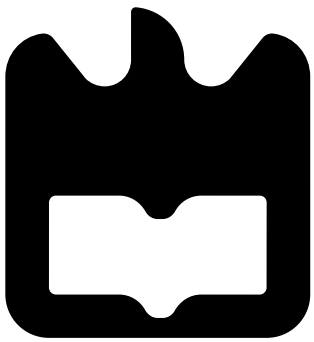
Universidade de Aveiro
Departamento de
Electrónica, Telecomunicações e Informática
2012

Ricardo Mateus Gonzaga

Sistema de informação de armazenamento e procura de proteínas



Ricardo Mateus Gonzaga

Sistema de informação de armazenamento e procura de proteínas

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia de Computadores e Telemática, realizada sob a orientação científica do Doutor José Luís Guimarães Oliveira, Professor Associado do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro

o júri / the jury

presidente / president

Doutor Joaquim Manuel Henriques de Sousa Pinto
Professor Auxiliar da Universidade de Aveiro

vogais / examiners committee

Doutor António Manuel de Jesus Pereira
Prof. Coordenador do Dep. de Engª Informática da Esc. Sup. de Tecnologia e Gestão do Inst. Politécnico de Leiria

Doutor José Luis Guimarães Oliveira
Professor Associado da Universidade de Aveiro (orientador)

Doutor Miguel Monsanto Pinheiro
Biocant (co-orientador)

agradecimentos /
acknowledgements

Depois de seis anos de trabalho é com muita satisfação que agradeço em primeiro lugar aos meus pais e irmãos por todos os sacrifícios feitos para permitir que eu alcançasse este objetivo e também agradecer todo o apoio e entusiasmo que me deram nos momentos mais complicados.

Quero também agradecer ao meu orientador Doutor José Luís Guimarães Oliveira e ao Doutor Miguel Monsanto Pinheiro que me acompanharam no desenrolar deste projeto, mostrando-se sempre disponíveis e cujos pareceres foram fundamentais para a conclusão do trabalho.

A todos os meus amigos e colegas que durante seis anos me acompanharam e ajudaram e pelos momentos de convívio e de descontração, que quebraram tantas dificuldades, ao longo desta etapa.

Palavras-chave

Bioinformática, biotécnologia, sequenciação, proteínas, InterPro, InterProScan, domínios funcionais.

Resumo

Com a redução dos custos de sequenciação, muitos grupos de investigação têm efetuado sequenciações maciças dos organismos com que trabalham. Estas sequenciações revelam a combinação de nucleótidos (A,C,T,G) que compõe os genomas, desvendando o código para a construção das proteínas, que por sua vez, são usadas nos mais diversos processos como regulação, crescimento, manutenção, entre outros mecanismos.

As sequências proteicas poderão ter um elevado potencial biotecnológico. Por exemplo, quando se sequenciam organismos que vivem em ambientes extremos, tais como, com elevadas concentrações de produtos tóxicos, grandes profundidades ou elevados níveis de radiação, as proteínas que os protegem contra esses agentes poderão ser importantes em diversas áreas da biotecnologia.

Atualmente existem vários métodos que permitem reconhecer as propriedades específicas de cada proteína, desde a análise da sequência, passando pela análise estrutural ou até mesmo funcional. Ferramentas como o InterProScan permitem conhecer de uma forma rápida e simples as mais valias funcionais de uma sequência e perceber se uma dada proteína pode ter ou não potencial para que seja produzida num ambiente laboratorial.

O objetivo deste trabalho é desenvolver um sistema que permita organizar e catalogar todas as proteínas pertencentes a um determinado grupo de investigação sendo capaz de publicar características funcionais sem desvendar as suas sequências genéticas. A publicação e pesquisa destas proteínas será baseada nos seus grupos funcionais, revelados recorrendo à ferramenta InterProScan. Este sistema é constituído por dois componentes, uma aplicação local e um site web, e destina-se a grupos de investigação que pretendam divulgar as proteínas sequenciadas nos seus laboratórios e a entidades que necessitem de proteínas com propriedades específicas.

Keywords Bioinformatics, biotechnology, sequencing, proteins, InterPro, InterProScan, functional domains.

Abstract

With the decrese in the cost of genome sequencing, many research groups have made massive sequencing of organisms with which they work. These sequencing reveals the combination of nucleotides (A, C, T, G) that makes up the genome, revealing the code for constructing the proteins which are used in several processes such as regulation of growth, maintenance, and other mechanisms.

Proteins may have a high biotechnological potential. For example when sequencing organisms that live in extreme environments such as high concentrations of toxic products, high deep or high levels of radiation, the proteins that protect against these agents may be important in areas of biotechnology.

Currently there are several methods for recognizing the specific properties of each protein since the sequence, structural or even functional analysis. Tools, such as InterProScan, allows us to quickly know the protein function and to know if a sequence of a given protein may or may not have the potential to be produced in a laboratory environment for commercial use.

The objective of this work was to develop a system that organizes and catalogs all the proteins belonging to a particular research group, and to develop a database and a web site that allows the storage and search of proteins in several areas of interest. The publication and research of these proteins will be based on their functional groups using the InterProScan tool. This system consists in two components, a local application and a web site and was designed for research groups that wish to publish the proteins in their laboratories and to entities that require proteins with specific properties that were sequenced.

Conteúdo

Lista de Figuras

Lista de acrónimos

ADN Ácido Desoxirribonucleico

API Application Programming Interface

CATH Class Architecture Topology Homologous superfamily

CRC Cyclic Redundancy Check

CSS Cascading Style Sheets

EBI European Bioinformatics Institute

GO Gene Ontology

GUI Graphical User Interface

HAMAP High-quality Automated and Manual Annotation of microbial Proteomes

HTML HyperText Markup Language

HTTP Hypertext Transfer Protocol

JAWS Java Web Start

JDBC Java Database Connectivity

JRE Java Runtime Environment

MVC Model-view-controller

PANTHER Protein ANalysis THrough Evolutionary Relationships

PIRSF Protein Information Resource SuperFamily

REST REpresentational State Transfer

Scop Structural Classification of Proteins

SDK Software Development Kit

SMART Simple Modular Architecture Research Tool

SSL Secure Socket Layer

UML Unified Modeling Language

UniProtKB UniProt Knowledgebase

XML Extensible Markup Language

Capítulo 1

Introdução

O trabalho corrente está inserido no âmbito das atividades do grupo de Bioinformática da Universidade de Aveiro.

Este grupo tem, entre outros, o objetivo desenvolver aplicações na área da biotecnologia e da Biomedicina de forma a solucionar alguns dos problemas com que os investigadores se deparam no seu trabalho do dia-a-dia.

1.1 Enquadramento

O grande avanço da ciência e tecnologia permite a sequenciação de proteínas em larga escala. Atualmente existem já centenas de genomas sequenciados e com o aparecimento de novas técnicas de sequenciação esse número tenderá a aumentar a um ritmo elevado. Estas grandes quantidades de informação são trabalhadas em laboratórios onde biólogos estudam e analisam as sequências reveladas de forma a encontrar propriedades com potencial biotecnológico. Cada organismo sequenciado tem características e propriedades únicas que poderão ser reproduzidas num meio externo e trazer grandes vantagens ao mundo em que vivemos. Por exemplo, ao sequenciar organismos que vivem em ambientes extremos, tais como, com elevadas concentrações de produtos tóxicos, grandes profundidades ou elevados níveis de radiação, as proteínas que lhes dão imunidade contra esses agentes poderão ser utilizadas noutras áreas tecnológicas.

De modo a documentar as diversas propriedades existentes em sequências proteicas foram surgindo e evoluindo ao longo do tempo vários métodos que reconhecem assinaturas funcionais dentro de uma sequência de aminoácidos. Estas assinaturas podem ser famílias de proteínas, domínios ou grupos funcionais, que descrevem determinada propriedade ou função proteica. Ao longo do tempo surgiram várias técnicas de reconhecimento, cada uma especializada numa determinada área científica, que resultaram em vários repositórios de assinaturas.

InterPro é uma base de dados de assinaturas funcionais que tem diversas bases de dados como fonte de informação. A intenção do InterPro é juntar num único sítio os vários métodos e fontes de classificação de sequências proteicas, criando entradas que representem assinaturas equivalentes fontes em distintas.

Tal como a Biologia Molecular também a Informática foi uma das ciências que mais progressos teve nas últimas décadas. Hoje em dia um computador tem capacidade de processar enormes quantidades de informação em segundos, processo que se fosse realizado manualmente poderia demorar anos. Tem então todo o sentido desenvolver ferramentas informáticas que auxiliem os biólogos, nos seus centros de investigação, a analisar e caracterizar as proteínas com que trabalham. Desta forma

será muito mais eficiente distinguir proteínas com interesse e potencial biotecnológico.

1.2 Objetivos

O objetivo deste projeto é o desenvolvimento de um sistema aplicacional que permita analisar as proteínas sequenciadas nos centros de investigação e divulgar estas proteínas com base nas suas assinaturas funcionais.

Esta aplicação será constituída por duas partes distintas: uma aplicação local utilizada nos centros de investigação, e um site web para utilização do público geral. A aplicação local será utilizada pelos biólogos como uma biblioteca das proteínas sequenciadas e disponibilizará funcionalidades que permitam a análise das assinaturas InterPro existentes em cada proteína. As análises das proteínas serão realizada com recurso à ferramenta InterProScan e esta aplicação tem também como objetivo facilitar a submissão e gestão destas análises. Finalmente esta aplicação possibilitará a publicação de proteínas no site web utilizando para isso os resultados das análises InterProScan, ou seja, a publicação de proteínas será baseada nas funcionalidades e características das proteínas e não na sua sequência de aminoácidos. Pretende-se que a aplicação local seja uma ferramenta autónoma que auxilie os biólogos no seu trabalho do dia-a-dia.

Já a parte web do sistema tem como objetivo permitir o armazenamento e pesquisa de proteínas nas mais diversas áreas de interesse. Este sistema terá como público alvo todas as entidades que procurem proteínas com propriedades específicas.

1.3 Estrutura da dissertação

Este documento está dividido em quatro grandes capítulos: "Análise e divulgação de proteínas", "Requisitos", "Modelo Proposto e Implementação"e finalmente "Descrição da Aplicação". Cada um destes capítulos está dividido em várias subsecções descritas de seguida.

- **Capítulo 2 - Análise e divulgação de proteínas**

  Este capítulo introduz o contexto biológico onde a aplicação se insere. É apresentada uma pequena explicação sobre o processo de sequenciação de proteínas e as técnicas atualmente utilizadas neste processo. É introduzido o conceito de domínio funcional, apresentado o principal repositório utilizado pelos biólogos para analisar as proteínas com que lidam e qual o potencial que estes domínios poderão vir a ter. Finalmente são apresentadas as aplicações já existentes dentro da mesma área, avaliando as suas vantagens e os seus pontos fracos.

- **Capítulo 3 - Requisitos**

  Todos os requisitos da aplicação são descritos neste capítulo. Começa-se por fazer uma análise de todos os requisitos funcionais, ou seja, as funcionalidades a serem disponibilizadas aos utilizadores. Esta análise é dividida em duas secções, uma correspondente à aplicação local e outra correspondente à aplicação web. São também analisados os requisitos não funcionais e estruturais, ou seja, todas as questões relacionadas com desempenho, usabilidade, segurança e qual a organização estrutural do sistema. Finalmente são apresentados alguns *mockups*, que pretendem aproximar as características dos sistema às expectativas dos utilizadores. Foram desenvolvidos *mockups* para as principais interfaces do sistema.

- **Capítulo 4 - Modelo proposto e implementação**

Aqui são descritos todos os detalhes relativos à implementação do sistema. Começa-se por descrever as tecnologias utilizadas, desde linguagem de programação, bases de dados até às frameworks e serviços. É também apresentada a arquitetura adotada tanto para a aplicação local como para o site web. O capítulo termina com todos os detalhes do sistema de gestão e armazenamento de dados do sistema, desde os diretórios criados pela aplicação local até às bases de dados e modelação de dados dos dois módulos da aplicação.

- **Capítulo 5 - ProtSpread - descrição da aplicação**

  Este capítulo descreve o resultado final obtido com o desenvolvimento deste projeto. São apresentadas algumas figuras que ilustram o aspeto final da aplicação e são exemplificadas algumas situações de utilização habitual, tanto para a aplicação local como para a parte web.

Capítulo 2

Análise e divulgação de proteínas

Proteínas são compostos orgânicos que exercem funções muito importantes em todos os organismos vivos, funções essas que vão desde função estrutural, hormonal, defesa, energética entre outras. Do ponto de vista químico, as proteínas são incomparavelmente as moléculas com maior complexidade estrutural e com as funções mais sofisticadas conhecidas [1]. Existem na natureza milhares de diferentes tipos de proteínas, com diferentes tamanhos e formas, que podem cooperar em conjunto para realizar uma determinada função e podem-se associar de forma a criar proteínas mais complexas.

As proteínas assumem funções como a divisão celular, metabolismo, a entrada e saída de matéria e de informação das células, ou até funções estruturais ou mecânicas. As funções de uma proteína dependem da sua estrutura terciária. Esta estrutura surge da existência de sequências particulares de aminoácidos na sequência polipeptídea, que dobram a sequência linear de forma a criar domínios que dão à cadeia uma estrutura tridimensional [2]. A figura 2.1 demonstra as várias estruturas existentes numa proteína.

Tipicamente uma proteína contém entre duzentos a trezentos aminoácidos mas existem proteínas muito mais pequenas (designados peptídos) e algumas muito maiores. Existem vinte tipos diferentes de aminoácidos nas proteínas, cada um com diferentes propriedades químicas. Uma molécula proteica é formada a partir de uma cadeia formada por estes aminoácidos ligados entre si por ligações covalentes. Esta cadeia de aminoácidos é traduzida a partir da sequência de um gene, que por sua vez faz parte do genoma de um organismo. Cada tipo de proteína tem uma sequência de aminoácidos única, existindo milhares de proteínas diferentes na natureza[1].

Algumas das características presentes em proteínas têm muito interesse e podem ser usadas em

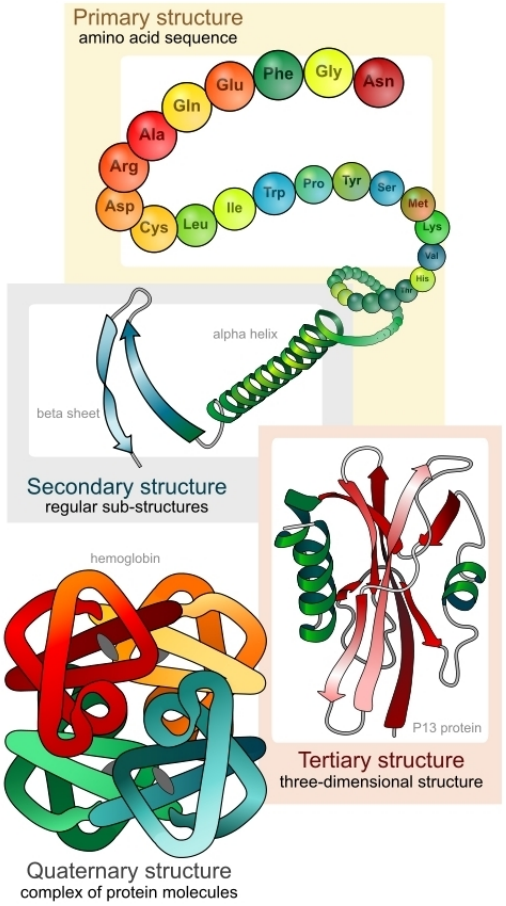


Figura 2.1: Estrutura de uma proteína. Fonte: Wikipedia

processos exteriores aos organismos onde estão presentes. São usadas nas mais variadas áreas, como saúde, alimentação, cosmética entre outras. Podem ser utilizadas para melhorar processos produtivos, redução de custos ou melhoria de qualidade dos produtos, tornando-as, por vezes, muito valiosas. No entanto,para que isso seja possível, é preciso reproduzir essas proteínas em ambientes exteriores.

2.1 Sequenciação de proteínas

Tal como foi referido, as proteínas são constituídas por sequências de aminoácidos. As moléculas de Ácido Desoxirribonucleico (ADN) existem em todas as células de organismos vivos, sendo codificadas através de uma combinação de quatro nucleótidos ou bases: adenina, timina, citosina e guanina, também frequentemente designadas pelas suas iniciais A, T, C e G. Na cadeia genética, um conjunto de três nucleótidos corresponde a um codão, e cada codão codifica um aminoácido.

Com a evolução tecnológica surgiram novas técnicas para analisar proteínas, que fornecem grandes quantidades de informação e permitem aos cientistas estudar proteínas de uma forma muito mais profunda. A sequenciação de proteínas é uma destas técnicas, e é utilizada para determinar a sequência de nucleótidos e consequentemente de aminoácidos que compõem uma determinada proteína. Atualmente existem diversos métodos utilizados no processo de sequenciação de proteínas. Iremos discutir a espectrometria de massa e a reação de degradação de Edman, mas existem mais técnicas que podem ser utilizadas.

Espectrometria de massa é uma técnica analítica para determinar a composição elementar da amostra de uma molécula. Tem como princípio de funcionamento introduzir uma amostra num instrumento que lança uma substância carregada eletrões para produzir iões, ou átomos eletricamente carregados. Os iões atravessam um campo magnético que dobra as suas trajetórias de modos diferentes, dependendo de suas massas. A partir da trajetória produzida é então possível identificar os elementos e isótopos presentes na amostra e desvendar a sequência de codões que dá origem à proteína [3].

O outro método referido é reação de degradação de Edman [4] que permite identificar os aminoácidos presentes numa proteína sem destruir as ligações peptídicas entre outros resíduos de aminoácidos. Este método consiste numa sequência de reações químicas alcalinas e acidas que separa o grupo aminoácido que pode então ser extraído e identificado usando cromatografia ou eletroforese. Este procedimento é então repetido sequencialmente para identificar cada aminoácido seguinte.

Atualmente estas técnicas têm vindo a ser substituídas por sequenciação.

2.2 Domínios funcionais

A literatura está repleta de definições que tentam encontrar o conceito correto de domínio funcional numa proteína. Da forma mais simples, domínio funcional é uma região de uma proteína que tem um determinado propósito e pode existir independentemente da proteína e é por isso considerada uma região autónoma da proteína.

Tal como uma proteína também um domínio funcional é formado por uma sequência de aminoácidos, em que o seu comprimento pode variar entre os quarenta e trezentos e cinquenta aminoácidos [1]. Uma proteína pode possuir vários domínios funcionais, funcionando estes como blocos de construção para uma proteína mais complexa como é o caso da proteína *Pyruvate kinase* representada da figura 2.2. Da mesma forma o mesmo domínio funcional pode existir em diferentes proteínas, com ligeiras diferenças, e muitas vezes encontram-se na mesma proteínas diferentes domínios funcionais sobrepostos.

