	27746060	-1	13	core	2
21	ENSG00000063515	homo_sapiens	ensembl_havana	Gene	GSC2	GRCh38	22	19148576	19150283	-1	2	core	1
24	ENSG00000227351	homo_sapiens	havana	Gene	NANOGP6	GRCh38	10	99788564	99789410	-1	2	core	1
26	ENSG00000228670	homo_sapiens	havana	Gene	NANOGP2	GRCh38	2	222452400	222453126	-1	4	core	1
28	ENSG00000231750	homo_sapiens	havana	Gene	NANOGP10	GRCh38	0	43407665	43408559	-1	3	core	1
29	ENSG00000231809	homo_sapiens	havana	Gene	NANOGP9	GRCh38	0	65772741	65773632	-1	4	core	1
30	ENSG00000270552	homo_sapiens	havana	Gene	DUXAP3	GRCh38	10	42747544	42748687	-1	1	core	1

Tabela 4: Cluster 1 gerado pelo algoritmo Simple-Kmeans

Cluster 2:

Instância	ID	species	source	type	display_name	assembly_name	seq_region_name	start	end	strand	version	db_type	transcript
4	ENSG00000123407	homo_sapiens	ensembl_havana	Gene	HOXC12	GRCh38	12	53954834	53958956	1	3	core	1
10	ENSG00000259849	homo_sapiens	havana	Gene	VENTXP1	GRCh38	0	26558337	26561052	1	1	core	1
15	ENSG00000270222	homo_sapiens	havana	Gene	DUXAP11	GRCh38	16	59655602	59656336	1	1	core	1
16	ENSG00000250482	homo_sapiens	havana	Gene	DUX4L51	GRCh38	5	31249879	31250987	1	3	core	1
18	ENSG00000206195	homo_sapiens	havana	Gene	DUXAP8	GRCh38	22	15784959	15829984	1	10	core	6
25	ENSG00000282308	homo_sapiens	havana	Gene	DPRXP3	GRCh38	14	35222174	35222746	1	1	core	1

Tabela 3: Cluster 2 gerado pelo algoritmo Simple-Kmeans

Caracterização dos clusters:

Cluster 0:

strand= 1 Em 100% das instâncias.

source= ensembl_havana Em 66,7% das instâncias.

Cluster 1:

strand= -1 Em 100% das instâncias.

transcript= 1 Em 61,1% das instâncias.

Cluster 2:

strand= 1 Em 100% das instâncias.

source= havana Em 83,3% das instâncias.

transcript= 1 Em 83,3% das instâncias.

Figura 25: Caracterização dos Clusters gerados pelo Simple K-Means

4.2.3 Caso de estudo 2

O caso de estudo 1 contém analises dos dados na base de dados do Ensembl, NCBI e Kegg utilizando algoritmos de clustering Simple K-Means e MDBC.

4.2.3.1 Conjunto dos dados analisados

O conjunto dos dados de análise deste caso de estudo são constituídos por 22 identificadores de genes da família *14-3-3 phospho-serine/phospho-threonine binding proteins*, uma família de moléculas reguladoras conservadas que são expressas em todas as células eucarióticas. Os genes reais utilizados para os testes estão listados na Tabela 6. É de referir que o Portal Web só necessita de receber o identificador de uma base de dados, neste caso foi do Ensembl, fazendo automaticamente a conversão dos identificadores para as outras bases de dados.

Gene Name	Ensembl Gene ID	Kegg Gene ID	NCBI Gene ID
AANAT	ENSG00000129673	hsa:15	15
AKT1	ENSG00000142208	hsa:207	207
ARRB2	ENSG00000141480	hsa:409	409
BAD	ENSG00000002330	hsa:572	572
DAB2IP	ENSG00000136848	hsa:153090	153090
DDIT4	ENSG00000168209	hsa:54541	54541
EEF1G	ENSG00000254772	hsa:1937	1937
HDAC7	ENSG00000061273	hsa:51564	51564
IRS2	ENSG00000185950	hsa:8660	8660
KCNH1	ENSG00000143473	hsa:3756	3756
KIF13B	ENSG00000197892	hsa:23303	23303
PI4KB	ENSG00000143393	hsa:5298	5298
PPP1R12A	ENSG00000058272	hsa:4659	4659
PRKCE	ENSG00000171132	hsa:5581	5581
PRKCZ	ENSG00000067606	hsa:5590	5590
RPTOR	ENSG00000141564	hsa:57521	57521
SIK1	ENSG00000142178	hsa:150094	150094
SRPK2	ENSG00000135250	hsa:6733	6733
SYNPO2	ENSG00000172403	hsa:171024	171024
TBC1D22A	ENSG00000054611	hsa:25771	25771
TBC1D22B	ENSG00000065491	hsa:55633	55633
YWHAB	ENSG00000166913	hsa:7529	7529

Tabela 4: Genes da família 14-3-3 phospho-serine utilizada no caso de estudo 2

4.2.3.2 Metodologia

Para assegurar máxima eficiência e de modo a assegurar que o tempo de execução dependa apenas do hardware da máquina, várias restrições foram aplicadas. Em primeiro lugar, foi assegurado que a máquina estava a correr o seu sistema operativo com as mínimas aplicações requeridas. Em seguida a conexão de rede foi verificada, de modo a ter certeza que a velocidade de conexão à internet era a referida na tabela 1.

A experiência foi então executada com genes das várias bases de dados disponíveis no portal *web*, Ensembl, Kegg e NCBI. Apenas foi necessário introduzir os identificadores de uma base de dados. O portal *web* encarrega-se de converter esses identificadores para as restantes bases de dados. A experiência foi realizada até se obter os resultados dos algoritmos de *clustering Simple K-means* e *Make Density Based Clusterer*. Esta experiência utilizou através do portal *web*, dados de genes do Ensembl, Kegg e NCBI, a utilização duma ferramenta de conversão de genes e a utilização da ferramenta weka.

4.2.3.3 Resultados

Os resultados gerados através do portal *web* pela ferramenta do weka são apresentados nas figuras 26 e 27. Estes resultados são estatísticas que servem para dividir as instâncias por *clusters*.

