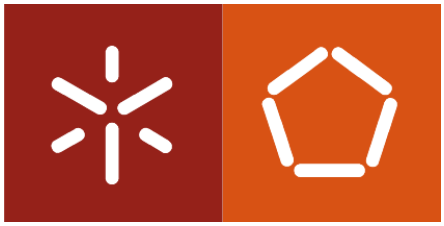

Universidade do Minho

Departamento de Informática

Mestrado Integrado em Engenharia Informática

Bioinformática

Unidade Curricular de Algoritmos para

Análise de Sequências Biológicas

2015/2016

Visualizador multidimensional de alterações genéticas em casos de estudo de cancro

André Geraldes a67673

Emanuel Queiroga a61003

Hélder Gonçalves a64286

Luís Brito a54056

Abstract

The following report describes the creation and development of a tool that allows the visualization of genetic alterations in cancer case studies. The reading of this document is advised to whoever wishes to know in detail the architecture of the application as well as to anyone that has doubts regarding any functionality offered by the tool.

Accordingly, the notion and knowledge that the group has of cancer is introduced followed by the efforts being made worldwide regarding prevention and early detection of the disease. The group also enumerates the motives that lead it to pick this theme in particular accompanied by the proposed objectives expected to be achieved at the end of the development. This report then illustrates the architecture of the tool assisted by an image that easily explains the different components that compose it. Following this is the detailed explanation of the components described before divided in two categories: back end and front end. The first regards the components of the application that the user cannot see which in this case are the PHP and Python parts of the code (where all the pertinent calculations are executed) while the second contains the HTML language module that the application uses to interact with and show data to the user. It is then made a brief description of the final product with images of the different website pages and its description.

Lastly, the group does a small self evaluation stressing some strong aspects of the application as well as enumerating some difficulties encountered along the way accompanied by some improvements and functionalities to add in order to give a more professional look and use to the tool.

It should be noted that the report is written in Portuguese as it is the developers native language as well as the language in which the course for which this project was done is taught.

Resumo

Este relatório descreve a criação e desenvolvimento de uma ferramenta para visualização de alterações genéticas em casos de estudo de cancro. É aconselhada a leitura deste documento a quem pretenda saber em pormenor como funciona a aplicação criada ou em caso de dúvida sobre alguma funcionalidade disponível.

Deste modo, é feita uma introdução onde o grupo fala da sua noção e conhecimento da doença em si e dos esforços feitos a nível mundial na prevenção da doença. É também referida a motivação do grupo na escolha do trabalho assim como os objetivos propostos no desenvolvimento da ferramenta. De seguida é explicitado o funcionamento da ferramenta com auxílio a uma imagem que ilustra a arquitetura da mesma e como se interligam todos os componentes presentes, divididos em duas categorias: *back end* e *front end*. A primeira diz respeito a todos os componentes que o utilizador não vê, que neste caso é o código *PHP* e *Python* que corre no servidor remoto enquanto a segunda contêm o módulo da linguagem *HTML* que a ferramenta usa para mostrar os dados e interagir com utilizador. Seguidamente é feita uma descrição da ferramenta com imagens do resultado final e uma pequena descrição sobre as várias páginas do *website*.

Por fim, o grupo auto avalia-se salientando aspetos fortes do produto final e enumerando algumas dificuldades encontradas seguidas de ideias e funcionalidades importantes a adicionar à ferramenta de modo a tornar toda a aplicação mais profissional e eficiente.

Índice

1 Introdução

1.1 Contextualização

No âmbito da disciplina de Algoritmos de Análise de Sequências Biológicas foi pedido aos alunos do Mestrado Integrado em Engenharia Informática que desenvolvessem um projeto, a escolher entre várias opções, de modo a serem avaliados na componente prática da disciplina.

Deste modo, o grupo escolheu desenvolver um visualizador multidimensional de alterações genéticas em estudos de cancro por motivos explicitados na secção seguinte, *Motivação*.