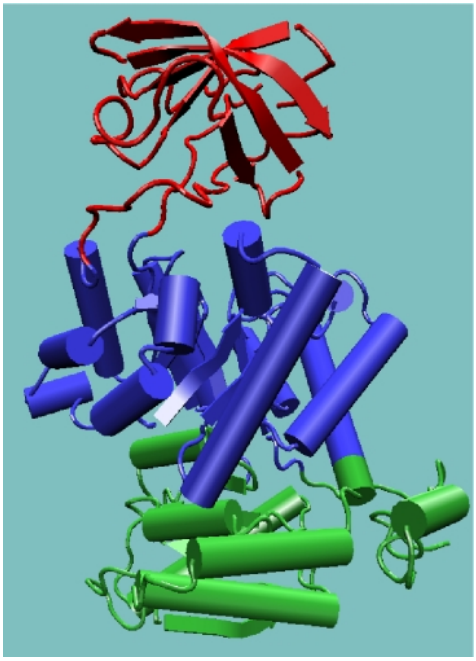


Figura 2.2: Representação da proteína Pyruvate kinase, constituída por três domínios funcionais. Fonte: wikipedia.org

Devido a serem completamente independentes, estáveis e com funcionalidades bem definidas é de todo o interesse que estes domínios sejam estudados em profundidade, de forma a se conseguir obter mais conhecimento acerca das suas capacidades e da relação existente entre estes domínios funcionais e dos organismos onde estão presentes. Com o desenvolvimento da ciência e da tecnologia e com a redução dos custos de sequenciação o conhecimento acerca dos domínios funcionais aumentou a grande ritmo o que levou a que surgissem vários sistemas de classificação que tentar definir e classificar domínios em algumas bases de dados. Estes sistemas variam tanto no tipo de dados que classificam como na quantidade de classificação manual que incorporam [5].

Structural Classification of Proteins (Scop) [6] e Class Architecture Topology Homologous superfamily (CATH) [7] são ambos sistemas de classificação de domínios que analisam as várias estruturas presentes em proteínas. O Scop é um sistema de classificação completamente manual baseado na semelhança entre aminoácidos e estruturas terciárias. Já o CATH utiliza um cruzamento de procedimentos automáticos e manuais para obter uma classificação hierárquica. Tal como o Scop ou o CATH, ao longo do tempo surgiram vários repositórios de domínios de proteínas e locais funcionais. Estes repositórios utilizam diferentes técnicas de reconhecimento, resultando também diferentes tipos de informação em bases de dados. Cada base de dados está mais vocacionada para uma área específica de aplicação, tendo todas elas pontos fortes e pontos fracos.

2.2.1 InterPro

De forma a homogeneizar os vários sistemas de classificação de domínios funcionais surgiu um novo repositório, designado InterPro (`http://www.ebi.ac.uk/interpro/`) [8], que visa criar uma camada de abstração por cima de todos os repositórios já existentes, de forma a obter uma caracterização única, não-redundante de um domínio, família ou local funcional. Quando diferentes assinaturas coincidem com o mesmo conjunto de proteínas na mesma região da sequência, presume-se que elas

descrevam o mesmo domínio funcional e são colocados numa única entrada InterPro por um administrador. Agrupar assinaturas equivalentes de diferentes fontes desta forma traz benefícios óbvios, atribuindo nomes e anotações consistentes [9].

Atualmente a base de dados InterPro integra modelos preditivos, ou assinaturas, de múltiplas fontes: Pfam [10], PRINTS [11], PROSITE [12], SMART [13], ProDom [14], PIRSF [15], SUPERFAMILY [16], PANTHER [17], CATH [7], Gene3D [18], TIGRFAMs [19] e HAMAP [20].

Cada uma destas fontes é especializada num ramo biológico obtendo-se vários tipos de informação sobre as anotações existentes. Assim, um dos objetivos do repositório InterPro é agrupar os pontos fortes de cada fonte e conseguir uma visão mais global e completa sobre domínios funcionais e anotações de proteínas. Ao mesmo tempo, com toda a informação reunida é também possível traçar as varias relações biológicas existentes, criando uma hierarquia do tipo "pai-filho"[9]. Por exemplo, se uma assinatura corresponde apenas a um subconjunto de proteínas em comparação com outra assinatura, é provável que esta assinatura seja mais funcional ou taxonomicamente específica do que a outra. Neste caso, as assinaturas seriam consideradas relacionadas, a assinatura correspondente ao subconjunto seria denominada uma descendente e a outra como sendo o seu pai.

Depois de criada uma entrada no InterPro são executados procedimentos automáticos que têm como objetivo adicionar alguma informação útil para os utilizadores, como seja, a existência de correspondências com a base de dados UniProt Knowledgebase (UniProtKB) [21], ou a existência de alguma estrutura tridimensional no repositório PDB.

Com toda a informação processada, os utilizadores têm a possibilidade de consultar todas as correspondências de assinaturas catalogadas e navegar através de uma interface web em vários formatos gráficos e textuais, assim como interagir com as várias correspondências existentes para cada entrada.

No entanto, o conteúdo do repositório InterPro é baseado em assinaturas de proteínas conhecidas existentes em repositórios públicos como o UniProtKB, o que por um lado torna este repositório pouco útil para procurar domínios funcionais em sequências privadas e recentes como as que são obtidas em unidades de sequenciação. Para isso foi desenvolvida uma ferramenta que possibilita encontrar domínios funcionais dada uma sequência proteica, denominada InterProScan.

2.2.2 InterProScan

InterProScan (`http://www.ebi.ac.uk/Tools/pfa/iprscan/`) [22] é uma ferramenta que combina vários métodos de reconhecimento de proteínas apenas num recurso. Esta ferramenta, dada uma determinada sequência proteica, pesquisa por domínios funcionais servindo-se para isso das várias fontes do repositório InterPro. Este processo é vital para conhecer as propriedades de uma proteína como, por exemplo, prever a sua função ou funções, qual a sua área de aplicação ou as relações da proteína em estudo com as já conhecidas. Esta é uma ferramenta muito completa visto que utiliza uma conjunto muito variado de fontes o que leva a que se fique a conhecer os pontos fortes de determinada proteína. Por exemplo, as expressões regulares utilizadas no base de dados PROSITE são ideais para procurar pequenos motivos mas não são tão eficazes na pesquisa de membros de famílias muito divergentes. Por outro lado os *fingerprints* existentes no repositório PRINTS são excelentes para determinar sub-famílias específicas, mas não são fiáveis para encontrar pequenos domínios. Os métodos baseados em modelos de Markov e perfis são os melhores para identificar membros de super-famílias divergentes mas não obtêm bons resultados a determinar membros específicos de sub-famílias. Idealmente todas os métodos devem ser utilizados para obter os melhores resultados e é exatamente isto que o InterProScan faz.

A ferramenta tem como parâmetro de entrada uma sequência proteica, que pode ser em diversos formatos (FASTA, texto, UniProtKB/Swiss-Prot, ...). Depois de submetida a sequência o InterProS-

can executa as diferentes aplicações de pesquisa de domínios que retornam uma lista de resultados de cada uma das base de dados que as compõem. Finalmente, todos os resultados obtidos são processados e combinados de forma a obter o resultado final. A figura 2.3 representa de forma simplificada o funcionamento desta ferramenta.

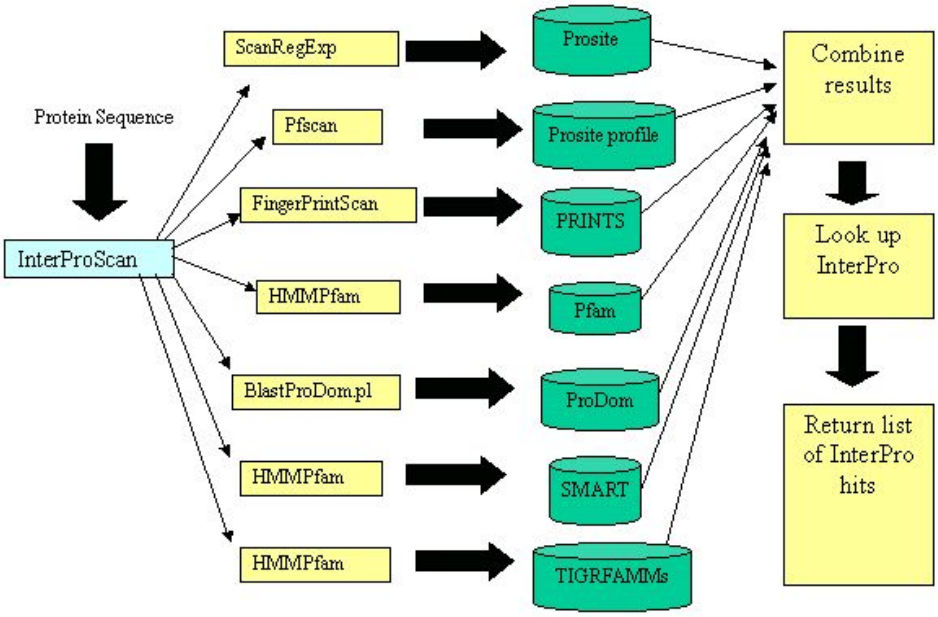


Figura 2.3: Esquema simplificado do funcionamento da ferramenta InterProScan. Fonte: `http://www.ebi.ac.uk`

É importante referir que o InterProScan não funciona apenas como um agregador de resultados de várias ferramentas, mas também consulta dados de fontes externas e remove alguma possível redundância.

Esta ferramenta é livre tanto para o uso académico como comercial e encontra-se disponível de várias formas [23]. É possível obter uma instalação local da aplicação para uso pessoal mas também é possível aceder ao serviço remotamente quer através de *webservices* ou mesmo utilizando o *website* da aplicação (`http://www.ebi.ac.uk/Tools/pfa/iprscan/`).

Qualquer que seja a forma como se utilize a ferramenta os resultados obtidos são disponibilizados em vários formatos, desde imagens até texto simples passando por XML ou tabelas. Para se perceber os resultados mais facilmente o formato mais adequado será a imagem disponibilizada. Com esta imagem é muito fácil perceber quais os domínios InterPro detetados, desde as famílias, até às entradas individuas de cada aplicação. Analisando, por exemplo, o resultado representado na figura 2.4, podemos perceber que a sequência analisada faz parte da super-família *Rhodopsin-like GPCRs*. É também membro da família *opsin* e da família *rodopsina*, uma subfamília da família *opsina*.

As assinaturas que descrevem a mesma família, ou domínio, são agrupadas numa única entrada InterPro e cada uma tem um identificador único. No resultado é também representado graficamente o local onde a assinatura se encontra na proteína, informações relativas à origem desse domínio e uma breve descrição das suas características.

Estes resultados fornecem-nos diferentes níveis de informação sobre a sequência analisada. Pode-se depois obter informações mais detalhadas sobre a sua estrutura, função, proteínas onde existe, consultando a documentação da super-família a que pertence.

Figura 2.4: Exemplo do resultado de uma análise InterProScan em imagem. Fonte: `http://www.ebi.ac.uk`

2.3 Potencial biotecnológico

Atualmente a redução dos custos de sequenciação permite que muitos grupos de investigação efetuem sequenciações maciças dos organismos com que trabalham. Por vezes estas sequências poderão ter um elevado potencial biotecnológico. Por exemplo, quando se sequenciam organismos que vivem em ambientes extremos, tais como, elevadas concentrações de produtos tóxicos, grandes profundidades ou elevados níveis de radiação as proteínas que os protegem contra esses agentes poderão ser importantes em áreas da biotecnologia.

Estas proteínas, muitas delas específicas de um organismo particular, podem ser utilizadas em meios exteriores aos organismos. Com a sequência genética revelada é possível reproduzir essas proteínas em ambientes exteriores podendo ser posteriormente isoladas e utilizadas em outros meios. São utilizadas nas mais variadas áreas, como saúde, alimentação, cosmética entre outras. Assumem funções que vão desde o melhoramento de processos produtivos, redução de custos ou melhoria da qualidade dos produtos tornando-as muito valiosas.

Com toda a tecnologia e ferramentas existentes torna-se cada vez mais fácil analisar as características específicas de cada proteína desde a análise da sequência, passando pela análise estrutural ou até mesmo funcional. Ferramentas como o InterProScan permitem conhecer de uma forma rápida e simples as mais valias funcionais de uma sequência e perceber se uma dada proteína pode ter ou não potencial para que seja produzida num ambiente laboratorial para uso comercial.

A análise dos domínios de uma proteína também permite conhecer a constituição de cada grupo

funcional individualmente, e estudar estas regiões para futuras utilizações ou mesmo construção de novas proteínas artificiais. Por exemplo, é possível introduzir funcionalidade numa proteína através da incorporação de domínios funcionais com uma atividade biológica específica [24]. Embora estas proteínas modificadas possam imitar as funções das proteínas naturais, elas podem também ser construídas para ter novas funções.

É portanto de todo o interesse que os investigadores consigam saber todo o potencial das proteínas desvendadas nos seus centros de investigação e assim valorizar todo o trabalho por eles realizado.

2.4 Aplicações existentes

Atualmente já existem algumas aplicações que lidam com anotações funcionais, algumas delas apenas para visualização ou geração de estatísticas a partir destes dados, outras passando pela utilização dos serviços disponíveis para anotação de sequências ou apenas para gerir grandes quantidades de informação. De seguida apresentamos a descrição de algumas das aplicações utilizadas que apresentam funcionalidades relacionadas com anotações funcionais, algumas das quais utilizam anotações InterPro ou matérias relacionadas como funcionalidade central.

2.4.1 IPRStats

IPRStats [25] é uma ferramenta para visualização de resultados InterProScan. De uma forma simplificada a aplicação usa os resultados InterProScan como entrada e constrói gráficos e tabelas que permitem a visualização das funções das sequências analisadas.

É uma ferramenta ideal para quando se pretende analisar o genoma completo de um determinado organismo, criando uma visualização estatística completa dos resultados. Desta forma tem-se uma noção inicial muito completa acerca das potenciais funções das sequências analisadas.

A ferramenta obtém os resultados InterProScan a partir de um ficheiro Extensible Markup Language (XML). De seguida a informação é interpretada e colocada numa base de dados relacional. Finalmente os resultados de cada análise são lidos e representados em vários tipos de gráficos de forma estatística. É também disponibilizada uma tabela, onde cada entrada é um tipo de assinatura (por exemplo CATH) acompanhado por um ou mais termos Gene Ontology (GO), se o InterProScan foi executado com a opção GO ativa. Os resultados podem ainda ser exportados em HyperText Markup Language (HTML) para visualização num browser.

A aplicação pode ser instalada em dois modos: uma versão local e uma versão web para utilização remota. A figura 2.5 demonstra uma utilização comum da aplicação na sua versão web.

Como se pode ver, o IPSRStats é uma aplicação que permite de forma muito simples traçar o perfil de um determinado genoma e perceber quais os domínios onde o organismo se insere.

2.4.2 JIPS - Interface gráfica em Java para resultados InterProScan

JIPS é uma aplicação que facilita a análise de grandes quantidades de resultados InterProScan [26]. Tal como referido anteriormente, na secção 2.2.2, a ferramenta InterProScan pode ser utilizada através de uma interface *web*, através de um *web-service*, ou mesmo numa versão instalada localmente, versão esta que tem a vantagem de se poderem executar análises a grandes quantidades de proteínas apenas de uma vez. A visualização deste tipo de resultados pode ser um processo demorado e muito maçador. Para facilitar este tipo de tarefas existe a aplicação JIPS que, entre as suas características, conta com funcionalidades como:

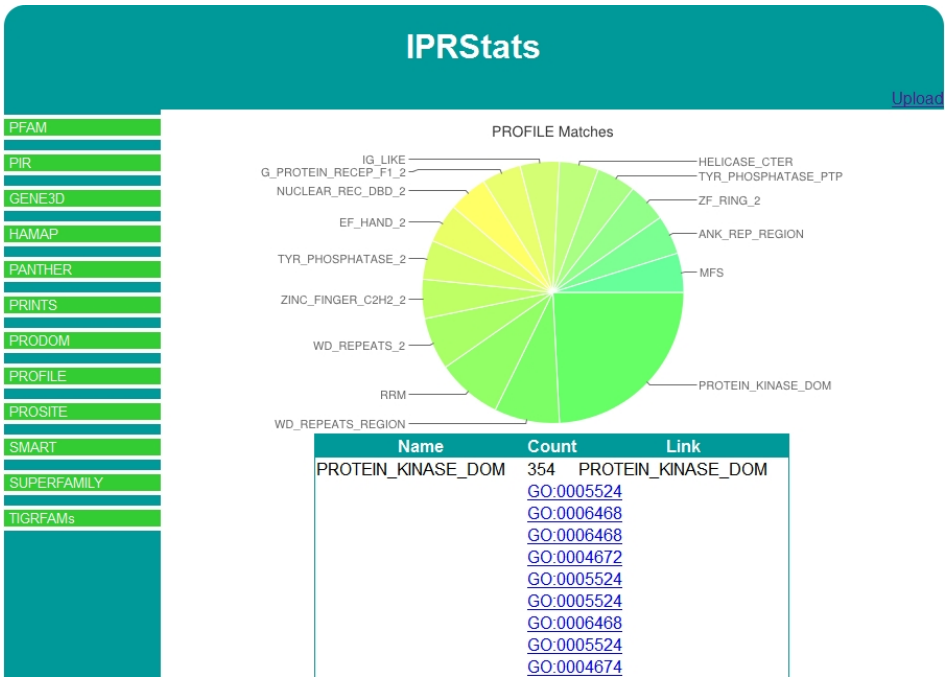


Figura 2.5: Exemplo do resultado de uma interpretação estatística realizada com a aplicação IPRStats

---

- interface para simplificar a execução de análises a grandes quantidades de proteínas,
- mecanismo para sinalizar a existência de novas assinaturas nas análises,
- ferramenta para auxiliar na comparação de ortólogos e em novas análises.

Apesar de possuir algumas ferramentas muito úteis como visualização de alinhamentos de proteínas, ou comparação de ortólogos, a característica mais poderosa e que distingue esta aplicação de todas as aplicações semelhantes é a capacidade de gerir grandes quantidade de análises InterProScan. Funcionalidades como o agendamento para realizar novas análises periodicamente ou indicar ao utilizador a existência de novos resultados, dão aos biólogos uma grande facilidade em gerir grandes quantidades de informação e permite reduzir substancialmente o tempo necessário para realizar esta verificação.

A figura 2.6 ilustra um exemplo da visualização do resultado de uma análise InterProScan para o gene "VARV-Bsh-B1R". Neste exemplo é possível verificar que foram encontradas 5 assinaturas em bases de dados distintas, correspondendo a 3 entradas InterProScan.

Para terminar, a aplicação JIPS é uma ferramenta poderosa que permite aos biólogos lidar com grandes quantidades de sequências e gerir de forma simples as análises InterProScan realizadas a estas sequências.

2.4.3 GenBeans

GenBeans (`http://www.geneinfinity.org/genbeans/`) é uma aplicação cujo objetivo é integrar o máximo de ferramentas especializadas numa única aplicação criando um contexto muito apropriado para a área em estudo.

De uma forma geral esta aplicação funciona como uma biblioteca digital para gerir os vários ficheiros que armazenam sequências num computador. A aplicação tem várias vantagens quando se

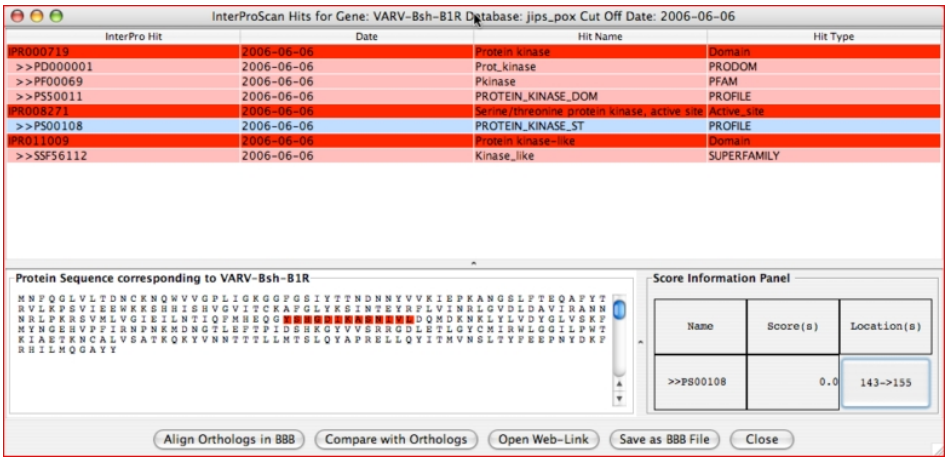


Figura 2.6: Exemplo da interface da aplicação JIPS, mostrando que a sequência selecionada tem 5 assinaturas em 3 entradas InterPro. A zona da sequência em destaque corresponde à entrada PS00108.

pretende lidar diretamente com as sequências, fornecendo vários instrumentos para lidar com este tipo de dados. Por exemplo o gestor de sequências reconhece automaticamente uma grande quantidade de formatos de ficheiros contendo sequências (FASTA, GenBank, EMBL, INSD, UniProt, etc) e apresenta todas as sequências reconhecidas com uma interface muito acessível e simples de operar. O editor de sequências, apresentado na figura 2.7, permite visualizar e editar as sequências de uma forma muito dinâmica utilizando, por exemplo, vários esquemas de cores para evidenciar as características das proteínas com que se está a trabalhar e até algumas anotações que ajudam a conhecer a sequência em análise. Apresenta também algumas ferramentas que permitem visualizar a tradução da sequência ao mesmo tempo que se edita a sequência de nucleótidos.