Com base no algoritmo de *clustering Simple K-Means*, é possível agruparmos os dados de cada gene, ou seja, cada instância em diferentes *clusters*. As tabelas 7, 8 e 9 mostram que a ferramenta encontrou três *clusters* de genes dentro do conjunto de dados. Com este agrupamento de instâncias em *clusters*, é possível fazermos uma caracterização de cada *cluster*. Neste caso, foi escolhido o resultado com menor erro até um número de *clusters* = 10% dos dados.

Observe-se que o cluster 0 é caracterizado por ter o exon<15 em todas as suas instâncias, trancripts< 5 em 85,7% das instâncias e chromosome>10 em 71,4%. Por outro lado, o cluster 1 é caracterizado por ter o exon>15 em 81,9% das instâncias e pelos geneWeight<10.000 em 63,6% das instâncias. Já o cluster 2, tem o atributo strand=1 em 75% das instâncias e chromosome<5 em 100% das instâncias.

Os resultados obtidos estão de acordo com o esperado uma vez que para cada *cluster* foi possível fazer uma caracterização que o distinguisse dos outros *clusters*.

Clustering Model: Simple K-Means

Instances: 22

Number of iterations: 2

Within cluster sum of squared errors: 16.315373969917946

Initial starting points (random):

Cluster 0: 72273920,72276036,1,4,3,10,607729,10,3,6396,10

Cluster 1: 80644819,80966371,1,13,11,17,607130,17,34,10536,17

Cluster 2: 45651345,46167990,1,13,15,2,176975,2,32,29919,2

Final cluster centroides:

Attribute	Full Data (22.0)	Cluster 0 (7.0)	Cluster 1 (11.0)	Cluster 2 (4.0)
start	75430844.4091	90463641	64398960.9091	79461075
end	75583294.4091	90651228	64564684.9091	79690586.75
strand	0	-0.1429	-0.0909	0.5
version	12.3182	8.8571	14	13.75
transcript	9.8182	3.1429	12.3636	14.5
chromosome	10.7273	9.8571	14.4545	2
int	475086.7273	537631.5714	521070.1818	239178.75
chrLoc	10.7273	9.8571	14.4545	2
exonCount	18.6364	7.4286	24.9091	21
geneWeight	21900.5455	10132.8571	31094.8182	17209.75
chrSort	10.7273	9.8571	14.4545	2

Time taken to build model (full training data) : 0.01 seconds

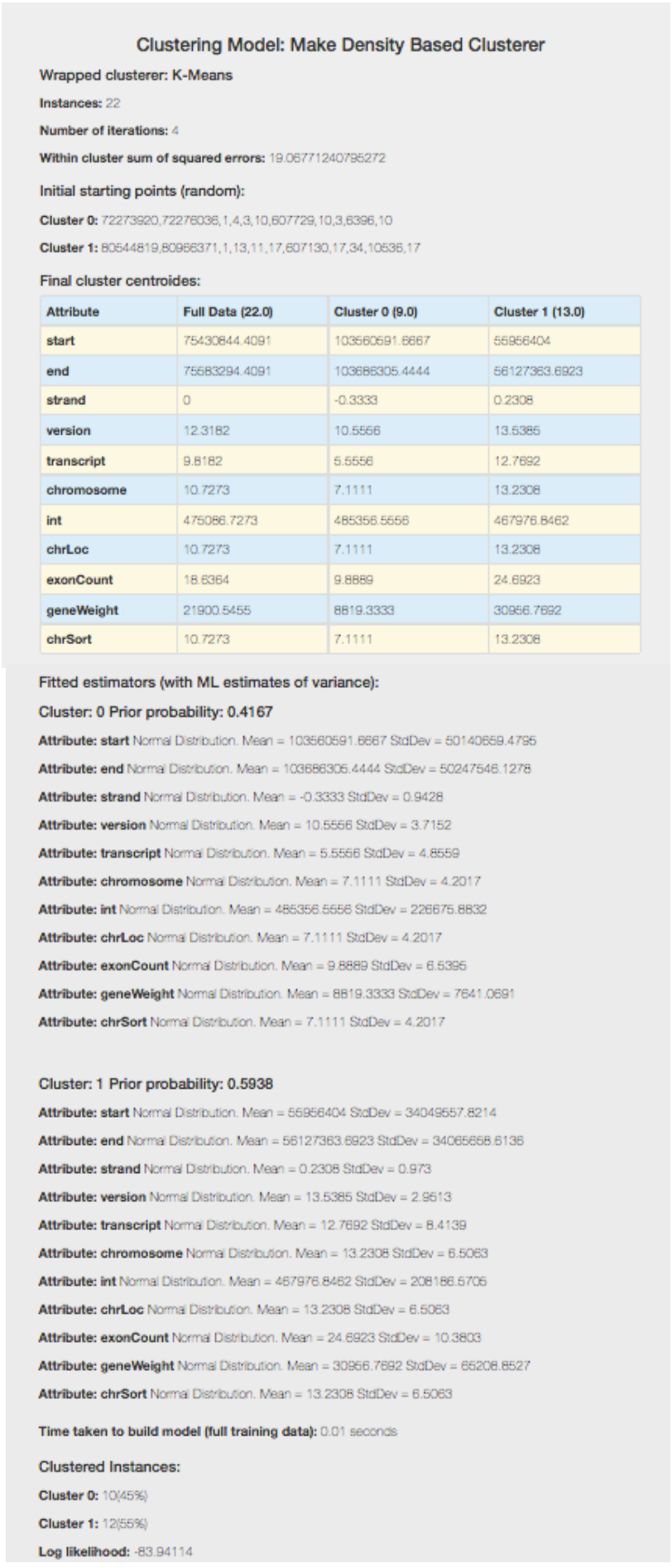
Clustered Instances:

Cluster 0: 7 (32%)

Cluster 1: 11 (50%)

Cluster 2: 4 (18%)

Figura 26: Output do algoritmo de clustering Simple K-Means



Clustering Model: Make Density Based Clusterer

Wrapped clusterer: K-Means

Instances: 22

Number of iterations: 4

Within cluster sum of squared errors: 19.0677124079272

Initial starting points (random):

Cluster 0: 72273920,72276036,1,4,3,10,607729,10,3,6396,10

Cluster 1: 80544819,80966371,1,13,11,17,607130,17,34,10536,17

Final cluster centroides:

Attribute	Full Data (22.0)	Cluster 0 (9.0)	Cluster 1 (13.0)
start	75430844.4091	103560591.6667	55956404
end	75583294.4091	103686305.4444	56127363.6923
strand	0	-0.3333	0.2308
version	12.3182	10.5556	13.5385
transcript	9.8182	5.5556	12.7692
chromosome	10.7273	7.1111	13.2308
int	475086.7273	485356.5556	467976.8462
chrLoc	10.7273	7.1111	13.2308
exonCount	18.6364	9.8889	24.6923
geneWeight	21900.5455	8819.3333	30956.7692
chrSort	10.7273	7.1111	13.2308

Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.4167

Attribute: start Normal Distribution. Mean = 103560591.6667 StdDev = 50140659.4795

Attribute: end Normal Distribution. Mean = 103686305.4444 StdDev = 50247546.1278

Attribute: strand Normal Distribution. Mean = -0.3333 StdDev = 0.9428

Attribute: version Normal Distribution. Mean = 10.5556 StdDev = 3.7152

Attribute: transcript Normal Distribution. Mean = 5.5556 StdDev = 4.8559

Attribute: chromosome Normal Distribution. Mean = 7.1111 StdDev = 4.2017

Attribute: int Normal Distribution. Mean = 485356.5556 StdDev = 226675.8832

Attribute: chrLoc Normal Distribution. Mean = 7.1111 StdDev = 4.2017

Attribute: exonCount Normal Distribution. Mean = 9.8889 StdDev = 6.5395

Attribute: geneWeight Normal Distribution. Mean = 8819.3333 StdDev = 7641.0691

Attribute: chrSort Normal Distribution. Mean = 7.1111 StdDev = 4.2017

Cluster: 1 Prior probability: 0.5938

Attribute: start Normal Distribution. Mean = 55956404 StdDev = 34049557.8214

Attribute: end Normal Distribution. Mean = 56127363.6923 StdDev = 34065668.6136

Attribute: strand Normal Distribution. Mean = 0.2308 StdDev = 0.973

Attribute: version Normal Distribution. Mean = 13.5385 StdDev = 2.9513

Attribute: transcript Normal Distribution. Mean = 12.7692 StdDev = 8.4139

Attribute: chromosome Normal Distribution. Mean = 13.2308 StdDev = 6.5063

Attribute: int Normal Distribution. Mean = 467976.8462 StdDev = 208186.5705

Attribute: chrLoc Normal Distribution. Mean = 13.2308 StdDev = 6.5063

Attribute: exonCount Normal Distribution. Mean = 24.6923 StdDev = 10.3803

Attribute: geneWeight Normal Distribution. Mean = 30956.7692 StdDev = 65208.8527

Attribute: chrSort Normal Distribution. Mean = 13.2308 StdDev = 6.5063

Time taken to build model (full training data): 0.01 seconds

Clustered Instances:

Cluster 0: 10(45%)

Cluster 1: 12(55%)

Log likelihood: -83.94114

Figura 27: Output do algoritmo de clustering MDBC

Cluster 0:

Instância	ID Ensembl	start	end	strand	version	transcript	ID Kegg	chromosome	int	chrLoc	exonCount	geneWeight	chrSort
1	ENSG00000129673	76453351	76470117	1	9	4	hsa:15	17	600950	17	8	4037	17
4	ENSG00000002330	64269830	64284704	-1	13	7	hsa:572	11	603167	11	3	23629	11
6	ENSG00000168209	72273920	72276036	1	4	3	hsa:54541	10	607729	10	3	6396	10
7	ENSG00000254772	62559601	62574086	-1	9	4	hsa:1937	11	130593	11	10	4389	11
9	ENSG00000185950	109752698	109786568	-1	8	1	hsa:8660	13	600797	13	2	21014	13
10	ENSG00000143473	210678315	211134115	-1	11	2	hsa:3756	1	603305	1	12	10426	1
21	ENSG00000065491	37257772	37332970	1	8	1	hsa:55633	6	616880	6	14	1039	6

Tabela 7: Cluster 0 gerado pelo algoritmo Simple-Kmeans

Cluster 1:

Instância	ID Ensembl	start	end	strand	version	transcript	ID Kegg	chromosome	int	chrLoc	exonCount	geneWeight	chrSort
2	ENSG00000142208	104769349	104795751	-1	15	19	hsa:207	14	164730	14	16	254213	14
3	ENSG00000141480	4710489	4721499	1	17	20	hsa:409	17	107941	17	17	24623	17
5	ENSG00000136848	121567057	121785530	1	16	10	hsa:153090	9	609205	9	27	7074	9
8	ENSG00000061273	47782722	47833132	-1	17	30	hsa:51564	12	606542	12	32	7830	12
11	ENSG00000197892	29067279	29263124	-1	12	6	hsa:23303	8	607350	8	42	1830	8
13	ENSG00000058272	79773563	79935460	-1	16	4	hsa:4659	12	602021	12	31	9088	12
16	ENSG00000141564	80544819	80966371	1	13	11	hsa:57521	17	607130	17	34	10536	17
17	ENSG00000142178	43414515	43427128	-1	7	2	hsa:150094	21	605705	21	14	2533	21
18	ENSG00000135250	105110704	105399308	-1	16	15	hsa:6733	7	602980	7	23	3580	7
20	ENSG00000054611	46762617	47175699	1	13	11	hsa:25771	22	616879	22	30	1316	22
22	ENSG00000166913	44885676	44908532	1	12	8	hsa:7529	20	601289	20	8	19420	20

Tabela 5: Cluster 1 gerado pelo algoritmo Simple-Kmeans

Cluster 2:

Instância	ID Ensembl	start	end	strand	version	transcript	ID Kegg	chromosome	int	chrLoc	exonCount	geneWeight	chrSort
12	ENSG00000143393	151291797	151327715	-1	16	13	hsa:5298	1	602758	1	15	5715	1
14	ENSG00000171132	45651345	46187990	1	13	15	hsa:5581	2	176975	2	32	29919	2
15	ENSG00000067606	2050470	2185395	1	16	26	hsa:5590	1	176982	1	30	30019	1
19	ENSG00000172403	118850688	119061247	1	10	4	hsa:171024	4	0	4	7	3186	4

Tabela 9: Cluster 2 gerado pelo algoritmo Simple-Kmeans

Caracterização dos clusters:

Cluster 0:

chromosome > 10 Em 71,4% das instâncias.

transcript < 5 Em 85,7% das instâncias.

exonCount < 15 Em 100% das instâncias.