A pergunta que inevitavelmente se coloca quando se discute um tema tão sensível como o cancro é como podemos fazer progressos na prevenção e deteção precoce do mesmo. Esta é uma das questões mais discutidas a nível mundial e todas as semanas surgem ideias diversas e inovadoras acerca deste tema. O cancro é o nome dado a um conjunto de doenças que causado pela divisão descontrolada de células do corpo humano e consequente propagação para os tecidos adjacentes. Assim, este fenómeno pode começar em qualquer parte do corpo humano, constituído por milhares de milhões de células pois, normalmente, as células humanas crescem e dividem-se em novas células à medida que o corpo necessita. Quando uma célula envelhece ou fica danificada, morre e é substituída por uma nova que toma o seu lugar. No entanto, quando o cancro se desenvolve, este processo metódico é quebrado e, à medida que as células se tornam mais anormais, células velhas ou danificadas sobrevivem quando não deviam e novas células são formadas desnecessariamente.

1.2 Motivação

Depois de lida atentamente a contextualização, consegue-se perceber o que levou o grupo a escolher este projeto. Certamente todos nós já ouvimos falar de um caso de um conhecido nosso que sofre desta doença e, tendo-se apresentado a hipótese de desenvolver uma ferramenta nesta área, quisemos dar a nossa contribuição no estudo das mutações a causam. Foi-nos também dito que não existem muitas ferramentas que permitam apresentar a informação de forma simples e intuitiva. Assim, achamos que o desenvolvimento desta ferramenta tal como foi pedido, auxiliado de algum trabalho futuro, pode realmente contribuir para um melhor entendimento das mutações que causam esta angustia na vida de tantas pessoas.

1.3 Objetivos

Perante o domínio do problema, i.e., um visualizador multidimensional de alterações genéticas em estudos de cancro, o grupo propõe-se a criar um *website* que permita visualizar os dados, mutações genéticas em pacientes que sofram da doença, em forma de matriz bidimensional com as seguintes funcionalidades:

- Permitir organizar linhas e colunas de diferentes formas;
- Visualizar com diferentes cores os vários tipos de mutações;
- Fazer o agrupamento da informação segundo a importância de cada mutação;
- Permitir o utilizador introduzir um ficheiro de configuração.

Para além disso, pretende-se que a ferramenta cumpra as funcionalidades que dela são esperadas de forma rápida e eficaz.

Concretizados os objetivos anteriores, tornar-se-á possível ao utilizador ter uma representação bastante intuitiva dos dados para assim conseguir procurar padrões na informação.