No entanto, esta aplicação não tem ferramentas bastante importantes na área biotecnológica como, por exemplo, alinhamento de sequências ou análise de assinaturas nas várias bases de dados existentes. Este facto limita muito a utilização da aplicação, tornando-a até pouco útil no contexto tratado nesta tese.

2.4.4 Blast2GO

Este pacote de aplicações é, de todas as soluções analisadas, a mais semelhante à que se pretende desenvolver neste projeto. A aplicação Blast2GO [27] é uma solução integrada orientada para a área da biologia, para gestão e anotação funcional de sequências de ADN ou proteínas, baseado no vocabulário GO. A solução está acessível através da tecnologia *Java Web Start* (`http://www.blast2go.com`) e é destinada principalmente para apoiar a investigação em laboratórios experimentais onde o apoio da bioinformática não está muito desenvolvido. Como se pode perceber pela figura 2.8 a aplicação utiliza uma interface muito simples e de fácil utilização para que os utilizadores se concentrem apenas no seu trabalho de pesquisa sem necessidade de realizar tarefas de configuração.

Basicamente, a aplicação Blast2GO utiliza serviços locais ou remotos, como alinhamentos BLAST, para encontrar sequências semelhantes a uma ou mais sequências. Contudo, a aplicação não é um mero agregador de serviços tendo também uma componente de *data mining* já que disponibiliza uma grande variedade de gráficos e estatísticas que permitem uma mais fácil e eficiente interpretação dos dados. Por exemplo, é possível perceber quais as assinaturas mais frequentes ou qual a distribuição das espécies das sequências em análise.

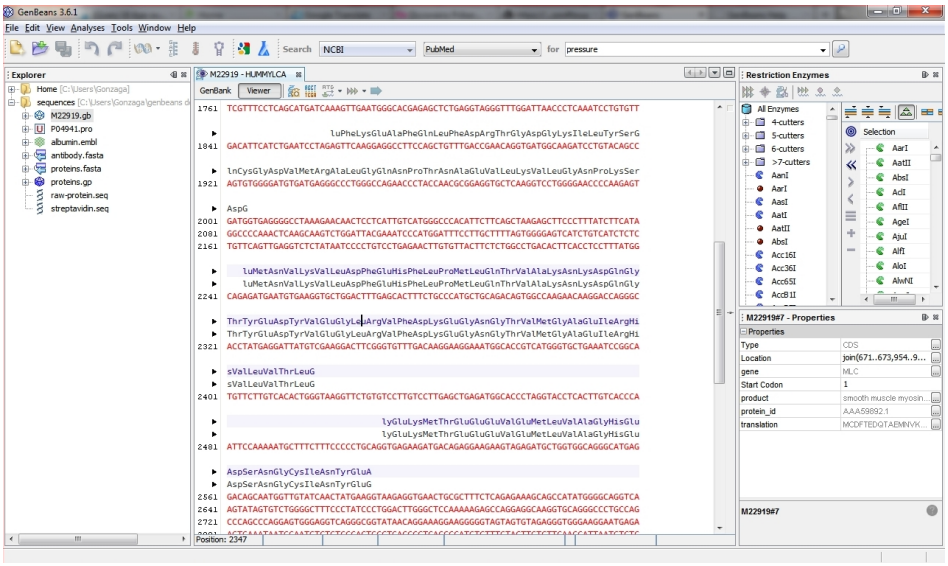


Figura 2.7: Demonstração da interface da aplicação GenBeans, enquanto se utiliza o editor de sequências.

2.5 Sumário

Este capítulo teve como objetivo fazer uma pequena introdução do contexto em que este projeto se insere. Foi introduzido o conceito de proteína, sequências de aminoácidos e quais os métodos atuais utilizados no processo de sequenciação de proteínas. Esta é a base de todo o projeto e estas noções devem estar bem presente durante todo o trabalho.

Um outro conceito apresentado nesta introdução foi a noção de domínio funcional. Apesar de não ser consensual, domínio funcional é definido como uma região de uma proteína que tem um determinado propósito e pode existir independentemente da proteína. Cada domínio funcional tem uma função específica, tal como o nome indica, mas não é específica de uma proteína, ou seja, o mesmo domínio pode existir em proteínas distintas e domínios distintos podem partilhar a mesma região de uma proteína. Este é um dos conceitos mais importantes neste projeto, pois o sistema de divulgação de proteínas será baseado nestas regiões.

Foi também apresentado o repositório de domínios funcionais InterPro e a ferramenta de análise InterProScan. Este repositório pretende agregar e uniformizar todos os repositórios de domínios funcionais, que utilizam diferentes estratégias de reconhecimento de regiões funcionais. Uma entrada InterPro pode corresponder a múltiplas assinaturas funcionais de diferentes fontes, como PANTHER ou CATH.

São os domínios funcionais que distinguem uma proteína com potencial biotecnológico. Por exemplo, uma proteína que tenha a capacidade de evitar que o gelo forme cristais muito grandes, pode ser muito interessante do ponto de vista biotecnológico, para permitir vida a temperaturas negativas. Estas características estão associadas à existência de domínios funcionais numa sequência proteica, sequências estas que podem ser reproduzidas em ambientes exteriores podendo ser posteriormente isoladas e utilizadas em outros meios. O conhecimento das sequências genéticas de alguns organismos é então um conhecimento com muito potencial e que deve ser valorizado.

Sendo o objetivo deste projeto oferecer uma ferramenta simples para analisar e publicar proteínas, são também analisadas algumas ferramentas já existentes atualmente dentro da mesma área. São ana-

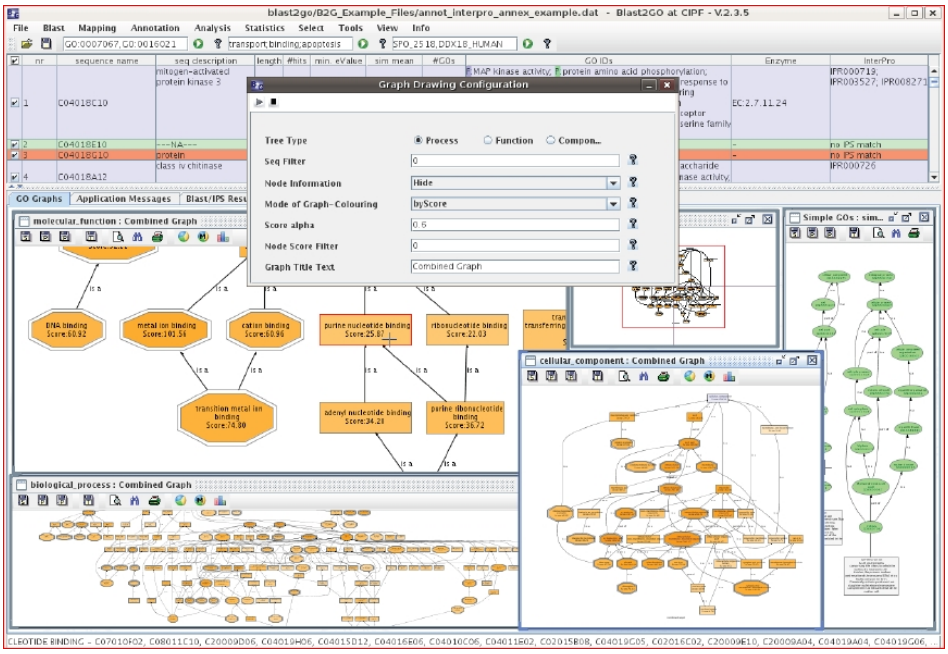


Figura 2.8: Exemplo da interface gráfica da aplicação Blast2GO.

lisadas quatro aplicações relacionadas com a gestão e manipulação de proteínas, desde visualização e interpretação de grandes quantidades de resultados InterProScan, biblioteca local de proteínas, ou ferramentas de análise de assinaturas funcionais em sequencias proteicas.

Capítulo 3

Requisitos

A motivação da realização deste projeto partiu de uma carência existente nos laboratórios de biologia molecular, mais especificamente em projetos de sequenciação de proteínas, em divulgar o trabalho desenvolvido nos seus centros. Para colmatar esta limitação o primeiro passo a fazer é o levantamento de requisitos que a aplicação deve cumprir.

O levantamento de requisitos para a aplicação informática a desenvolver foi realizado em conjunto com os investigadores do BIOCANT, cuja experiência com a manipulação de dados proteicos e análises de resultados InterProScan foi fundamental para definir as funcionalidades da aplicação. Para que esta tarefa fosse feita da forma mais completa foram realizadas algumas reuniões, o que resultou num número de requisitos a serem tidos em conta no desenvolvimento da aplicação. No entanto, devido ao contexto experimental e pedagógico deste projeto, muitos requisitos foram sendo redefinidos ao longo do desenvolvimento da aplicação.

Neste capítulo são apresentados alguns esboços que permitiram formular um plano para a construção da aplicação assim como todos os requisitos definidos para a realização do projeto de uma forma detalhada.

3.1 Requisitos funcionais

Os requisitos funcionais definem as funcionalidades do sistema de software a desenvolver. Assim definindo todos os requisitos funcionais obtém-se uma descrição completa da aplicação e conseguem-se objetivos claros para a implementação.

Tal como já foi referido a aplicação tem dois tipos distintos de utilizadores. Por um lado os biólogos dos centros de investigação e por outro os potenciais interessados nas proteínas divulgadas. Os biólogos interagem com os dois módulos da aplicação, utilizando a aplicação local para gerir, analisar e publicar as proteínas com que lidam diariamente, e a parte web para se registarem na aplicação e consultarem as suas publicações, estatísticas, etc. Já os consumidores ou clientes, apenas utilizam a parte web da aplicação para fazer pesquisas das proteínas pretendidas e eventualmente entrar em comunicação com a entidade que representa essas proteínas.

Os requisitos funcionais podem ser divididos em dois grupos separando as funcionalidades para cada uma das duas faces da aplicação. De forma a facilitar a análise de requisitos da aplicação, construiu-se um diagrama de casos de uso para cada um dos módulos e realizou-se uma análise cuidada de quais os processos que cada uma das ações implica para que seja completamente funcional.

3.1.1 Aplicação local

Analisando em primeiro lugar a **aplicação local**, pretende-se uma ferramenta informática de utilização simples e que permita analisar e publicar proteínas através das assinaturas reveladas pela ferramenta InterProScan.

Casos de utilização

A figura 3.1 representa os casos de utilização, em notação Unified Modeling Language (UML), que definem os requisitos básicos do sistema.

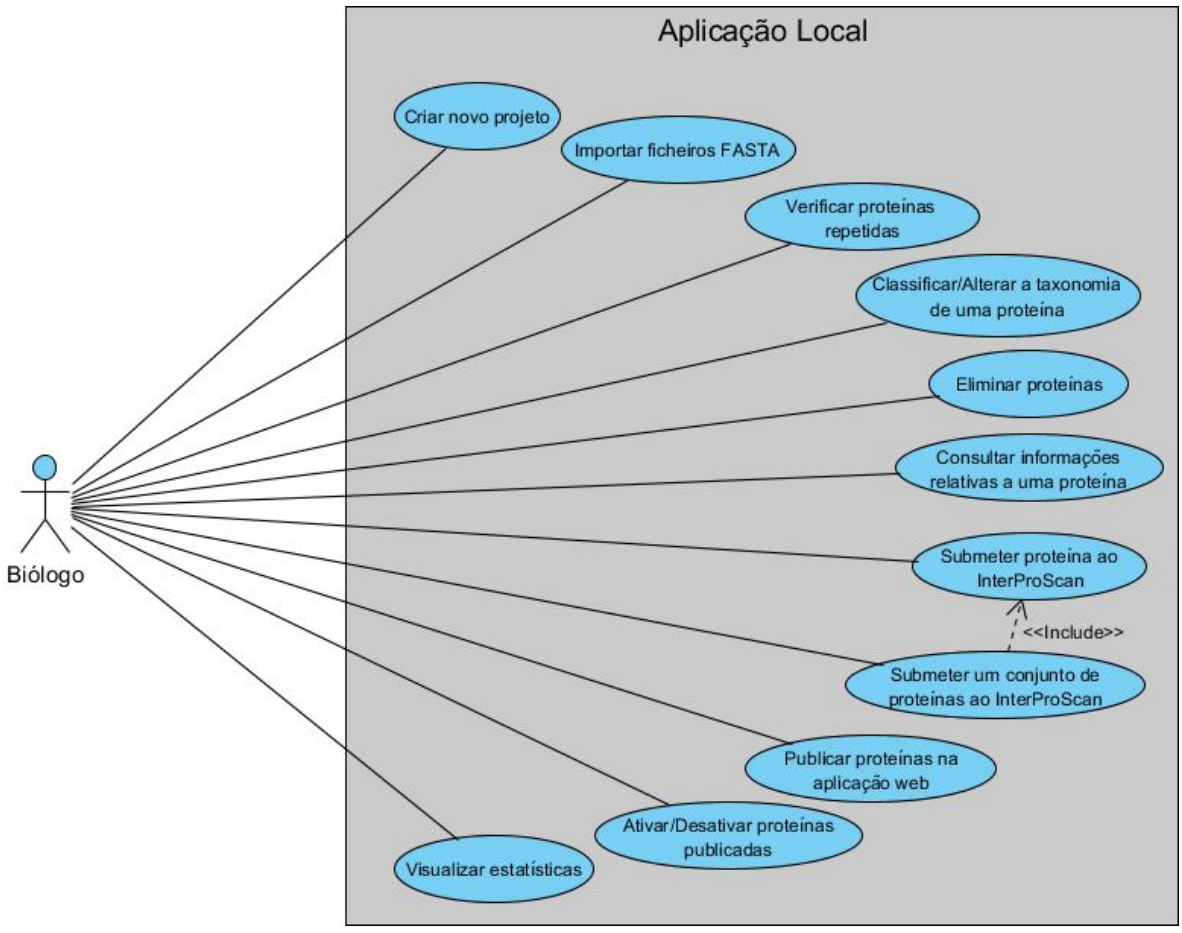


Figura 3.1: Diagrama dos casos de uso a serem suportados pela aplicação local.

Tal como referido o único ator deste diagrama é o biólogo. O acesso à aplicação é garantido com o sistema de login utilizando um nome de utilizador e password para cada utilizador. Esta ação não é considerada um caso de utilização visto que não está no contexto do problema tratado e apenas serve para aceder á aplicação.

Para obter um *nome de utilizador* para aceder à aplicação o utilizador tem de se registar primeiro na aplicação local. Este processo garante a sincronização entre os utilizadores da aplicação local e da aplicação web e é também uma forma de proteger os dados presentes na aplicação.

Uma vez realizado o login o utilizador acede à sua informação pessoal, tal como proteínas armazenadas e análises, e pode utilizar todas as funcionalidades disponibilizadas.

Descrição dos casos de utilização

De seguida são descritos em pormenor todos os casos de uso referentes à aplicação local.

- **Criar novo projeto**

  O utilizador pode criar múltiplos projetos para organizar as suas coleções de proteínas.

- **Importar ficheiros FASTA**

  Funcionalidade que permite importar ficheiros com sequências proteicas no formato FASTA. Este processo associa automaticamente as proteínas importadas a um projeto e permite escolher uma taxonomia a aplicar a todas as proteínas. Durante o processo de importação deverá ser possível organizar as proteínas existentes no ficheiro por tamanho ou nome, e filtrar as proteínas por texto contido no nome.

- **Verificar proteínas repetidas**

  A aplicação permite verificar rapidamente se existem proteínas repetidas, mesmo para diferentes projetos. Considera-se a mesma proteína todas as entradas que tenham a mesma sequência.

- **Classificar/Alterar a taxonomia de uma proteína**

  Classificar ou alterar a taxonomia de cada proteína existente na aplicação.

- **Eliminar proteínas**

  Permitir eliminar proteínas existentes na aplicação.

- **Consultar informações relativas a uma proteína**

  Deverá ser possível consultar toda a informação relativa as proteínas carregadas no sistema. As informações vão desde tamanho, sequência, análise InterProScan ou a informação se está ou não publicada.

- **Submeter proteína ao InterProScan**

  Possibilidade de submeter proteínas ao InterProScan nos servidores remotos. Este processo passa por submeter uma análise de uma proteína, verificar o estado da submissão e transferir o resultado quando a análise estiver concluída. Visto que uma análise InterProScan pode ser um processo demorado, estes três passos não são realizados em simultâneo e são consideradas tarefas independentes. A aplicação deve disponibilizar um sistema que permita gerir a submissão de análises InterProScan de forma transparente para o utilizador.

- **Submeter um conjunto de proteínas ao InterProScan**

  Realizar análises InterProScan a um grande número de proteínas de uma só vez. Este processo permite ao utilizador selecionar um conjunto de proteínas e iniciar uma análise InterProScan de uma vez só, o que permite poupar muito tempo.

- **Importar análises InterProScan a partir de resultados XML**

  Permitir ao utilizador importar resultados de análises InterProScan no formato XML. Estes resultados podem ser obtidos a partir de análises realizadas em instalações locais da aplicação InterProScan.

- **Publicar proteínas na aplicação web**

  Permitir publicar proteínas na aplicação web para posterior consulta dos interessados. Apenas devem ser publicadas as assinaturas e características reveladas pela análise InterProScan e não a sequência completa.

- **Ativar/Desativar proteínas publicadas**

  Possibilidade tonar ou não visíveis as proteínas publicadas na aplicação web. O biólogo pode ter interesse em desativar uma publicação, ou em voltar a torná-la visível no web site.

- **Visualizar estatísticas**

  Capacidade de visualizar algumas estatísticas acerca das pesquisas realizadas na aplicação web sobre as proteínas publicadas.

Apesar do funcionamento desta aplicação ser muito baseado na utilização de ferramentas remotas e na comunicação com o site web, deve ser possível utilizar a aplicação em modo *offline* como uma biblioteca de proteínas.

3.1.2 Aplicação Web

Analisando agora a parte Web do projeto, pretende-se construir um web site para interação com os potenciais interessados nas proteínas publicadas. Será também a partir deste web site que os utilizadores poderão descarregar a aplicação local.

Casos de utilização

Para identificar as funcionalidades do portal desenhou-se um diagrama de casos de uso, na notação UML, representado na figura 3.2.