Cluster 1:

exonCount > 15 Em 81,9% das instâncias.

geneWeight < 10000 Em 63,6% das instâncias.

Cluster 2:

strand = 1 Em 75% das instâncias.

chromosome < 5 Em 100% das instâncias.

Figura 28: Caracterização dos Clusters gerados pelo Simple K-Means

4.2.4 Caso de estudo 3

4.2.4.1 Conjunto dos dados analisados

O conjunto dos dados de análise deste caso de estudo são constituídos por oito identificadores de proteínas. Os identificadores das proteínas podem ser encontrados na tabela 11.

PDB Protein ID
1hhb
2hhb
3hhb
4hhb
1hv4
3MU6
3QV1
3SSF

Tabela 6: Proteínas utilizadas no caso de estudo 3

4.2.4.2 Metodologia

Para realizar a experiência inicialmente foram introduzidos os identificadores das proteínas

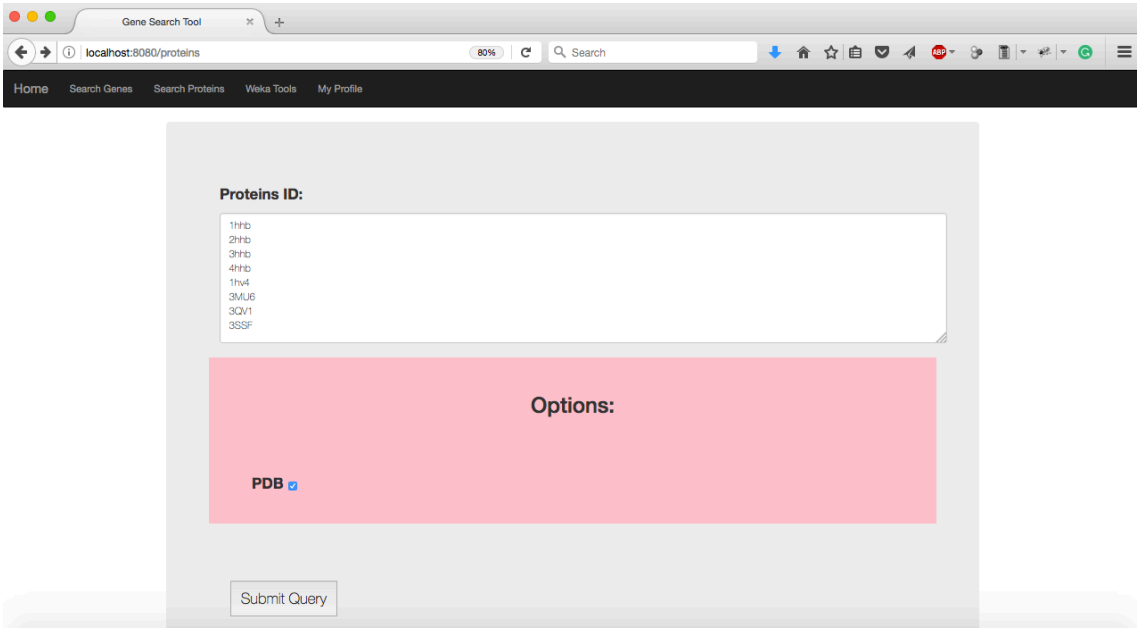


Figura 29: Página de pesquisa de proteínas

no Portal Web. Em seguida foi selecionada a base de dados do PDB para se fazer a pesquisa.

4.2.4.3 Resultados

Após a pesquisa feita é mostrada ao utilizador uma lista das proteínas pesquisadas com uma breve descrição de cada uma, como é mostrado na figura 30. A informação detalhada pode ser mostrada ao utilizador carregando no id da proteína, ver exemplo na figura 31, ou então é possível fazer o download de toda a informação das proteínas.