2 Desenvolvimento

2.1 Arquitetura

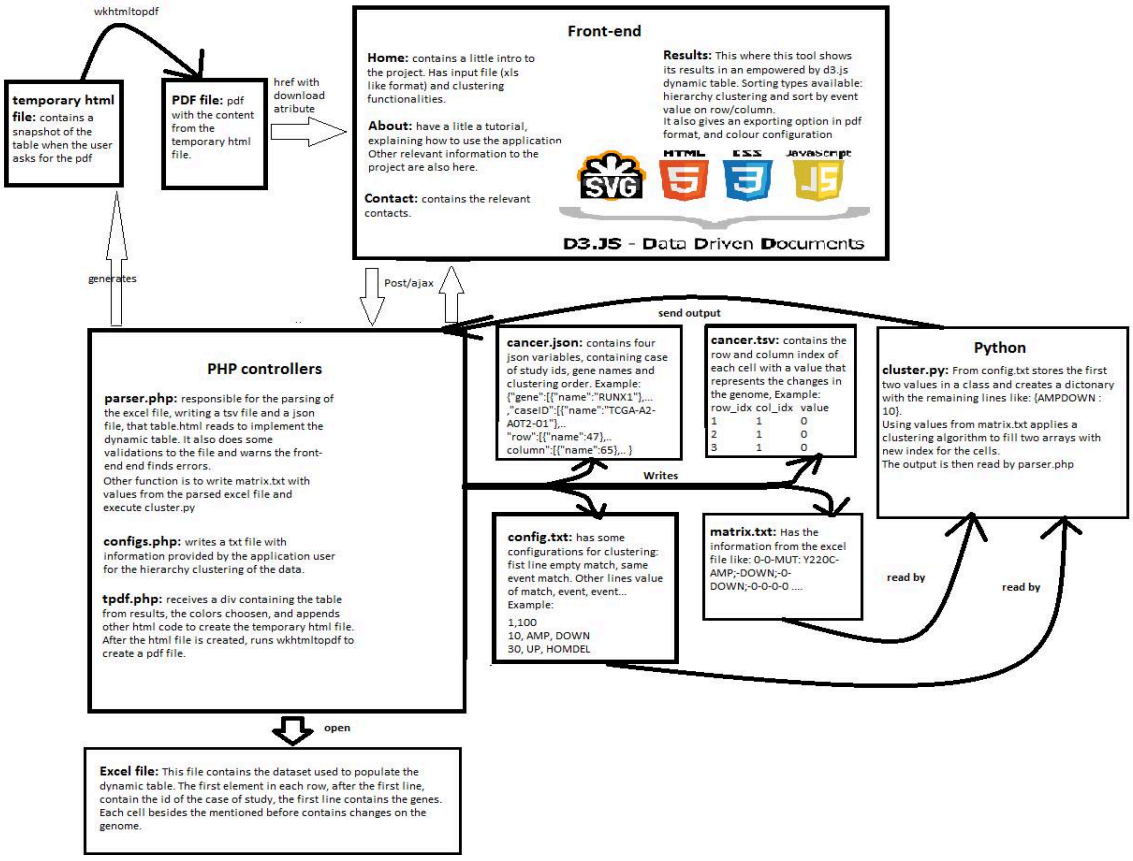

Ilustração 1 Arquitetura da ferramenta.

2.2 Back End

Para se poder trabalhar nos dados foi preciso fazer *parse* e transformações aos mesmo para os tornar usáveis na nossa ferramenta. Para isso recorremos a scripts em python e php. O python foi utilizado para a parte de algoritmia na análise aos dados e clustering, o php foi utilizado para o tratamento dos dados e criação dos ficheiros necessários.

PHP

Foram criados dois scripts em PHP, um parser.php e um tpdf.php, este ultimo serve apenas para criar uma pagina html temporária que é usada pela aplicação wkhtmltopdf que transforma o html em pdf, este script é invocado quando se carrega no botão PDF na página de resultados.

O script parser.php é chamado quando é inserido o ficheiro Excel com os dados, utiliza a biblioteca PHPExcel para aceder aos dados, nesta parte de leitura dos dados é feito o seu *parsing* para gerar os ficheiros necessários. Primeiro cria um ficheiro

chamado matrix.txt com as células do Excel que contém os eventos de cada caso de teste, i.e mutações, amps, etc., este txt é utilizado pelo python para acesso aos dados, poderia ter sido usada uma biblioteca em python para trabalhar com Excel como a openpyxl mas como poderia haver problemas de dependências e versões no servidor remoto decidimos utilizar o python sem bibliotecas que necessitavam de instalação. Após isto, o php corre o script python parser.py para realizar o clustering, que retorna 2 arrays com índices, e guarda os seus resultados em forma de string numa variável que é separada para se obter os índices, volta depois ao ficheiro Excel onde são depois retirados os nomes dos genes da primeira linha e os nomes dos caseIDs da primeira coluna, estas informações são necessárias para a criação dos resultados, para isso é criado um ficheiro JSON com os nomes dos genes, caseIDs e os arrays resultantes do clustering. E por fim são utilizadas expressões regulares para perceber o que contém cada célula do ficheiro Excel, para se poder atribuir um valor a cada tipo que será usado para a indentificação nos resultados, por exemplo um "AMP" terá valor '1' e uma mutação do tipo frameshift será C, caso exista mais do que um na mesma célula de Excel é feita concatenação e fica por exemplo '1C', estes valores são depois escritos num ficheiro TSV que em que cada linha contém os índices do valor e o valor. Este script verifica também se o ficheiro inserido está em formato Excel, mas não é feita nenhuma validação aos dados contidos no ficheiro Excel. Estes ficheiros criados pelo PHP serão utilizados pela biblioteca D3.js para a apresentação dos dados.

