
Universidade de Aveiro
2010
Departamento de Electrónica, Telecomunicações e Informática

Joel Perdiz Arrais

Sistemas de informação para DNA *microarrays*

Universidade de Aveiro
2010
Departamento de Electrónica, Telecomunicações e Informática

Joel Perdiz Arrais

Sistemas de Informação para DNA *microarrays*

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

Apoio financeiro da Fundação para a Ciência e a Tecnologia

À minha família, em especial, à Joana.

o júri

presidente

Professora Doutora Ana Maria Vieira Silva Viana Cavaleiro
Professora Catedrática do Departamento de Química da Universidade de Aveiro

Professor Doutor Francisco José Moreira Couto
Professor Auxiliar do Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa

Professor Doutor Miguel Pereira Rocha
Professor Auxiliar do Departamento de Informática da Escola de Engenharia da Universidade do Minho

Professor Doutor Armando José Formoso Pinho
Professor Associado com Agregação do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Professor Doutor José Luís Guimarães Oliveira
Professor Associado do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

agradecimentos

Como todos os trabalhos, esta dissertação deve muito ao apoio prestado por pessoas e instituições a quem não posso deixar de reconhecer publicamente.

O meu primeiro obrigado dirige-se ao Professor Doutor José Luís Oliveira, meu orientador de tese, por ter acreditado em mim e no meu trabalho e me ter oferecido oportunidades únicas de desenvolver competências numa área até então desconhecida. Agradeço ainda a sua disponibilidade para discussões, sapiência na proposta de soluções e ajuda no desenvolvimento de espírito crítico, porque apontou caminhos mas deixou sempre a liberdade de os aceitar ou rejeitar.

Uma palavra de reconhecimento endereço-a ao Professor Doutor Manuel Santos por me ter disponibilizado o seu laboratório e, acima de tudo, o seu conhecimento e o seu tempo.

Devo mencionar a preciosa colaboração da Doutora Laura Carreto pelo constante apoio, pelas acuradas sugestões e pelas oportunidades de aprendizagem.

Merecem o meu apreço a Doutora Helen Parkinson do EBI e o Doutor Kevin Robertson do ScGTI pela oportunidade que me ofereceram e pela generosidade com que me acolheram.

Determinantes foram também todos os colegas que ao longo destes cinco anos passaram pelo laboratório, sobretudo àqueles com quem tive o privilégio de trabalhar mais directamente.

Estou ainda reconhecido a duas instituições, a Fundação para a Ciência e a Tecnologia e a Universidade de Aveiro, pelo apoio financeiro e logístico.

Não posso deixar de lembrar, com muita amizade, o Quim e o Carlos por terem sido os meus primeiros professores de informática. Ao Quim que acompanhou todo o meu percurso académico agradeço também os pontuais, mas sábios, conselhos que, nos momentos mais turbulentos e decisivos, ajudaram na escolha do melhor caminho. Gostaria ainda de agradecer à Né por me ter ensinado a gostar de matemática.

A todos os meus amigos que sabem o quanto lhes estou grato por tolerarem as minhas ausências sem deixarem de reclamar a minha presença.

Por último, devo um especial reconhecimento à minha família. Agradeço, assim,

à Fernanda e ao António pela infinita disponibilidade e compreensão;

à minha irmã, Inês, e ao meu *irmão*, João, pelo constante apoio, paciência e estímulo;

à minha mãe e ao meu pai pelo apoio, material e emocional, incondicional e inabalável;

à Joana… simplesmente por tudo.

palavras-chave Sistemas de informação, integração de dados, LIMS, DNA *microarrays*

resumo

O projecto de sequenciação do genoma humano veio abrir caminho para o surgimento de novas áreas transdisciplinares de investigação, como a biologia computacional, a bioinformática e a bioestatística. Um dos resultados emergentes desde advento foi a tecnologia de DNA *microarrays*, que permite o estudo do perfil da expressão de milhares de genes, quando sujeitos a perturbações externas. Apesar de ser uma tecnologia relativamente consolidada, continua a apresentar um conjunto vasto de desafios, nomeadamente do ponto de vista computacional e dos sistemas de informação. São exemplos a optimização dos procedimentos de tratamento de dados bem como o desenvolvimento de metodologias de interpretação semi-automática dos resultados.

O principal objectivo deste trabalho consistiu em explorar novas soluções técnicas para agilizar os procedimentos de armazenamento, partilha e análise de dados de experiências de *microarrays*. Com esta finalidade, realizou-se uma análise de requisitos associados às principais etapas da execução de uma experiência, tendo sido identificados os principais défices, propostas estratégias de melhoramento e apresentadas novas soluções.

Ao nível da gestão de dados laboratoriais, é proposto um LIMS (*Laboratory Information Management System*) que possibilita a gestão de todos os dados gerados e dos procedimentos realizados. Este sistema integra ainda uma solução que permite a partilha de experiências, de forma a promover a participação colaborativa de vários investigadores num mesmo projecto, mesmo usando LIMS distintos.

No contexto da análise de dados, é apresentado um modelo que facilita a integração de algoritmos de processamento e de análise de experiências no sistema desenvolvido. Por fim, é proposta uma solução para facilitar a interpretação biológica de um conjunto de genes diferencialmente expressos, através de ferramentas que integram informação existente em diversas bases de dados biomédicas.

keywords Information systems, data integration, LIMS, DNA microarrays

abstract

The sequencing of the human genome paved the way for the emergence of new transdisciplinary research areas, such as computational biology, bioinformatics and biostatistics. One example of such is the advent of DNA microarray technology, which allows the study of the expression of thousands of genes when subjected to an external disturbance. Despite being a well-established technology, it continues to present a wide range of challenges, particularly in terms of computing and information systems. Examples include the optimization of procedures for processing data as well as the development of methodologies for semi-automated interpretation of results.

The main objective of this study was to explore new technical solutions to streamline the procedures for storing, sharing and analyzing the data from microarray experiments. To this end, it was performed an analysis of the key steps from the experiment, having been identified the major deficits, proposed strategies for improving and presented new solutions.

Regarding the management of laboratory data we propose a LIMS (Laboratory Information Management System) that allows the storage of all data generated and procedures performed in the laboratory. This system also includes a solution that enables the sharing of experiments in order to promote collaborative participation of several researchers in the same project, even using different LIMS.

In the context of data analysis, it is presented a model that allows the simplified integration of processing and analysis algorithms in the developed system. Finally, it is proposed a solution to facilitate the biological interpretation of a set of differentially expressed genes, using tools that integrate information from several public biomedical databases.

Índice Geral

Índice de Figuras

Índice de Tabelas

Índice de Acrónimos

AADM	*Affymetrix Analysis Data Model*
ADF	*Array Description File*
AJAX	*Asynchronous Javascript And XML*
BASE	*BioArray Software Environment*
cDNA	*complementary DNA*
CGH	*Comparative Genomic Hybridization*
ChIP	*Chromatin Immunoprecipitation*
Cibex	*Center for Information Biology gene Expression database*
DAG	*Directed Acyclic Graph*
DAS	*Distributed Annotation System*
DNA	*Deoxyribonucleic Acid*
EBI	*European Bioinformatics Institute*
FDR	*False Discovery Rate*
FTP	*File Transfer Protocol*
FWER	*Familywise Error Rate*
Gens	*Genomic Name Server*
GEO	*Gene Expression OmniBus*
GO	*Gene Ontology*
GOBO	*Global Open Biology Ontologies*
HTML	*HyperText Markup Language*
IDF	*Investigation Design File*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
LAD	*Longhorn Array Database*
LIMS	*Laboratory Information Management System*

LOWESS	*Locally Weighted Linear Regression*
MADAM	*Microarray Data Manager*
MAGE	*Microarray and Gene Expression*
MGED	*Microarray Gene Expression Database Group*
MIAME	*Minimum Information about a Microarrays Experiment*
Mind	*Microarray Information Database*
MINIML	*MIAME Notation in Markup Language*
mRNA	*messenger RNA*
NCBI	*National Center for Biotechnology Information*
NCGR	*National Center for Genome Resources*
OBO	*Open Biomedical Ontologies*
OMIM	*Online Mendelian Inheritance in Man*
PCR	*Polymerase Chain Reaction*
PDB	*Protein Data Bank*
QUEXT	*Query Expansion Tool*
REST	*Representative State Transfer*
RNA	*RiboNucleic Acid*
SAM	*Significance Analysis for Microarrays*
SCE	*Saccharomyces Cerevisiae*
ScGTI	*Scottish Centre for Genomic Technology and Informatics*
SDRF	*Sample and Data Relationship Format*
SGBD	Sistema de Gestão de Base de Dados
SNP	*Single Nucleotide Polymorphism*
SOAP	*Simple Object Access Protocol*
SOFT	*Simple Omnibus Format*
UMLS	*Unified Medical Language System*
URL	*Uniform Resource Locator*
XML	*eXtensible Markup Language*

Capítulo 1

1 Introdução

À imagem do que sucedeu noutras áreas em décadas anteriores, as ciências da vida encontram-se actualmente no centro de uma revolução computacional. De facto, avanços significativos têm sido obtidos principalmente devido à possibilidade de aplicar ferramentas computacionais no armazenamento e processamento dos dados. Um dos mais notáveis projectos dos últimos anos resultou na sequenciação do genoma humano, que ficou concluído antes da data prevista e sem ultrapassar o orçamento estabelecido precisamente devido à evolução na capacidade de processamento. Apesar de em 2001 a finalização deste projecto de 10 anos ter constituído um importante marco, a tecnologia de sequenciação disponível em 2009 possibilita a realização da mesma tarefa em algumas semanas. O desafio actualmente proposto pela *X Prize Foundation* consiste em sequenciar 100 genomas humanos em 10 dias [1-3].

O resultado do projecto de sequenciação oferece uma vista estática do funcionamento das células humanas, na medida em que mostra o que estas são potencialmente capazes de produzir. No entanto, estas fazem parte de um sistema bastante dinâmico em que cada célula assume uma função específica no organismo e em que factores externos influenciam o seu funcionamento. Com efeito, o homem possui mais de 20.000 genes dos quais apenas uma pequena fracção se encontra activa em cada instante [4]. A compreensão dos mecanismos subjacentes é no entanto dificultada por vários factores, incluindo o facto de os genes não operarem de forma isolada na célula mas sim numa complexa rede de interacções que apenas recentemente se começou a desvendar. O estudo destas interacções é essencial no estabelecimento de relações entre genótipos e fenótipos o que se torna especialmente útil na compreensão dos mecanismos de funcionamento de doenças genéticas, oferecendo potencialidades no desenvolvimento de novos métodos de diagnóstico e tratamento.

Os *microarrays* de DNA são actualmente a tecnologia mais relevante no estudo da dinâmica do funcionamento das células. Através da medição do mRNA (*messenger*

RiboNucleic Acid) presente na célula é possível obter uma estimativa dos níveis de expressão do gene correspondente [5]. Interessante é o facto de esta tecnologia ser escalável a dezenas de milhar de genes, possibilitando a monitorização de todo o genoma humano com um único teste. Apesar de já existirem alguns resultados concretos, muitos dos objectivos continuam ainda por cumprir, nomeadamente a sua aplicação em ambientes clínicos [6]. Antes que tal seja possível, é necessária uma maior estabilização da tecnologia, assim como uma normalização e uma sistematização dos procedimentos de tratamento de dados. É ainda necessário o desenvolvimento de novas metodologias de interpretação dos resultados, que, frequentemente, implicam a comparação dos dados locais com informação de vários domínios disponíveis em bases de dados públicas.

A questão que se coloca é como podem as tecnologias de sistemas de informação ser usadas para agilizar os actuais procedimentos de armazenamento, partilha e análise de dados de experiências de *microarrays*. Esta questão é a principal motivação para o trabalho desenvolvido e apresentado neste documento.

1.1 Enquadramento

A tecnologia dos *microarrays* possibilita o estudo dos perfis de expressão de milhares de genes em simultâneo. Esta tecnologia já se encontra bastante difundida ao nível da investigação, existindo inúmeros estudos publicados com base em dados de *microarrays* [5, 7].

As oportunidades criadas pelos *microarrays* apresentam, no entanto, algumas ameaças, tendo em consideração as dificuldades em gerir de forma eficiente os dados gerados. Um *microarray* típico contém cerca de 20.000 elementos, sendo necessário, para cada um, armazenar a sequência do gene correspondente. Cada estudo pode incluir mais de uma centena de *microarrays,* sem considerar sequer a descrição das amostras e dos protocolos usados. Esta elevada quantidade de dados criou novos desafios nos procedimentos de armazenamento e análise. Em muitos casos, a capacidade de lidar com os dados de forma eficiente é o factor condicionante, quando o investigador pretende obter resposta às suas questões. Os sistemas de gestão de dados laboratoriais para *microarrays* surgem, neste contexto, como uma solução para esta necessidade. O objectivo destes sistemas é o de servirem de repositório de todos os procedimentos aplicados, assim como de todos os dados obtidos [8].

O desenvolvimento destes sistemas tem beneficiado da criação de normas e ontologias específicas para *microarrays*. As normas tem como principal objectivo indicar que dados devem ser registados, como devem ser armazenados e, não menos importante, como podem ser partilhados. As ontologias, por sua vez, permitem capturar de forma estruturada a complexidade semântica de um determinado domínio.

A capacidade de extrair conhecimento válido de um estudo com *microarrays* é considerada, por vários investigadores, como a questão mais delicada de todo o processo. A existência de vários factores experimentais que podem influir na qualidade dos dados, a intrínseca relação existente entre a questão biológica a endereçar e o desenho experimental assim como as implicações que estes têm na escolha dos métodos de análise a usar são os principais desafios apontados. Deste modo, apesar de já existirem ferramentas que possibilitam a análise de dados de *microarrays,* muitas tarefas e decisões continuam a recair sobre o investigador.

A interpretação do resultado da análise de um estudo de *microarrays* nem sempre é directa. Isto acontece porque é necessário compreender as relações existentes entre o conjunto de genes obtidos e os processos biológicos em que estes estão envolvidos. Um método tipicamente aplicado consiste na comparação dos resultados locais com entidades biológicas existentes em bases de dados públicas. O acesso a estes dados é, no entanto, dificultado por vários factores, como a sua heterogeneidade, dimensão, fragmentação e dispersão.

1.2 Objectivos

Este trabalho tem como objectivo principal propor um modelo de integração de ferramentas que facilite a gestão de informação de um laboratório de *microarrays.* Pretende-se um sistema de informação que possa ser usado no armazenamento, partilha e análise de dados de experiências de *microarrays* (Figura 1.1).

O primeiro passo consiste, portanto, em identificar todas as etapas de execução de uma experiência. Foi dado especial ênfase aos passos potencialmente geradores de dados, remetendo para segundo plano os exclusivamente associados ao trabalho laboratorial. Para cada um dos elementos foram identificados os principais défices, tendo sido estudadas estratégias de melhoramento.

Nos últimos anos, várias ferramentas foram apresentadas com o intuito de possibilitarem o armazenamento dos dados de *microarrays*. Estes sistemas, designados de LIMS (*Laboratory Information Management System*), possibilitam o armazenamento dos dados. O sistema proposto pretende colmatar várias das lacunas encontradas nos pré-existentes, assim como responder aos novos requisitos identificados. Pretende-se, ainda, fazer uso das normas e ontologias.

A partilha de dados é essencial ao processo de investigação, podendo esta ocorrer a vários níveis. Até ao momento, esta partilha tem tendencialmente sido realizada através da submissão em repositórios públicos. Este facto está associado à obrigatoriedade de disponibilizar os resultados de forma a proceder à sua publicação. E é também um dos propósitos do presente trabalho colaborar com esses repositórios, de forma a assegurar a compatibilidade do sistema com os mesmos. Paralelamente, são, ainda, apresentadas

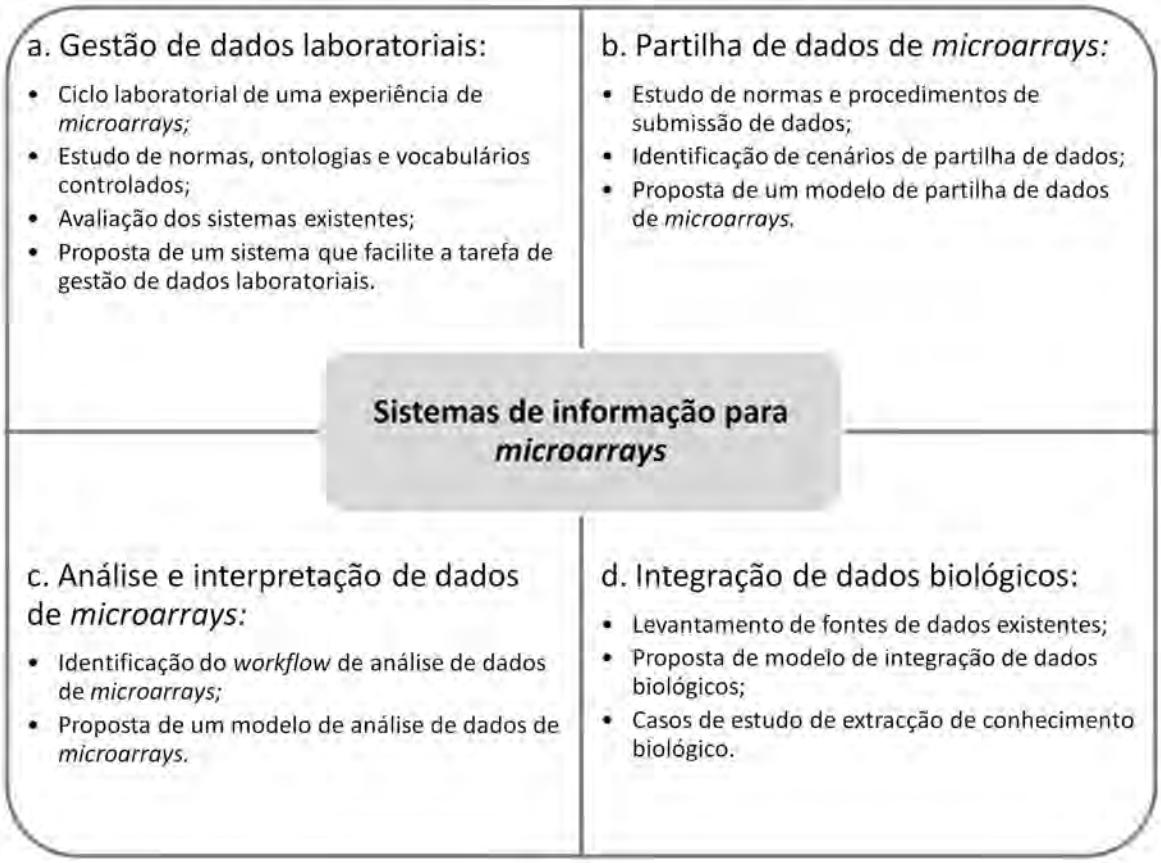


Figura 1.1: Objectivos gerais e específicos da tese.

soluções que possibilitam a partilha de estudos antes mesmo da sua conclusão, de forma a promover a participação colaborativa de vários investigadores num mesmo projecto.