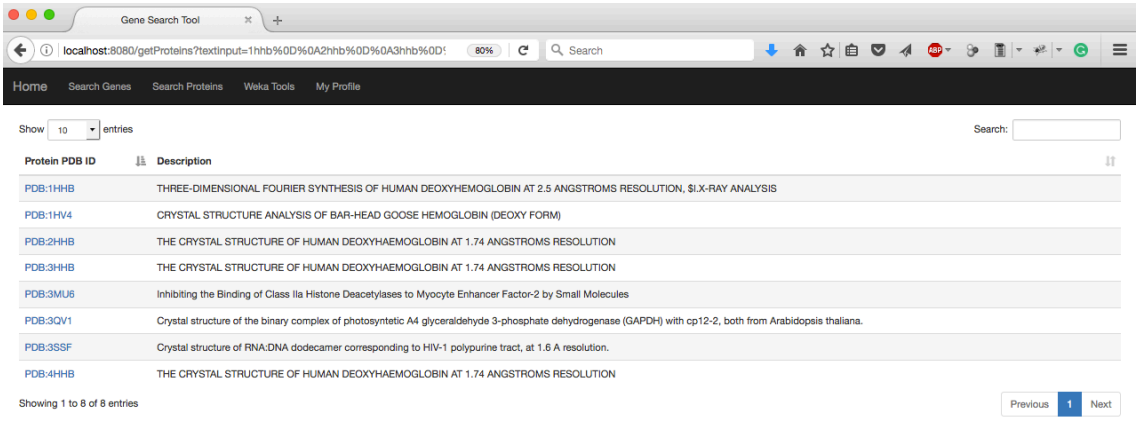


Figura 30: Página com lista das proteínas pesquisadas

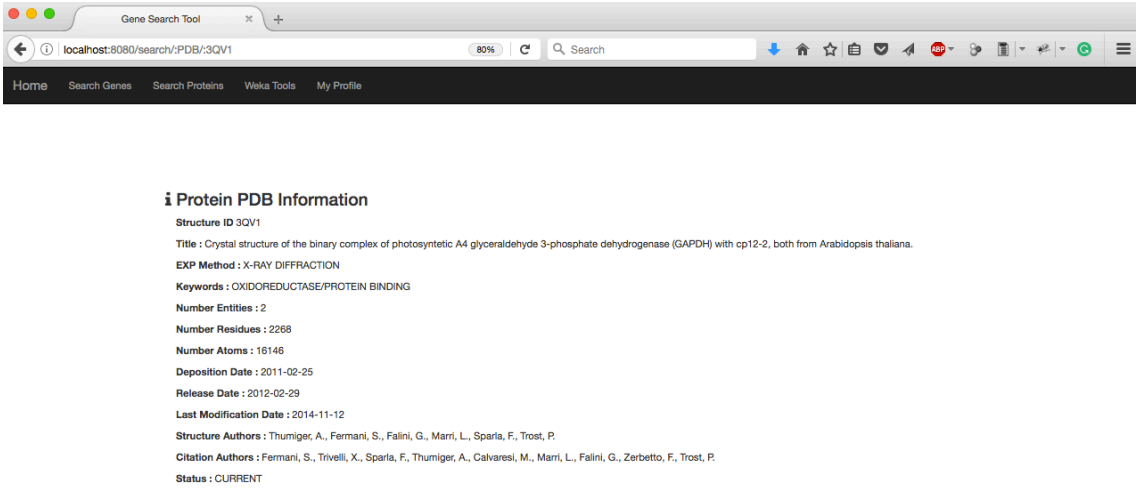


Figura 31: Página com informação detalhada de uma proteína

4.2.5 Comparação da Pesquisa de genes

Como em qualquer ferramenta recentemente desenvolvida, é essencial avaliar a nova ferramenta é uma melhoria em relação aos métodos anteriormente disponíveis. Como tal, a ferramenta deve ser avaliada do ponto de vista do utilizador, nomeadamente em termos de simplificação e eficiência das tarefas a realizar. Esta comparação é baseada no caso de estudo 2 descrito na secção 4.2.3.

4.2.5.1 Simplificação

A Figura 32 mostra como o Portal Web simplificou o processo de pesquisas de genes.

O Portal Web simplifica as tarefas de pesquisa de genes, fazendo automaticamente as conversões dos identificadores necessárias para a pesquisa, procurando ao mesmo tempo informação nas diferentes bases de dados. Gera *datasets* em formato weka, e corre algoritmos de *clustering*.

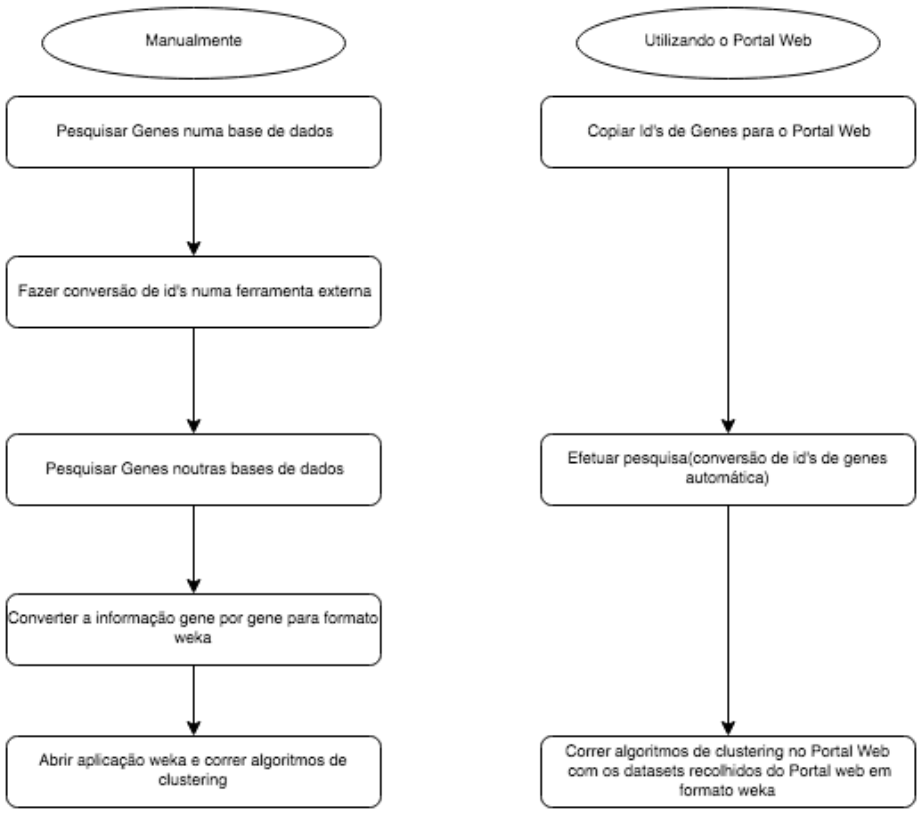


Figura 32: Comparação entre pesquisa manual e pelo Portal Web

Ao realizar uma pesquisa manual, era necessário converter os identificadores dos genes para as respetivas bases de dados. As pesquisas nas diferentes bases de dados têm de ser realizadas em diferentes websites com os diferentes motores de pesquisa de genes de cada website. Além disso, necessitava-se de fazer uma conversão de cada gene para um ficheiro em formato weka, de modo a utilizar a ferramenta weka para análise dos dados com algoritmos *clustering*.

4.2.5.2 Eficiência

É fundamental para a aplicação desenvolvida realizar as tarefas num tempo inferior ao que está atualmente disponível. Desta forma é comparado o tempo médio de uma pesquisa manual e de uma pesquisa no Portal Web para o mesmo conjunto de dados, nas bases de dados do ensembl, kegg e NCBI.

Para fazer uma pesquisa manual de um gene nas diferentes bases de dados e a conversão para depois este ser analisado, demora cerca de quinze minutos. Ou seja, para todo o conjunto de dados do caso de estudo (22 genes), eram necessárias uma média de cinco horas e trinta minutos.

Por outro lado, uma pesquisa para o mesmo conjunto de dados através do portal leva menos tempo a pesquisar e a gerar *datasets* para ser analisado. A tabela 11 mostra a comparação entre os tempos da pesquisa manual e a pesquisa através do portal web.