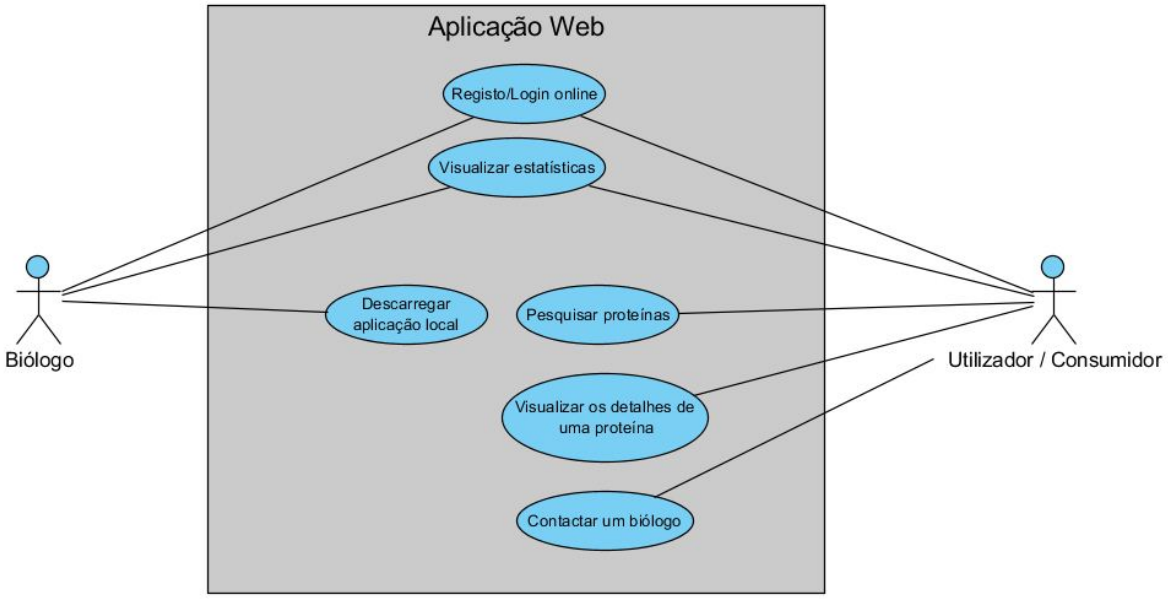


Figura 3.2: Diagrama dos casos de uso a serem suportados pela aplicação web.

Para a aplicação web existem dois tipos de possíveis atores. Por um lado o biólogo/investigador também representado no diagrama de casos de uso da aplicação local (figura 3.1), e um novo ator designado utilizador/consumidor. Este ator, representa todas as entidades que interagem com o sistema de forma a pesquisar e encontrar as proteínas pretendidas.

Descrição dos casos de utilização

De seguida são descritos em pormenor todos os casos de uso referentes à aplicação web.

- **Registo/Login online**

  Qualquer tipo de utilizador tem a possibilitar de se registar na aplicação. Este registo é necessário para utilizar a aplicação local e permite o acesso a algumas funções na aplicação web.

- **Visualizar estatísticas**

  A aplicação permite visualizar um conjunto de estatísticas de quais as assinaturas mais procuradas.

- **Descarregar aplicação local**

  A aplicação local está disponível para descarregar no site web. Desta forma consegue-se uma maior integração entre as duas partes da aplicação. Assim para que os biólogos possam utilizar a aplicação local nos seus terminais devem primeiro transferi-la do site web onde está disponível.

- **Pesquisar proteínas**

  Esta é a funcionalidade central da aplicação web destinada aos potenciais interessados pelas proteínas. Graças à pesquisa de proteínas pelas suas assinaturas os consumidores podem encontrar as proteínas que satisfazem as suas necessidades facilmente.

- **Visualizar os detalhes de uma proteína**

  Visualizar em pormenor os detalhes da publicação de uma proteína, como o biólogo/centro de investigação que publicou aquela entrada, as assinaturas InterProScan existentes, data de publicação, ...

- **Contactar um biólogo**

  Um consumidor pode contactar um biólogo de forma a mostrar interesse em alguma das proteínas publicadas, ou para fazer algumas questões que sejam do seu interesse.

Serviços utilizados pela aplicação local

Apesar de estes serem os casos de utilização diretamente relacionados com os utilizadores, existem outros casos não mencionados, que são fundamentais para o funcionamento do sistema. Muitas das funcionalidades da aplicação local utilizam serviços disponibilizados pela aplicação web. Estes serviços também podem ser vistos como requisitos apesar de não servirem utilizadores mas sim um sistema aplicacional.

Os serviços que devem ser disponibilizados de forma a garantir todas as funcionalidades são os seguintes:

- verificação/autenticação de utilizadores registados

- publicação de proteínas através das suas assinaturas InterProScan
- Ativar/Desativar a visibilidade de uma proteína publicada
- Sincronização de estatísticas de pesquisa

3.2 Requisitos não funcionais

Definidas todas as funcionalidades do sistema, faltam definir as questões relacionadas com desempenho, usabilidade, segurança, disponibilidade ou manutenção.

Em relação a este tema a questão mais importante a ter em conta é a segurança dos dados tratados. Apesar da aplicação ter como objetivo principal a divulgação de proteínas esta divulgação deve-se basear apenas nas assinaturas InterProScan que uma proteína contém e nunca revelar a sequência completa. Desta forma as sequências importadas apenas permanecem na aplicação local, que é executada nos computadores dos laboratórios ou centros de investigação dos biólogos. De forma a garantir o acesso aos dados a aplicação está protegida com o sistema de nome de utilizador e palavra chave, e assim apenas quem conhecer estes dados consegue aceder à aplicação.

Ao nível de desempenho, a aplicação deve ser construída de forma a que a sua execução seja fluída. Um cuidado a ter em conta neste aspeto, é o facto de que a aplicação deverá gerir grandes quantidades de informação, como proteínas, ou análises InterProScan. Este tipo de dados deve ser gerido de forma cuidada de forma a não tornar a aplicação com um funcionamento pesado e desagradável para o utilizador.

Uma questão importante a ter em conta é que a aplicação deve ser desenvolvida de forma a poder ser executada em diferentes sistemas operativos. Desta forma é necessário ter em conta a linguagem de programação a utilizar, assim como bibliotecas e dependências que possam existir, para que seja possível utilizar a aplicação no máximo de ambientes possíveis.

Visto que este é um trabalho implementado de raiz, deve ser tomado especial cuidado para que seja construída de forma estruturada e organizada. Isto para que seja possível uma posterior continuação do desenvolvimento do trabalho, tanto implementando novas funcionalidades como melhorando as já existentes de forma simples rápida. Esta é também uma forma de facilitar a manutenção da aplicação e de detetar possíveis falhas ou comportamentos anormais.

3.3 Requisitos estruturais

Tal como referido, a grande motivação no desenvolvimento deste projeto é a possibilidade dos biólogos divulgarem o trabalho desenvolvido nos seus centros de investigação. Este objetivo implica diretamente a existência de duas entidades na utilização da aplicação, ou seja, por um lado os biólogos que utilizam a aplicação no seu dia a dia para catalogar e divulgar as proteínas com que trabalham, e por outro lado os potenciais interessados que pretendem encontrar proteínas dentro das suas necessidades, com características e funções específicas. Outro dos requisitos já referidos é a possibilidade da aplicação ser utilizada de modo autónomo, e obviamente em vários centros de investigação simultaneamente. De forma a suportar estes requisitos a aplicação foi projetada em dois módulos independentes: uma aplicação local, a ser utilizada pelos biólogos, e uma aplicação web, que permitirá o armazenamento e pesquisa das proteínas publicadas.

A figura 3.3 demonstra qual a organização adotada para a implementação da aplicação, o que não define completamente a sua arquitetura que será mais aprofundada no capítulo 4.

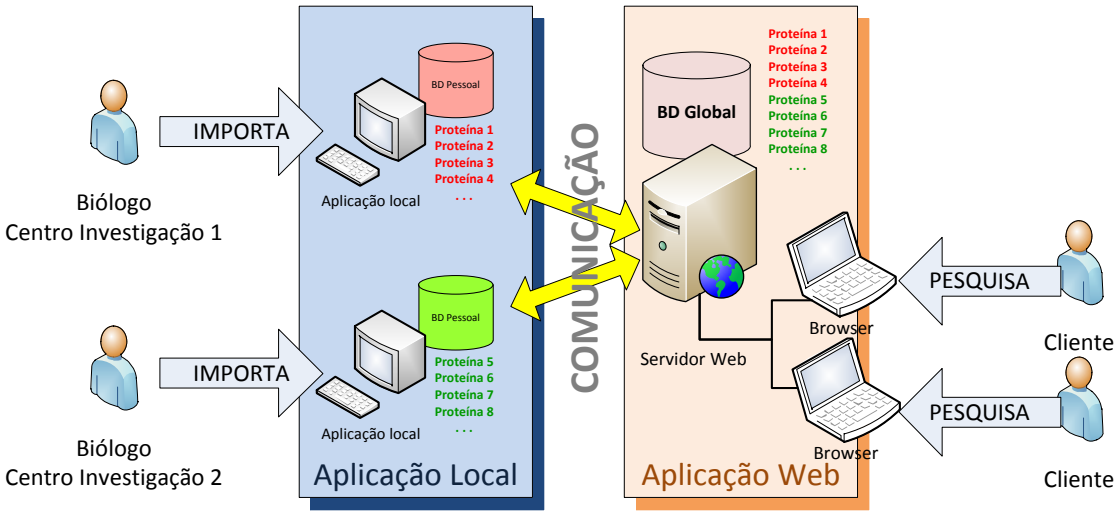


Figura 3.3: Representação da organização completa da aplicação, onde são facilmente visíveis os dois lados funcionais com diferentes utilizadores.

Como se pode observar na figura existem dois módulos distintos na aplicação. Por um lado uma aplicação local que será usada nos centros de investigação e que permite aos biólogos gerir, catalogar e analisar as sequências proteicas com que trabalham. Esta aplicação mesmo estando fortemente relacionada com a aplicação web funciona de forma autónoma apenas como uma biblioteca de proteínas, facilitando assim a sua utilização.

Do outro lado está a aplicação web que servirá para pesquisar pelas proteínas publicadas pelos biólogos. Esta plataforma terá uma base de dados própria e sincronizada com as várias bases de dados existentes nas aplicações locais. Estes mecanismos de sincronização são assegurados com comunicação entre as duas entidades, comunicação esta que permite por exemplo a publicação de proteínas, o registo de utilizadores ou a troca de informação estatística.

3.4 MockUps

De forma a conseguir ir de encontro às expetativas dos biólogos para o sistema a construir foram construídos alguns modelos visuais da interface da aplicação. Estas interfaces, ou *mockups* [1], são uma forma rápida de mostrar aos utilizadores como será a aplicação sem ter de implementar toda a funcionalidade subsequente. Foram construídos modelos pouco detalhados, representando apenas uma utilização mais comum da aplicação de forma não limitar muito o seu desenvolvimento e permitir um posterior ajuste e inclusão de novas funcionalidades. Esta técnica permitiu identificar algumas questões importantes numa fase inicial do projeto e ter em conta todos estes aspetos durante todo o desenvolvimento.

Para a aplicação local foram desenvolvidos *mockups* para as interfaces mais utilizadas. Em primeiro lugar foi construída a interface para o ecrã de login, representado na figura 3.4.

Pretende-se uma interface muito simples que apenas permita ao utilizador inserir os seus dados de

[1] *mockup* é uma representação da interface que não cumpre nenhuma finalidade a não ser demonstrar uma proposta para a aparência final do sistema, sem a capacidade de simular seu comportamento

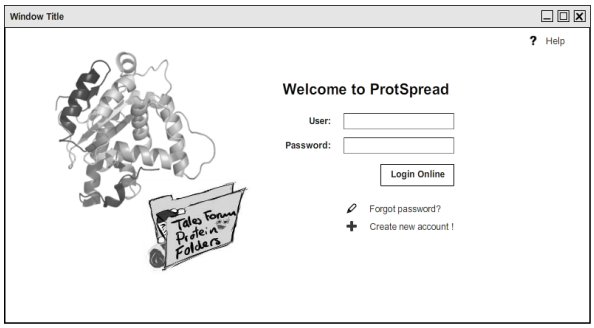


Figura 3.4: Mockup do ecrã de login da aplicação local.

autenticação. Existe também uma opção para o caso de ser um novo utilizador, que abre a aplicação web no *browser* para se proceder à criação de um novo registo.

Foram também desenvolvidos mais dois ecrãs do funcionamento normal da aplicação. O primeiro representado na figura 3.5 demonstra a forma como se pretende lidar com grandes quantidades de proteínas. Pretende-se visualizar toda a informação na forma de uma tabela, por ser uma forma compacta de mostrar grandes quantidades de informação. O utilizador pode também rapidamente reordenar as proteínas pela coluna que desejar ou pesquisar pela proteínas desejadas numa caixa de texto. É também desta forma que será possível iniciar análises InterProScan para uma grande quantidade de proteínas, sendo apenas necessário selecionar quais as proteínas a analisar e iniciar o processo.

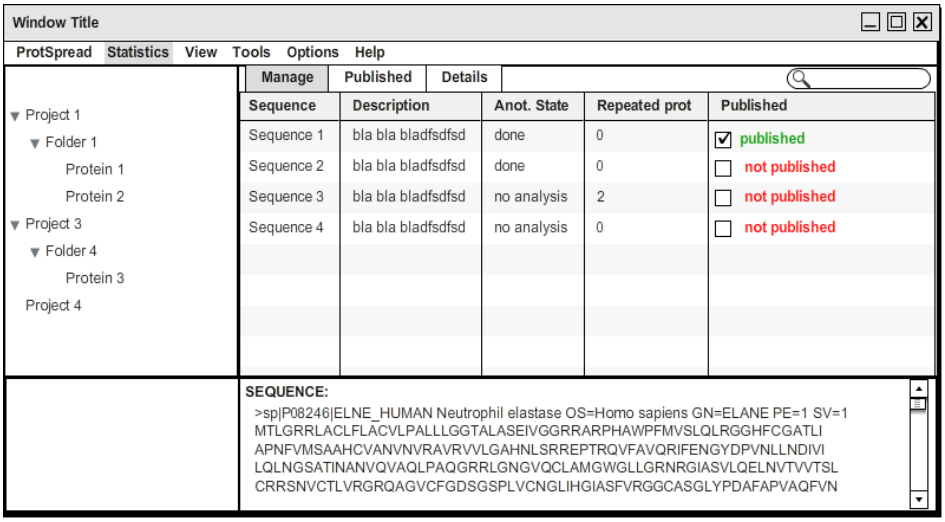


Figura 3.5: Mockup que representa a interface de gestão de múltiplas proteínas em simultâneo.

Foi também desenhado um mockup que representa a forma como se poderá visualizar a informação detalhada relativa a uma proteína (Figura 3.6). Nesta interface o utilizador pode consultar todas as informações de uma sequência proteica como a análise InterProScan, o tamanho da sequência, o organismo a que pertence, realizar uma análise InterProScan usando o serviço web ou publicar a proteína no site web.

Finalmente foram desenvolvidos alguns mockups da aplicação web. Pretendia-se uma interface simples e limpa, de forma a que se tenha uma navegação intuitiva e agradável. Os aspetos mais importantes a ter em conta são a estrutura do web site, a forma como os resultados das pesquisas

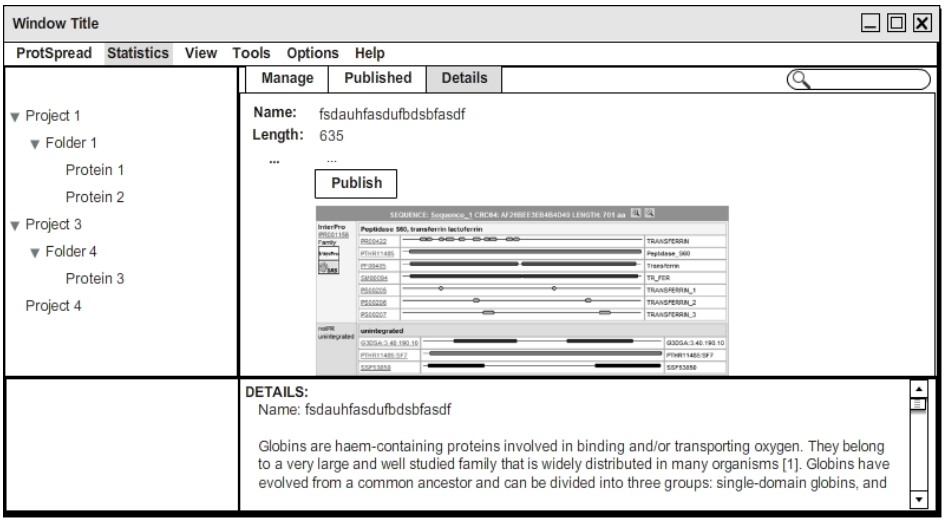


Figura 3.6: Mockup que representa a interface de visualização dos detalhes de uma proteína.

são representados e a visualização em detalhe de uma publicação. Assim foram construídos alguns mockups para projetar a forma como estas ações serão tratadas.

Em primeiro lugar, optou-se por construir um layout muito simples que se mantém inalterado qualquer que seja a página onde o utilizador se encontre. No topo da página existe uma barra que tem como finalidade mostrar se o utilizador se encontra autenticado ou não, e as opções para login/logout, ou caso seja um novo utilizador criar uma nova conta. Do lado esquerdo permanece um menu que permite ao utilizador navegar facilmente e aceder às várias funções disponíveis no site. É também neste menu que o utilizador poderá descarregar a aplicação local para o seu computador. Na figura 3.7 está projetada a *home page* da aplicação web onde é possível visualizar os vários componentes existentes e a sua organização.

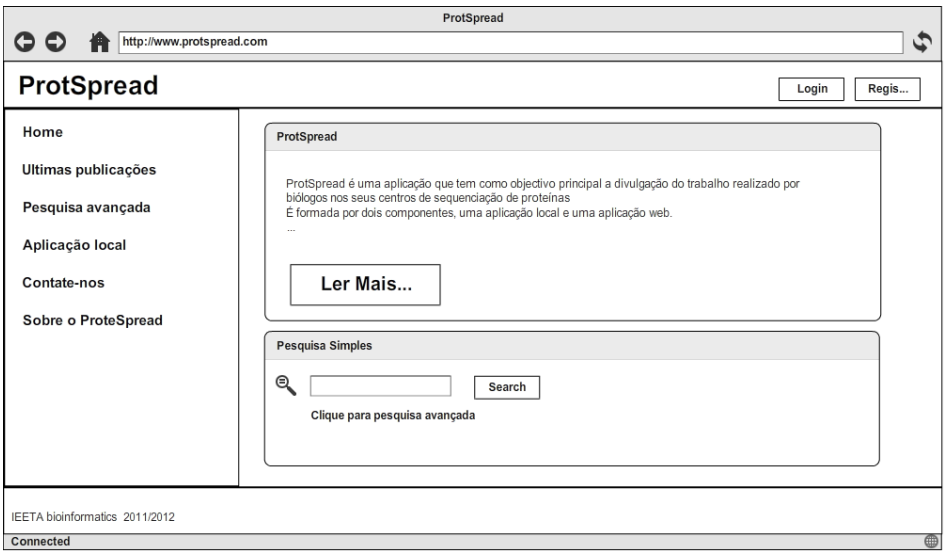


Figura 3.7: Mockup da *home page* da aplicação web, onde é possível visualizar a organização dos vários elementos da aplicação.

Ao realizar uma pesquisa, ou ao consultar as ultimas proteínas publicadas na aplicação o utilizador vai obter uma lista de resultados. Esta funcionalidade revela apenas o mínimo de informação sobre a proteína sendo que o utilizador pode obter uma vista detalhada de uma publicação clicando num botão disponível para o efeito.