Após o armazenamento dos dados, é essencial proceder à sua análise. Esta pode ser realizada a vários níveis. Existem actualmente diversas ferramentas que já possibilitam a execução da análise dos dados. Não é meta deste trabalho contribuir com novos métodos de análise, mas, antes, oferecer um *workflow* que integre os principais métodos existentes.

Outro objectivo específico deste trabalho consiste na extracção de conhecimento a partir dos resultados obtidos. É, neste sentido, proposta e implementada uma plataforma que possibilita a integração de dados biológicos. São ainda apresentados exemplos concretos do uso desta plataforma na extracção de conhecimento de experiências de *microarrays*.

1.3 Organização do documento

Este documento encontra-se dividido em sete capítulos (Figura 1.2). O primeiro e o último são essencialmente de síntese do trabalho realizado, sendo no segundo introduzidos os principais conceitos biológicos usados no documento. Os restantes capítulos, do terceiro ao sexto, expõem as principais contribuições deste trabalho. Sendo o propósito do trabalho

endereçar questões associadas com o fluxo de dados de *microarrays*, tal como a Figura 1.2 ilustra, existe um paralelismo evidente entre este fluxo e a organização do documento.

No capítulo dois, é apresentada uma introdução dos conceitos de biologia molecular e da tecnologia dos *microarrays*, sendo dado especial ênfase aos aspectos ligados à expressão génica, aos mecanismos de regulação da expressão e à tecnologia de *spotted microarrays*. São discutidas as principais motivações para o uso desta tecnologia e exposto o funcionamento dos principais tipos de *microarrays de DNA*, as suas limitações, assim como as suas principais aplicações presentes e futuras.

No capítulo três, propõe-se uma plataforma de gestão de dados laboratoriais. Em primeiro lugar, faz-se uma análise do procedimento de realização de experiências de *microarrays*, do fluxo de informação gerado e das normas, das ontologias e dos sistemas de gestão de dados existentes. Tendo em consideração as lacunas identificadas e os resultados do estudo realizado, é proposto o sistema Mind para gestão de dados de *microarrays*.

O capítulo quatro centra-se na definição de um modelo de partilha de dados de *microarrays*. É realizado um levantamento dos principais repositórios de armazenamento centralizado de dados e dos principais processos de normalização existentes. Estes têm como principais objectivos evitar a dispersão dos resultados por vários servidores, assegurar a correcta formatação dos dados e garantir a disponibilidade e perpetuidade das experiências. É apresentada a arquitectura de um sistema distribuído que possibilita que diferentes investigadores participem num mesmo projecto mesmo utilizando sistemas distintos.

O capítulo cinco é dedicado ao estudo das diferentes fases da análise de uma experiência de *microarrays*. São analisados os procedimentos de pré-processamento e de normalização de dados, de selecção de genes diferencialmente expressos e de detecção de padrões, assim como a análise funcional. É ainda proposto um conjunto de ferramentas sobre o sistema Mind que privilegiam a flexibilidade na adição de novos algoritmos ou de novos métodos de visualização.

No capítulo seis, são endereçadas as ferramentas que dão resposta ao problema de interpretar biologicamente um conjunto de genes diferencialmente expressos. Como suporte a estas ferramentas é ainda desenvolvida uma plataforma de integração de dados biológicos que facilita a tarefa de acesso a fontes de dados dispersas. Pese o seu propósito inicial pretende-se cumulativamente que esta plataforma seja suficientemente genérica e extensível de modo a poder ser utilizada noutros domínios.

Por fim, o capítulo sete é dedicado às conclusões e aos resultados obtidos, sendo ainda levantadas questões que incentivam ao desenvolvimento e ao aprofundamento do trabalho aqui apresentado, no futuro.

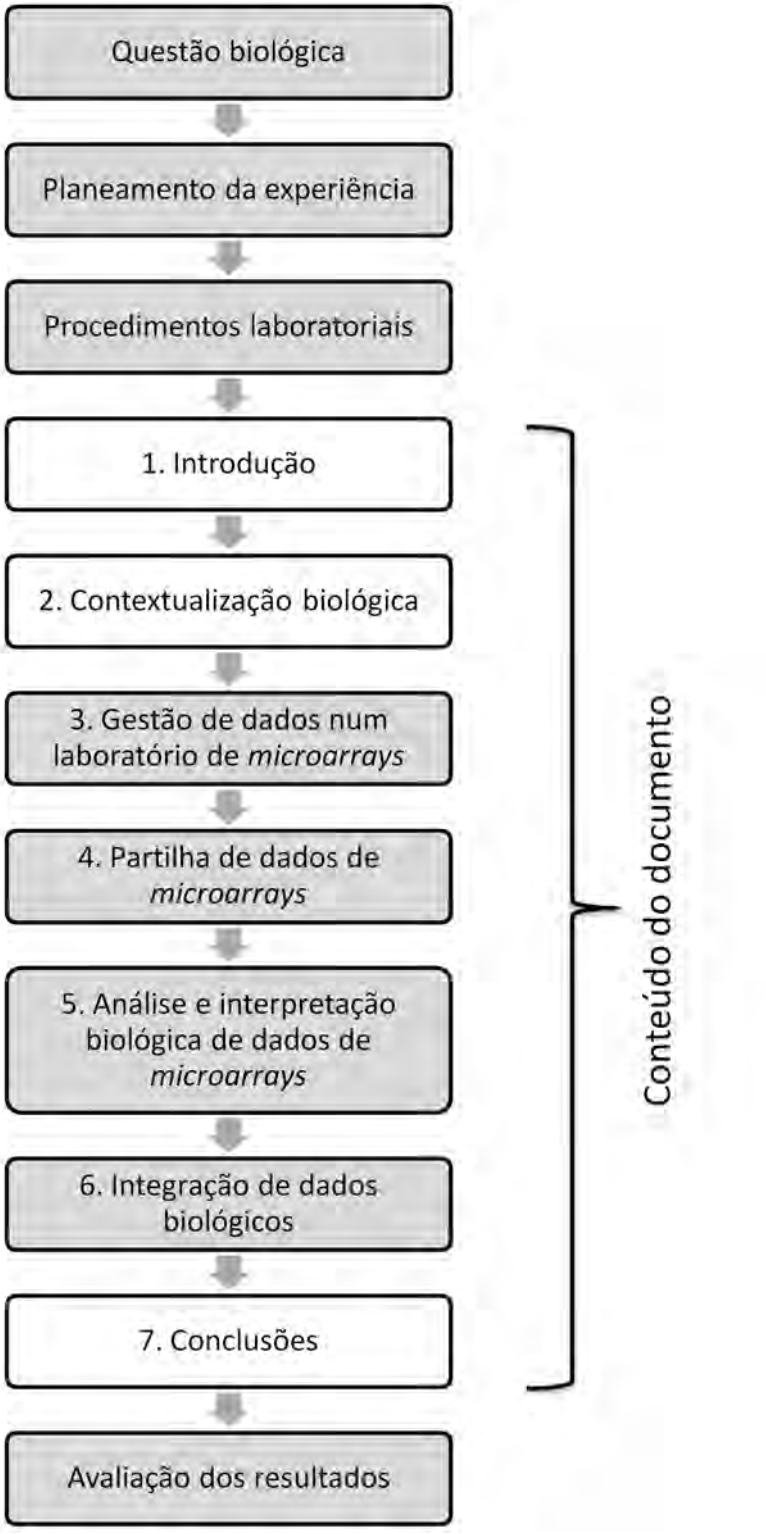


Figura 1.2: Paralelismo entre o fluxo de uma experiência de *microarrays* e a organização do documento. A sombreado encontra-se o ciclo de uma experiência, definindo a chaveta a organização do documento.

Capítulo 2

2 Biologia molecular e *microarrays* de DNA

Apesar da elevada diversidade de seres vivos na natureza, existe um elemento que é comum a todos: a célula. É a célula, enquanto unidade funcional e estrutural base de todos os organismos, a responsável pelo desempenho das suas funções vitais. E, ainda que, em organismos multicelulares as células colaborarem para um objectivo comum, cada uma pode ser considerada autónoma, visto possuir toda a maquinaria e informação genética necessária ao seu funcionamento. Esta informação encontra-se armazenada no núcleo da célula sob a forma de genes. Uma só célula possui milhares de genes, estando cada um associado a processos biológicos específicos. No entanto, esta relação não é necessariamente unívoca, pois, em vários casos, os genes participam em complexas redes de interacções. Os métodos tradicionais de biologia molecular tratam um gene por experiência, limitando, por isso, a capacidade de compreensão do completo funcionamento destas redes de interacções [5, 9].

A tecnologia de *microarrays* de DNA pretende ser uma solução para este problema. Esta permite, com um único teste, a monitorização dos níveis de expressão génica de todo o genoma num dado instante, facilitando a compreensão da interacção de milhares de genes [5, 10]. Esta tecnologia tem sido usada com sucesso como ferramenta laboratorial, contudo, existem várias expectativas quanto à sua aplicação na área médica. Um *microarray* pode ser utilizado como auxiliar na escolha do melhor tratamento para uma determinada doença, na detecção de agentes patogénicos, assim como para detectar genes que possam estar relacionados com uma doença específica, introduzindo deste modo a medicina preditiva [11, 12].

Este capítulo tem por objectivo introduzir os conceitos base de biologia molecular necessários para a compreensão deste documento. Não é propósito apresentar um estudo exaustivo, sendo fornecidas referências para informação mais detalhada. Propositadamente, é dado maior ênfase aos aspectos ligados à expressão génica, aos mecanismos de regulação da expressão e à tecnologia dos *microarrays*. É exposto o

funcionamento de um *microarray* de DNA, as principais tecnologias disponíveis, as suas limitações, assim como as principais aplicações presentes e futuras.

2.1 A célula como constituinte básico

A célula é a unidade estrutural e funcional base de todos os organismos vivos (com excepção dos vírus). Existe uma elevada diversidade no tipo de células existentes, no que se refere à sua morfologia, dimensão, agilidade e habitat. Por exemplo, o tamanho típico de um eritrócito é de 9 μm, enquanto a célula do ovo não fecundado de uma avestruz pode chegar aos 120 mm (aproximadamente 13.000 vezes maior!). Outro exemplo da elevada diversidade existente é encontrado nas células procariontes, que tanto podem ser encontradas no oceano, a mais de dez quilómetros de profundidade, como na atmosfera, a 60 quilómetros de altitude [9, 13].

Apesar das diferenças evidentes, todas as células possuem uma composição química comum e contêm a informação genética necessária para o desempenho de todas as suas funções. No que se refere à sua organização, é possível encontrar organismos unicelulares, como as bactérias, constituídas por uma única célula, ou organismos multicelulares, como os humanos, constituídos por um número elevado de células. Em ambos os casos, cada célula, por si, possui uma elevada complexidade funcional e estrutural.

Existem dois tipos de células: procariontes e eucariontes (Figura 2.1). As procariontes são as estruturalmente mais simples e, em termos evolutivos, foram as primeiras a aparecer. São caracterizadas pela ausência de núcleo e de muitos dos organelos presentes nas células eucariontes. Pese existirem dois tipos de células procariontes, bactéria e *archaea*, o primeiro é a forma mais comum e a mais estudada. Numa célula procarionte típica, existem três regiões distintas: 1) os apêndices, designados de flagelos; 2) o envelope celular, que consiste numa cápsula, parede celular e membrana plasmática; e 3) a região do citoplasma, que contém o genoma celular, os ribossomas e os vários tipos de inclusões.

As células eucariontes incluem fungos, animais e plantas, assim como alguns seres unicelulares. Embora estas possuam uma dimensão dez vezes superior quando comparadas com as procariontes, a maior e mais significativa diferença entre ambas consiste no facto das eucariontes possuírem compartimentos membranares, nos quais as diferentes actividades metabólicas ocorrem. Um destes compartimentos, o núcleo, é responsável por armazenar os cromossomas e é também onde ocorrem os processos de replicação e a transcrição do DNA.

O surgimento das células eucariontes pode ser considerado como um avanço significativo na evolução da vida. Pois com quanto os seres eucariontes usem o mesmo código genético e processos metabólicos do que os procariontes, o seu superior nível de complexidade foi essencial ao desenvolvimento de organismos multicelulares.

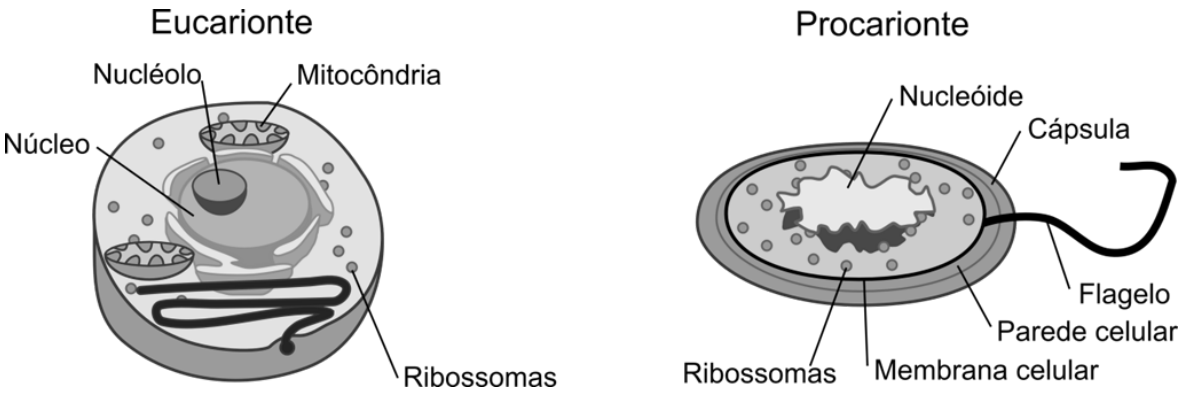


Figura 2.1: Célula eucarionte e célula procarionte. Esta figura ilustra uma célula humana (eucarionte) e uma bactéria típica (procarionte). No esquema da esquerda, encontram-se em evidência as principais estruturas internas das células eucariontes, nomeadamente a mitocôndria, os ribossomas o nucleólo e o núcleo. O esquema da direita apresenta a estrutura típica de uma bactéria, incluindo a cápsula, a parede celular, a membrana celular, o flagelo, os ribossomas e o nucleóide. (Adaptação da imagem celltypes.svg, obtida de *wikimedia commons*)

2.2 Genoma

O genoma de um organismo consiste na sua sequência genética completa. No caso dos humanos, esta corresponde à sequência dos 23 cromossomas lineares que se encontram no núcleo da célula. Note-se que, sendo o ser humano diploide, visto cada célula somática possuir duas cópias de cada cromossoma, considera-se que este possui dois genomas completos. Comparativamente, as bactérias apresentam uma organização do genoma bastante diferente. Não possuindo um verdadeiro núcleo, o genoma compreende tipicamente, um único cromossoma circular, localizado no nucleóide. Em ambos os casos, porém, ao genoma de um organismo acresce, ainda, informação de outros elementos genéticos não cromossomais, tais como, a mitocôndria, os vírus, os plasmídeos e os transposões [9, 13].

Ao nível químico, o genoma consiste numa cadeia de quatro tipos de subunidades designadas nucleótidos: adenina, citosina, guanina e timina. Por sua vez, cada nucleótido é constituído por três componentes: grupo fosfato, pentose (desoxirribose para DNA e ribose para RNA) e bases azotadas. Uma vez que o grupo fosfato e a pentose se mantêm constantes a todos os nucleótidos é a base azotada que dá o nome ao nucleótido. A cadeia de nucleótidos é, então, formada através da ligação existente entre cada grupo fosfato e a pentose do nucleótido anterior.

A forma mais comum de organização do DNA na célula é sob a forma de uma estrutura em hélice dupla, em que duas sequências se emparelham formando uma espiral. Nesta estrutura existem regras de emparelhamento que especificam que o nucleótido guanina (G) emparelha unicamente com o nucleótido citosina (C) e que o nucleótido adenina (A) emparelha com o nucleótido timina (T). O emparelhamento entre guanina a citosina formam três pontes de hidrogénio, enquanto o emparelhamento entre adenina e timina

forma apenas duas pontes de hidrogénio. É deste modo assegurada a complementaridade e estabilidade das duas faixas na dupla hélice do DNA. A Figura 2.2 ilustra a estrutura em hélice dupla de organização do DNA, a composição química das bases, assim como as regras de emparelhamento.