Python

Nesta linguagem, aprendida integralmente nas aulas, é feito o agrupamento de dados para permitir uma visualização pertinente dos dados recebidos. De modo a fazer tal agrupamento, analisamos o problema e decidimos aplicar um algoritmo de *clustering*. No entanto, existem muitos algoritmos diferentes em que cada um tem os seus pontos fortes e fracos, obtendo assim diferentes resultados para os mesmos dados.

De modo a prosseguirmos foi necessário decidir o tipo de clustering a usar, a forma como vão sendo agrupados os elementos e a função de distância. Para nos apoiar nesta decisão, tivemos a ajuda do professor assim como de literatura encontrada na internet (ver Bibliografia), o que facilitou bastante a escolha e nos permitiu compreender a melhor maneira de abordar o problema. É de notar que foi também dado um algoritmo de *clustering* hierárquico (UPGMA) nas aulas da cadeira. Assim, decidimos que o melhor seria implementar um algoritmo hierárquico pois é o mais intuitivo e foi já aprofundado nas aulas.

De seguida, de modo a calcular a distância colunas (ou linhas) da matriz, foi necessário decidir como vão sendo agrupadas as distâncias entre células. Para tal, encontramos três formas simples e intuitivas de resolver o problema, sendo elas *complete-linkage*, *single-linkage* e *average-linkage clustering*. As anteriores, para uma dada função de similaridade, calculam a distância entre duas listas de valores de maneira diferente. A primeira diz que a distancia entre duas listas é a maior distância entre dois valores das respetivas listas. Esta abordagem é relevante pois poderia agrupar a informação pelas mutações mais importantes. A segunda sugere definirmos a distância como a distância mínima presente o que neste caso não faz sentido pois existem muitas células vazias, tornando assim este método impossível. Por fim, o

average-linkage clustering faz uma média de todas as distâncias presentes, sendo dada importância à ocorrência de mutações iguais, assim como células vazias na mesma posição. O grupo achou que este método era o melhor para resolver o nosso problema pelas razões descritas em cima. Coincidentemente, o algoritmo dado na aula (UPGMA) é um algoritmo deste tipo, o que facilitou imenso a implementação do mesmo pois já o tínhamos implementado num trabalho de casa.

Para completar o algoritmo faltava então decidir a função de similaridade a usar. Neste caso não foi possível aceder a literatura pois estamos a usar um conjunto de dados único e, portanto, a similaridade entre as células é única para este problema. Deste modo, e como tinha sido definido nos objetivos, deixamos o utilizador introduzir 3 tipos de valores: o valor de igualdade em caso de células vazias, o valor em caso de existir exatamente o mesmo tipo de mutações nas duas células e valores para mutações que tenham a mesma relevância, i.e., se dois eventos diferentes acontecerem, terão o valor definido.

Decidido então o algoritmo a usar, verificamos que é igual ao que foi dado nas aulas. A implementação foi então fácil, sendo apenas mais complicado introduzir os diferentes valores na função de similaridade.

2.3 Front End

Para a interação do utilizador com a ferramenta foi utilizado HTML, CSS e JavaScript, tornado possível disponibilizar a ferramenta online para a sua utilização poder ser feita por outras pessoas. A estrutura base foi retirada da Framework Bootstrap, sendo bastante simples apenas com uma barra de navegação cinza, foi depois alterada e adaptada por nós.

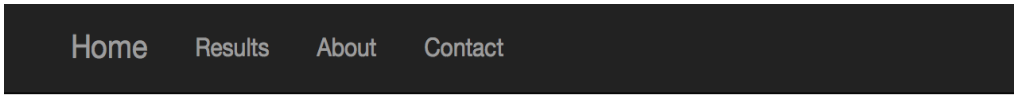

Ilustração 2 Barra de navegação.

A ferramenta está organizada em quatro páginas distintas, página inicial, página dos resultados, uma página com um tutorial e uma página com os contactos.

A página inicial apresenta uma breve descrição do tema, uma imagem ilustrativa dos resultados e uma área onde é possível carregar o ficheiro Excel e escrever as configurações para o clustering.