Nesta funcionalidade é possível visualizar toda a informação relativa à proteína, à análise InterProScan e à publicação. De forma a facilitar a visualização de uma análise InterProScan apenas são mostradas as entradas InterPro. Para visualizar em mais detalhe as assinaturas das bases de dados, basta o utilizador clicar na assinatura InterProScan para surgirem todas as entradas relacionadas com determinada assinatura. Por exemplo, é possível verificar na figura 3.8, que a proteína 1 contém as assinaturas InterProScan IPR000719, IPR002219, IPR008271 e entradas sem correspondência InterProScan. Selecionando, por exemplo, a assinatura IPR002219, é possível verificar quais as assinaturas correspondentes e quais as bases de dados a que correspondem, neste caso, Pfam, PROSITE, PROFILE e SMART.

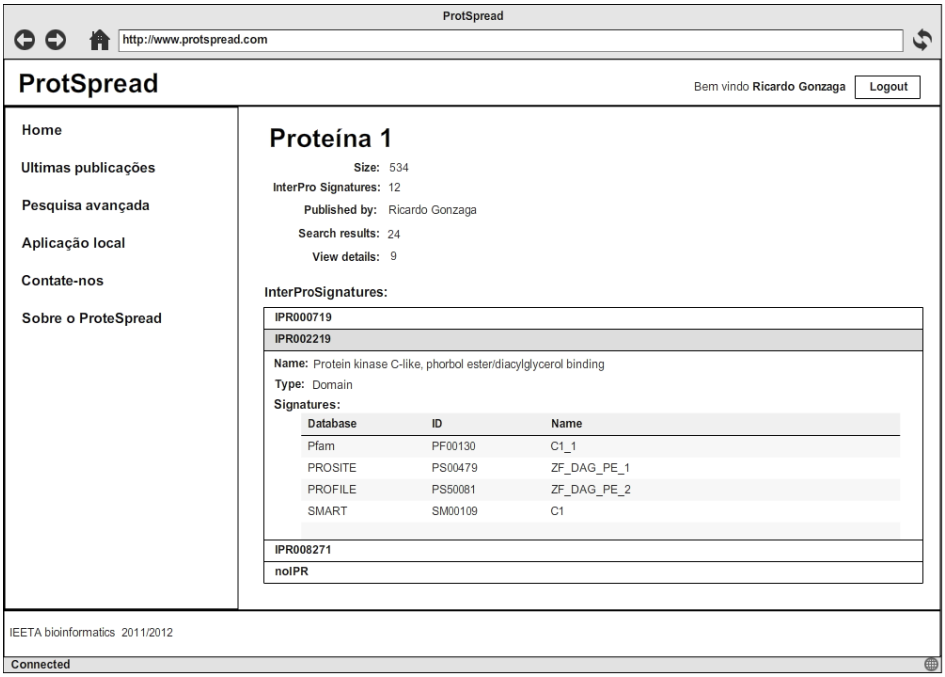


Figura 3.8: Mockup da forma de visualização dos detalhes da publicação de uma proteína, onde é possível visualizar todas as assinaturas publicadas.

3.5 Sumário

Neste capítulo foram apresentados todos os requisitos definidos para a aplicação a desenvolver. A aplicação tem dois tipos de utilizadores e o sistema foi definido analisando os possíveis casos de uso para cada um destes atores (Biólogo e Consumidor). Foram analisados os requisitos funcionais implementando um diagrama de casos de uso para cada uma das partes da aplicação, primeiro para a aplicação local e em segundo lugar para o site web. A descrição detalhada destes casos de uso define todos os requisitos do sistema.

Também foram analisados os requisitos não funcionais, ou seja, os aspetos que descrevem as qua-

lidades do sistema como por desempenho ou segurança dos dados. Os dados tratados na aplicação podem ser confidenciais e por isso devem ser protegidos e com acesso reservado. Apesar da aplicação ter como objetivo principal a divulgação de proteínas, esta divulgação será baseada apenas nos domínios funcionais InterPro e a sequência nunca será divulgada, já que esta informação é o resultado do trabalho dos biólogos nos seus centros de investigação.

A estrutura do sistema também é apresentada neste capítulo. Tal como já referido a solução é composta por dois módulos, uma aplicação local e um site web. Estes dois componentes são aqui apresentados e definidos com mais rigor tal como a forma como as duas partes interagem.

Finalmente são apresentados alguns *mockups* que pretendem definir a apresentação do sistema. Esta é uma forma fácil de ir de encontro às expectativas dos utilizadores e facilita o início do desenvolvimento da aplicação. São apresentados modelos tanto para a aplicação local como para a aplicação web.

Capítulo 4

Modelo Proposto e Implementação

De forma a suportar os requisitos descritos no capítulo anterior, foi desenhada uma arquitetura adequada. Este capítulo descreve a solução adotada para a construção da aplicação ProtSpread, tanto para a parte local da aplicação como para web. São descritas todas as tecnologias utilizadas desde as linguagens de programação, bases de dados, serviços web utilizados até às *frameworks* de suporte à aplicação.

É também especificada a arquitetura de todo o sistema tanto a nível local como para a aplicação web, e são descritos os vários módulos implementados e a sua funcionalidade. Finalmente são apresentadas e descritas todas as estruturas que suportam o armazenamento e gestão de informação na aplicação.

4.1 ProtSpread - Especificações das tecnologias propostas

4.1.1 Java - Linguagem de programação utilizada

Java é uma linguagem de programação disponibilizada de forma gratuita sob a forma de um pacote de desenvolvimento de aplicações Java ou Software Development Kit (SDK) [1]. Nos últimos anos o Java teve um grande aumento no número de utilizadores, e tornou-se uma linguagem madura, com um grande número de funcionalidades e com bom desempenho. Com a sua evolução, o Java desenvolveu um número de Application Programming Interfaces (APIs) e uma documentação muito completa que ajudam no desenvolvimento de aplicações.

Outra das vantagens de utilizar Java é sua arquitetura *Java 2 Enterprise Edition*, que permite que a linguagem seja utilizada tanto do lado do cliente como do servidor (Local e Web). Desta forma é possível a reutilização de código em ambas as partes e uma maior integração e facilidade de comunicação.

Relativamente às outras linguagens Java perde ao nível de desempenho, devido à maquina virtual interpretar *bytecode* para a execução de programas. No entanto, esta diferença de desempenho tem sido substâncialmente reduzida ao longo dos anos com o lançamento de novas versões.

Finalmente outra das razões que levaram à escolha desta linguagem foi o suporte nativo para o desenvolvimento de interfaces gráficas utilizando a tecnologia Swing. Esta é uma API é muito completa e permite obter uma aparência muito semelhante nas aplicações, independentemente do sistema operativo utilizado.

[1] Disponível para download em http://www.oracle.com/technetwork/java/javase/downloads/index.html (visitado a 05-06-2012)

4.1.2 BioJava

BioJava é uma framework Java para processamento de dados biológicos [28]. Esta biblioteca pretende aproveitar as capacidades da linguagem Java e disponibilizar um conjunto de métodos de forma a encapsular conceitos básicos de bioinformática. Fornece funcionalidades como processamento de dados biológicos, análise estatística, ferramentas para leitura de formatos *standards* de ficheiros e pacotes para a manipulação de sequências e de estruturas 3D.

BioJava é um projeto maduro utilizado em inúmeras aplicações e divulgado em mais de 50 estudos já publicados. Esta ferramenta dispõe de objetos para representação de sequências, alinhamentos, anotações, algoritmos de comparação de sequências ou representações gráficas de sequências. Na base do BioJava está um alfabeto simbólico que representa sequências como uma lista de referências a *tokens* de objetos derivados de um alfabeto. As listas de símbolos são armazenadas, sempre que possível, sob a forma comprimida de até 4 símbolos por *byte* de memória.

Além dos símbolos fundamentais de um alfabeto (A,C,G,T) todos os alfabetos BioJava contêm implicitamente outros símbolos que representam todas as possíveis combinações dos símbolos fundamentais.

4.1.3 Java Web Start

Visto que o sistema desenvolvido é composto por dois módulos, uma parte local e uma parte web, é necessário que o utilizador instale a aplicação local na sua máquina para que a possa utilizar. O Java Web Start (JAWS) é uma tecnologia que permite ao utilizador descarregar e iniciar uma aplicação Java a partir de um browser comum, bastando para isso fazer um simples *click*.

Desta forma são automatizados vários processos, facilitando assim a utilização do sistema:

- fornece uma forma fácil de executar a aplicação local a partir do web site integrando assim os dois sistemas;
- garante que a aplicação executa sempre a sua última versão, realizando atualizações automaticamente no início de cada execução;
- elimina processos instalação ou *upgrades* complicados.

Atualmente o JAWS está incluído no pacote Java Runtime Environment (JRE) desde a versão 5.0. Isto significa que quando um utilizador instala a tecnologia Java, automaticamente instala o JAWS. Quando se inicia uma aplicação Java com esta tecnologia pela primeira vez, o JAWS descarrega a aplicação e as bibliotecas necessárias para um sistema de *cache* local. Posteriormente, nas execuções seguintes, é verificado se existe uma versão mais recente da aplicação e caso seja necessário é descarregada a nova versão, ou se não houver *updates* a aplicação arranca instantaneamente.

4.1.4 Swing Application Framework - Plataforma de desenvolvimento de aplicações *Desktop*

O objetivo do *Swing Application Framework* é definir uma infraestrutura que é comum à maioria das aplicações locais com utilização de Graphical User Interfaces (GUIs). Esta framework visa fornecer um ponto de partida para aplicações que utilizem interfaces Swing, tornando o seu desenvolvimento mais fácil e mais simples, mesmo para programadores com pouca experiência neste tipo de aplicações.

As principais vantagens na utilização desta framework são:

- gestão do ciclo de vida da aplicação, especialmente *arranque* e *encerramento*.
- controlo e carregamento de recursos, como mensagens, imagens, cores, etc.
- facilidade para lidar com ações
- execução de tarefas em *backgroung*
- guarda o estado da sessão entre execuções

Algumas destas funcionalidades são imprescindíveis para o bom funcionamento da aplicação. A gestão do ciclo de vida permite criar as ligações a bases de dados necessárias, configurar toda a GUI da aplicação e criar o diretório de dados da aplicação antes do arranque do interface de utilizador. Da mesma forma durante o encerramento da aplicação é possível desligar as ligações às bases de dados assim como terminar todas as tarefas em execução.

Outra grande vantagem é a capacidade de executar e gerir tarefas em *background*. A aplicação tem funcionalidades que exigem algum tempo de processamento como execução de *web-services*, que bloqueariam a interface de utilizador caso não fossem executadas em segundo plano. Esta framework não só permite executar estas tarefas de forma a manter uma execução fluída como permite gerir a sua execução, enviando mensagens sobre o estado da tarefa ou qual o seu progresso.

Finalmente as outras funcionalidades não são fundamentais para o bom funcionamento do sistema, mas melhoram a sua qualidade oferecendo ao utilizador uma melhor experiência de utilização.

4.1.5 Java DB - Base de dados embutida

Uma das tecnologias integradas na linguagem Java é uma base de dados relacional de código fonte aberto desenvolvida pela *Apache Software Foundation*. Esta base de dados, também designada por Apache Derby [2], tem a capacidade de ser integrada com a aplicação, não sendo necessário a existência de um servidor. Desta forma é possível utilizar a mesma tecnologia para base de dados em ambas as partes da aplicação,quer na aplicação local, quer na página web. Mais uma vez também é possível reutilizar código entre os dois módulos, visto que parte das bases de dados são comuns.

A utilização de Java DB tem várias vantagens como, por exemplo, ser construída na mesma linguagem da aplicação, acesso através de um driver Java Database Connectivity (JDBC) integrado e tamanho muito reduzido. Outro aspeto que decidiu a utilização desta base de dados é o facto de ser suportada numa grande quantidade de sistemas operativos e de ter uma licença de utilização não proprietária, o que permite a sua utilização mesmo para fins comerciais.

4.1.6 *Play!* Framework - Plataforma de desenvolvimento web

Play! [3] é uma framework de construção de aplicações web na linguagem Java.

A sua arquitetura facilita em muito o desenvolvimento de aplicações web devido à separação entre os dados e a apresentação da aplicação.

A grande vantagem da utilização desta tecnologia é o facto de ser baseada na linguagem Java, tal como a parte da aplicação local. Desta forma é possível a partilha de código entre as duas partes e consegue-se uma maior integração. Esta característica também permite manter a utilização das mesmas ferramentas e bibliotecas nas duas partes do sistema, visto que é compatível com qualquer biblioteca Java padrão.

[2] http://db.apache.org/derby/

[3] http://www.playframework.org/

A utilização desta framework torna-se muito simples visto que não é necessário qualquer tipo instalação ou configuração. A *Play!* inclui vários componentes prontos a ser utilizados como por exemplo um módulo de autenticação que pode ser integrado facilmente na aplicação, ou um sistema de validação de *inputs*.

Devido à sua forma de funcionamento, a *Play!* permite um desenvolvimento rápido de aplicações, visto que não necessita do processo de compilação, instalação e reiniciar o servidor em cada alteração, ao contrário da maioria das frameworks do mesmo género. Nesta caso basta apenas editar os ficheiros, guardar e atualizar o browser para obter o novo resultado. Este processo permite poupar muito tempo de desenvolvimento ao mesmo tempo que se torna mais entusiasmante.

Finalmente a escolha desta framework também se deveu, mais uma vez, a ser possível a sua utilização numa aplicação comercial sem qualquer restrição. A versão utilizada (1.2.4) está no momento em modo de manutenção, o que significa que quaisquer erros que possam existir serão corrigidos e a compatibilidade da sua API será mantida.

4.1.7 Twitter Bootstrap

O Twitter Bootstrap [4] é um pacote de ferramentas para criação de sites e aplicações web, que contém um conjunto de elementos baseados em HTML, Cascading Style Sheets (CSS) e JavaScript, tais como modelos de tipografia, botões, formulários e componentes de navegação. É portanto uma ferramenta com grande utilidade para desenvolver o site web do sistema.

É uma tecnologia que ainda não é completamente compatível com HTML 5, mas é compatível com todos os principais *browsers* atualmente utilizados. Tem a vantagem de permitir uma forma de desenvolvimento simples, sem no entanto limitar na construção de componentes complexos.

Outra das vantagens de utilizar este pacote é permitir que a aplicação seja utilizada em qualquer tipo de plataforma (computador, dispositivos móveis ou tablets), já que os designs construídos através de Bootstrap se adaptam ao tipo de plataforma onde são executados. Mais uma vez, o facto de ser abrangido por uma licença *Open-source* foi um fator muito importante na escolha desta plataforma.

4.2 Arquitetura

Tal como já referido, o sistema é constituído por dois blocos que podem ser consideradas duas aplicações distintas. Do ponto de vista estrutural a aplicação local é considerada um cliente enquanto que a aplicação web é considerada um servidor, tendo portanto arquiteturas distintas. No entanto, é importante tentar combinar o máximo de componentes nos dois sistemas, de forma a simplificar e reutilizar algumas funcionalidades em ambas as partes.

Desta forma teve-se o cuidado de separar as várias funcionalidades por blocos, assim como separar a parte de processamento da parte de apresentação, para que alguns dos módulos sejam comuns nas duas arquiteturas e assim simplificar o processo de desenvolvimento do sistema.

Além disso, a separação em blocos facilita a criação de novas funcionalidades e a expansão do sistema numa fase futura.

4.2.1 Aplicação local

A aplicação local é constituída por seis blocos principais divididos em três camadas lógicas (Figura 4.1).

[4] http://twitter.github.com/bootstrap

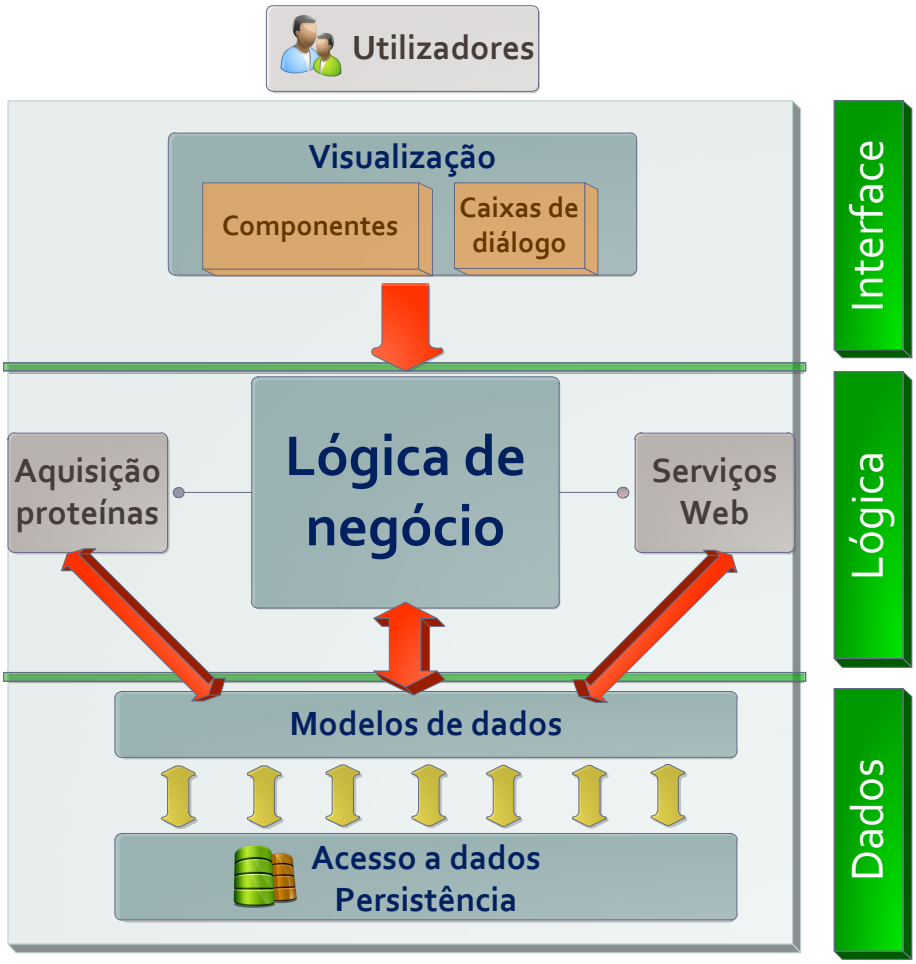


Figura 4.1: Arquitetura da aplicação local, onde é possível distinguir as três camadas lógicas e os vários módulos que as compõem.

Os blocos definidos na arquitetura são:

Acesso a dados/Persistência – Parte da aplicação responsável por garantir a persistência dos dados entre execuções. Nesta camada são implementados todos os processos de acesso à base de dados local, fornecendo assim uma camada de abstração a todos os módulos que se encontrem nos níveis acima na arquitetura da aplicação. Existem dois tipos de acessos distintos: acesso à base de dados de utilizadores locais, ou acesso à base de dados pessoal de um utilizador. Este último acesso funciona com base no nome de utilizador e password do utilizador de forma a garantir a segurança dos dados. Este módulo tem também a responsabilidade de construir e gerir o diretório de dados e definições necessárias para o funcionamento da aplicação. Todos estes detalhes são explicados mais à frente no relatório.

Modelos de dados – Neste módulo estão alojados todos os modelos de dados utilizados pela aplicação. Cada tipo de dados é representado com um conceito de alto nível em que as suas características específicas são encapsuladas de forma a conseguir uma manipulação mais simples. Para permitir uma maior facilidade em lidar com estes dados cada objeto tem um conjunto de mé-

todos que permite obter e guardar informação de uma forma direta. Por exemplo, uma análise InterProScan é representada num único objeto, onde é possível obter uma lista das entradas que constituem uma assinatura, ou por exemplo obter apenas as assinaturas de um tipo específico.

Aquisição proteínas – Este é o módulo da aplicação responsável por importar novas proteínas para a aplicação. A sua funcionalidade consiste em interpretar os ficheiros no formato FASTA e transformar todas as sequências proteicas existentes no ficheiro em objetos de dados do tipo "Proteína".