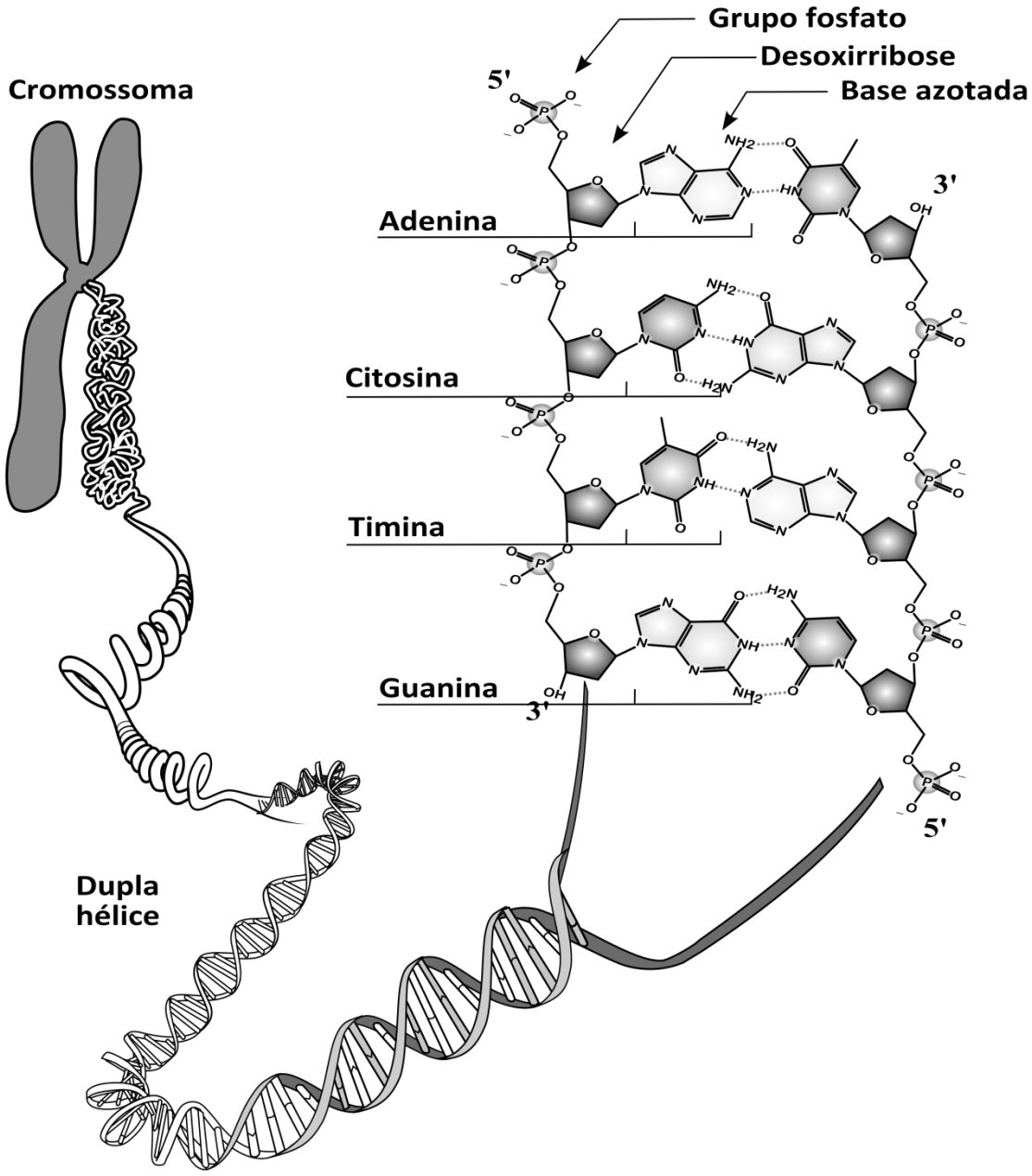


Figura 2.2: Representação esquemática do modelo de dupla hélice do DNA. Em destaque encontra-se a composição química dos quatro tipos de nucleótidos (adenina, citosina, timina e guanina), constituída pelo grupo fosfato, pentose e base azotada.

Apesar da organização do DNA poder variar de acordo com o ciclo celular, o seu estado mais usual é enquanto cromossoma. Cada cromossoma consiste em duas cadeias de DNA duplicadas unidas num ponto designado de centrómero. O processo de condensação dos cromossomas é realizado com o auxílio de uma proteína designada cromatina que obriga ao enrolamento da molécula de DNA possibilitando a sua compactação.

Devido à composição química dos resíduos das pentoses das bases, as faixas de DNA possuem direccionalidade. Num extremo, os polímeros de DNA contêm um grupo hidroxilo exposto na desoxoribose, também conhecido como a extremidade 3' da molécula. No outro extremo, contêm um grupo fosfato exposto, que dá pelo nome de extremidade 5'. A direccionalidade do DNA é de extrema importância em vários processos celulares, tais como a sua replicação. De facto, a síntese de novos ácidos nucleicos na célula ocorre na direcção 5'->3', pois os novos monómeros são adicionados através de reacções de desidratação que usam o grupo 3' hidroxilo como nucleófilo.

2.3 Proteínas

As proteínas, alternativamente denominadas de polipeptídeos, são compostos orgânicos formados por aminoácidos dispostos sob uma cadeia linear e unidos por ligações peptídicas entre o grupo carboxil e amina dos resíduos dos aminoácidos adjacentes. A sequência de aminoácidos de uma proteína é definida pela sequência do gene, codificada no código genético.

As proteínas são parte integrante de todos os organismos, participando em todos os processos que ocorrem dentro das células. Algumas proteínas, tais como a actina e a miosina, possuem funções estruturais e mecânicas, enquanto outras, no citoesqueleto, asseguram a forma das células. Outras ainda possuem um papel relevante na sinalização celular, na resposta imunitária e no ciclo celular.

Cada conjunto de três bases do código genético especifica univocamente um aminoácido. No entanto, em vez de existirem 64 aminoácidos (o correspondente a 4^3 – tripletos de quatro tipos de bases) existem, no máximo, 22 aminoácidos (20 aminoácidos padrão mais selenocistina e pirrolisina). Isto deve-se ao facto da terceira base possuir menos significância do que as restantes e de, consequentemente, existirem vários aminoácidos a serem codificados com variantes na terceira base. A Tabela 2.1 resume a correspondência entre tripletos e aminoácidos.

Após a tradução dos tripletos em aminoácidos, a proteína não fica de imediato activa. É necessário efectuar-se primeiro o processamento pós-translacional, que altera as suas propriedades químicas e físicas, o enrolamento e a estabilidade. As proteínas podem também interagir entre si de forma a desempenhar determinada função, sendo comum estas associarem-se sob a forma de complexos proteicos.

Tabela 2.1: Correspondência entre tripletos e aminoácidos.

1ª base no codão	2ª base no codão: U	2ª base no codão: C	2ª base no codão: A	2ª base no codão: G	3ª base no codão
U	Fenilalanina (Fen)	Serina (Ser)	Tirosina (Tir)	Cisteína (Cis)	U
					C
	Leucina (Leu)		STOP	STOP	A
				Triptofano (Trp)	G
C	Leucina (Leu)	Prolina (Pro)	Histidina (His)	Arginina (Arg)	U
					C
			Glutamina (Glu)		A
					G
A	Isoleucina (Ile)	Treosina (Tre)	Asparagina (Asn)	Serina (Ser)	U
					C
	Metionina (Met)		Lisina (Lis)	Arginina (Arg)	A
					G
G	Valina (Val)	Alanina (Ala)	Ácido aspártico (Asp)	Glicina (Gli)	U
					C
			Ácido glutâmico (Glu)		A
					G

2.4 Do gene à proteína

O processo de usar a informação codificada sob a forma de genes para gerar proteínas envolve três etapas: transcrição da informação genética, processamento e migração dessa informação do núcleo para o citoplasma e tradução da mensagem em proteínas [9, 13].

A primeira etapa, designada transcrição, corresponde à síntese de uma cadeia de RNA (*RiboNucleic Acid*) a partir do DNA que contém a informação que lhe serve de molde. Este processo inicia-se com a fixação da enzima RNA-polimerase sobre a região promotora correspondente ao gene a transcrever. De seguida, a RNA-polimerase desliza ao longo da cadeia de DNA provocando a sua abertura, iniciando-se, então, a transcrição da informação. De acordo com a regra da complementaridade das bases, dá-se a polimerização de ribonucleótidos, formando-se a cadeia de RNA. A síntese de RNA a partir de nucleótidos livres faz-se no sentido 5' para 3', sendo apenas uma das cadeias de DNA utilizada como molde. Após a passagem da RNA-polimerase, a molécula de DNA reconstitui-se, pelo restabelecimento de ligações de hidrogénio existentes entre as bases complementares. À semelhança do DNA, a molécula de RNA obtida consiste numa cadeia de polinucleótidos, apresentando, no entanto, diferenças na sua estrutura que possibilitam que a sua forma estável seja a de uma cadeia simples. Outra diferença encontrada no RNA consiste no facto de as bases de timina terem sido substituídas por bases de uracilo.

Nas células eucariontes, antes de sair do núcleo na forma de mRNA (*messenger RNA*) o pré-mRNA sofre algumas transformações. A unidade de transcrição na cadeia de DNA, que serviu de molde à criação do RNA, é composta por regiões codificantes, chamadas de exões, intercaladas por regiões não codificantes, denominadas de intrões (Figura 2.3). O

processamento pós-transcrição consiste na remoção, por acção de enzimas, dos intrões e da união de exões. De seguida, o mRNA processado migra do núcleo para o citoplasma. A migração vai possibilitar a ocorrência da tradução, a terceira e última fase.

A tradução corresponde à transformação da mensagem contida no mRNA na sequência de aminoácidos que constituem a cadeia polipeptídica. Neste processo estão envolvidos, para além do mRNA, que contém a informação genética, ribossomas, que consistem em sistemas de leitura do mRNA, assim como tRNAs, pequenas moléculas de RNA que transportam os aminoácidos para junto dos ribossomas. O processo de tradução principia com a ligação do mRNA e de um tRNA iniciador, que transporta usualmente a metionina (correspondente ao codão UTG) à pequena subunidade de um ribossoma, acoplando-se de seguida a este sistema a grande subunidade ribossomal. Após esta fase, a tradução progride com a ligação de um segundo tRNA que transporta um aminoácido que se vai ligar à metionina dando lugar à primeira ligação peptídica. De seguida, o ribossoma avança três bases, o correspondente ao tripleto traduzido, repetindo-se este processo ao longo de toda a cadeia de mRNA.

Quando é encontrado um tripleto de finalização (UAA, UAG, UGA), que não possui nenhum anticodão complementar, a síntese é obrigada a terminar. Por fim, as unidades do ribossoma separam-se, ficando livres para iniciar outro processo.

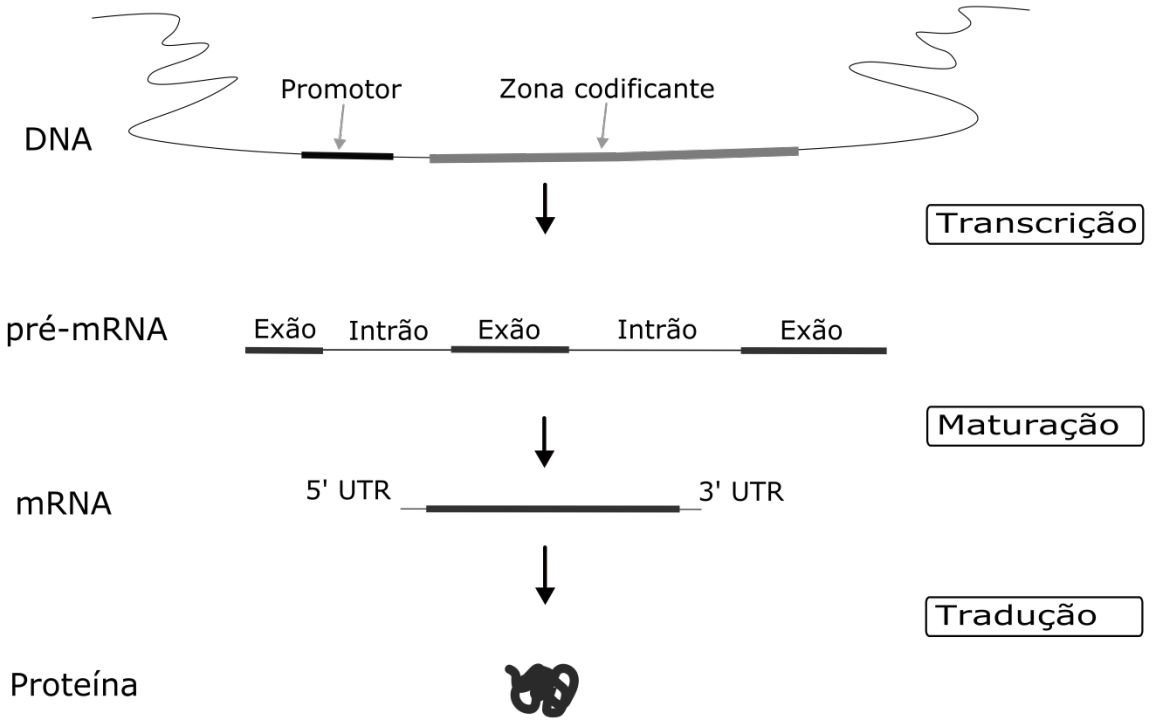


Figura 2.3: Mecanismo de síntese proteica.

Após a síntese das proteínas, estas não possuem de imediato actividade biológica, necessitando de sofrer vários tipos de alterações. Estas moléculas vão desempenhar funções variadas nas células, podendo ser funcionais, enquanto enzimas, ou estruturais, ao serem integradas em estruturas celulares, como a membrana plasmática.

2.5 Regulação da expressão génica

O ser humano possui um número estimado de 20.000 genes codificantes de proteínas dos quais apenas uma pequena porção se encontra activa num determinado momento. O conjunto de genes que se encontra activo depende de vários factores, tais como a função ou estado de desenvolvimento da célula ou o conjunto de estímulos externos que estão a ser aplicados sobre ela.

Os mecanismos responsáveis pela regulação da expressão génica actuam a vários níveis. O primeiro ocorre na transcrição, ao permitir que um mRNA apenas seja transcrito quando necessário. Tal é conseguido com o uso de factores de transcrição que tanto podem actuar como promotores quanto como repressores. O segundo mecanismo consiste na modificação pós-transcricional, permitindo que apenas uma parte dos mRNAs passem para a tradução. A outros níveis, as células regulam a expressão génica, através de DNA *folding* e modificações químicas dos nucleótidos e também por meio de mecanismos de *feedback*, nos quais as proteínas resultantes de um determinado gene permitem que a célula termine a produção da proteína [14].

2.6 Ferramentas biomoleculares

Muitas das descobertas realizadas nos últimos anos ficaram a dever-se ao desenvolvimento de novas técnicas e ferramentas laboratoriais das quais se destacam as apresentadas de seguida.

Reacção em cadeias de polimerase

A reacção em cadeia de polimerase (*Polymerase Chain Reaction* - PCR) é uma técnica usada para amplificar em várias ordens de grandeza, um determinado segmento de DNA. O procedimento compreende a aplicação de vários ciclos térmicos, em que, numa primeira fase, a temperatura é elevada de forma a obrigar à desnaturação da cadeia de dupla hélice em duas cadeias simples e, numa segunda fase, a temperatura é reduzida para que os *primers* se anelem com a fita e a DNA polimerase possa sintetizar uma nova molécula. O ciclo é então repetido, sendo o DNA previamente criado usado no novo processo de replicação, fazendo com que o número de sequências disponíveis cresça exponencialmente [15, 16].

O uso desta técnica revela-se especialmente útil quando a quantidade inicial de DNA disponível não é suficiente para condução da experiência.

Electroforese em gel

A electroforese em gel possibilita a separação de DNA, RNA e proteínas, através da aplicação de um campo eléctrico. Depositadas num suporte de gel, as partículas vão oferecer uma resistência ao movimento proporcional à sua dimensão. Deste modo, a mistura inicial é fraccionada numa série de diferentes bandas, arranjadas de acordo com a sua massa molecular [9].

Southern blot

O *Southern blot* consiste numa técnica que possibilita verificar a presença de uma determinada sequência de DNA numa amostra. Após a marcação (radioactiva ou fluorescente) da sequência a detectar, esta hibrida com a amostra de forma a remover as sequencias não ligadas, prossegue-se com a lavagem da membrana. Por fim, obtém-se uma imagem com as localizações a que a sonda se ligou e com as suas respectivas intensidades [17].

Tendo como base a técnica de *Southern blot,* foram desenvolvidas técnicas derivadas tais como o *Northern blot* [18], que possibilita o estudo do perfil de expressão do mRNA, ou o *Western blot* [19], que permite a detecção de proteínas.

2.7 Motivação ao uso de *microarrays*

A principal motivação para o desenvolvimento dos *microarrays* encontra-se no facto de, uma vez conhecida a sequência de um organismo se pretender obter uma perspectiva global da expressão do seu genoma sob a influência de uma determinada condição externa [5, 20].

Técnicas anteriores, como o *Northern blot,* já permitiam a compreensão do funcionamento das células e das associações existentes entre genes e respectivos fenótipos. No entanto, o princípio de funcionamento destas implica que, devido a questões de escalabilidade, apenas um conjunto limitado de genes possa ser testado por experiência, o que apresenta um *throughput* muito reduzido, dificultando a percepção do funcionamento do sistema como um todo. A tecnologia dos *microarrays* pretende, exactamente, endereçar esta questão, possuindo, como principal vantagem a possibilidade de monitorizar simultaneamente a actividade de milhares de genes. Foi outra importante motivação o crescente interesse em perceber as redes de relações biomoleculares a um nível global. Cada tipo particular de células é caracterizado por um diferente padrão nos seus níveis de expressão, ou seja, cada célula produz um conjunto específico de proteínas em quantidades bem definidas. A capacidade de usar *microarrays* para interrogar milhares de genes em simultâneo abre inúmeras possibilidades.

De forma genérica, os estudos de *microarrays* podem ser divididos em duas categorias: estudos em que as amostras são usadas de forma a fornecer informação sobre os genes e

estudos em que os genes são usados para fornecer informações sobre as amostras. No primeiro caso, pretende-se estudar os diferentes padrões genéticos associados às diferenças conhecidas nas amostras. Exemplo disto é a possibilidade de estudar os efeitos da indução de um choque térmico numa levedura (*Saccharomyces Cerevisiae*) de forma a estudar o perfil genético associado. O segundo caso resume-se à utilização do conhecimento existente sobre perfis de expressão na identificação das doenças associadas. Esta possibilidade é especialmente útil na área médica, nomeadamente na identificação de padrões de expressão que se podem associar a doenças genéticas, potenciando a aceleração do seu diagnóstico [6, 21, 22].

2.8 *Microarrays* de DNA

Genericamente um *microarray* consiste num substrato (de vidro, plástico ou *nylon*) no qual sondas (*probes*), constituídas por moléculas de DNA, se encontram em posições bem definidas, designadas *spots*. Um *microarray* típico contém dezenas de milhar de *spots*, em que cada um contém vários milhares de sondas com comprimento variável entre as dezenas e as centenas de nucleótidos.