Conjunto de dados	**Pesquisa e Análise no Portal Web**	**Pesquisa e Análise Manual**
22	7min	≈5h30

Tabela 7: Comparação entre os tempos de pesquisa e análise de genes

4.3 Conclusões e Resumos

Após a realização dos casos de estudo foi possível concluir que com o Portal Web conseguimos uma melhor eficiência uma vez que o tempo de execução é muito menor do que o disponível atualmente e veio simplificar este processo uma vez que o utilizador não necessita de realizar tantas etapas para atingir o objetivo.

Capítulo 5

Conclusões e Trabalho Futuro

Neste capítulo são apresentadas as conclusões sobre o Portal Web e as previsões do trabalho futuro.

5.1 Conclusão

Este projeto culminou com a implementação de um website de pesquisa de genes, o Portal Web para enriquecimento da informação genómica e proteómica, com o objetivo de tornar esta pesquisa de genes numa tarefa mais fácil e mais rápida. Este website realiza pesquisas nas bases de dados do Kegg, NCBI e Ensembl, fazendo ao mesmo tempo a conversão de id's entre bases de dados. Além disso o portal web possui uma base de dados própria na qual vão sendo armazenados os dados sobre os genes à medida que vão sendo pesquisados, para que em pesquisas futuras o tempo de pesquisa ser bastante reduzido.

No final da implementação e da fase de testes, é notório que a pesquisa de genes utilizando o Portal Web tornou-se numa tarefa mais simples e menos demorada. Não sendo necessário estar sempre a mudar de website em website nem de precisar de recorrer a outras ferramentas para a conversão dos id's, tal como era o principal objetivo definido desta dissertação.

5.2 Trabalho Futuro

Apesar deste projeto estar de acordo com o que inicialmente foi previsto, existem sempre alguns aspetos que podem ser melhorados.

5.2.1 Extensão a mais websites de genes e proteínas

No Portal web apenas são feitas pesquisas nas três maiores bases de dados de genes e proteínas: Kegg, NCBI e Ensembl. No futuro espera-se que seja estendida a outras bases de dados, e que esteja sempre em constante atualização caso surjam outras bases de dados importantes.

5.2.2 Optimizações da base de dados

As otimizações de bases de dados são importantes pois podem fazer com que as pesquisas demorem menos tempo, e o utilizador não tenha de esperar muito tempo até a pesquisa sobre os seus genes seja concluída. Além disso estas otimizações podem melhorar o funcionamento geral do website.

Referências

[GENO1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. Genome Biology, 11(10):R106, 2010.

[GNBK] http://www.ncbi.nlm.nih.gov/genbank/, online em Julho de 2016.

[ENSBL] http://ensemblgenomes.org/ , online em julho de 2016.

[PDB] http://www.rcsb.org/pdb/home/home.do, online em julho de 2016.

[SWIPROT] http://www.ebi.ac.uk/uniprot, online em julho de 2016.

[GO] http://geneontology.org/, online em julho de 2016.

[DTMIN] Jiawei Han, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.

[GENO2] http://www.news-medical.net/life-sciences/What-is-Genomics.aspx, online em julho de 2016.

[WEKA] http://www.cs.waikato.ac.nz/ml/weka/, online em julho de 2016.

[RAPM] http://rapidminer.com/, online em julho de 2016.

[CLUST] Nuno A Fonseca, Vítor Santos Costa, and Rui Camacho. Conceptual clustering of multi-relational data. In Inductive Logic Programming, pages 145–159. Springer, 2012.

[PSQL] https://www.postgresql.org/, online em julho de 2016.

[MON] https://www.mongodb.com/, online em julho de 2016.

[DJAN] https://www.djangoproject.com/, online em julho de 2016.

[NODE] https://nodejs.org/en/, online em julho de 2016.

[ANGU] https://angularjs.org/, online em julho de 2016.

[UIKIT] http://getuikit.com/, online em julho de 2016.

[BSTR] http://getbootstrap.com/, online em julho de 2016.

[Rib16] Leandro Ribeiro. O que é uml e diagramas de caso de uso: Introdução prática à uml. Disponível em http://www.devmedia.com.br/o-que-e-uml-e-diagramas-de-caso-de-uso-introducao-pratica-a-uml/23408, janeiro 2017.

[Robert] Robert Lyons. A molecular biology glossary. DNA Sequencing Core, University of Michigan, July 1998, Disponível em http://seqcore.brcf.med.umich.edu/doc/educ/dnapr/mbglossary/mbgloss.html, janeiro 2017.

[Ensembl] HubbardT.; et al. (January 2002). *"The Ensembl genome database project"*. Nucleic Acid Res. 30 (1): 38–41. Doi: https://doi.org/10.1093/nar/30.1.38, online em janeiro 2017

[Kegg] Kanehisa M, Goto S (2000). *"KEGG: Kyoto Encyclopedia of Genes and Genomes"*. Nucleic Acids Res. 28 (1): 27–30. Doi: https://doi.org/10.1093/nar/28.1.27, online em janeiro 2017

[NCBI] re3data.org: NCBI; editing status 2016-01-19; re3data.org - Registry of Research Data Repositories. Doi: http://doi.org/10.17616/R37G7F, online em janeiro 2017

[Genomics] Gomase, V., Tripathi, A., & Tagore, S. (2009). Genomics: new aspect of cancer research. International Journal of Systems Biology, 1(1), 1–19. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Genomics+:+New+aspect+of+cancer+research#0

Appendix A

Exemplo da informação extraída do Kegg, Ensembl, NCBI

Exemplo de um objeto JSON retornado pela API do Ensembl ao pedido get do Portal Web.

Pedido GET:

GET http://rest.ensembl.org/lookup/id/ENSG00000227059?expand=1

Resposta:

[{ source: 'ensembl_havana',
object_type: 'Gene',
logic_name: 'ensembl_havana_gene',
version: 6,
species: 'homo_sapiens',
description: 'anomalous homeobox [Source:HGNC Symbol;Acc:HGNC:40024]',
display_name: 'ANHX',
assembly_name: 'GRCh38',
biotype: 'protein_coding',
end: 133236095,
seq_region_name: '12',
db_type: 'core',
strand: -1,
id: 'ENSG00000227059',
Transcript: [[Object], [Object]],
start: 133218312,
sequenceGene: 'ACCCCGCACCGCACACCCCAGAAACCCCAGGTCGTCCGGGACTCC
TCGGACCCGCAGATGCCACGGACACCAGATCCCCCACGGACCCCTCAGCTCCCCCGGACTCCGCGGTCG
CATCGGGGGCTGAGGGGCGCCGGCCCCCGGGACGCCTTGTGGGCGGGGCCTCGCGGGATTGGCTGCGA
GCCT…' }]

Exemplo da informação em texto retornada pela API do Kegg ao pedido get do Portal Web.