Serviços web – Visto que a aplicação utiliza muitos serviços web, é neste módulo que se encontra grande parte da funcionalidade da aplicação. Aqui estão implementadas todas as funcionalidades relacionadas com a utilização de serviços externos. Em primeiro lugar todas as ferramentas InterProScan que são disponibilizadas sob a forma de *web-services*, e também todos os serviços de sincronização com a aplicação web, por exemplo consulta de utilizadores ou publicação de proteínas.

Lógica de negócio – Este módulo é a parte do sistema que se encarrega das tarefas relacionadas com os processos de negócio, ou seja tarefas que realizam entradas de dados, consultas de dados, geração de resultados, etc. Como se pode também verificar pelo diagrama de arquitetura (Figura 4.1) este é o módulo que interliga os vários componentes da camada lógica da aplicação, e comunica com a camada de dados e interface.

Visualização – Módulo onde estão implementados todos os componentes gráficos da aplicação local. Está dividido nos dois sub-módulos seguintes:

- **Componentes** - Painéis gráficos utilizados na janela principal da aplicação.
- **Caixas de diálogo** - Janelas independentes usadas para obter ou

4.2.2 Aplicação web

A arquitetura utilizada na aplicação web é a arquitetura da framework utilizada para o desenvolvimento do web site. A *Play!* foi construída segundo o modelo de desenvolvimento Model-view-controller (MVC) (Figura 4.2), atualmente considerado uma arquitetura padrão.

Este modelo separa a lógica da aplicação da interface de utilizador permitindo desenvolver, editar e testar cada parte separadamente. O objetivo deste modelo é mapear os pedidos à aplicação, processar as tarefas necessárias e produzir um modelo gráfico para representar os resultados obtidos. No entanto, neste tipo de arquitetura este processo torna-se muito simples.

Os vários componentes que fazem parte desta arquitetura são:

Modelo – O modelo representa os dados da aplicação e as regras lógicas que permitem operar esses dados. Este é no fundo o núcleo da aplicação onde está definido o comportamento da aplicação, e onde a parte responsável pelo armazenamento, manipulação e geração de dados. Funciona como um encapsulamento de dados e de comportamento independente da apresentação.

Vista – A vista apresenta o conteúdo de um modelo. Recebe dados da aplicação através do modelo e define a forma como estes dados devem ser apresentados. É nesta camada que existem os elementos de visualização como HTML ou XML. É a camada que fornece a interface de utilizador e é portanto responsável por receber as entradas de dados, e da mesma forma apresentar resultados.

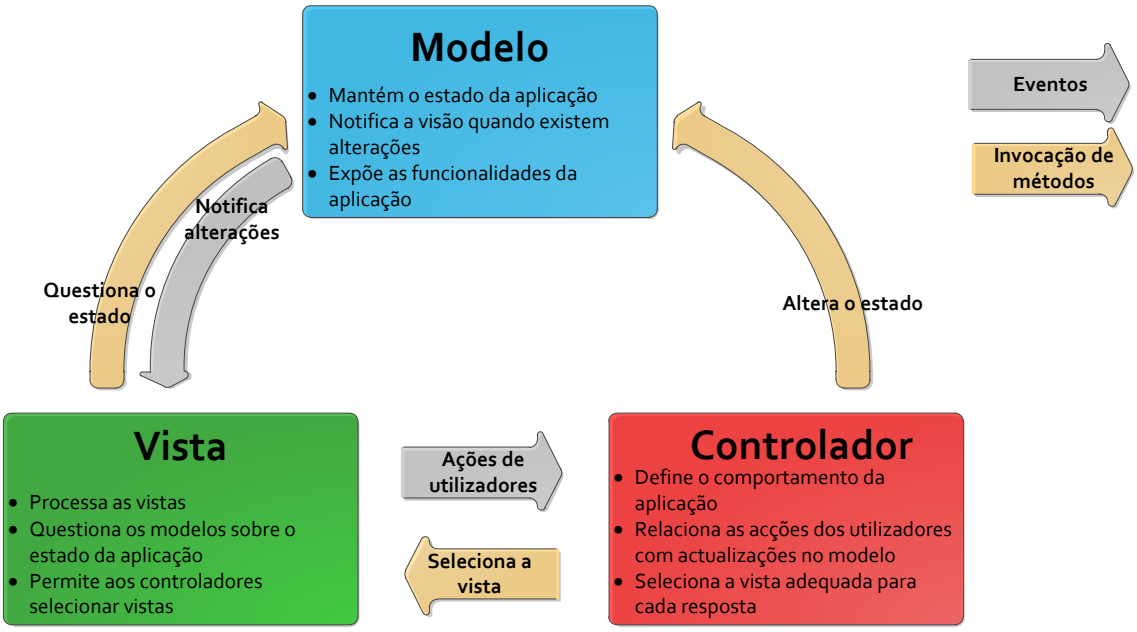


Figura 4.2: Diagrama que ilustra a relação existente entre os vários componentes na arquitetura MVC.

Controlador – O controlador tem a responsabilidade de converter as interações do utilizador com a vista em ações a serem realizadas pelo modelo. Este tipo de interação, no caso da aplicação web, são métodos Hypertext Transfer Protocol (HTTP) como por exemplo GET ou POST. As ações executadas pelo modelo incluem consultar ou alterar o estado do modelo.

4.3 Estruturas de dados - modelo proposto

Visto que o sistema gere uma grande quantidade de dados, foi necessário criar os sistemas que possibilitem guardar e gerir estes dados de forma eficiente e estruturada. Esta estruturação tem de ser implementada de forma a suportar todos os requisitos do sistema, e assim há vários aspetos a ter em conta, tais como:

- o sistema é constituído por duas partes, que apesar de funcionarem de forma independente, devem manter sincronização de dados, tais como publicações de proteínas.
- a aplicação local corre em computadores locais e os dados manipulados também devem ser guardados localmente.
- a aplicação local tem vários utilizadores e os dados dos vários utilizadores não devem ser partilhados.

De forma a conciliar todos estes requisitos optou-se por criar um sistema que utiliza vários diretórios e bases de dados de forma a estruturar toda a informação da forma mais organizada e lógica possível. Para a aplicação local é criado um diretório, utilizando a propriedade **"Home directory"** do sistema operativo, onde estão guardadas todas as informações dos utilizadores, assim como os seus

dados pessoais. Para a aplicação web o sistema de gestão de dados não é tão complexo. Neste caso toda a informação é mantida numa base de dados relacional. Todos estes detalhes são discutidos nas próximas secções.

4.3.1 Diretório da aplicação local

Em primeiro lugar foi necessário criar um local onde fosse possível armazenar localmente todo o conteúdo relacionado com a aplicação. Assim, aquando o primeiro arranque, a aplicação verifica se na pasta pessoal do utilizador, já existe um diretório dedicado à aplicação. No caso de não existir é criado um diretório com a estrutura representada na figura 4.3.

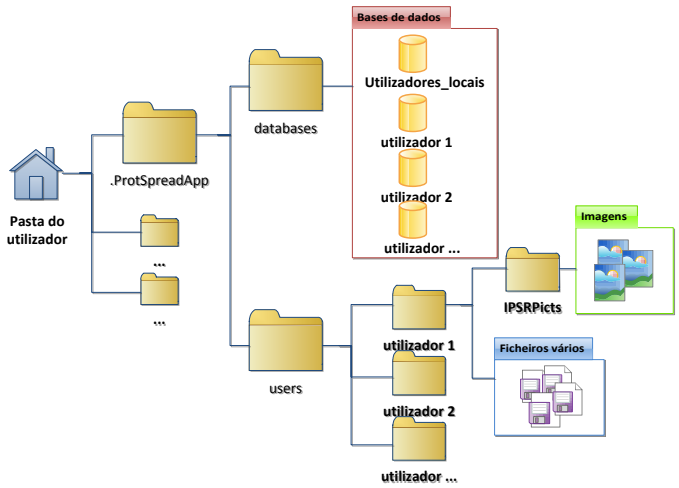


Figura 4.3: Representação da estrutura de diretórios criada pela aplicação local na pasta pessoal do utilizador

No topo da hierarquia encontra-se uma pasta com o nome **".ProtSpreadApp"** que é o diretório local da aplicação. Dentro desta pasta existem duas outras pastas denominadas **"databases"** e **"users"**.

O primeiro diretório, designado "databases", é o local onde são criadas e armazenadas todas as bases de dados utilizadas pela aplicação. É criada uma base de dados para gestão de utilizadores e uma base de dados para cada utilizador da aplicação, onde são armazenados dados relativos a um utilizador, como proteínas importadas, análises, publicações. Na pasta **"users"** é criada uma pasta para cada utilizador. Em cada uma das pastas aqui existentes serão guardados vários ficheiros complementares á base de dados dos utilizadores, como por exemplo imagens que representam graficamente as análises InterProScan.

Esta organização modular facilita não só a organização de conteúdos, mas facilita também a implementação de funcionalidades do tipo exportar/importar conteúdos, como projetos ou contas pessoais. Por exemplo, se um utilizador pretender exportar os seus projetos, em que estão armazenadas as proteínas e análises, para posteriormente carregar a mesma informação num outro local, toda a informação que será necessária guardar está devidamente separada da aplicação e poderá ser facilmente exportada.

4.3.2 Bases de dados

Tal como já referido na secção anterior a aplicação serve-se de várias bases de dados para armazenar toda a informação necessária ao seu funcionamento. No total, todo o sistema utiliza três modelos/arquiteturas de bases de dados, duas na aplicação local e uma na aplicação web.

A aplicação local utiliza uma base dados dedicada apenas à gestão de utilizadores, e para cada um destes utilizadores registados inicia uma base de dados independente para armazenamento de toda a informação pessoal (proteínas, análises, publicações, etc).

Gestão de utilizadores

Esta base de dados tem um modelo muito simples, de apenas uma tabela, e serve para gerir os utilizadores no processo de login local. De forma a garantir a sincronização entre os dois módulos da aplicação, local e web, o *login* local apenas é realizável se o utilizador já estiver registado no site web, como está representado com pormenor no diagrama de atividade representado na figura 4.4.

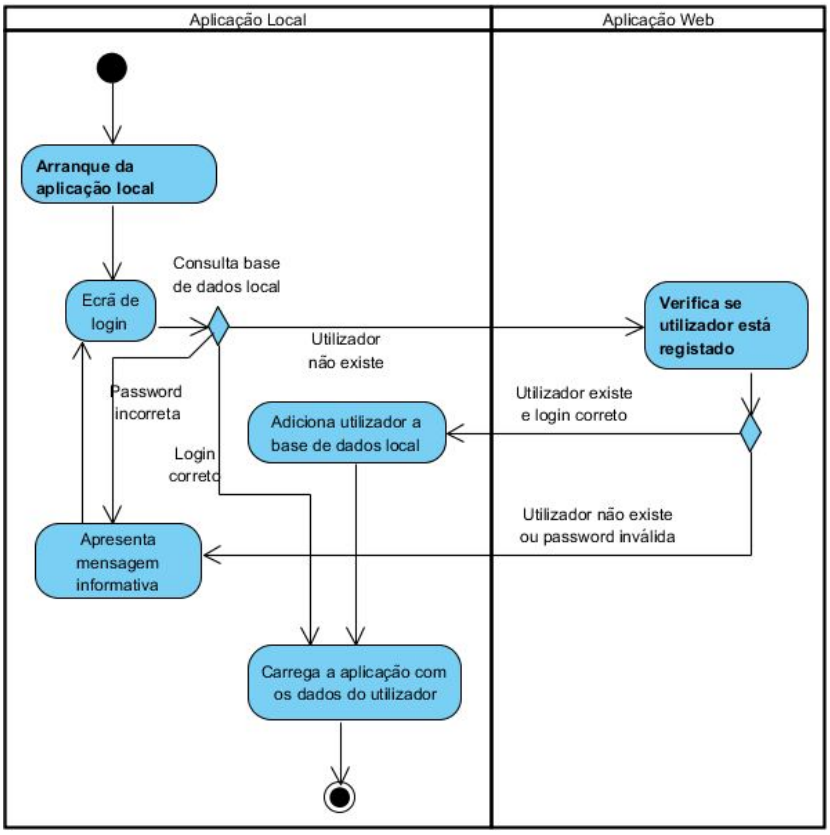


Figura 4.4: Diagrama de sequência que representa com pormenor o processo de login local

Aquando do arranque da aplicação o utilizador é interrogado para indicar qual o seu nome de utilizador e a sua password. Com estes dados a aplicação consulta a base de dados de utilizadores locais e verifica se o nome de utilizador já se encontra ou não registado localmente. Caso o utilizador já se encontre registado, é apenas verificado se a password está correta e o processo continua normalmente. Caso o utilizador não exista, a aplicação local comunica com a aplicação web e questiona se existe algum utilizador registado com aquelas credenciais (utilizador e password). Caso os dados estejam

corretos o novo utilizador é adicionado à base de dados de utilizadores locais, não só as credenciais mas todos os detalhes relativos ao utilizador como email, ou instituição de investigação. Esta comunicação é realizada a partir de um serviço web disponibilizado pela aplicação web, descrito na secção 4.4.

Como é possível verificar existe um processo que garante a correspondência entre os utilizadores registados na aplicação web e os utilizadores registados localmente. Desta forma é sempre possível relacionar as análises InterProScan e as respetivas publicações *online* com o mesmo identificador de utilizador.

Proteínas e análises InterProScan

Para a gestão de proteínas e respetivas análises InterProScan foi criado um modelo de base de dados que pudesse ser partilhado pelos dois módulos da aplicação. O modelo utilizado para representar este tipo de informação está representado na figura 4.5.

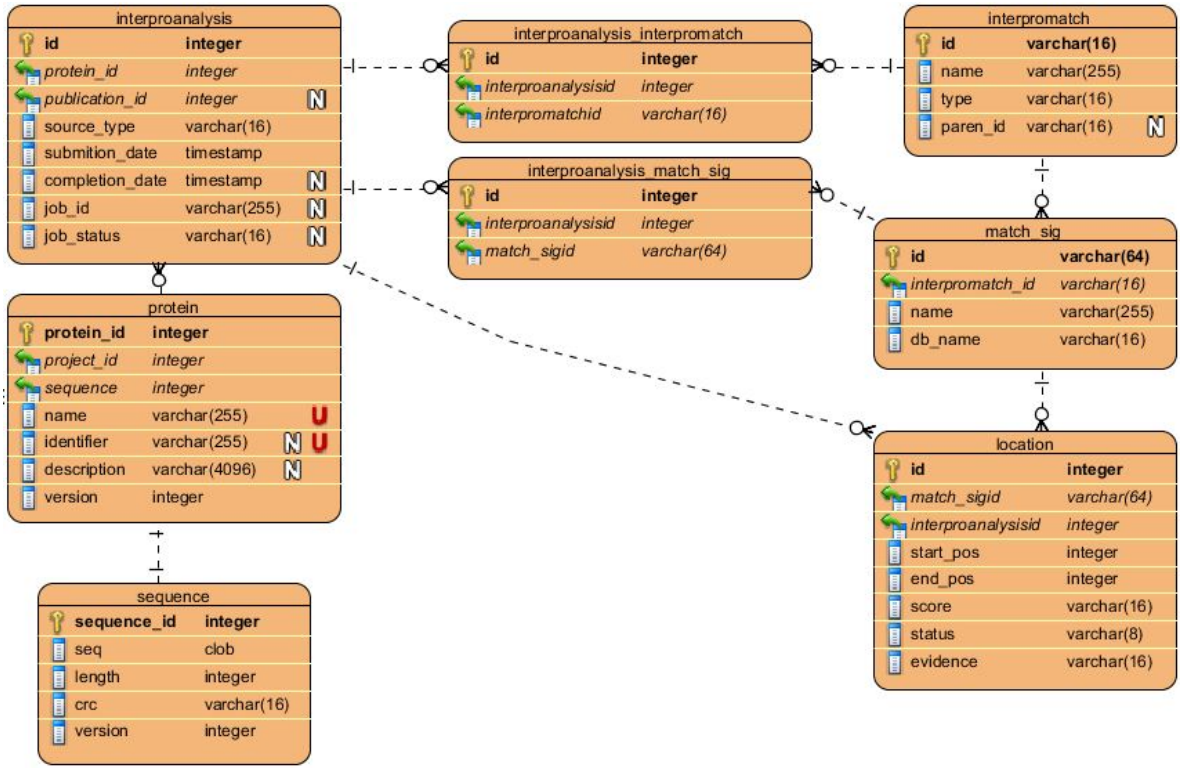


Figura 4.5: Modelo de base de dados utilizado para armazenar a informação relativa às proteínas e às respetivas análises InterProScan

É importante salientar que este modelo não é o modelo completo das bases de dados utilizadas, mas sim um fragmento comum às duas plataformas. Em cada um dos módulos da aplicação este esquema está adaptado de forma a suportar todas as funcionalidades e requisitos específicos de cada aplicação.

De seguida, são descritas as várias tabelas que compõem este esquema, e as relações entre elas:

protein

Entidade central do modelo. Cada entrada nesta tabela corresponde a uma proteína importada

pelo utilizador. Na aplicação local uma proteína está sempre relacionada a um projeto e a uma sequência, enquanto que na aplicação web, estas relações não existem e uma proteína está relacionada apenas com um utilizador. São guardados alguns dados referentes à proteína como o seu nome e uma pequena descrição.

sequence
Entidade onde é armazenada a sequência proteica correspondente a uma proteína. Esta entidade apenas existe na aplicação local e armazena a sequência de aminoácidos que compõem uma proteína. Alguns dados calculados são também armazenados para permitir uma maior performance da aplicação, como é o caso do tamanho da sequência ou o seu código Cyclic Redundancy Check (CRC).

interproanalysis
Tabela onde são registadas todas as análises InterProScan realizadas. Estas análises podem ter como fonte uma análise utilizando o *web-service* online, ou estar associado a uma importação de um resultado de uma análise através de um ficheiro XML. Esta tabela não contém toda a análise mas apenas armazena algumas informações como a data da análise ou o seu tipo.

interproanalysis_interpromatch / interproanalysismatch_sig
Tabelas que contém as associações das assinaturas InterPro e assinaturas das bases de dados específicas com uma análise InterProScan.

interpromatch
Registo de todas as assinaturas InterProScan existentes na aplicação. São guardadas as informações mais importantes acerca de cada entrada, e um identificador que permite aceder a toda a informação detalhada no site do European Bioinformatics Institute (EBI).

match_sig
Registo de todas as assinaturas de bases de dados especializadas como por exemplo CATH, Simple Modular Architecture Research Tool (SMART) ou High-quality Automated and Manual Annotation of microbial Proteomes (HAMAP). Neste caso é registado o nome da assinatura e a base de dados a que corresponde, e está sempre relacionada com uma assinatura InterPro. Tal como na tabela *interpromatch* é guardado um identificador que permite aceder aos detalhes da assinatura no site da base de dados correspondente.

location
Tabela que armazena o local correspondente a cada assinatura, ou seja, a sua localização na cadeia de aminoácidos. Ao contrário das tabelas *interpromatch* e *match_sig* a cada assinatura de uma análise corresponde um entrada. É também aqui que é armazenada a pontuação associada a cada assinatura assim como o seu estado.

4.4 Serviços web

Como já foi referido várias vezes ao longo deste documento, o sistema implementado é composto por dois componentes, a aplicação local e a aplicação web. Estes dois componentes apesar de autónomos, disponibilizam serviços em que a partilha de informação é fundamental.

Os serviços web implementados são a forma de comunicação entre as duas aplicações, e são implementados utilizando a framework *Play!* já referida na secção . Com esta tecnologia torna-se fácil integrar estes serviços web na aplicação web, devido à sua natureza REpresentational State Transfer (REST) e à arquitetura MVC.