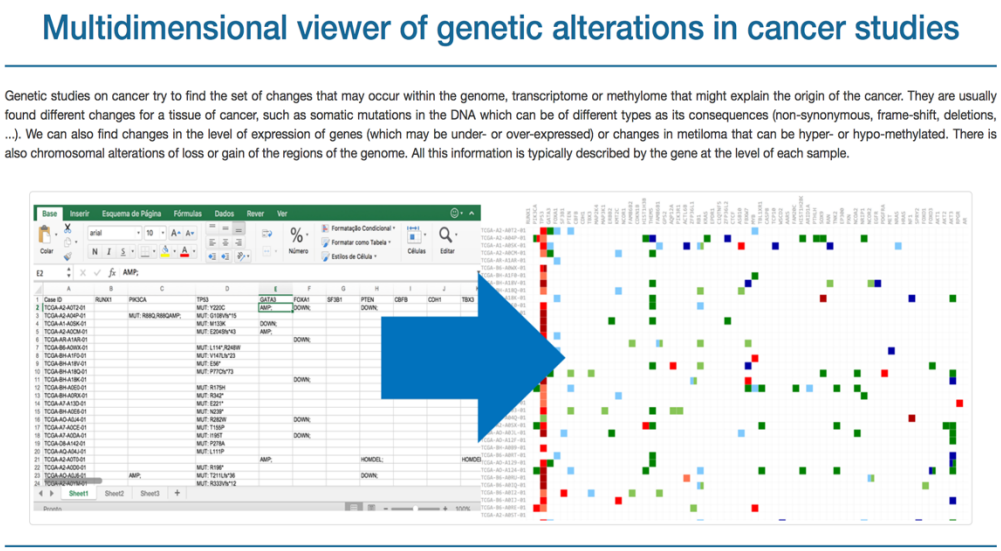

Ilustração 3 Texto inicial e imagem.

Ilustração 4 Área para inserção do ficheiro.

Na página resultados, é apresentado no inicio o esquema de cores padrão, a possibilidade de configurar algumas cores e a opção para gerar um ficheiro pdf. Após isto é possível escolher a maneira como os dados estão ordenados.

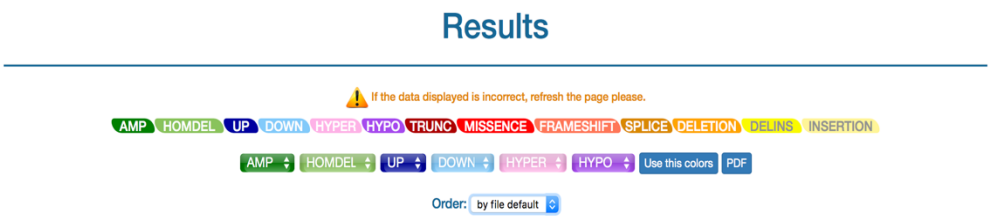

Ilustração 5 Inicio da página de resultados.

São depois apresentados os resultados, para esta parte recorremos à biblioteca D3.js que serve para visualização de dados recorrendo a SVG, HTML e CSS.