Pedido GET:

GET http://rest.kegg.jp/get/hsa:84792

Resposta:

```
ENTRY       84792           CDS       T01001
NAME        FAM220A, ACPIN1, C7orf70, SIPAR
DEFINITION  (RefSeq) family with sequence similarity 220 member A
ORGANISM    hsa  Homo sapiens (human)
POSITION    7p22.1
MOTIF       Pfam: FAM220
DBLINKS     NCBI-ProteinID: NP_001032240
NCBI-GeneID: 84792
            OMIM: 616628
            HGNC: 22422
            HPRD: 14427
            Ensembl: ENSG00000178397
            Vega: OTTHUMG00000122091
            UniProt: Q7Z4H9
AASEQ       259
            MRDRRGPLGTCLAQVQQAGGGDSDKLSCSLKKRMPEGPWPADAPSWMNKPVVDGNSQSEA
            LSLEMRKDPSGAGLWLHSGGPVLPYVRESVRRNPASAATPSTAVGLFPAPTECFARVSCS
            GVEALGRRDWLGGGPRATDGHRGQCPKGEPRVSRLPRHQKVPEMGSFQDDPPSAFPKGLG
            SELEPACLHSILSATLHVYPEVLLSEETKRIFLDRLKPMFSKQTIEFKKMLKSTSDGLQI
            TLGLLALQPFELANTLCHS
NTSEQ        780
            atgagggacagaagagggcccctcggcacctgcctggcacaagtgcagcaggccggagga
            ggtgactcggacaaactatcatgcagccttaagaaaagaatgccggagggcccttggcct
            gcagatgcaccctcctggatgaataagcctgtggttgatggaaattcacaaagtgaggca
            ttatcactggaaatgagaaaggatccgagcggggctggcctctggcttcacagtggcggc
            ccagtgcttccatatgtgagagaatcagtaagaagaaatccagcctcagcagccactccg
            agcacagccgtgggtttgttccctgctccaacagagtgttttgctcgggtgtcctgcagt
            ggtgttgaagctctggggcggcgagactggctgggaggagggcccagggccactgacggc
            cacagaggacagtgccccaaaggagagcctcgggtgtcacgactgccacgccatcaaaaa
            gtgccggaaatgggaagttttcaggatgacccaccaagtgcttttcccaagggtctgggc
            tctgagttggaacccgcttgcctgcactccatcctgtctgcaacgctgcacgtgtatccc
            gaagtgctcctgagtgaggagacaaaacgcattttccttgaccgtttaaagcccatgttt
            tcaaagcaaacaatagaattcaagaaaatgcttaaaagcacctcagatggtctgcagata

                acactggggttactggctctgcaacctttgaattagcaaatacattatgccatagttaa
```

Exemplo da informação formato xml retornada pela API do NCBI ao pedido get do Portal Web.

Pedido GET:

GET www.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=gene&id=84792

Resposta:

```
<?xml version="1.0" encoding="UTF-8" ?><!DOCTYPE eSummaryResult PUBLIC "-//NLM//DTD esummary gene 20150202//EN"
```

"https://eutils.ncbi.nlm.nih.gov/eutils/dtd/20150202/esummary_gene.dtd"><eSummaryResult><DocumentSummarySet status="OK"><DbBuild>Build170127-0300m.1</DbBuild><DocumentSummary uid="84792">
<Name>FAM220A</Name>
<Description>family with sequence similarity 220 member A</Description>
<Status>0</Status>
<CurrentID>0</CurrentID>
<Chromosome>7</Chromosome>
<GeneticSource>genomic</GeneticSource>
<MapLocation>7p22.1</MapLocation>
<OtherAliases>ACPIN1, C7orf70, SIPAR</OtherAliases>
<OtherDesignations>protein FAM220A|STAT3-interacting protein as a repressor|acrosomal protein 1</OtherDesignations>
<NomenclatureSymbol>FAM220A</NomenclatureSymbol>
<NomenclatureName>family with sequence similarity 220 member A</NomenclatureName>
<NomenclatureStatus>Official</NomenclatureStatus>
<Mim><int>616628</int></Mim>
<GenomicInfo><GenomicInfoType><ChrLoc>7</ChrLoc>
<ChrAccVer>NC_000007.14</ChrAccVer>
<ChrStart>6348958</ChrStart>
<ChrStop>6329408</ChrStop>
<ExonCount>2</ExonCount>
<GeneWeight>715</GeneWeight>
<Summary></Summary>
<ChrSort>07</ChrSort>
<ChrStart>6329408</ChrStart>
<ScientificName>Homo sapiens</ScientificName>
<CommonName>human</CommonName>
<TaxID>9606</TaxID>

Appendix B

Exemplo de um ficheiro em formato arff extraído do Portal Web

@relation GenesfromEnsembl.Homeobox

@attribute ID string
@attribute species string
@attribute source string
@attribute type string
@attribute logic_name string
@attribute description string
@attribute display_name string
@attribute assembly_name string
@attribute seq_region_name numeric
@attribute start numeric
@attribute end numeric
@attribute strand numeric
@attribute version numeric
@attribute db_type string
@attribute transcript numeric

@data
ENSG00000227059,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,anomalo
ushomeobox[Source:HGNCSymbol;Acc:HGNC:40024],ANHX,GRCh38,12,133218312,13323
6095,-1,6,core,2

ENSG00000178928,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,tetrapeptiderepeathomeobox1[Source:HGNCSymbol;Acc:HGNC:32174],TPRX1,GRCh38,19,47801243,47819051,-1,8,core,3

ENSG00000119608,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,prosperohomeobox2[Source:HGNCSymbol;Acc:HGNC:26715],PROX2,GRCh38,14,74852871,74871940,-1,12,core,3