Foram implementados dois serviços, que estão disponíveis para serem usados pela aplicação local, quer para consultar ou publicar informação. Os serviços implementados, são descritos de seguida:

getUserIdentifier – Este serviço tem como função permitir à aplicação local verificar se um dado nome de utilizador se encontra registado na aplicação e da mesma forma verificar se a palavra chave associada é correta. Para isso o serviço tem como parâmetros de entrada o nome de utilizador e a password que se pretendem verificar, e estes dados são enviados no pedido HTML como parâmetros do endereço do serviço e processados pela framework.

Caso o utilizador esteja registado e a password associada esteja correta, é enviada uma resposta afirmativa com todas as informações relativas ao utilizador, como contactos, grupo de investigação e inclusive um identificador único para o utilizador em todo o sistema. No caso de os dados estarem incorretos, a resposta informa qual a anomalia encontrada, por exemplo, utilizador não encontrado, ou palavra chave incorreta. O serviço funciona com base no método GET HTTP e a resposta é enviada em formato XML.

publishProteinAnalysis – Este é o serviço responsável pela publicação de proteínas no site web e é também o serviço web mais complexo. Esta funcionalidade recebe como entrada os detalhes da proteína a publicar (nome, identificador e descrição), a análise InterProScan correspondente e a informação do utilizador relativo à proteína. Tal como no serviço anterior as informações relativas à proteína e do utilizador são enviados como parâmetros no endereço do serviço, mas já a análise InterProScan é enviada no corpo da mensagem na forma binária. Para isso é utilizada a tecnologia de serialização do Java e um objeto que representa toda a análise InterProScan é enviado na forma binária. Esta método traz algumas vantagens como simplicidade e robustez na transmissão de informação entre as duas entidades. Outra grande vantagem conseguida com a utilização deste método é o facto de se poder partilhar os mesmos objetos Java nas duas entidades, ou seja, a estrutura de dados que representa uma análise InterProScan é a mesma tanto na aplicação local como na aplicação web. Desta forma, quando a aplicação local pretende publicar uma proteína, apenas necessita de serializar o objeto que representa a análise em forma binária, e colocar estes dados no corpo da mensagem HTTP. Da mesma forma, quando a aplicação web recebe o pedido, lê os bytes do corpo da mensagem e converte estes dados para um objeto Java que contém toda a informação relativa à análise que poderá ser facilmente tratada.

O serviço tem como resposta a informação se a publicação da proteína foi conseguida com êxito ou não, enviando um mensagem HTTP com a resposta adequada.

Este serviços web foram implementados com um sistema muito simples de segurança, já que ainda se encontram numa fase muito experimental. No entanto, a inclusão de formas de segurança mais realistas não será complicado bastando para isso utilizar um canal de comunicação seguro como Secure Socket Layer (SSL).

4.5 Sumário

Este capítulo teve como finalidade apresentar os aspetos relativos à implementação do sistema descrito. O objetivo não é descrever com muito pormenor técnico todos os detalhes da construção da aplicação, mas sim apresentar as linhas gerais da estrutura do sistema.

Todo o sistema está fortemente relacionado com a linguagem Java. São apresentadas as principais tecnologias utilizadas, desde frameworks a bases de dados ou bibliotecas, e quais as razões que levaram a escolher estas tecnologias. É também discutida a arquitetura do sistema e são definidos os vários blocos funcionais e quais as interações existentes entre eles.

Outro assunto abordado neste capítulo é a solução para lidar com toda a informação no sistema, nomeadamente as bases de dados utilizadas e o diretório da aplicação local. Muitas das opções tomadas a este nível tiveram em conta a reutilização de código e a integração nas duas aplicações.

Finalmente foram apresentados os serviços web construídos para permitir a comunicação entre as duas aplicações, e o seu funcionamento. Estes serviços são imprescindíveis quer para assegurar a sincronização entre as duas partes do sistema assim como permitir a publicação de proteínas com as respetivas análises InterProScan.

Capítulo 5

ProtSpread - descrição da aplicação

No presente capítulo é feita uma descrição e apresentação do sistema de informação ProtSpread. Neste sentido, o capítulo está formalmente sub-dividido em duas grandes partes: a aplicação local, e a aplicação web, mesmo sendo estas duas aplicações muito interligadas.

Para isto vão ser demonstradas algumas das funcionalidades dos dois módulos do sistema e serão analisados os resultados obtidos.

5.1 Aplicação Local

5.1.1 Iniciar aplicação e login

Tal como referido no sub-capítulo 4.4, de forma a facilitar a distribuição e instalação da aplicação local, esta está disponível no site web. Para iniciar a sua utilização é apenas necessário fazer um *click* e a última versão da aplicação é automaticamente transferida, instalada e executada localmente, tal como ilustra a figura 5.1. Na primeira execução da aplicação é criado um atalho no ambiente de trabalho do utilizador, assim como no menu de aplicações de forma a facilitar o posterior acesso. É também criado o diretório da aplicação na pasta pessoal do utilizador.

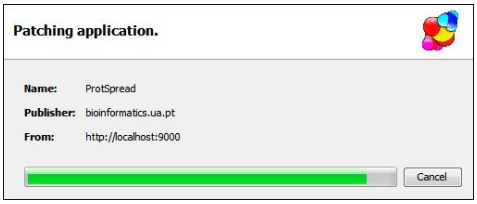


Figura 5.1: Download e instalação automática da aplicação utilizando a tecnologia JAWS

Com este tipo de instalação assegura-se que é sempre executada a última versão da aplicação e evitam-se processos de configuração complicados.

Terminado o processo de transferência e instalação da aplicação, surge o ecrã de login, representado na figura 5.2. É neste processo que o utilizador tem de fazer autenticação para entrar na aplicação e para isso deve de estar registado na aplicação web, e é com as credenciais desse registo que poderá fazer o login local. Caso ainda não tenha realizado o registo online o utilizador pode aceder diretamente ao site web através da aplicação, clicando em *"Create new account!"*.

Ao realizar o login com sucesso a aplicação carrega toda a informação do utilizador e surge o layout principal da aplicação onde o utilizador tem acesso a todo o conteúdo. Nesta interface o utili-

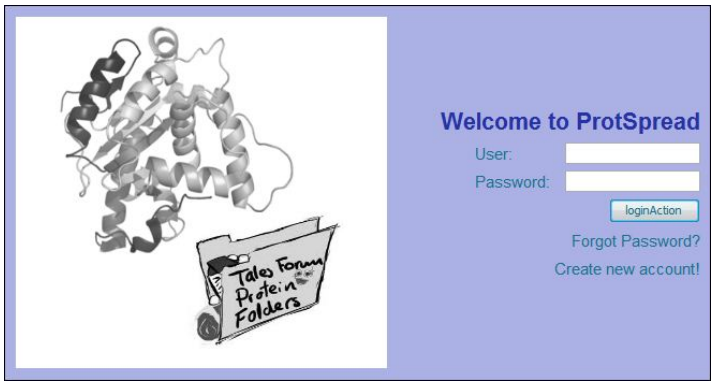


Figura 5.2: Ecrã de login da aplicação local

zador tem a possibilidade de gerir toda a sua informação, criar projetos, importar proteínas, realizar análises.

5.1.2 Criar projetos e importar proteínas

De forma a facilitar a gestão de grandes quantidades de informação, o utilizador deve criar vários projetos com contextos distintos. A criação de um novo projeto consiste apenas em atribuir um nome e uma descrição para o projeto pretendido, através da interface representada na figura 5.3.

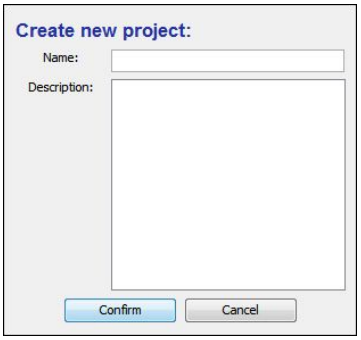


Figura 5.3: Interface de criação de um novo projeto

Ao clicar em **"Confirm"** o novo projeto é automaticamente adicionado ao gestor de projetos, ficando disponível para armazenar novas proteínas. Todos os projetos relacionados com o utilizador são listados na forma de árvore no lado direito do layout da aplicação e cada projeto tem associado um conjunto de proteínas que são também listadas da mesma forma.

Para adicionar novas proteínas a um projeto basta carregar no botão **"Import Proteins"**. As proteínas a importar devem estar num ficheiro FASTA, e ao carregar estas proteínas para o sistema, tanto a sequência como toda a informação contida no ficheiro é carregada para a base de dados.

Tal como referido, o ficheiro que contém as sequências que se pretendem importar deve estar no formato FASTA. Este é um formato baseado em texto para representar sequências de nucleótidos. Uma sequência em formato FASTA começa com uma descrição numa linha única, iniciada pelo símbolo >, seguida por linhas de aminoácidos em sequência, tal como demonstrado de seguida:

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
```

```
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

Ao iniciar o processo de importação de proteínas surge um diálogo que pede ao utilizador a localização do ficheiro a importar. Ao selecionar o ficheiro pretendido, surge um novo diálogo onde são listadas todas as proteínas encontradas no ficheiro associadas à sua descrição e com a indicação de qual o tamanho de cada proteína, tal como ilustra a figura 5.4.

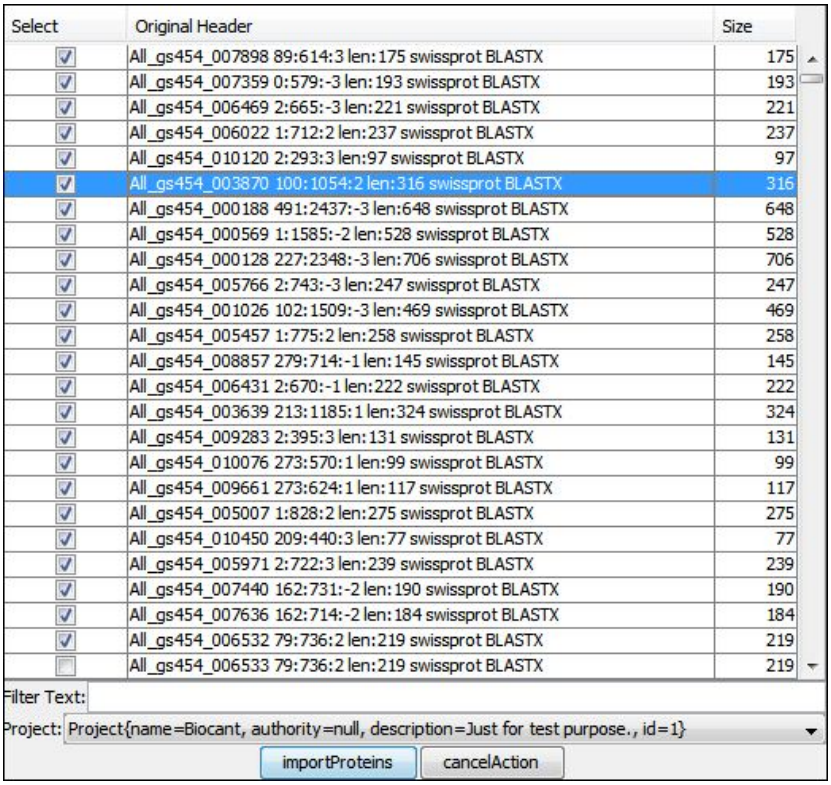


Figura 5.4: Interface de importação de proteínas

Neste diálogo o utilizador tem a possibilidade de selecionar quais as proteínas que pretende importar para a aplicação e qual o projeto ao qual as proteínas devem ser associadas. Existe também uma caixa de texto que permite filtrar as proteínas visíveis pelo texto presente no seu *header*.

Ao clicar em **"Import Proteins"** a aplicação inicia o processo de importação e as proteínas selecionadas ficam disponíveis na aplicação para serem visualizadas e analisadas. Ao selecionar cada uma das proteínas importadas a aplicação mostra todas as informações relativas a essa proteína assim como o nome, tamanho, sequência de aminoácidos, ou análises InterProScan realizadas.

5.1.3 Realização e visualização de análises InterProScan

Ao selecionar uma proteína no gestor de projetos, a aplicação mostra todos os detalhes dessa proteína, inclusive se existe alguma análise InterProScan associada (Figura 5.5).

Caso não exista nenhuma análise associada o utilizador tem a possibilidade de iniciar uma nova análise utilizando o serviço web, ou importar uma análise já realizada a partir de um resultado no formato XML.

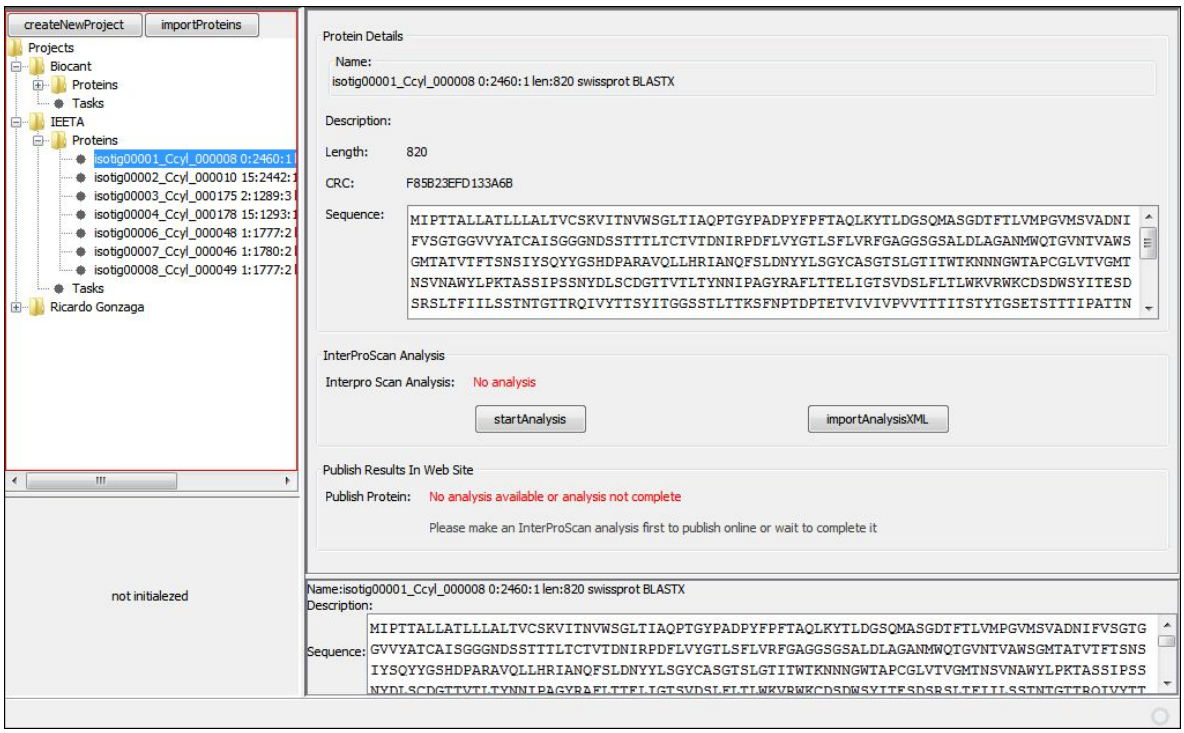


Figura 5.5: Interface principal da aplicação local

A realização de uma análise é um processo demorado e que é composto por várias fases. Assim ao iniciar uma nova análise, a aplicação insere apenas o registo na base de dados que existe uma análise a ser executada mas que ainda não está disponível para consulta. Esta análise fica anotada como **"RUNNING"**. Regularmente o utilizador verifica o estado da análise e caso a análise já esteja concluída a aplicação faz o download do resultado para a base de dados e simultaneamente transfere a imagem disponibilizada pelo serviço web, que facilita a visualização do resultado graficamente. Esta imagem não é armazenada na base de dados mas sim no diretório correspondente aos dados do utilizador (ver secção 4.3.1), e pode ser visualizada utilizando a aplicação, como ilustra a figura 5.6.

5.1.4 Publicação de proteínas

Finalmente, mas não menos importante, o utilizador tem a possibilidade de publicar as proteínas pretendidas no site web. Para publicar uma proteína, com base nos seus domínios InterProScan, basta o utilizador clicar no botão **"Publish Analysis"**. Ao clicar neste botão a aplicação utiliza um serviço web disponibilizado pela aplicação web, que permite enviar um objeto Java *serializado* [1] com toda a informação necessária. Ao publicar uma proteína ela fica disponível para pesquisa e consulta na aplicação web com todos os detalhes da análise.

[1] Serialização - processo de guardar um objeto ou transmiti-lo por uma ligação de rede em forma binária.

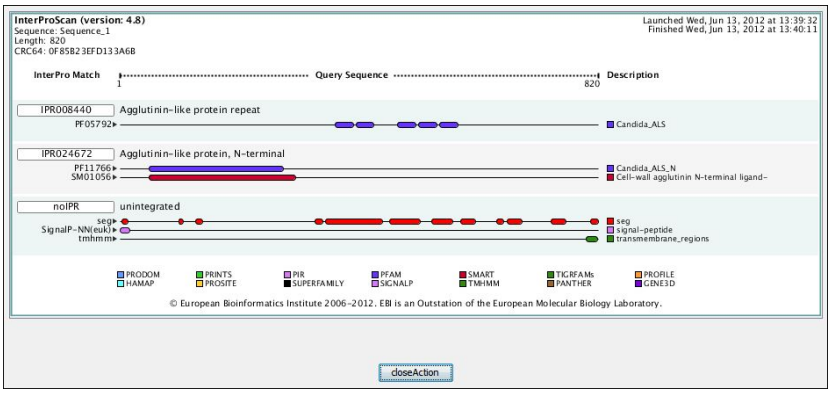


Figura 5.6: Visualização de um resultado InterProScan

5.2 Aplicação Web

5.2.1 Estrutura do site

A aplicação web consiste num conjunto de páginas e serviços web onde são publicadas as proteínas estudadas na aplicação local, para posterior pesquisa e consulta. Todas as páginas existentes na aplicação mantêm sempre o mesmo *layout* para manter a consistência em toda a aplicação. Este *layout* (Figura 5.7) é muito simples de forma a tornar a sua utilização intuitiva e com um aspeto limpo e agradável. Assim cada página tem no seu topo uma barra de estado que permite ao utilizador em qualquer altura aceder à *home page*, verificar se tem ou não o *login* ativo e realizar o *login/logout*. É também nesta barra que os novos utilizadores encontram o *link* para realizar um novo registo.

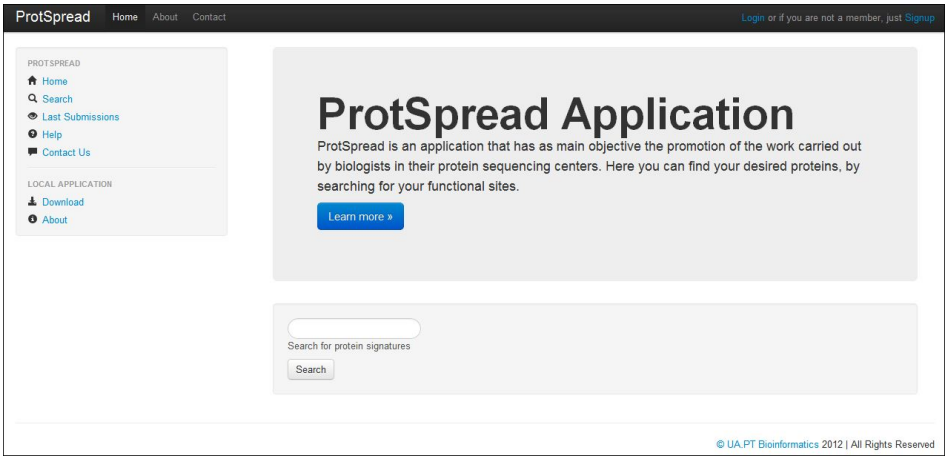


Figura 5.7: *Home page* da aplicação web, onde é possível visualizar o *layout* adotado para a aplicação

Do lado esquerdo de cada página surge um menu de navegação que permite aos utilizadores aceder a todas as funcionalidades da aplicação. Este menu encontra-se dividido em duas partes, sendo que uma corresponde à aplicação web e a outra à aplicação local.