De acordo com a regra de emparelhamento das bases, as sondas permitem a identificação da presença e da abundância de alvos (*targets*), correspondentes às sequências complementares. Os alvos ligam-se por hibridação às sondas do *microarray* com as quais partilham um nível satisfatório de complementaridade da sequência. Após decorrer tempo suficiente para que a hibridação ocorra o excesso de amostra é retirado da superfície do *microarray*. Finalmente, através da medição dos níveis de intensidade dos *spots,* obtém-se uma estimativa da abundância das sequências alvo na solução [23]. A Figura 2.4 ilustra um *microarray* de DNA, incluindo a sua organização em *spots,* onde as sondas, através de hibridação competitiva, emparelham como as sequências alvo.

Deste modo, é possível a quantificação do tipo e da quantidade de mRNA transcrito que se encontra presente numa colecção de células. Note-se, no entanto, que, no uso dos *microarrays,* é assumido que o número de moléculas de mRNA presente na célula corresponde à quantidade de produto final produzido, o que nem sempre se verifica, tendo em consideração a influência de factores pós-transcrição.

São dois os principais métodos de construção de *microarrays* de DNA: deposição de fragmentos e síntese *in situ*. Apesar de, neste documento, ser dado especial ênfase à tecnologia baseada na deposição de fragmentos de sequências de DNA, existe uma segunda abordagem que consiste na síntese *in situ* dos oligonucleótidos. Nesta abordagem, são três as tecnologias disponíveis: fotolitografia, impressão baseada em jacto de tinta e síntese electroquímica. Uma revisão detalhada das tecnologias de construção de *microarrays* de DNA encontra-se em [24-26].

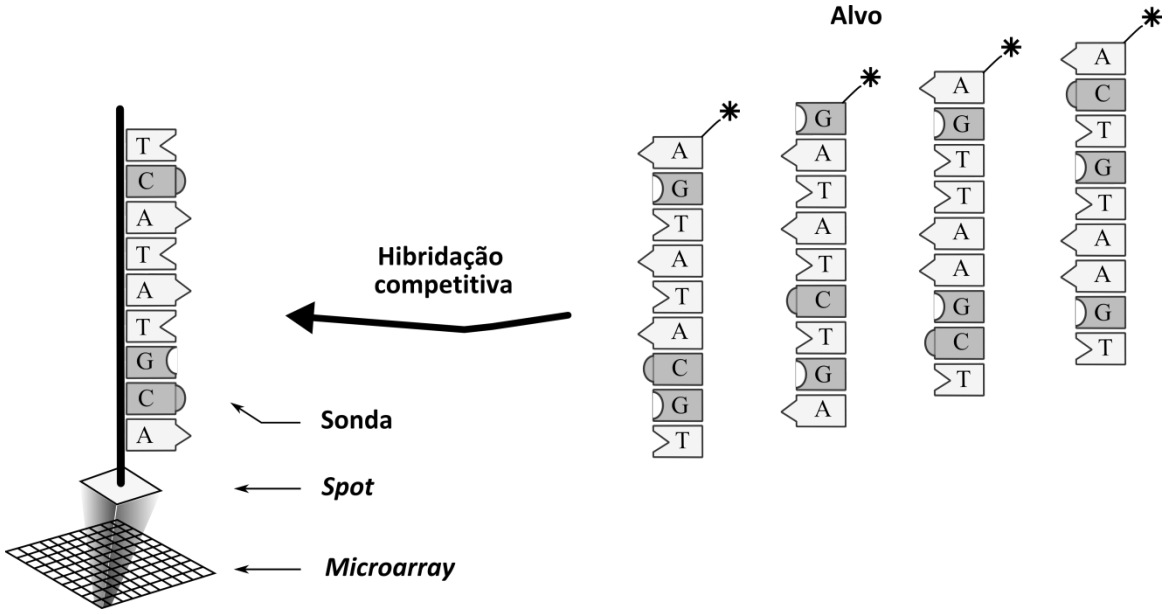


Figura 2.4: Esquema de funcionamento de um DNA *microarray*.

2.8.1 *Spotted* microarrays

A primeira tecnologia a ser usada na construção de *microarrays* foi a baseada na deposição de segmentos de DNA sobre um substrato [20].

Nesta tecnologia as sondas podem ser constituídas por oligonucleótidos, cDNA ou por pequenos segmentos de produtos de PCR correspondentes a mRNAs. Após síntese das sondas, estas são depositadas em localizações específicas no *microarray*, através do uso de um braço robótico que possui um conjunto de agulhas e que é sucessivamente mergulhado em placas com a solução de sondas e nos *spots* correspondentes no *microarray*.

Apesar de ter sido esta a primeira tecnologia disponível para construção de *microarrays* continua a ser amplamente usada, especialmente, no âmbito de projectos de investigação. A principal vantagem traduz-se na elevada flexibilidade oferecida, na medida em que podem ser incluídas sondas produzidas localmente no laboratório, específicas para a sequência alvo a identificar. Podem ainda ser integradas sondas provenientes de colecções de oligonucleótidos que, mesmo possuindo um elevado custo de aquisição, permitem a obtenção de um menor custo unitário relativamente às plataformas comerciais.

2.8.2 Fotolitografia

Uma das tecnologias mais interessantes é a disponibilizada pela empresa Affymetrix, cujas sondas de oligonucleótidos são sintetizadas num chip de silício por meio de um processo designado de fotolitografia [27]. Num primeiro passo, o chip é imerso numa solução, contendo um precursor para um dos quatro nucleótidos. A fixação do nucleótido ao *chip* é

conseguida através da incidência de um feixe luminoso. É, no entanto, utilizada uma máscara para assegurar que a luz apenas atinge as sondas cujo primeiro nucleótido corresponde à base imersa. De seguida o *chip* é lavado de forma a remover todos os nucleótidos ainda presentes e para que possa proceder-se ao próximo passo. O processo de síntese continua suportado pela aplicação de sucessivas imersões, lavagens e aplicações de máscaras.

Devido ao facto da plataforma da Affymetrix consistir numa solução chave na mão, em que todo o processo é meticulosamente controlado e em que não são admitidas alterações na organização ou constituição das sondas, os resultados são mais consistentes do que os obtidos com a tecnologia dos *spotted microarrays*.

Os custos de fabrico e a frequência de erros ampliam-se com o aumento do tamanho da sonda empregue e, por esse motivo, as sondas são de reduzida dimensão, com um máximo de 25 nucleótidos. Consequentemente é no intuito de se encontrar correspondência para os 25 oligonucleótidos, as condições de hibridação não são tão rígidas como as usadas nos *spotted microarrays*.

De modo a evitar situações de hibridação cruzada, eventualmente obtidas pela reduzida dimensão das sondas usadas, a Affymetrix aplica vários pares de sondas para cada transcrito alvo. Cada par de sondas perfaz-se de uma sequência de 25 oligonucleótidos com complementaridade completa com o exão do gene alvo (*perfect macth*) e de uma outra de 25 oligonucleótidos, que difere da anterior num único nucleótido localizado na posição central (*mismatched*). O objectivo é o de que as sondas que possuem um nucleótido errado não hibridem com o transcrito alvo sem erros, mas hibridem com muitos dos transcritos alvo com os quais as restantes sondas sem erros também erradamente hibridam. Assim, o valor de intensidade do *mismatched* subtraído ao do *perfect macth* deve dar uma estimativa mais realística da intensidade correspondente à hibridação do transcrito alvo.

Como consequência do menor comprimento das sondas, quando comparadas com as usadas nos *microarrays* de cDNA, as actuais implementações necessitam de 11 a 16 pares de sondas para cada gene a identificar.

2.8.3 *Microarrays* de jacto de tinta

Várias empresas, tais como a Protogene e a Agilent, desenvolveram uma tecnologia de construção de *microarrays in situ,* alternativa à proposta pela Affymetrix [28, 29]. Neste caso, a síntese dos oligonucleótidos é realizada através do uso da tecnologia da impressão de jacto de tinta, em que estes são projectados de forma a construir a sequência. Em cada passo da síntese, gotas da base a sintetizar são disparadas contra o *spot* desejado no substrato, do mesmo modo que ocorre numa impressora de jacto de tinta convencional. A diferença é que, neste caso, são disparadas bases de A, C, T e G, em alternativa à tinta.

A principal vantagem desta tecnologia relativamente às anteriores, está no facto do processo de sintetização dos oligonucleótidos ser inteiramente controlado por parâmetros fornecidos ao computador. Deste modo é uma tecnologia bastante flexível, podendo o operador personalizar a construção, de acordo com as suas necessidades.

2.8.4 Electroquímica

O método de síntese electroquímica usa pequenos eléctrodos embebidos no substrato para gerir os sítios de reacção individuais. Sucessivas soluções, contendo bases específicas, são espalhadas sobre a superfície, sendo que os eléctrodos são activados nas posições desejadas de forma a que as sequências sejam construídas base a base. Este tecnologia, apesar de não se encontrar tão difundida quanto as restantes, é aplicada por empresas como a CombiMatrix [30].

2.9 Áreas de aplicação dos *microarrays* de DNA

A capacidade dos *microarrays* detectarem a presença de dezenas de milhar de sequências genéticas em paralelo despertou o interesse em várias áreas da biologia e da medicina [6, 12, 31, 32]. Como consequência, ao longo dos últimos anos várias variantes à tecnologia base foram exploradas possibilitando a descoberta de novas funcionalidades. Apesar da dificuldade em definir uma lista fechada, os três principais modos de operação dos *microarrays* são: monitorização da expressão génica, detecção de mutações e polimorfismos e *tiling microarrays*.

2.9.1 Monitorização da expressão génica

A monitorização da expressão génica consiste em detectar a presença e a abundância de sequências alvo numa determinada solução. O uso dos *microarrays* de DNA possibilita a quantificação simultânea da expressão de milhares genes, promovendo o estudo da expressão diferencial como resposta a diferentes tratamentos, estímulos ambientais ou alterações pato-fisiológicas. Estes *microarrays* são constituídos por DNA complementar à sequência de mRNA correspondente ao gene a ser detectado.

O resultado dos estudos realizados traduz-se em catálogos de perfis de expressão que associam condições experimentais a conjuntos de genes regulados positiva ou negativamente. A monitorização da expressão pode ainda ser aplicada ao estudo de uma série temporal, tal como o ciclo celular. Os catálogos obtidos apresentam-se como um valioso recurso no diagnóstico de várias doenças, com especial foco na oncologia molecular, ao possibilitar a detecção de diferentes tipos e estados de desenvolvimento de doenças cancerígenas [6, 28, 33].

Os *microarrays* podem ainda ser usados para acelerar o desenvolvimento de novos fármacos, assim como para optimizar os já existentes, através da eliminação de efeitos adversos. Por exemplo, se um determinado gene se encontrar sobre-expresso num determinado tipo de cancro, o uso de *microarrays* possibilita verificar se um fármaco candidato reduz o seu nível de expressão, forçando o atrofiamento deste. Este processo implica um conhecimento da correlação entre a actividade de genes e a actividade de compostos candidatos [21].

2.9.2 Detecção de mutações e polimorfismos

Um SNP (S*ingle Nucleotide Polymorphism*) é uma sequência do genoma que se mantém constante, a menos de uma variação num único nucleótido. Estas variações, bastante comuns no genoma humano, mantêm-se conservadas dentro de populações, pelo que podem ser usados como marcadores genótipos [34].

Apesar do princípio de funcionamento do *microarray* se manter, a constituição das sondas e dos alvos difere da dos microarrays de *monitorização* da expressão. Na construção de um *microarray* de SNPs, as sondas são um segmento da zona conservada, em que um polimorfismo se encontra na zona central. Por sua vez, as sequências alvo são constituídas por segmentos de DNA do genoma completo do organismo.

São várias as aplicações dos *microarrays* de SNP's, no entanto, uma das mais relevantes resulta na avaliação da susceptibilidade de determinadas doenças genéticas. Tal é conseguido através do estudo da sequência associada a determinada doença e da capacidade de encontrar polimorfismos. Para várias doenças, tais como diabetes e artrite reumatóide já se conhecem realizações bem sucedidas [35]. A identificação da propensão de um indivíduo para contrair uma determinada doença pode ser conseguida antes mesmo de esta possuir manifestações, o que possibilita a toma de fármacos específicos de forma a minorar ou, mesmo, evitar os seus efeitos.

2.9.3 *Tiling microarrays*

Os *tiling microarrays* são constituídos por sondas desenhadas para representar densamente uma região genómica de interesse. Ao contrário dos *microarrays* de DNA típicos que apenas avaliam o produto resultante da expressão génica, neste caso, como toda a região genómica está a ser analisada, é possível a detecção de novos genes ou de regiões até então desconhecidas. Os *tiling microarrays* possuem várias funcionalidades, nomeadamente, o mapeamento do transcriptoma, a análise ChIP-on-chip ou a hibridação comparativa do genoma [36].

Hibridação genómica comparativa

Os *microarrays* de CGH (*Comparative Genomic Hybridization*) possibilitam a comparação de duas moléculas de DNA pertencentes a duas células distintas ou a indivíduos próximos. Esta técnica permite uma elevada resolução na detecção de remoções, duplicações, inversões e transladações existentes entre as duas moléculas. É especialmente útil no estudo das diferenças existentes no DNA de uma célula normal e no de uma célula cancerígena.

ChIp-on-chip

A técnica de imunoprecipitação da cromatrina (C*hromatin Immunoprecipitation* - *ChIp*) possibilita o estudo de interacções entre proteínas e sequências no genoma. A técnica ChIp-on-chip utiliza como alvos produtos da imunoprecipitação da cromatrina e aplica-os sobre um *microarray* (também designado chip) em que as sondas consistem em sequências representativas do genoma a analisar.

Mapeamento do transcriptoma

Outro uso popular dos *tiling microarrays* reside na detecção das sequências expressas. Transpondo as limitações inerentes aos tradicionais métodos de sequênciação do cDNA, tais como a dificuldade em identificar moléculas de RNA raras ou o não reconhecimento de genes apenas activos durante um período temporal limitado, os *tiling microarrays* apresentam-se como uma promissora alternativa. Neste caso concreto, os *microarrays* possuem uma elevada resolução e sensibilidade, pois mesmo moléculas de reduzidas dimensões podem ser detectadas. De facto, várias sequências não codificantes apenas foram identificadas com o uso dos *tiling microarrays* [37].

2.10 Desafios no uso dos *microarrays* de DNA

Apesar dos esforços dispendidos na optimização da tecnologia, a utilização de *microarrays* ainda apresenta alguns desafios. De seguida são detalhados os principais, sendo dado especial ênfase aos associados com a tecnologia *spotted*. Uma análise mais detalhada das actuais limitações encontra-se em [38].

Ruído

Devido à sua natureza, os *microarrays* tendem a apresentar dados com bastante ruído. De Com efeito, realizando a mesma experiência, com os mesmos métodos, materiais e condições, é possível que após a digitalização e processamento da imagem, os valores obtidos sejam distintos. Na origem desta situação está o facto do ruído ser cumulativo a todos os passos e de não ser possível reproduzir com precisão todas as condições de uma experiência.

O desafio surge, quando se pretende comparar o resultado de diferentes experiências, pois a variação encontrada num determinado gene pode ser genuína ou simplesmente devida ao ruído introduzido. A forma usualmente aplicada para reduzir estes efeitos é a utilização de várias réplicas da experiência e o cálculo de valores médios.

Normalização

O objectivo da normalização é o de eliminar diferenças sistemáticas entre experiências e artefactos existentes nos dados. A normalização é essencial para que duas experiências possam ser directamente comparadas.

A necessidade de normalização pode advir de várias fontes, tais como o recurso a diferentes quantidades de mRNA (levando a intensidades médias distintas), a não linearidade dos marcadores usados ou a aplicação de diferentes níveis de saturação.

Desenho experimental

Ainda que muitas vezes negligenciado, o desenho experimental é essencial ao sucesso da experiência. Porém, para a mesma questão biológica podem existir várias combinações de desenhos experimentais possíveis, o que pode conduzir a diferentes valores de expressão finais. A escolha do desenho experimental é ainda influenciada por factores como a quantidade de material genético disponível ou o número de *microarrays* usados.

Elevado número de genes

Uma das principais vantagens dos *microarrays* reside no facto de possibilitarem questionar os níveis de expressão de um elevado número de genes. Devido à quantidade de genes envolvidos, o método de investigação assemelha-se muitas vezes a colocar a questão biológica e esperar pelo resultado dos genes marcados como diferenciadamente expressos. Se bem que promissor, este método apresenta desafios às actuais ferramentas estatísticas.

Significância dos resultados

Quando utilizados para caracterizar condições específicas, é crucial averiguar a significância existente entre os grupos considerados. Mais uma vez os métodos estatísticos tradicionais não podem ser directamente usados, pois numa experiência de *microarrays* existe um elevado número de variáveis (genes no *microarrays*) e um reduzido número de amostras (hibridações).

Factores biológicos

Pesem as mais-valias introduzidas pelos *microarrays*, estes não invalidam a utilização das técnicas de biologia molecular já usadas. O funcionamento do *microarray* é baseado na medição da quantidade de mRNA de um determinado gene, partindo-se do princípio que a quantidade de proteínas obtidas está directamente relacionada com este valor inicialmente obtido. Embora genericamente se assuma este princípio, a existência de efeitos pós-transcrição podem-no invalidar.

Mesmo assumindo que todos os genes que efectivamente se encontram como diferencialmente expressos efectivamente o são, é necessário converter esta informação para conhecimento biológico sendo ainda fundamental perceber em que processos regulatórios os genes estão envolvidos e de que forma estão relacionados.

2.11 Sumário

Este capítulo centrou-se na célula, a estrutura base de todos os organismos vivos, assim como nas suas principais partes integrantes. É descrita a unidade central de armazenamento da informação genética, o DNA, e a forma como este é transcrito em mRNA, passa do núcleo para o citoplasma e é traduzido em proteínas.

A compreensão do funcionamento destes processos tem sido possível graças ao desenvolvimento de novas ferramentas biomoleculares das quais se destacam os *microarrays*. Os *microarrays* são aqui apresentados como uma solução para a análise em larga escala dos níveis de expressão dos genes da célula. Existem actualmente diversas tecnologias alternativas usadas na construção de *microarrays* e, apesar de neste trabalho nos centrarmos na tecnologia de *spotted* DNA *microarrays*, foi apresentada uma descrição sumária das restantes tecnologias disponíveis.

São vários os modos de operação dos microarrays de DNA, sendo os principais a monitorização da expressão génica, a detecção de mutações e polimorfismos e os *tiling microarrays*. Alguns dos métodos apresentados possuem interesse ao nível da sua aplicação em ambientes clínicos. No entanto, sendo uma tecnologia relativamente recente, existem ainda bastantes condicionantes ao seu uso, nomeadamente na validação da qualidade dos resultados. Nos últimos anos, tem sido realizado um esforço colectivo para endereçar as questões existentes e para promover e generalizar o uso desta tecnologia.