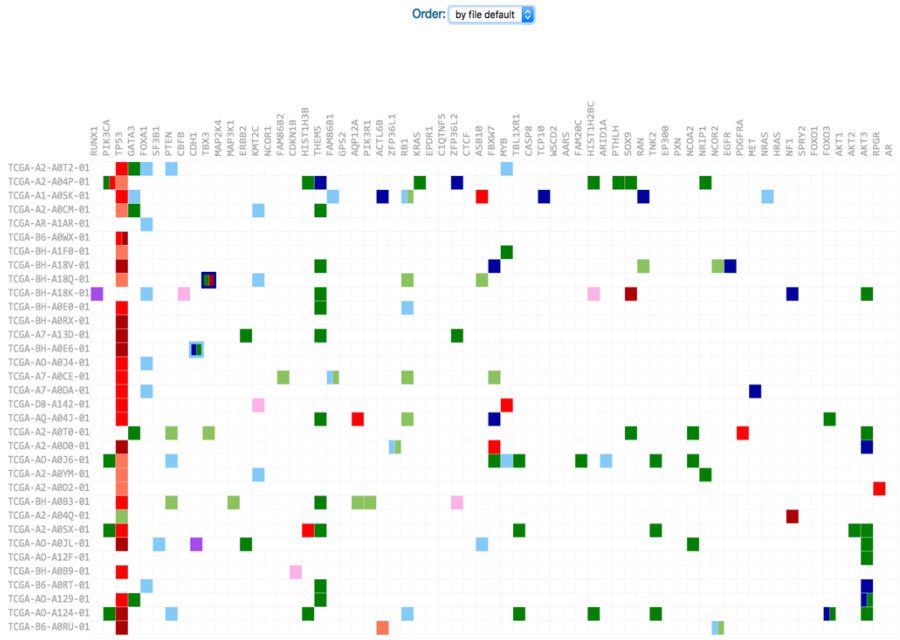

Ilustração 6 Exemplo de resultados.

Na página "About" encontra-se um tutorial sobre como utilizar a ferramenta e referencias a algumas ferramentas e bibliotecas utilizadas para a realização deste trabalho. Nesta página o utilizador tem também acesso a um ficheiro Excel de exemplo para utilizar na página.

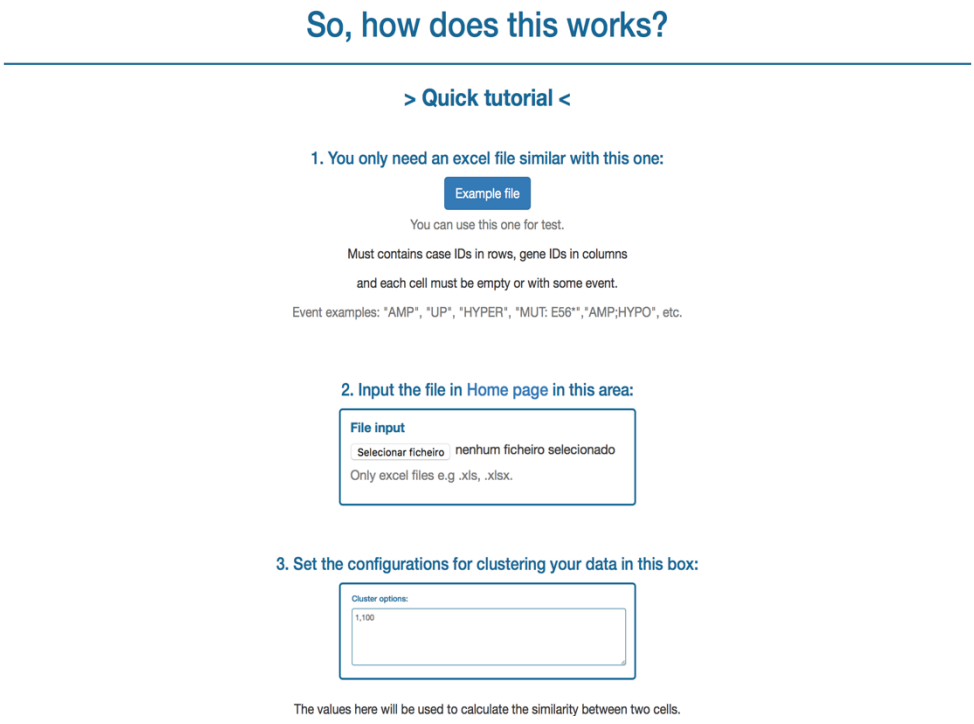

Ilustração 7 Primeira parte do tutorial.

Ilustração 8 Segunda parte do tutorial.

Ilustração 9 Ultima parte da página "About".

Por fim temos a página “Contact” que contém algumas informações nossas e respetivos contactos.