O menu relativo à aplicação local apenas pretende ser uma forma de facilitar o acesso a este módulo do sistema, e é também uma forma de integrar as duas partes que constituem a aplicação. Assim apenas são disponibilizadas duas funcionalidades, a primeira que permite ao utilizador transferir e instalar a aplicação local no seu computador (utilizando Java Web Start), e uma segunda apenas

informativa que tem como objetivo ser uma pequena descrição desta aplicação.

5.2.2 Registo de utilizadores

O primeiro passo para utilizar a aplicação no seu todo é criar um novo registo online. Para isso a aplicação disponibiliza um formulário simples, onde o utilizador deve introduzir as suas informações. Este formulário tem um sistema de validação que facilita o utilizador a introduzir as informações corretamente e caso existam erros em algum campo estes são sinalizados e surge uma mensagem informativa que descreve individualmente cada um dos erros. Este processo está ilustrado na figura 5.8.

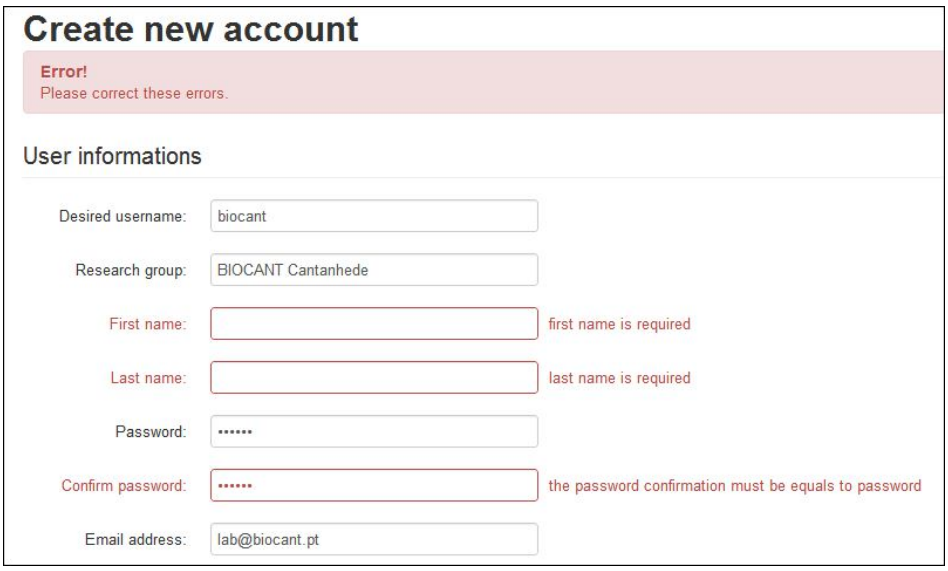


Figura 5.8: Registo de um novo utilizador na aplicação web, onde é possível visualizar o sistema de validação

5.2.3 Pesquisa e visualização de publicações

Depois de efetuado o registo online os vários utilizadores estão em condições de publicar as suas proteínas através da aplicação local. Estas proteínas ficam disponíveis para consulta na aplicação web através das suas assinaturas InterPro. Todas estas assinaturas podem ser visualizadas na aplicação com detalhe, incluindo todas as entradas de bases de dados específicas correspondentes a cada assinatura.

Para facilitar os utilizadores a encontrarem as proteínas adequadas às suas necessidades, a aplicação disponibiliza um sistema de pesquisa de proteínas. Esta pesquisa funciona com base nos identificadores InterPro.

Ao realizar uma pesquisa a aplicação mostra ao utilizador uma listagem dos resultados obtidos.

Esta listagem, representada na figura 5.9, mostra todas as proteínas obtidas no resultado da pesquisa e contém alguns detalhes como o autor da publicação, ou o número total de assinaturas existentes na proteína. Com esta informação o utilizador pode perceber de melhor forma quais os resultados com mais interesse. Em cada entrada da listagem existe um botão que permite ao utilizador visualizar com todo o detalhe o conteúdo de cada publicação.

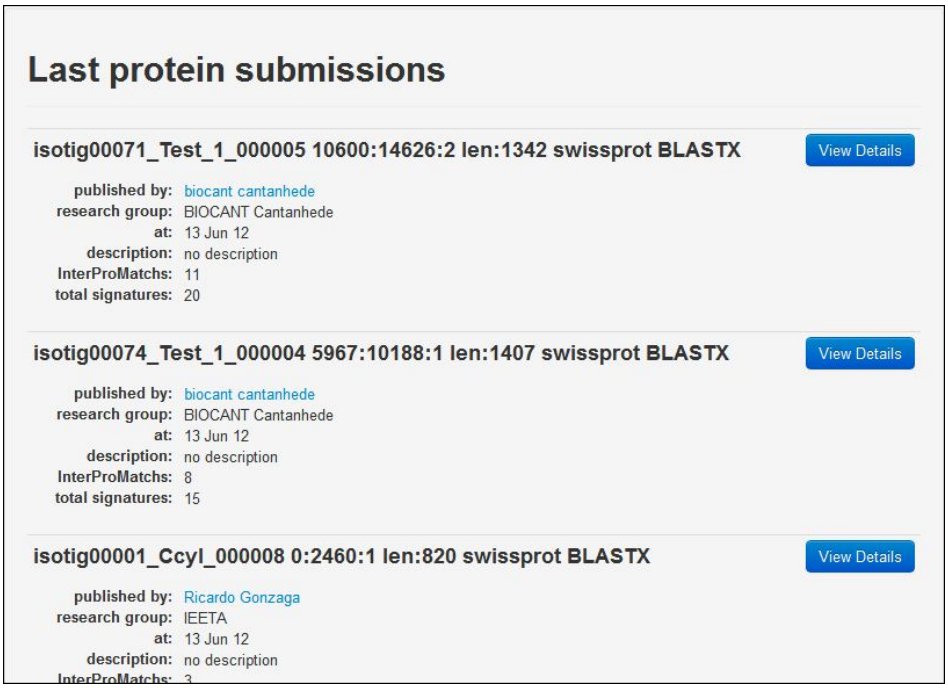


Figura 5.9: Listagem de proteínas na aplicação web, onde é possível verificar alguns detalhes da publicação como o autor da publicação ou o número de assinaturas correspondentes

Nesta visualização é possível analisar detalhadamente todas as assinaturas associadas à proteína publicada, não só as assinaturas InterPro como também todas as assinaturas que são representadas pela assinatura InterPro. Visto que esta funcionalidade implica a visualização de uma grande quantidade de informação, foi adotado um método em que o utilizador pode navegar pelos resultados individualmente. Todas as assinaturas são mostradas na forma de 'acordeão'podendo o utilizador facilmente selecionar cada uma das assinaturas para visualizar com mais pormenor a informação relativa a cada entrada. Nesta visualização é possível perceber quais as bases de dados que contêm assinaturas existentes na proteína publicada e o número de correspondências de cada assinatura. Existe também um botão que permite abrir a página do InterPro que contém toda a informação da assinatura, como está demonstrado na figura 5.10.

5.3 Sumário

Neste capítulo foram analisadas as principais funcionalidades de todo o sistema desenvolvido. Para isso começou-se por demonstrar a forma de obter e instalar a aplicação local e a forma como esta funciona. Foi seguido o processo normal de utilização do sistema começando pela criação de projetos e a importação de proteínas para posterior análise. Por fim, demonstrou-se a forma de realizar e publicar análises InterProScan utilizando a aplicação local para que fiquem disponíveis no site web.

Para a aplicação web, começou-se por ilustrar o processo de registo de novos utilizadores. Foi então explicado o *layout* principal do site e as principais funcionalidades como pesquisa e visualização de publicações na aplicação.

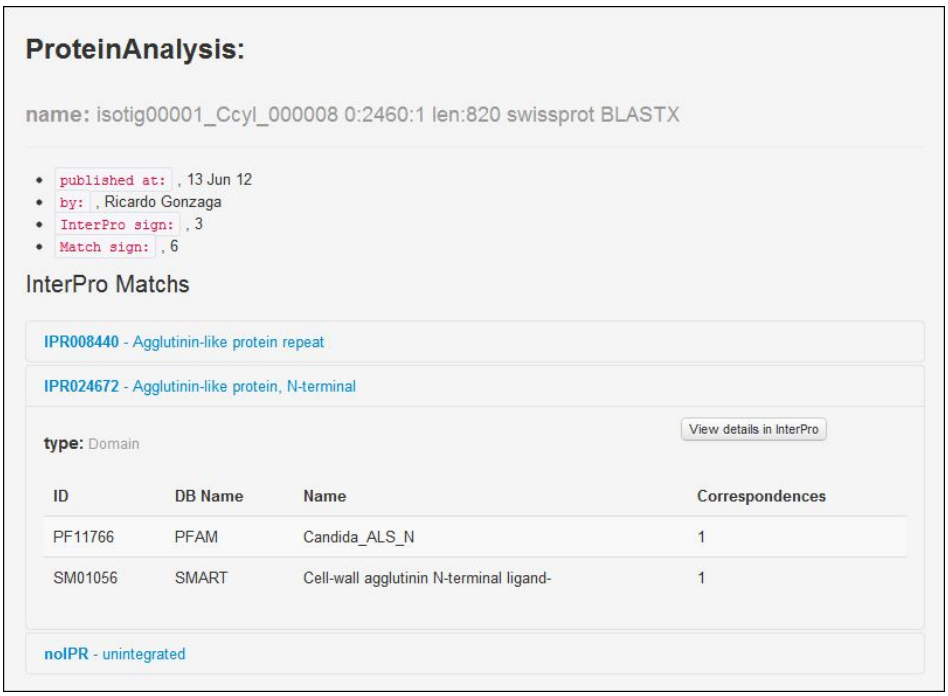


Figura 5.10: Visualização em detalhe de uma publicação na aplicação web

Capítulo 6

Conclusões

O resultado desta tese foi o planeamento e construção de um sistema aplicacional que permite a publicação e divulgação de proteínas com base nas suas assinaturas funcionais InterPro. Este sistema é constituído por duas partes distintas, uma aplicação local e um site web. A aplicação local pode ser utilizada de forma autónoma como uma biblioteca de proteínas nos centros de investigação onde são sequenciadas as proteínas. Esta ferramenta permite gerir grandes quantidades de proteínas, analisar análises InterProScan e publicar os resultados destas análises no site web.

Do outro lado, foi construído um site web que permite que potenciais interessados em proteínas com características específicas pesquisem por essas proteínas. Esta pesquisa baseia-se nas assinaturas e características funcionais das proteínas permitindo aos biólogos divulgar as suas proteínas sem revelar a sua sequência.

Analisando todo o trabalho realizado para a conclusão desta tese, existem vários pontos importantes que devem ser tidos em conta. O primeiro desafio foi compreender completamente todo o contexto biológico em que o sistema se insere, tal como a sequenciação de proteínas ou as assinaturas InterPro. Esta matéria é muito complexa e de difícil compreensão para alguém inexperiente nesta área. Todos estes aspetos foram descritos no primeiro capítulo deste documento que deve ser lido com muita atenção de forma a compreender todo o desenvolvimento do trabalho.

A utilização de um grande número de tecnologias, algumas delas bastante recentes, também foi um ponto muito positivo no desenvolvimento deste projeto. Este facto não só permitiu que a aplicação ficasse tecnologicamente avançada como foi uma mais valia ao nível de aquisição de novos conhecimentos e experiência ao nível de engenharia de software.

O grande objetivo do trabalho foi conseguido, já que as duas plataformas ficaram operacionais e os requisitos definidos foram quase todos implementados. A aplicação local cumpre a sua função permitindo a realização de análises InterProScan e a publicação destes resultados na plataforma web. Na aplicação web é possível pesquisar por identificadores InterPro assim como por assinaturas específicas ou por termos existentes nos nomes das assinaturas. No entanto, o trabalho ainda se encontra numa fase muito experimental com poucas funcionalidades.

Algumas funcionalidades planeadas no início do desenvolvimento do projeto não foram implementadas, como a existência de estatísticas de pesquisas, ou estatísticas de assinaturas mais publicadas. Outro objetivo não realizado foi o desenvolvimento de tarefas automáticas como, por exemplo, uma pesquisa automática que informa os utilizadores quando existirem proteínas dentro de alguns parâmetros definidos. Estes objetivos não foram conseguidos não por dificuldades técnicas, mas sim devido à duração limitada do projeto.

O trabalho futuro passa por implementar um sistema de pesquisa mais completo e que permita

especificar de forma mais detalhada o tipo de proteína pretendida, e desenvolver o sistema de estatísticas de forma a perceber quais os motivos mais procurados, tanto por utilizador como globalmente, ou até ao nível de proteínas publicadas. Seria também muito interessante que se integrassem mais ferramentas de análise de proteínas na aplicação local, como alinhamento de proteínas, ou pesquisa automática de informações nos repositórios disponíveis.

Bibliografia

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Molecular Biology of the Cell 5th Edition, 2008.

[2] C Branden and J Tooze. *Introduction to Protein Structure*, volume 85 of *Introduction to Protein Structure Series*. Garland Publishing, 1999.

[3] Leonard J Foster and Matthias Mann. Protein Identification and Sequencing by Mass Spectrometry. *Sort*, pages 363–369, 2006.

[4] P Edman. Method for determination of the amino acid sequence in peptides. *Acta Chemica Scandinavica*, 4(4):283–293, 1950.

[5] Elon Portugaly, Nathan Linial, and Michal Linial. EVEREST: a collection of evolutionary conserved protein domains. *Nucleic Acids Research*, 35(Database issue):D241–D246, 2007.

[6] Loredana Lo Conte, Bart Ailey, Tim J P Hubbard, Steven E Brenner, Alexey G Murzin, and Cyrus Chothia. SCOP: a structural classification of proteins database. *Nucleic Acids Research*, 25(1):236–239, 2000.

[7] C A Orengo, F M Pearl, J E Bray, A E Todd, A C Martin, L Lo Conte, and J M Thornton. The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Research*, 27(1):275–279, 1999.

[8] R Apweiler, T K Attwood, A Bairoch, A Bateman, E Birney, M Biswas, P Bucher, L Cerutti, F Corpet, M D R Croning, R Durbin, L Falquet, W Fleischmann, J Gouzy, H Hermjakob, N Hulo, I Jonassen, D Kahn, A Kanapin, Y Karavidopoulou, R Lopez, B Marx, N J Mulder, T M Oinn, M Pagni, F Servant, C J A Sigrist, and E M Zdobnov. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, 29(1):37–40, 2001.

[9] Sarah Hunter, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, Robert D Finn, Julian Gough, Daniel Haft, Nicolas Hulo, Daniel Kahn, Elizabeth Kelly, Aurélie Laugraud, Ivica Letunic, David Lonsdale, Rodrigo Lopez, Martin Madera, John Maslen, Craig McAnulla, Jennifer McDowall, Jaina Mistry, Alex Mitchell, Nicola Mulder, Darren Natale, Christine Orengo, Antony F Quinn, Jeremy D Selengut, Christian J a Sigrist, Manjula Thimma, Paul D Thomas, Franck Valentin, Derek Wilson, Cathy H Wu, and Corin Yeats. InterPro: the integrative protein signature database. *Nucleic acids research*, 37(Database issue):D211–5, January 2009.

[10] Robert D Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, and Alex Bateman. The Pfam protein families database. *Nucleic Acids Research*, 38(Database issue):D211–D222, 2010.

[11] T K Attwood, M J Blythe, D R Flower, A Gaulton, J E Mabey, N Maudling, L McGregor, A L Mitchell, G Moulton, K Paine, and P Scordis. PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Research*, 30(1):239–241, 2002.

[12] Nicolas Hulo, Amos Bairoch, Virginie Bulliard, Lorenzo Cerutti, Edouard De Castro, Petra S Langendijk-Genevaux, Marco Pagni, and Christian J A Sigrist. The PROSITE database. *Nucleic Acids Research*, 34(Database issue):D227–D230, 2006.

[13] Ivica Letunic, Tobias Doerks, and Peer Bork. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Research*, 40(November 2011):1–4, 2011.

[14] F Corpet, J Gouzy, and D Kahn. The ProDom database of protein domain families. *Nucleic Acids Research*, 26(1):323–326, 1998.

[15] Cathy H Wu, Anastasia Nikolskaya, Hongzhan Huang, Lai-Su L Yeh, Darren A Natale, C R Vinayaka, Zhang-Zhi Hu, Raja Mazumder, Sandeep Kumar, Panagiotis Kourtesis, Robert S Ledley, Baris E Suzek, Leslie Arminski, Yongxing Chen, Jian Zhang, Jorge Louie Cardenas, Sehee Chung, Jorge Castro-Alvear, Georgi Dinkov, and Winona C Barker. PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Research*, 32(Database issue):D112–D114, 2004.

[16] Martin Madera, Christine Vogel, Sarah K Kummerfeld, Cyrus Chothia, and Julian Gough. The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Research*, 32(Database issue):D235–D239, 2004.

[17] Huaiyu Mi, Betty Lazareva-Ulitsky, Rozina Loo, Anish Kejariwal, Jody Vandergriff, Steven Rabkin, Nan Guo, Anushya Muruganujan, Olivier Doremieux, Michael J Campbell, Hiroaki Kitano, and Paul D Thomas. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, 33(Database Issue):D284–D288, 2005.

[18] Jonathan Lees, Corin Yeats, Oliver Redfern, Andrew Clegg, and Christine Orengo. Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Research*, 38(Database issue):D296–D300, 2010.

[19] Daniel H Haft, Jeremy D Selengut, and Owen White. The TIGRFAMs database of protein families. *Nucleic Acids Research*, 31(1):371–373, 2003.

[20] Tania Lima, Andrea H Auchincloss, Elisabeth Coudert, Guillaume Keller, Karine Michoud, Catherine Rivoire, Virginie Bulliard, Edouard De Castro, Corinne Lachaize, Delphine Baratin, Isabelle Phan, Lydie Bougueleret, and Amos Bairoch. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Research*, 37(Database issue):D471–8, 2009.

[21] The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research*, 38(Database issue):D142–D148, 2010.

[22] E Quevillon, V Silventoinen, S Pillai, N Harte, N Mulder, R Apweiler, and R Lopez. InterProScan: protein domains identifier. *Nucleic acids research*, 33(Web Server issue):W116–20, July 2005.

[23] Zdobnov E.M. and Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro., 2001.

[24] Reza Razeghifard, Brett B Wallace, Ron J Pace, and Tom Wydrzynski. Creating functional artificial proteins. *Current protein peptide science*, 8(1):3–18, 2007.

[25] Ryan J Kelly, David E Vincent, and Iddo Friedberg. IPRStats: visualization of the functional potential of an InterProScan run. *BMC Bioinformatics*, 11(Suppl 12):S13, 2010.

[26] Aijazuddin Syed and Chris Upton. Java GUI for InterProScan (JIPS): A tool to help process multiple InterProScans and perform ortholog analysis. *BMC Bioinformatics*, 7:462, 2006.

[27] Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Tim D Williams, Shivashankar H Nagaraj, María José Nueda, Montserrat Robles, Manuel Talón, Joaquín Dopazo, and Ana Conesa. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 36(10):3420–3435, 2008.

[28] R C G Holland, T A Down, M Pocock, A Prlić, D Huen, K James, S Foisy, A Dräger, A Yates, M Heuer, and M J Schreiber. BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096–2097, 2008.