Capítulo 3

3 Gestão de dados num laboratório de *microarrays*

Os *microarrays* são, neste momento, uma das tecnologias mais promissoras na área da biologia molecular [21, 39, 40]. Com um único *microarray* é possível verificar o comportamento de milhares de genes para uma determinada condição experimental. Apesar do seu elevado potencial, a crescente quantidade de dados gerados criou novos desafios nos procedimentos de armazenamento e análise [8]. Em muitos casos, a capacidade de lidar com estes dados de forma eficiente é o factor condicionante na tarefa do investigador que pretende obter resposta às suas questões. O desenvolvimento de sistemas LIMS (*Laboratory Information Management System*) específicos para *microarrays* pretende dar cumprimento a este desafio [41, 42]. O principal objectivo destes sistemas é servirem de repositório de todos os procedimentos aplicados, assim como de todos os dados gerados no laboratório.

Para o desenvolvimento dos sistemas LIMS, foi decisiva a importância de normas e ontologias [42, 43]. As normas tem como principal finalidade indicar que dados devem ser registados, como devem ser armazenados e, não menos importante, como podem ser, depois, partilhados. As ontologias, por sua vez, permitem capturar, de forma estruturada, a complexidade vocabular de um determinado domínio. Através do seu uso, é possível descrever, em detalhe, um evento ou objecto de estudo. Deste modo, o uso combinado de normas e ontologias permite que diferentes grupos académicos ou comerciais consigam partilhar de forma eficiente dados entre os seus sistemas.

Encontram-se, actualmente, disponíveis vários sistemas de gestão de dados de *microarrays* [41]. Estes podem ser analisados de acordo com vários critérios e, pese a existência de características comuns, tais como o uso de normas, são vários os pontos em que divergem. Assim, a escolha de um sistema deve ter em consideração vários factores e, em especial, a necessidade concreta do laboratório.

Neste capítulo, é apresentada uma análise do fluxo de informação de uma experiência de *microarrays,* bem como um levantamento das normas, das ontologias e dos sistemas de

gestão de dados existentes. Pela avaliação realizada aos sistemas de gestão de dados, foram identificadas várias limitações que se reflectem ao nível das funcionalidades disponibilizadas, dos paradigmas de usabilidade aplicados e da facilidade de instalação e de manutenção. A contribuição deste capítulo consiste na proposta, desenvolvimento e avaliação do sistema Mind (*Microarray Information Database*), que possibilita a gestão de dados de *microarrays*. As principais vantagens do Mind residem na sua interface intuitiva e fácil de usar, na possibilidade de realizar trabalho colaborativo, no controlo automatizado da qualidade dos dados, na submissão directa em repositórios públicos e numa integração transparente com aplicações externas.

3.1 Ciclo de uma experiência de *microarrays*

Uma experiência de *microarrays* compreende uma série de iterações encadeadas em que a falha numa pode comprometer todo o estudo [21]. De forma a evitar esta situação, ou pelo menos a minimizar os seus efeitos, é essencial um correcto planeamento da experiência a realizar, assim como o registo de todos os procedimentos efectuados. Se, no primeiro caso, não existe uma necessidade específica de ferramentas informáticas, no segundo, o uso de ferramentas dedicadas é crucial [44]. Não obstante, a correcta modelação de um sistema capaz de lidar com este domínio, caracterizado não só pela sua dimensão mas também pela sua heterogeneidade, impõe uma completa compreensão de todas as iterações de uma experiência de *microarrays*. O esquema da Figura 3.1 contém as principais fases deste processo, assim como o seu encadeamento. Do topo para a base, as principais fases são: planeamento da experiência, execução dos procedimentos laboratoriais, obtenção e armazenamento dos dados e, por fim, análise e interpretação biológica dos resultados obtidos.

3.1.1 Desenho experimental

O objectivo do desenho experimental reporta-se à identificação das questões específicas a que a experiência se propõe responder e, ainda, quais os métodos, os *microarrays* e as amostras necessários [45, 46]. O planeamento do desenho experimental da experiência pode ser dividido em três fases: desenho conceptual, considerações estatísticas e considerações técnicas.

No passo inicial, desenho conceptual, é necessário definir se a experiência consiste numa comparação de classes ou na descoberta de novas classes. No primeiro caso, podem comparar-se dois sistemas biológicos, através do uso de uma determinada condição experimental (por exemplo, choque térmico ou efeito de um fármaco) e questionar quais os genes que se encontram diferencialmente expressos entre classes, ou estudar o efeito sobre uma série temporal e questionar quais os genes que partilham padrões de expressão. No caso da descoberta de classes, a questão é colocada ao contrário, pois o que se pretende

saber é se existem genes diferencialmente expressos que possam ser usados na definição de classes biológicas desconhecidas.

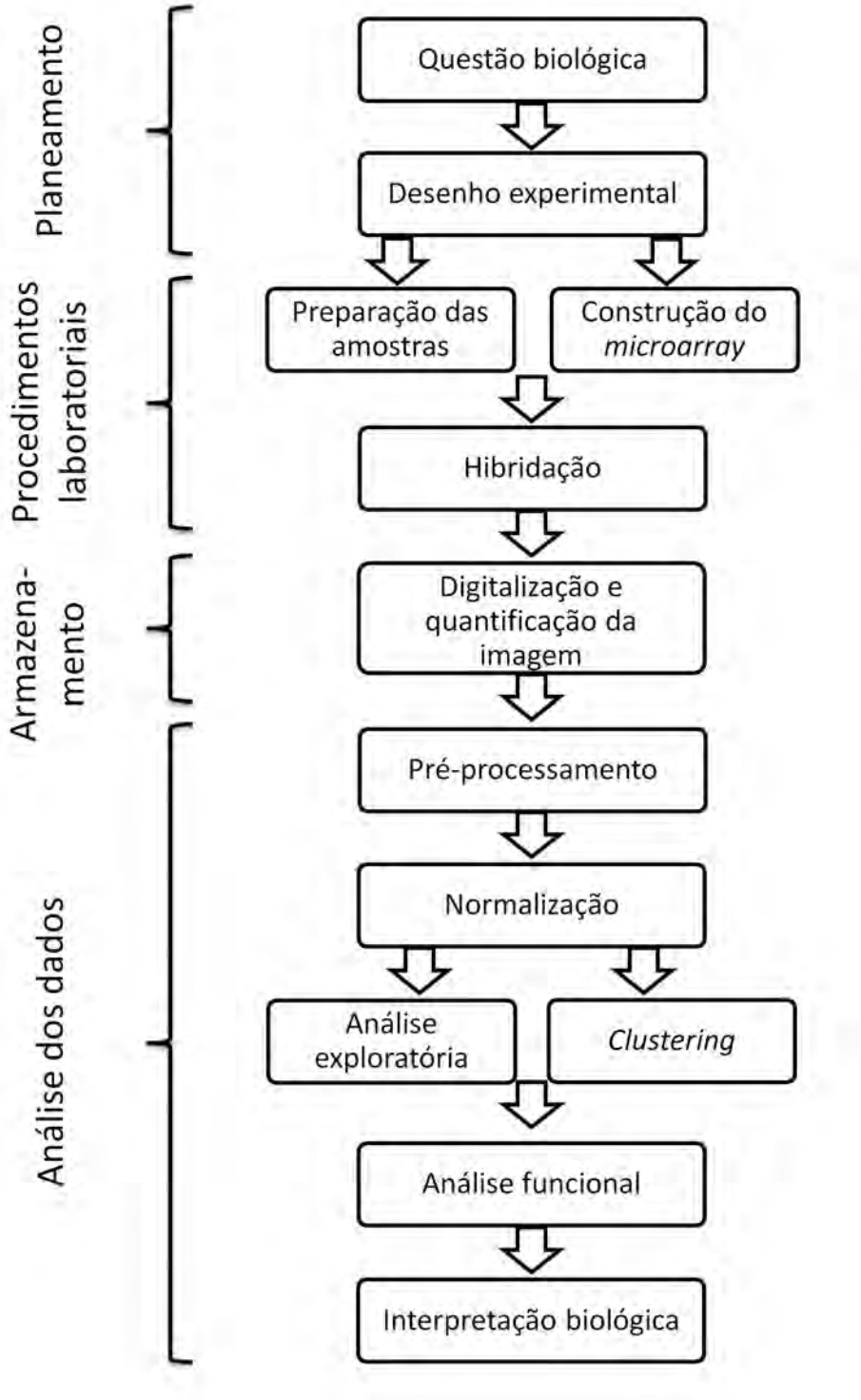


Figura 3.1: Fluxo de dados de uma experiência de *microarrays*.

Com o objectivo de obter dados que possuam significado estatístico, existem três aspectos que devem ser tidos em consideração: hibridação concorrente, replicação e *dye swapping*. A hibridação concorrente refere-se ao uso de duas amostras com dois marcadores fluorescentes ou radioactivos num mesmo *microarray*. Idealmente, a distribuição das amostras pelos *microarrays* deveria reflectir directamente as questões a endereçar. No entanto, de forma a minimizar os *microarrays* usados em alguns desenhos experimentais (por exemplo *loop design*), a resposta é obtida através da combinação de diferentes amostras, peso o prejuízo na qualidade dos resultados. A replicação das amostras, podendo esta ser técnica ou biológica, resulta num método simples e eficaz de diminuir a variabilidade dos dados e de aumentar a confiança nos mesmos. No entanto, tal acarreta um aumento considerável no número de *microarrays* necessários. Outro aspecto a ter em conta, o *dye swapping*, baseia-se na realização de hibridação complementar em que os marcadores usados em cada amostra são invertidos. O objectivo é o de eliminar a distorção que diferentes marcadores impõem.

A fase final da definição do desenho experimental implica a avaliação dos factores que limitam a experiência, podendo estes ser o número de *microarrays* ou a quantidade de RNA disponível. Apesar de não existir uma lista fechada de desenhos experimentais disponíveis, tal como a Figura 3.2 ilustra, os três mais comuns são: *reference*, *loop* e *balanced block*. No *reference* (Figura 3.2.a), todas as amostras são testadas contra uma única amostra de referência. No *loop design* (Figura 3.2.b), cada amostra é comparada com a próxima, de forma a criar uma forma circular. Por último, o *balanced block* (Figura 3.2.c) corresponde à situação em que para cada amostra existem múltiplos *microarrays*. De notar que cada desenho experimental apresenta as suas vantagens e desvantagens, sendo que a escolha do mesmo deve reflectir as condições específicas do estudo.

3.1.2 Construção do *microarray*

Empresas como a *Agilent* e a *Affymetrix* possuem catálogos bastante completos de *microarrays* para os organismos modelo mais estudados. Apesar de possuírem um custo unitário elevado, o uso destas plataformas apresenta como principal mais-valia a fiabilidade dos resultados. Como alternativa, existe a possibilidade de construção do próprio *microarray,* que, apesar de envolver um investimento inicial superior, possui, a médio prazo, várias vantagens, tais como flexibilidade e um menor custo por *microarray* [47].

Em ambos os casos, é necessária a obtenção de um ficheiro que descreva a organização e localização de todas as sondas no *microarray*, incluindo os controlos positivos e negativos. Este ficheiro é essencial para que seja possível a posterior associação dos valores de expressão à sonda correspondente.

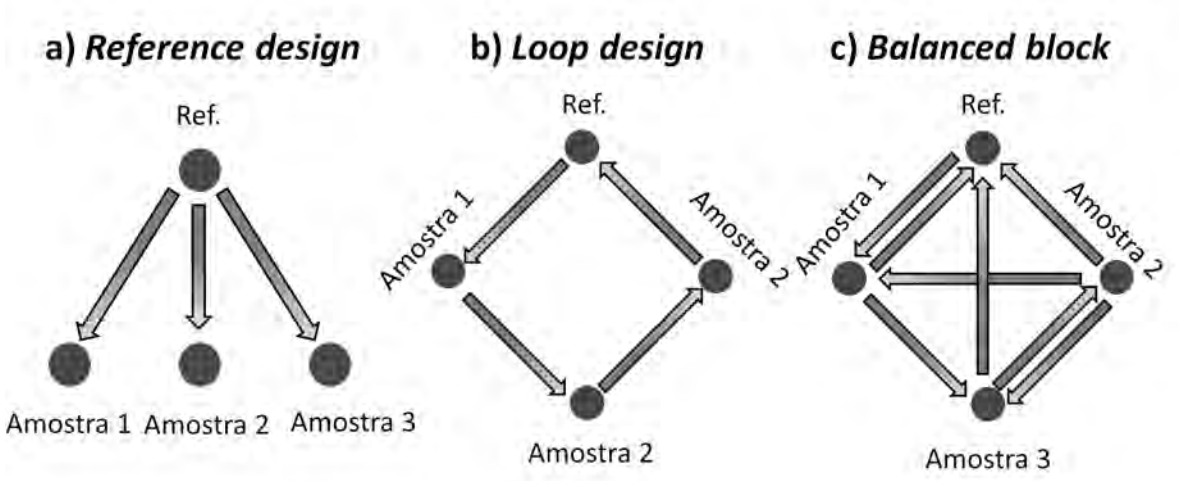


Figura 3.2: Representação esquemática dos três desenhos experimentais mais comuns: a) *reference*, b) *loop* e c) *balanced block*. Cada círculo representa uma amostra e cada seta um *microarray* no qual duas condições são simultaneamente testadas.

3.1.3 Preparação das amostras e hibridação

Após o planeamento da experiência e a construção, ou compra, do *microarray,* tem início o trabalho experimental [46]. Qualquer que seja o organismo usado e a condição experimental escolhida, parte-se sempre de uma amostra que constitui o objecto de estudo. Depois de sujeitar as células da amostra a duas condições experimentais distintas, o mRNA é extraído e etiquetado com dois marcadores distintos. A Cianina é tipicamente usada como marcador, podendo ser usada a sua variante *Cy3* (corresponde a verde, com intensidade máxima de excitação nos ~550nm e de emissão nos ~570nm) e *Cy5* (corresponde a vermelho, com intensidade máxima de excitação nos ~650nm e de emissão nos ~670nm). A escolha da Cianina justifica-se pela sua elevada eficiência de incorporação, elevada foto estabilidade e distinto espectro de emissão e excitação.

3.1.4 Obtenção dos dados

A obtenção dos dados é considerada o último passo da parte experimental. O procedimento principia com a digitalização do *microarray* de forma a obter-se uma imagem com os níveis de fluorescência dos *spots*. De seguida é realizada a segmentação da imagem, pela detecção dos sinais correspondentes aos *spots.* Por fim, é obtido para cada *spot* um conjunto de parâmetros que descrevem a sua morfologia e intensidade [38].

Aquisição da imagem

Neste passo, é obtida através da digitalização do *microarray*, uma imagem que reflecte a fluorescência de cada *spot*. Os parâmetros do *scanner* devem ser devidamente configurados, nomeadamente a resolução e a potência do laser. A escolha deste valor deve

ter em consideração a pertinência de se evitar a obtenção de valores de intensidade saturados. Deste modo, os valores de ganho devem ser ajustados durante o processo de digitalização, de forma a conseguir-se o melhor compromisso entre os dois canais [48]. O objectivo é o de maximizar o número de *spots* detectados minimizando, o número de pixéis saturados. Outro parâmetro a ser necessariamente ponderado é a resolução da imagem obtida. Se esta apresentar um baixo número de pixéis por *spot,* a validade dos resultados diminui, devido à possibilidade de passarem a ser validados valores fora da área do *spot*.

Segmentação da imagem

O processo de segmentação da imagem consiste na capacidade de, computacionalmente, se distinguir quais os pixéis que, na imagem, correspondem a *spots*. Esta tarefa é dificultada por vários factores que afectam a qualidade da imagem, tais como a irregularidade da forma dos *spots*, ou o deslocamento do *spot* relativamente à sua suposta localização. Algo que também não pode ser descorado é a existência de hibridação residual com o suporte do *microarray*, o que faz com que o valor da intensidade do fundo não seja nulo. Este valor deve pois, através de um dos vários métodos existentes, subtrair-se ao valor de intensidade do *spot* [38].

Obtenção dos valores de expressão

É, então, obtido, para cada *spot*, um conjunto de parâmetros que o caracterizam, dos quais se destacam o valor médio e mediano de intensidade e a sua área. Estes valores são organizados sob a forma de uma tabela, em que cada linha corresponde a um *spot* e cada coluna a uma característica. Terminado este passo, o conjunto de ficheiros tabulares, em que cada um corresponde a um *microarray*, é o material de trabalho para a condução do resto do estudo.

3.1.5 Análise dos dados

Os valores anteriormente obtidos vão ser usados como entrada nas ferramentas de análise de dados. Se bem que exista uma ampla oferta de ferramentas disponíveis, estas podem ser divididas em três classes: verificação da qualidade, pré-processamento e normalização [38, 48, 49]. A primeira diz respeito à verificação da qualidade elementar do resultado. Neste passo, vários erros sistemáticos, associados com o procedimento laboratorial, são detectados através do uso de um conjunto de ferramentas estatísticas e de métodos alternativos de visualização dos resultados. Após esta fase, é, normalmente, aplicado aos dados um conjunto de algoritmos que tem como objectivo remover os efeitos de hibridação basal, responsável por conduzir a valores de intensidade do fundo do *microarray*. É, ainda, realizada a normalização dos dados, de forma a que os valores produzidos sejam uniformes e passíveis de comparação. Só então, através do uso de ferramentas de visualização e de análise exploratória, a interpretação biológica dos dados é obtida. Caso o resultado alcançado não satisfaça a questão inicialmente enunciada, o desenho experimental necessita de ser reformulado, podendo o processo recomeçar.

3.2 Normas, ontologias e vocabulários controlados

Com o objectivo de partilhar dados de *microarrays*, seja entre investigadores, seja entre programas de computador, é necessário que quer os dados, quer os processos de transferência dos mesmos, obedeçam a normas e terminologias comuns [42, 43, 50]. De facto, o uso destas regras é indispensável à desambiguação na compreensão do conteúdo da mensagem. Por exemplo, os termos sonda (*probe)* e alvo (*target)* têm sido usados, por diferentes investigadores, de forma indistinta, referindo-se ora ao DNA que está no *microarray,* ora à amostra de DNA na solução a ser hibridada.