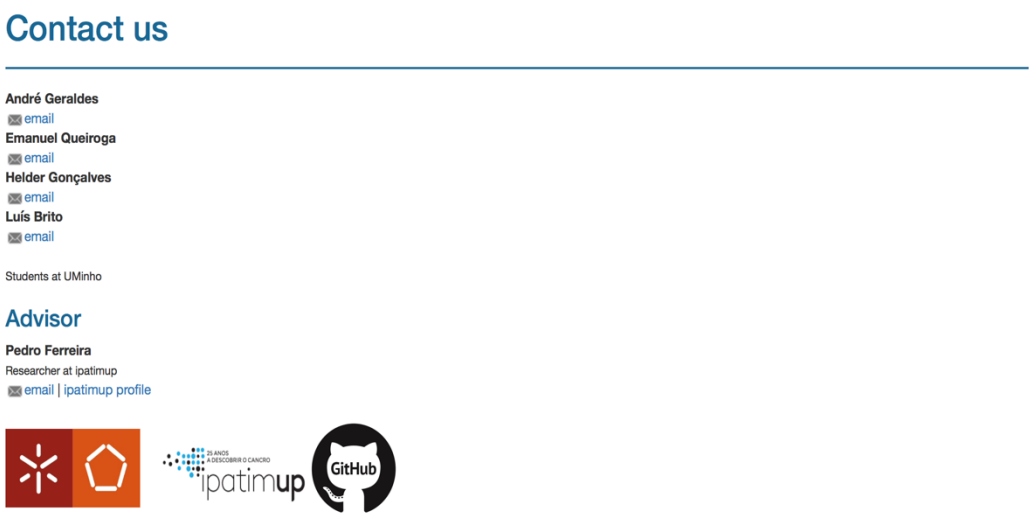

Ilustração 10 Página contactos.

Para correr a aplicação foi utilizada a aplicação MAMP para correr localmente e para colocar online utilizamos um servidor online, DigitalOcean, e um domínio grátis. cancerviewer.me

Para mais detalhes sobre como tudo foi feito, o código encontra-se todo detalhadamente comentado e está disponível para consulta, isto facilitará também uma possibilidade de continuação futura.

3. Conclusão e trabalho futuro

Após a realização do projeto, torna-se necessário e fundamental salientar alguns aspetos fortes e dificuldades encontradas ao longo da implementação de todo o processo. Neste projeto, o grupo utilizou conhecimentos adquiridos em anteriores unidades curriculares bem como conhecimentos adquiridos na disciplina de Algoritmos para Análise de Sequências Biológicas. No entanto, mesmo com estes conhecimentos, o grupo deparou-se com o uso de tecnologias e técnicas que nunca tinha usado, aumentando assim a dificuldade deste projeto pois foi necessário aprendermos a usar linguagens novas para alguns elementos do grupo assim como novas bibliotecas de software.

Sendo este projeto uma ferramenta de uso académico e o resultado de um projeto de grupo para avaliação prática de uma cadeira, o seu uso é limitado. Para um uso mais profissional da ferramenta seria necessário adicionar algumas funcionalidades e otimizações que dariam uma maior eficiência à aplicação. Entre funções que desejaríamos implementar com trabalho futuro destacam-se:

- Criar uma base de dados e implementar a noção de sessão;
- Dar mais liberdade no *input* de dados e configuração ao utilizador;
- Validar os dados recebidos na folha de cálculo;
- Melhorar a exportação para ficheiros PDF;
- Apresentar um dendrograma resultante do *clustering* hierárquico dos dados;
- Melhorar a função de similaridade de modo a otimizar os resultados;
- Mudar o local de *hosting* da aplicação, aumentando imenso a eficiência.

Aplicadas estas mudanças e implementadas estas funcionalidades, pensamos que a ferramenta teria um uso bastante aceitável e poderia auxiliar especialistas na área a melhor compreender as mutações que causam o cancro.

Por fim, o grupo acha que cumpriu os objetivos propostos pelo professor e que conseguimos usar o conhecimento adquirido na cadeira.

4. Bibliografia

Kramer, Barry. "The Importance of Cancer Prevention Research and Its Challenges." *National Cancer Institute*

"Hierarchical Clustering." *Wikipedia*. Wikimedia Foundation

UPGMA." *Wikipedia*. Wikimedia Foundation

"Hierarchical Clustering." *Documentation*. MathWorks