ENSG00000123407,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,homeoboxC12[Source:HGNCSymbol;Acc:HGNC:5124],HOXC12,GRCh38,12,53954834,53958956,1,3,core,1

ENSG00000174279,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,even-skippedhomeobox2[Source:HGNCSymbol;Acc:HGNC:3507],EVX2,GRCh38,2,176077472,176083913,-1,4,core,1

ENSG00000234493,homo_sapiens,havana,Gene,havana,Rhoxhomeoboxfamilymember1pseudogene1[Source:HGNCSymbol;Acc:HGNC:51580],RHOXF1P1,GRCh38,0,120010718,120015544,-1,2,core,2

ENSG00000274529,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,SEBOXhomeobox[Source:HGNCSymbol;Acc:HGNC:32942],SEBOX,GRCh38,17,28364268,28365244,-1,5,core,2

ENSG00000235602,homo_sapiens,havana,Gene,havana,POUclass5homeobox1pseudogene3[Source:HGNCSymbol;Acc:HGNC:9222],POU5F1P3,GRCh38,12,8133772,8134849,-1,5,core,1

ENSG00000237872,homo_sapiens,havana,Gene,havana,POUclass5homeobox1pseudogene4[Source:HGNCSymbol;Acc:HGNC:33310],POU5F1P4,GRCh38,1,155433178,155434262,1,4,core,1

ENSG00000259849,homo_sapiens,havana,Gene,havana,VENThomeoboxpseudogene1[Source:HGNCSymbol;Acc:HGNC:30900],VENTXP1,GRCh38,0,26558337,26561052,1,1,core,1

ENSG00000101883,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,Rhoxhomeoboxfamilymember1[Source:HGNCSymbol;Acc:HGNC:29993],RHOXF1,GRCh38,0,120109053,120115937,-1,4,core,1

ENSG00000128713,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,homeoboxD11[Source:HGNCSymbol;Acc:HGNC:5134],HOXD11,GRCh38,2,176104216,176109754,1,12,core,3

ENSG00000282757,homo_sapiens,havana,Gene,havana,doublehomeoboxB[Source:HGNCSymbol;Acc:HGNC:33345],DUXB,GRCh38,16,75694434,75700152,-1,2,core,2

ENSG00000231411,homo_sapiens,havana,Gene,havana,doublehomeobox4like31(pseudogene)[Source:HGNCSymbol;Acc:HGNC:51770],DUX4L31,GRCh38,0,10171590,10172725,-1,1,core,1

ENSG00000270222,homo_sapiens,havana,Gene,havana,doublehomeoboxApseudogene11[Source:HGNCSymbol;Acc:HGNC:51812],DUXAP11,GRCh38,16,59655602,59656336,1,1,core,1

ENSG00000250482,homo_sapiens,havana,Gene,havana,doublehomeobox4like51(pseudogene)[Source:HGNCSymbol;Acc:HGNC:51810],DUX4L51,GRCh38,5,31249879,31250987,1,3,core,1

ENSG00000258336,homo_sapiens,havana,Gene,havana,POUclass5homeobox1pseudogene3[Source:HGNCSymbol;Acc:HGNC:9222],DUX4L52,GRCh38,12,61600067,61600843,-1,2,core,1

ENSG00000206195,homo_sapiens,havana,Gene,havana,doublehomeoboxApseudogene8[Source:EntrezGene;Acc:503637],DUXAP8,GRCh38,22,15784959,15829984,1,10,core,6

ENSG00000150051,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,mohawkhomeobox[Source:HGNCSymbol;Acc:HGNC:23729],MKX,GRCh38,10,27672875,27746060,-1,13,core,2

ENSG00000148516,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,zincfingerE-boxbindinghomeobox1[Source:HGNCSymbol;Acc:HGNC:11642],ZEB1,GRCh38,10,31318495,31529814,1,21,core,25

ENSG00000063515,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,goosecoidhomeobox2[Source:HGNCSymbol;Acc:HGNC:4613],GSC2,GRCh38,22,19148576,19150283,-1,2,core,1

ENSG00000131721,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,Rhoxhomeoboxfamilymember2[Source:HGNCSymbol;Acc:HGNC:30011],RHOXF2,GRCh38,0,120158561,120165630,1,5,core,1

ENSG00000151650,homo_sapiens,ensembl_havana,Gene,ensembl_havana_gene,VENThomeobox[Source:HGNCSymbol;Acc:HGNC:13639],VENTX,GRCh38,10,133237404,133241929,1,7,core,1

ENSG00000227351,homo_sapiens,havana,Gene,havana,Nanoghomeoboxpseudogene6[Source:HGNCSymbol;Acc:HGNC:23104],NANOGP6,GRCh38,10,99788564,99789410,-1,2,core,1

ENSG00000282308,homo_sapiens,havana,Gene,havana,divergent-pairedrelatedhomeoboxpseudogene3[Source:HGNCSymbol;Acc:HGNC:32169],DPRXP3,GRCh38,14,35222174,35222746,1,1,core,1

ENSG00000228670,homo_sapiens,havana,Gene,havana,Nanoghomeoboxpseudogene2[Source:HGNCSymbol;Acc:HGNC:23100],NANOGP2,GRCh38,2,222452400,222453126,-1,4,core,1

ENSG00000231697,homo_sapiens,havana,Gene,havana,Nanoghomeoboxpseudogene5[Source:HGNCSymbol;Acc:HGNC:23103],NANOGP5,GRCh38,9,100175178,100176054,1,3,core,2

ENSG00000231750,homo_sapiens,havana,Gene,havana,Nanoghomeoboxpseudogene10[Source:HGNCSymbol;Acc:HGNC:23108],NANOGP10,GRCh38,0,43407665,43408559,-1,3,core,1

ENSG00000231809,homo_sapiens,havana,Gene,havana,Nanoghomeoboxpseudogene9[Source:HGNCSymbol;Acc:HGNC:23107],NANOGP9,GRCh38,0,65772741,65773632,-1,4,core,1

ENSG00000270552,homo_sapiens,havana,Gene,havana,doublehomeoboxApseudogene3[Source:HGNCSymbol;Acc:HGNC:32182],DUXAP3,GRCh38,10,42747544,42748687,-1,1,core,1