As normas são, então, essenciais ao desenho de programas de computador que possam ser integrados com outras aplicações. Fazendo uso de um conjunto de normas que especifiquem a representação e comunicação dos dados é possível que diferentes grupos académicos ou comerciais desenvolvam as suas aplicações, ou bases de dados, que suportem dados gerados por terceiros.

3.2.1 Normas na área da biologia molecular

Na área da biologia molecular, mais concretamente no caso dos *microarrays*, a necessidade de definição de normas foi reconhecida há muito por organizações como o EBI (*European Bioinformatics Institute*), o NCBI (*National Center for Biotechnology Information*) e o NCGR (*National Center for Genome Resources*) [42]. Estas têm liderado um esforço comum, na tentativa de estabelecer um repositório de dados de expressão génica que possa ser utilizado por toda a comunidade. No entanto entidades comerciais também propuseram formatos, tais como o AADM (*Affymetrix Analysis Data Model*), que têm por objectivo principal facilitar a troca de dados entre diferentes fontes de dados de expressão génica e o desenvolvimento de ferramentas de análise de dados. Estes esforços, no sentido de normalizar o formato dos dados de *microarrays*, foram consolidados no MGED (*Microarray Gene Expression Database Group*[1]), um consórcio que agrega organizações públicas e privadas com o propósito de definir normas que possam permitir que os repositórios de dados de expressão génica partilhem e troquem dados entre si [51]. O grupo MGED tem focado a sua intervenção em três áreas distintas:

- **O que armazenar:** específica que aspectos de uma experiência de *microarrays* necessitam de ser armazenados. Este é o objectivo da recomendação MIAME (*Minimum Information About a Microarrays Experiment*) [52].
- **Como descrever a experiência:** inclui os métodos experimentais e os dados resultantes. São utilizadas ontologias, nomeadamente o MGED-Ontology e GO

[1] http://www.mged.org

(*Gene Ontology*) [53], que consistem em vocabulários controlados com informações sobre relações entre genes, amostras e dados.

- **Como armazenar e partilhar os dados:** envolve as normas MAGE-OM *(Microarray Gene Expression - Object Model*) e MAGE-ML (*Microarray Gene Expression – Markup Language*) [54], que definem modelos de objectos e protocolos de comunicação, concretizando o MIAME e as ontologias apresentadas.

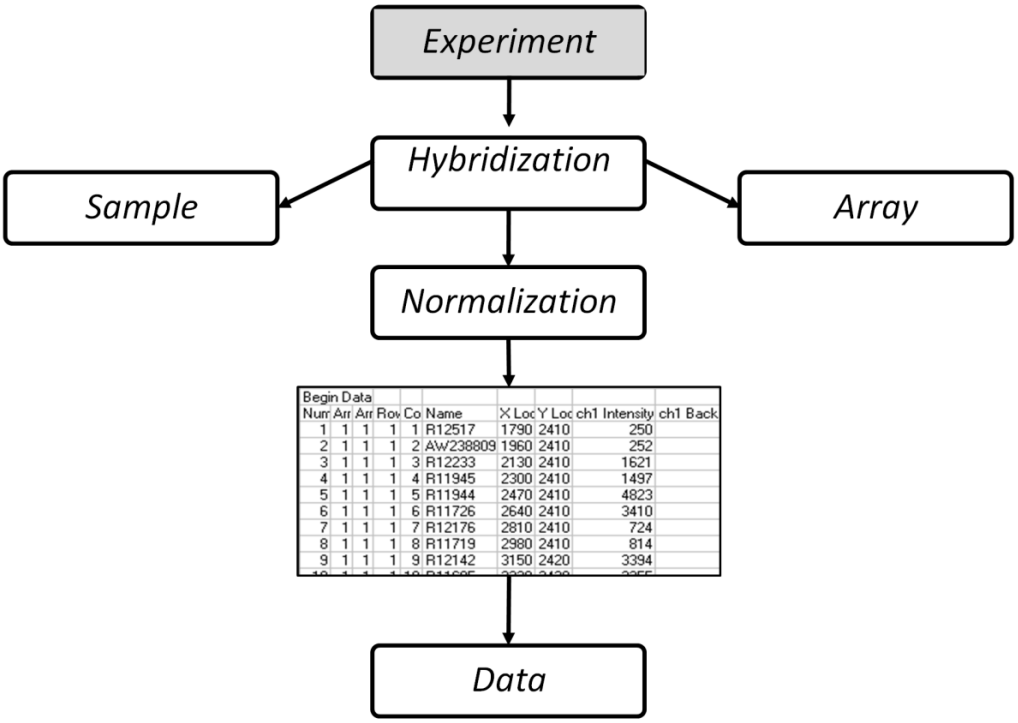
3.2.2 MIAME

A recomendação MIAME tem como finalidade principal responder à necessidade de uma anotação extensiva que possibilite a interpretação de uma experiência de *microarrays* [52]. Apesar de independente da plataforma, esta foi especialmente concebida para estudos de expressão genética.

A primeira motivação para criação do MIAME foi a necessidade de, para cada experiência, se registar informação suficiente, de modo que esta pudesse ser interpretada com o detalhe que permitisse a sua comparação com experiências similares e que possibilitasse a sua replicação. Por outro lado, o MIAME vinha viabilizar a pertinência da informação dever ser estruturada de forma a permitir a pesquisa, a análise e o processamento dos seus dados.

Na prática, obteve-se uma lista de recomendações que necessitavam de ser cumpridas para a correcta anotação da experiência. Como a Figura 3.3 ilustra, a recomendação MIAME encontra-se dividida em seis secções distintas:

- ***Experiment:*** Reporta os parâmetros globais da experiência;
- ***Array:*** Contém informação sobre o desenho do *microarray*, incluindo elementos relativos à localização e descrição de todas as sondas usadas;
- ***Sample:*** Descreve as amostras utilizadas, nomeadamente, a sua origem, os tratamentos aplicados e os protocolos de extracção e de marcação;
- ***Hybridization:*** Associa um *microarray* e um conjunto de amostras a uma experiência;
- ***Normalization:*** Traça os procedimentos de tratamento e de normalização dos dados obtidos;
- ***Data:*** Diz respeito às imagens, a especificações e a quantificações obtidas durante a experiência.



Begin Data

Num	Arr	Arr	Row	Co	Name	X Loc	Y Loc	ch1 Intensity	ch1 Back
1	1	1	1	1	R12517	1790	2410	250	
2	1	1	1	2	AW238809	1960	2410	252	
3	1	1	1	3	R12233	2130	2410	1621	
4	1	1	1	4	R11945	2300	2410	1497	
5	1	1	1	5	R11944	2470	2410	4823	
6	1	1	1	6	R11726	2640	2410	3410	
7	1	1	1	7	R12176	2810	2410	724	
8	1	1	1	8	R11719	2980	2410	814	
9	1	1	1	9	R12142	3150	2420	3394	

Figura 3.3: Estrutura hierárquica da recomendação MIAME constituída por seis classes.

3.2.3 MAGE-OM

A recomendação MIAME não fornece detalhes relativos ao armazenamento e à partilha de dados, pois apenas define genericamente a informação a armazenar. A informação técnica, sobre qual o formato em que os dados de *microarrays* devem ser armazenados, é definida pela norma MAGE-OM (MAGE - *Object Model*), proposta pelo grupo MGED [54]. O MAGE-OM é um modelo de dados centralizado, constituído por 132 classes, que se encontram distribuídas por 17 módulos. Cada classe representa um objecto ou evento, enquanto os módulos são usados para agrupar classes que partilham um mesmo propósito. O módulo *Array,* por exemplo, contém classes que descrevem *microarrays* individuais, incluindo informação detalhada do processo de construção.

Pese o modelo MAGE-OM ser do domínio público existe apenas uma única instância, o ArrayExpress [55]. A par do GEO [56] e do Cibex [57], o ArrayExpress é um dos mais relevantes repositórios públicos de estudos de expressão genética. Após a correcta inserção e validação dos dados, estes ficam publicamente disponíveis, podendo ser pesquisados por vários critérios, incluindo identificadores de genes, propriedades do estudo, investigador e organismo.

3.2.4 MGED-Ontology

A ontologia MGED consiste num conjunto de termos organizados de forma estruturada e que possibilitam uma correcta anotação de uma experiência de *microarrays* [44]. Esta foi desenvolvida com o objectivo de ser integrada nos restantes projectos do grupo MGED. Deste modo, existe uma consistência com a nomenclatura usada nas classes do MAGE-OM. Este desenvolvimento teve ainda como propósito a obtenção de uma ontologia que pudesse ser utilizada simultaneamente por investigadores, para anotarem os seus estudos, assim como por programadores, que desenvolvessem ferramentas bioinformáticas.

A separação entre ontologia e modelo de dados possui várias vantagens. Uma das mais relevantes é a possibilidade da ontologia poder ser actualizada, de forma a comportar novas técnicas ou métodos, sem que o modelo, em si, necessite de ser alterado.

Através da utilização de ontologias, é possível o enriquecimento semântico das experiências, sendo facilitada a sua interpretação, sem a existência de ambiguidade. Possibilita, ainda, o uso de critérios de pesquisa mais complexos, devido à forma estruturada dos dados.

3.2.5 Outras normas relacionadas

Apesar de abrangente, o conjunto de normas MAGE não consegue comportar toda a complexidade do domínio da biologia molecular. Na realidade, existem outros esforços, como o GOBO (*Global Open Biology Ontologies*) [58] e o GO (*Gene Ontology)* [53], na área da biologia, ou o UMLS (*Unified Medical Language System*) [59], usado na área médica.

O GOBO resulta de um esforço para transpor a metodologia usada na *Gene Ontology*, em que é a comunidade que constitui a maior fonte directa de informação, para outras áreas da biologia. Para tal, especialistas em ontologias estão a definir um conjunto de princípios que orientem o desenvolvimento de novas ontologias. O objectivo final é o de obter um conjunto interoperável de ontologias de referência na área biomédica.

Na área médica, e perspectivando o uso de *microarrays* como ferramenta de diagnóstico, informações sobre a morfologia e a patologia são fundamentais para a correcta interpretação de dados de expressão génica [60]. A título de exemplo, elementos sobre o estado do tumor ou referentes à medicação tomada são necessários para a compreensão dos valores de expressão de uma determinada amostra. Os dados destas amostras podem ser capturados através do UMLS [61] que é actualmente, na área médica, a mais completa colecção de termos incluindo informação relativa à sua classificação e ao seu significado.

3.3 Gestão de dados de *microarrays*

Devido à elevada quantidade de dados gerados numa experiência de *microarrays,* é imprescindível o uso de ferramentas dedicadas, para o correcto armazenamento e análise de dados [8]. Nos últimos anos, os sistemas de gestão laboratorial (LIMS - *Laboratory Information Management System*) foram re-desenhados com o fito de responder a esta nova necessidade. Estes sistemas, que remontam a meados da década de 60, tinham, originalmente, como principal objectivo substituir o caderno de laboratório, assim como permitir a anotação automática dos dados com informação proveniente de equipamentos laboratoriais. Presentemente estes sistemas são responsáveis pela gestão de todo o processo laboratorial, incluindo protocolos, amostras e materiais, de forma a permitir um fácil acesso, manuseamento e partilha da informação. Relativamente ao registo em papel, os LIMS, possuem vantagens notórias: um menor tempo de acesso, a possibilidade de existirem múltiplos acessos em simultâneo, ou a necessidade de um espaço físico inferior. Deve, ainda, ser tido em consideração que um LIMS é um sistema informático, logo torna-se possível ligá-lo directamente aos instrumentos laboratoriais, obtendo os dados de uma forma mais rápida e sem erros de transcrição manual.

3.3.1 Levantamento dos dados que necessitam de ser armazenados

No intuito de identificar os principais dados gerados numa experiência de *microarrays,* foi realizada uma avaliação exaustiva do fluxo de execução de uma experiência. Foram também considerados o tipo e a dimensão dos dados obtidos. Como resultado, foi obtida a seguinte lista:

- **Detalhes da experiência:** Aglomeram o conjunto da informação que descreve a experiência como um todo. Incluem título, lista de autores, condições experimentais, *microarrays*, amostras usadas e desenho experimental. Note-se que uma única experiência pode conter desde algumas dezenas até centenas de *microarrays*;
- **Protocolos:** Reportam, em detalhe, os procedimentos realizados nos diferentes passos da experiência;
- **Material biológico:** Indica qual a origem das amostras e qual o tratamento que lhes é aplicado;
- **Detalhes do *microarray*:** Descrevem a organização das sondas no *microarray*;
- **Imagens:** Correspondem à digitalização do *microarray* após a hibridação. Estas consistem em ficheiros TIFF com resoluções típicas de 7500 × 2200 pixéis e com aproximadamente 80 MB de dimensão;
- **Dados brutos:** Dizem respeito à saída do software de segmentação associado ao *scanner*. Para cada *spot* do *microarray,* este estima vários parâmetros que o

descrevem. Para um *microarray* com 10.000 *spots,* este ficheiro tem um tamanho aproximado de 6 MB;

- **Dados derivados:** Equivalem ao conjunto de ficheiros obtidos através da aplicação aos dados brutos de diferentes algoritmos de pré-processamento e normalização.

3.3.2 Avaliação de sistemas de gestão de dados laboratoriais

Os sistemas LIMS são, de facto, uma ferramenta imprescindível em qualquer laboratório de *microarrays*. O reconhecimento desta realidade fez com que, ao longo dos últimos anos, tenham sido propostas várias soluções, tanto no domínio público, quanto no comercial [41]. No entanto, a abrangência desta oferta apenas veio aumentar a dificuldade relativamente à escolha do sistema a usar. A principal questão é a pluralidade de parâmetros e de perspectivas sobre os quais os sistemas podem ser analisados, e, principalmente, o facto das características do sistema dependerem, em grande parte, das necessidades específicas do laboratório.

Com a finalidade de analisar a oferta de sistemas disponíveis, foi identificado um conjunto de parâmetros sobre os quais os sistemas LIMS mais usados pudessem ser avaliados. Os parâmetros seleccionados foram: tipo de instalação; licença de utilização; uso da norma MIAME; exportação para repositórios públicos; uso da ontologia MGED para anotação; trabalho colaborativo; partilha de dados entre utilizadores; requisitos de instalação, o que inclui o tipo de sistema operativo e o de sistema de gestão de base de dados necessários. Foi igualmente, especificado um conjunto de LIMS para *microarrays,* tendo em conta a sua popularidade: BASE, LAD, MARS, MaxD e MADAM.

O BASE (*BioArray Software Environment*[1]) [62] consiste num sistema *web,* desenvolvido na Universidade de Lund, Suécia, que se encontra disponível através da licença GNU. O BASE é apresentado como um sistema integrado de armazenamento e análise dos dados gerados por experiências de *microarrays*. Este sistema suporta a norma MIAME e a sua versão mais recente permite a exportação de dados para o ArrayExpress.

O sistema LAD (*Longhorn Array Database*[2]) [63] resulta numa versão de código aberto da SMD (*Stanford Microarray Database*). Este pretende ser uma solução simples, sem custos, aberta e eficaz para o armazenamento e análise de dados de *microarrays*. De acordo com os autores, as mais-valias do LAD são as provas já dadas do sistema que lhe serve de base e a possibilidade de armazenamento da imagem.

[1] http://base.thep.lu.se
[2] http://www.longhornarraydatabase.org

O MARS[1] (*Microarray Analysis, Retrieval, and storage System*) [64] consta de uma aplicação *web* que permite o armazenamento, visualização e análise de dados de *microarrays*. As principais características que diferenciam este sistema são o seu processo integrado de normalização, de anotação e de análise de dados e a possibilidade de disponibilizar *web services* para aceder aos dados do sistema através de aplicações externas tais como MatLab.

Ao contrário dos anteriores, baseados em *web*, o maxLoad2[2] [65] é um sistema monolítico, desenvolvido na Universidade de Manchester, que suporta a norma MIAME e permite a submissão de dados para repositórios públicos. Apesar do maxLoad2 apenas possibilitar o armazenamento dos dados, uma ferramenta adicional, o maxView, pode ser usada de forma a possibilitar a visualização e a análise dos dados. Uma funcionalidade interessante deste sistema é a possibilidade de se poder personalizar todas as interfaces de utilizador, uma vez que utiliza ficheiros de configuração para definir as interfaces. Estes ficheiros de meta-configuração são armazenados num servidor central, permitindo que alterações na aplicação sejam propagadas para os utilizadores, sem intervenção directa dos mesmos.

O sistema MADAM[3] (*Microarray DAta Manager*) [66] compõe-se de uma aplicação *desktop*, implementada em *Java*, que tem por objectivo auxiliar o investigador a registar todos os elementos necessários para a interpretação de estudos de expressão genética. Embora não possua funcionalidades de análise de dados, os autores disponibilizam um conjunto adicional de ferramentas que possibilitam esta mesma análise.

A Tabela 3.1 contém um levantamento realizado aos sistemas de gestão de dados laboratoriais [67].

3.3.3 Desafios na gestão de dados de *microarrays*

A análise apresentada permite verificar que a característica que categoriza, de forma mais evidente, os sistemas apresentados é o tipo de instalação: *desktop vs web*. As aplicações *desktop*, tais como o maxLoad2, defendem, como principal vantagem, a sua melhorada usabilidade, uma vez que o sistema é executado localmente e consequentemente, não existem atrasos causados pela transmissão dos dados. No entanto, este argumento parece perder a sua relevância à medida que os paradigmas *web* evoluem e tecnologias como o AJAX (*Asynchronous Javascript And XML*) permitem oferecer experiências de utilização semelhantes às aplicações *desktop*.

[1] http://genome.tugraz.at/software/MARS/MARS.html
[2] http://www.bioinf.manchester.ac.uk/microarray/maxd
[3] http://www.tm4.org/madam.html

Tabela 3.1: Resumo dos parâmetros usados na comparação dos sistemas de gestão de dados laboratoriais.

	BASE	**LAD**	**MARS**	**MaxD**	**MADAM**
Tipo de instalação	*Web*	*Web*	*Web*	Java Package	Java Package
Licença de utilização	GNU	OpenSource	Sem custos para fins académicos	OpenSource	Artistic License
MIAME	✓	✓	✓	✓	✓
Exportação para ArrayExpress	✓	✓	✓	✓	×
Ontologia MGED	✓	×	×	✓	×
Trabalho colaborativo	×	×	×	×	×
Partilha de elementos	×	×	×	×	×
Requisitos de instalação (S.O. /SGBD)	Linux / MySQL; PostgreSQL	Linux / PostgreSQL	Qualquer / Oracle	Qualquer / MySQL	Qualquer / MySQL

A favor das aplicações *web* existe ainda o argumento de que um servidor central, em alternativa a várias aplicações *desktop* instaladas em vários computadores, assegura um controlo melhorado sobre as políticas de segurança e de armazenamento dos dados. A opção por um servidor central pode ainda ser mais interessante, quando se pretende executar algoritmos computacionalmente complexos, uma vez que pode ser realizada uma melhor reserva dos recursos existentes e, consequentemente, conseguindo um menor tempo de resposta.

No que se refere à utilização de normas, conclui-se que todos os sistemas apresentados fazem uso da norma MIAME, o que reflecte a maturidade da mesma. Como consequência, verifica-se a existência de um conjunto comum de elementos capturados por todos os sistemas apresentados, tais como a descrição geral da experiência, o desenho do *microarray*, os detalhes do material biológico, os protocolos e a informação referente às hibridações. Relativamente à capacidade de exportar dados para repositórios públicos, e apesar da sua reconhecida importância, nem todos os sistemas existentes disponibilizam estas funcionalidades. O mesmo sucede com a possibilidade de usar a ontologia MGED para anotar as experiências.

As ferramentas de análise de dados são essenciais no processo de resposta à questão biológica. Dos sistemas apresentados a maioria inclui, pelo menos, ferramentas de controlo de qualidade e algumas funcionalidades de análise exploratória. Porém, coloca-se a questão de como gerir a crescente necessidade de novas ferramentas. Alguns sistemas adoptaram o uso de *plug-ins*, em que a adição de uma funcionalidade não implica alterações no código original. Outras optaram por criar ferramentas de análise independentes e especializadas.

Nenhum dos sistemas possibilita a conversão dos dados existentes no LIMS para formatos compatíveis com os das ferramentas de análise mais conhecidas. Esta funcionalidade permitiria, em caso de necessidade ou por preferência pessoal, a utilização de uma ferramenta externa.

Os sistemas apresentados tinham como denominador comum o facto de não possuírem custos de licença. No entanto, alguns possuem certas restrições. Por exemplo, o MARS e o MADAM apenas podem ser utilizados em contextos académicos, enquanto o BASE, que se rege pela licença GNU, pode ser usado e alterado, sendo que todas as alterações têm de ser disponibilizadas.

A escolha de uma ferramenta deve, ainda, ter em consideração factores de ordem técnica, nomeadamente, a facilidade de instalação, personalização e manutenção. De facto a customização de ferramentas abrangentes pode ser morosa, chegando, por vezes, a impossibilitar a actualização para novas versões, ou potenciando a introdução de erros que causem instabilidade no sistema.

3.4 Proposta, desenvolvimento e avaliação de um sistema de gestão de dados de *microrrays*

Conquanto seja elevada a oferta de soluções LIMS para *microarrays,* manifestas limitações podem ser identificadas nestes sistemas. São exemplo disso mesmo a falta de usabilidade, escalabilidade e facilidade de instalação e manutenção. Foi neste contexto, com o objectivo de colmatar estas lacunas, que o sistema Mind (*Microarray Information Database*) foi proposto [67, 68]. O Mind consiste numa plataforma que integra, de forma flexível, várias ferramentas que possibilitam a gestão e a análise de dados de *microarrays,* encontrando-se disponível, para uso não comercial, em http://bioinformatics.ua.pt/mind. A Figura 3.4 apresenta a página inicial do sistema.

As principais inovações deste sistema são a sua interface intuitiva e fácil de usar, a possibilidade de realização de trabalho colaborativo, o controlo de qualidade dos dados automatizado, a submissão directa a repositórios públicos, tais como ArrayExpress ou GEO (*Gene Expression Omnibus)* e uma integração transparente com aplicações externas.

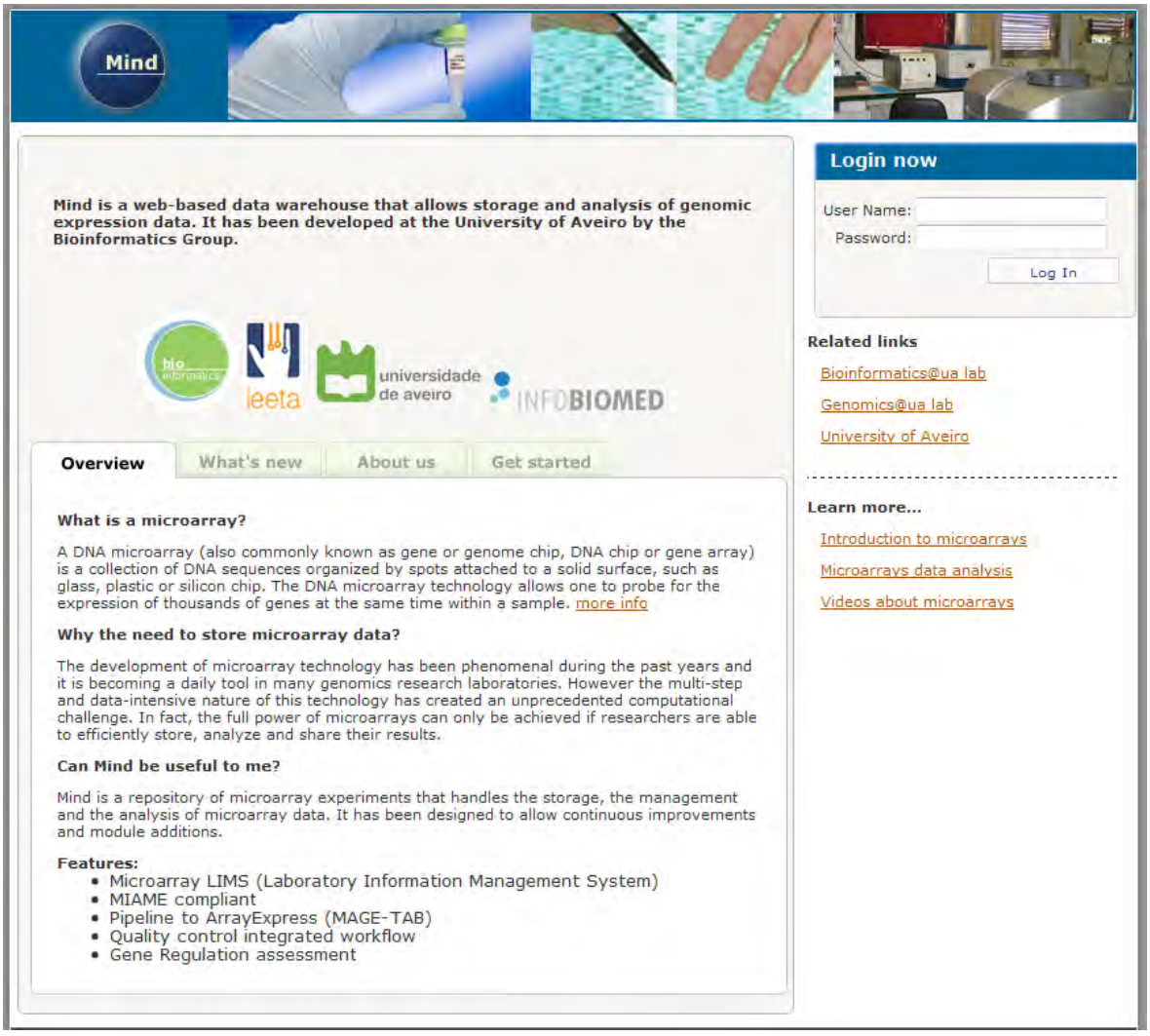


Figura 3.4: Página inicial do sistema Mind.

O desenvolvimento do Mind possuía um duplo propósito: primeiro, a obtenção de uma plataforma base, sobre a qual pudessem ser testadas algumas das ideias propostas nesta tese; segundo, a resposta a uma necessidade específica do laboratório de *microarrays* a ser instalado na Universidade de Aveiro. Isto permitiu o contacto directo com futuros utilizadores e, mais importante, a possibilidade de obter dados imprescindíveis a uma correcta validação do sistema.

O desenvolvimento do Mind pautou-se por um procedimento interdisciplinar, que contou com a participação activa dos futuros utilizadores do sistema de forma a obter-se um consenso quanto ao modelo de navegação e aos dados a armazenar e a processar. Este processo de desenvolvimento fez uso de uma metodologia que incluiu a análise de casos de estudo experimentais, que permitissem compreender o fluxo de informação numa experiência de *microarrays*. Como resultado, foram identificadas as tarefas mais comuns e foram propostos vários procedimentos de normalização, de modo a melhorar a usabilidade do sistema. Adicionalmente, foram, ainda, considerados os princípios de usabilidade de

interfaces, propostos por Jakob Nielsen e Ben Shneiderman [69]: facilidade de aprendizagem, eficiência, facilidade de memorização, tratamento de erros e satisfação.

De seguida, são dados a conhecer o modelo de navegação proposto, o paradigma de interacção implementado, o modelo de dados usado e a arquitectura do sistema. Por fim, são apresentados os principais indicadores de utilização e de validação do sistema Mind.

3.4.1 Modelo de navegação

O modelo de navegação visa definir os aspectos funcionais a considerar, assim como a sua localização na árvore de navegação (Figura 3.5). O correcto estabelecimento do modelo de navegação é essencial para garantir a consistência da navegação, de maneira a que esta reflicta o normal fluxo de dados gerados numa experiência.

No desenvolvimento do modelo proposto, foi dada especial atenção ao balanceamento entre o número de opções disponíveis por nível e o número de níveis existentes, tendo-se procurado progredir até ao limite imposto pela salvaguarda do normal fluxo da experiência. Com base em testes de usabilidade, que empregaram um protótipo funcional e que tiveram em consideração várias estruturas alternativas, foram estabelecidos quatro elementos de topo, sobre as quais todas as funcionalidades a implementar foram enquadradas: LIMS, análise de dados, ferramentas e repositório público.

LIMS

O elemento LIMS contém todos os instrumentos funcionais relacionados com a inserção e a pesquisa de dados numa experiência. Na análise realizada previamente aos sistemas LIMS existentes, e como consequência de todos os sistemas serem concordantes com a norma MIAME e da maioria deles possibilitar a exportação para o ArrayExpress, verificou-se uma uniformidade nas funcionalidades base oferecidas. Se bem que, neste campo, não existisse grande espaço para inovação, na medida em que os dados necessitam, efectivamente, de ser inseridos, as mais-valias do sistema proposto apresentam-se ao nível da organização e da facilidade de acesso às mesmas opções. Assim, e considerando os parâmetros que nortearam a análise efectuada anteriormente, conclui-se serem as principais funcionalidades disponibilizadas pelo Mind:

- Detalhes da experiência, incluindo a relação entre *microarrays* e amostras;
- Resultados da hibridação individual, integrando a associação à imagem das respectivas quantificações;
- Protocolos experimentais e de tratamento de dados;
- *Microrrays*, do esquema ao elemento físico;
- Processo de preparação das amostras biológicas, abrangendo: material biológico, amostras, aplicação de tratamento, extracção do material genético e marcação.

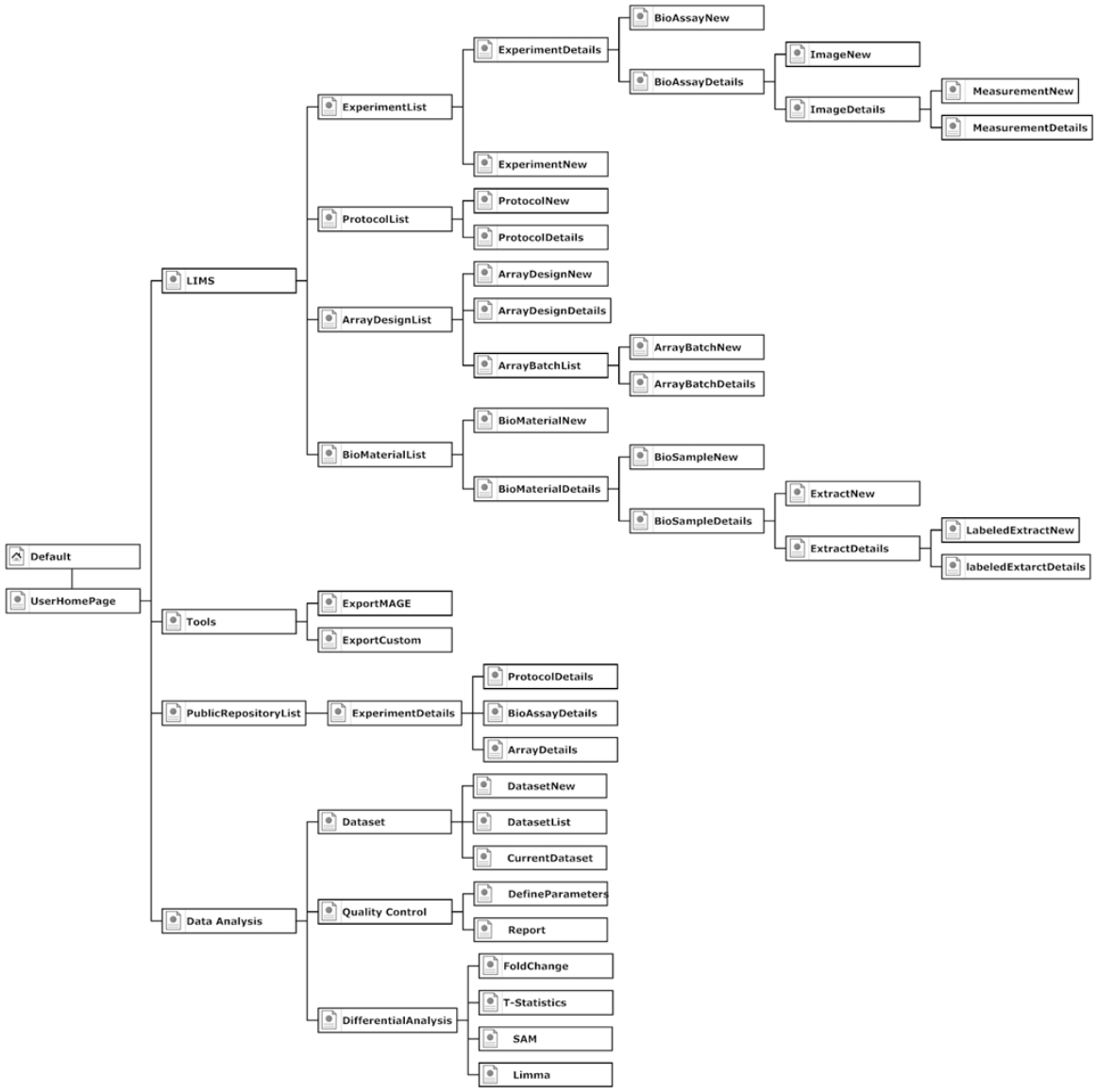


Figura 3.5: Modelo de navegação do sistema Mind.

Para integração de todas estas funcionalidades, a estratégia mais básica consistiria na utilização de um único nível. Porém, aproxima-se desta a implementada pelo sistema MaxD, anteriormente apresentado, e revela indesejáveis limitações. A interface principal deste sistema apresenta 22 pontos de entrada para as diversas funções disponíveis. Esta estratégia apresenta, no entanto, notórias lacunas, na medida em que, embora as opções se encontrem todas disponíveis em simultâneo, a escolha de algumas supõe, não raro, a selecção prévia de outras. Por exemplo, apenas faz sentido a aplicação de um tratamento a uma amostra caso esta já se encontre na base de dados. O MaxD encontra-se ainda condicionado por uma certa redundância ao nível da gestão dos protocolos. Existem cerca de 7 tipos de protocolos e, de acordo com as normas, os procedimentos de anotação são relativamente comuns. Em alternativa a existirem 7 entradas distintas para protocolos, no Mind propõe-se a existência de uma única entrada que integre todo o tipo de protocolos disponíveis.

Outra estratégia possível consiste na agregação das opções anteriores num conjunto limitado de formulários. No entanto, esta estratégia adoptada pelo sistema LAD, apresenta problemas óbvios ao requerer o preenchimento de longos formulários, o que, consequente, implica a impossibilidade de reutilização dos elementos inseridos.

No sistema Mind, é proposta uma metodologia que constitui uma melhoria, quando comparada com as anteriormente apresentadas. Foram definidos, dentro da categoria do LIMS, quatro elementos distintos, responsáveis pela agregação dos restantes: experiência, protocolo, *microarray* e material biológico. Estas entradas, embora interdependentes, pois a completa anotação de uma experiência necessita de uma relação com um conjunto de *microarrays* e de amostras biológicas, beneficiam de uma inserção independente. Esta opção tem como propósito evitar a selecção de elementos cuja inserção não seja válida, assim como a possibilidade de reutilizar elementos correspondentes a *microarrays*, a protocolos e a amostras, entre experiências, sem a efectiva necessidade de re-inserção de toda a informação.

Cada experiência está associada a um conjunto de *BioAssays*, relativos às hibridações realizadas, e, para cada um, existe uma *Image* que possui um ou mais *Measurements* associados. O elemento protocolo é, simplesmente, constituído pela possibilidade de listar, ver detalhes e criar um novo.

Para descrever o desenho do *microarray* e da sua instanciação, foi primeiro, criada uma entidade designada *ArrayDesign,* na qual é especificado o *layout* de cada *microarray*, o que viabiliza vários *ArrayBatch à posteriori,* e em associação com cada *ArrayDesign*. Esta divisão assenta no princípio que define que, após o estabelecimento do desenho do *microarray*, este permanece relativamente imutável nas impressões futuras e que em cada iteração de impressão é criado mais do que um *microarray*.

A descrição do material biológico corresponde a uma cascata de elementos que necessitam de ser inseridos de forma a anotar correctamente o *labelled extract* usado na hibridação. Esta é constituída por: *bioMaterial*, *bioSample*, *extract* e *labelled extract*. Esta divisão tem por objectivo possibilitar a reutilização de elementos sem necessidade de inserir todos os dados. Por exemplo, numa experiência com dois microarrays que usam o mesmo conjunto de amostras em *dye swap*, após inserção do primeiro elemento, é apenas necessário criar dois elementos *labelled extract* com os marcadores invertidos.

Repositório partilhado

O repositório permite a partilha de experiências já concluídas entre todos os utilizadores do sistema. As suas funcionalidades incluem a listagem de todas as experiências existentes, assim como a visualização dos seus detalhes, incluindo *arrays*, amostras biológicas e protocolos usados.

Tools

O elemento *tools* agrega todas as ferramentas adicionais existentes no sistema. Na modelação inicial, foi definida a exportação de dados para o formato MAGE-TAB, bem como a possibilidade de exportar para formatos de dados compatíveis com ferramentas externas de análise e de visualização de dados.

Análise de dados

Este elemento agrega ferramentas que possibilitam a análise dos dados inseridos. Após a criação do *dataset* com os dados a analisar, pode-se aceder às ferramentas de controlo de qualidade ou de identificação de genes diferencialmente expressos. A descrição detalhada do modelo de navegação proposto para a análise dos dados é apresentada no Capítulo 5.

3.4.2 Usabilidade e paradigma de interacção

Apesar do modelo de navegação anteriormente apresentado ter sido concebido de forma a promover a usabilidade do sistema, este, por si, não a garante, sendo necessária uma atenção especial na implementação do mesmo. Esta condicionante implicou a correcta organização das opções de navegação na interface, de modo a assegurar a sua acessibilidade.

Ao nível da usabilidade, uma das principais mais-valias desta aplicação consiste no paralelismo entre o fluxo de dados de uma experiência de *microarrays* e o respectivo mapeamento com as funcionalidades do Mind (Figura 3.6). A solução obtida possibilita com uma única selecção, o acesso directo aos elementos de topo, através do ícone  . A Figura 3.7 ilustra o resultado da selecção deste ícone sobre o elemento de topo *Experiment.* Por sua vez, o ícone  permite o acesso aos últimos dez elementos inseridos, tal como a Figura 3.8 ilustra para o caso particular das imagens.

O recurso à representação esquemática teve como principal propósito o aumento da facilidade de aprendizagem e de memorização na utilização do sistema. Estes pontos são de especial importância, pois são esperados utilizadores esporádicos que apenas se encontrem de visita ao laboratório. No entanto, uma vez que o paradigma usado minimiza o número de iterações, a eficiência dos utilizadores regulares no cumprimento das suas tarefas sai também beneficiada.

A barra superior, visível em todas as interfaces após autenticação, corresponde aos dados de primeiro nível de interacção. A presença constante desta barra possibilita um fácil acesso às principais opções do sistema independentemente da actual localização do utilizador. A sua disposição teve ainda em consideração futuras expansões do sistema, ao reservar espaço para mais elementos.

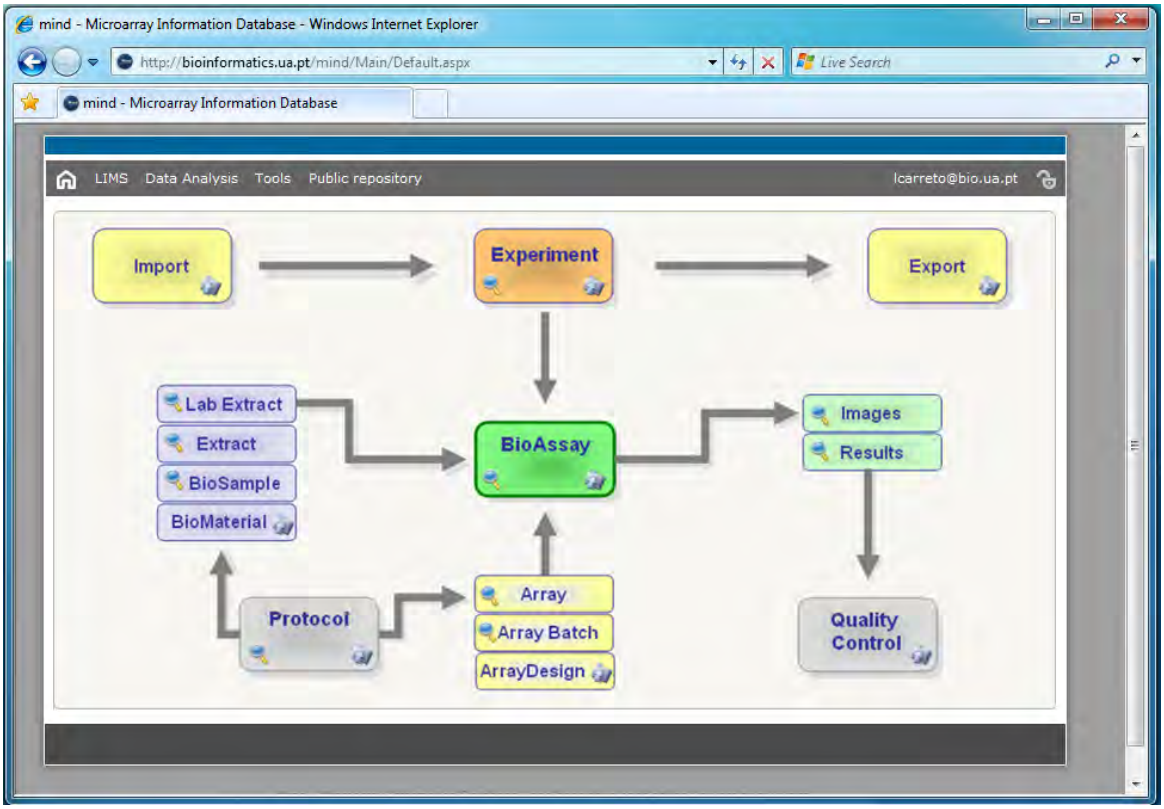


Figura 3.6: Interface principal após autenticação.

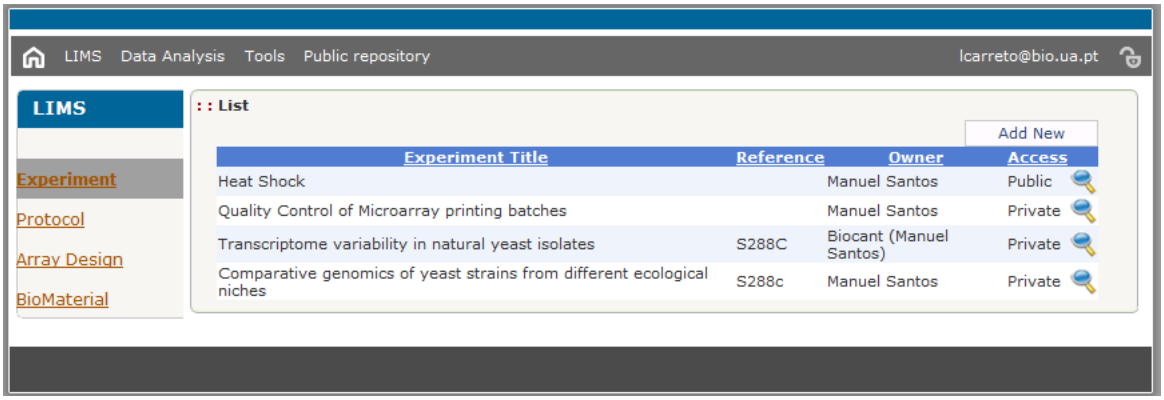


Figura 3.7: Exemplo de listagem de experiências.

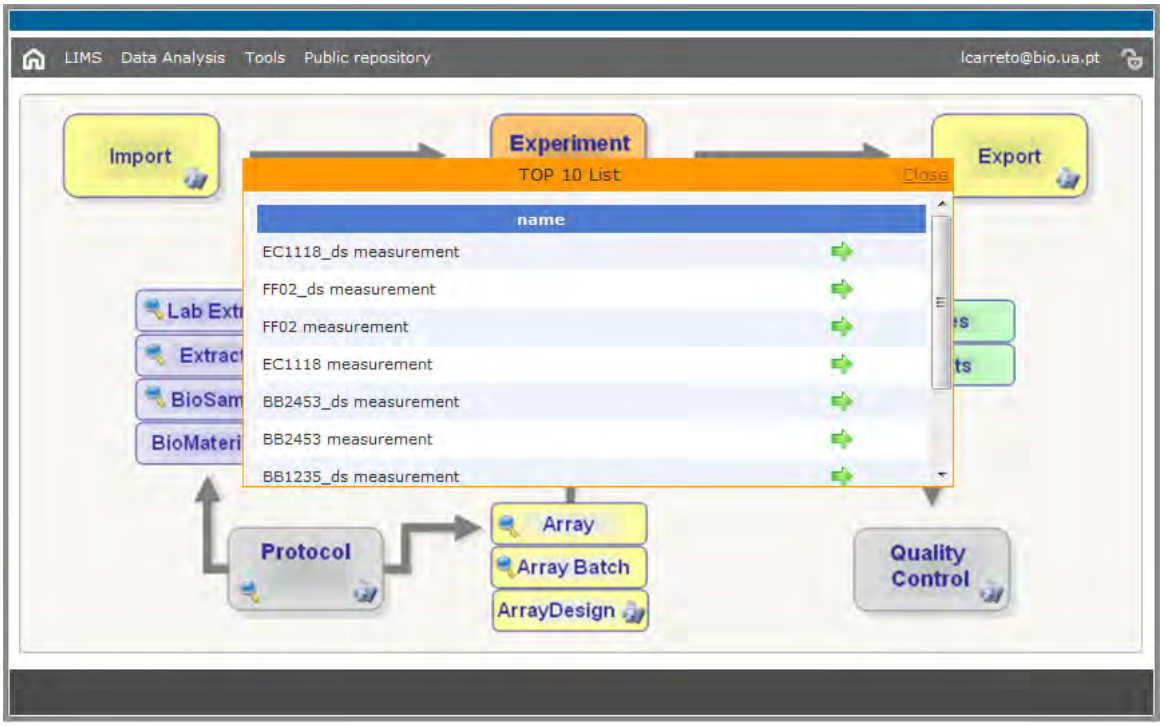


Figura 3.8: Exemplo da lista das dez últimas imagens inseridas.

Os elementos de nível 3 são dispostos de forma sistemática na barra lateral esquerda. Os restantes níveis não possuem um acesso directo na medida que estão relacionados com as entidades de nível superior que os englobam.

No caso concreto do material biológico, que se decompõe em quatro subníveis, de modo a facilitar a usabilidade, optou-se por incluir o acesso a todos os níveis a partir de uma única interface. Nesta é facultada a navegação entre todos os elementos da hierarquia, possibilitando a criação de novos elementos. Com a finalidade de evitar problemas de localização na hierarquia, é ainda disponibilizada, no topo do painel de navegação, informação relativa ao caminho.

Embora constitua a estrutura base de organização dos dados, a existência de três níveis a partir do *Experiment* pode ser considerada como penalizadora da usabilidade. Contornar este eventual problema, sem comprometer a estrutura e a organização do modelo de navegação, resultou na adopção de um conjunto de estratégias ao nível da implementação das interfaces.

A necessidade de anotar todos os procedimentos, se bem que essencial para a validade dos dados, tende a ser morosa. Para reduzir o tempo dispendido, utilizaram-se, então, mecanismos que permitiram a minimização do trabalho do utilizador. Foram consideradas a clonagem de elementos já existentes (*Samples*), o reaproveitamento de elementos que se revelassem constantes (*Protocols*), assim como a divisão de passos mais complexos em

cascatas de passos mais simples em que fosse realizado o reaproveitamento de elementos (*Biomaterial*).

Relativamente à gestão de erros, várias estratégias foram empregues. Em primeiro lugar, as incongruências ao nível da base de dados são evitadas através da obrigatoriedade de preenchimento de campos. Em segundo, todos os erros resultantes de acesso a componentes externos, como a base de dados ou o sistema de ficheiros, são capturados pelo programa e registados, sendo depois apresentada uma mensagem simplificada ao utilizador.

3.4.3 Modelo de dados

De forma a assegurar o correcto armazenamento dos dados, é usada uma base de dados relacional cujo modelo reflecte o fluxo de dados de uma experiência de *microarrays*. O desenvolvimento deste modelo foi fortemente inspirado na recomendação MIAME e no MAGE-OM. Apesar do MAGE-OM não ter sido concebido com o propósito de servir de suporte a um sistema LIMS, muitas das estratégias usadas puderam efectivamente ser aproveitadas. A principal diferença reside no facto deste apresentar várias concretizações ao nível das classes disponíveis. Assim, o modelo proposto no Mind possui cerca de 30 classes, enquanto o do MAGE-OM contém 137. Removendo grande parte da generalização encontrada no MAGE-OM, é possível obter um modelo bastante mais fácil de manter, sem perder a capacidade de armazenamento dos detalhes da experiência.

A Figura 3.9 apresenta o modelo de dados implementado no sistema Mind, agrupado em quatro secções: a) experiência; b) *microarray*; c) hibridação e d) dados relativos à amostra laboratorial.

A experiência, o elemento de topo, é constituída por um conjunto de *BioAssays,* que são usados para conduzir um determinado estudo. Neste, é armazenada uma descrição do estudo, a identificação dos responsáveis, assim como quais as condições experimentais que estão a ser testadas.

Cada *BioAssay* agrega um *Physical Array*, um conjunto de *Labelled Extracts* e, ainda, o protocolo de hibridação. Por sua vez, cada *Labelled Extract* é o resultado de uma cascata de procedimentos que começam com a definição do material biológico (*BioMaterial*) do qual uma amostra é retirada (*Sample*) e sobre a qual um ou mais tratamentos (*Treatments*) são aplicados. Finalmente, a amostra é marcada, de forma a obter-se o *Labelled Extract*.

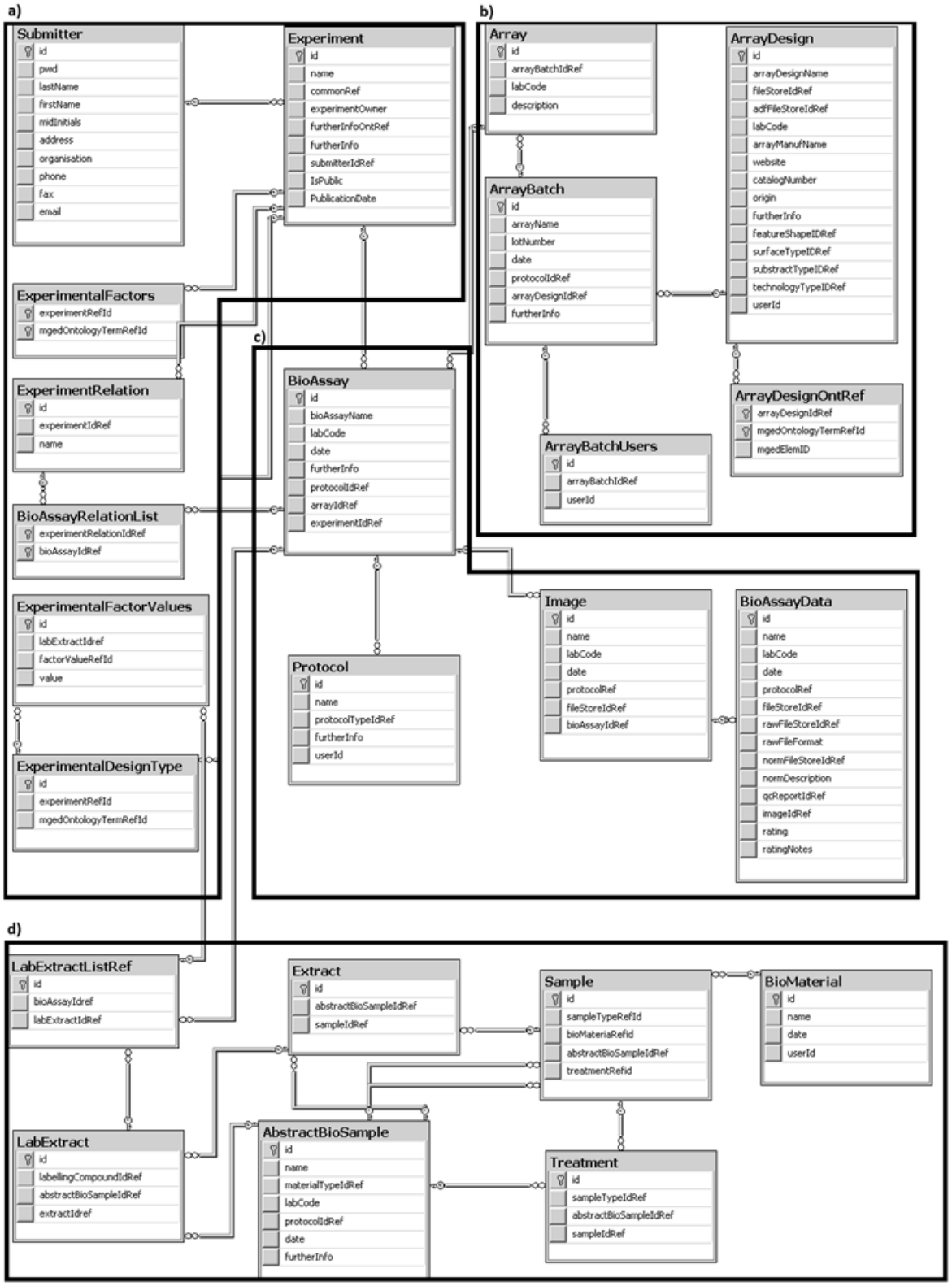


Figura 3.9: Modelo de dados do Mind agrupado em quatro secções: a) dados relativos à experiência; b) dados relativos ao *microarray*; c) dados associados com o resultado de cada hibridação, incluindo o *bioassay*, a imagem e os dados resultantes da quantificação da imagem; d) dados relativos à cascata de anotação de uma amostra laboratorial (*BioMaterial -> Sample -> Extract -> Labelled exract*).

O registo dos dados referentes ao *microarray* encontra-se dividido em três partes: *PhysicalArray*, *ArrayBatch* e *ArrayDesign*. O *ArrayDesign* consiste na descrição da organização das sondas de um *microarray*. É apenas um elemento conceptual, pois não tem correspondência com nenhum elemento físico. O *ArrayBatch* corresponde a um conjunto de *microarrays* que partilham o mesmo *ArrayDesign* e que pertencem à mesma série. O *PhysicalArray* constitui o *microarray* individual a ser usado numa hibridação.

Após a hibridação, uma ou mais imagens são obtidas e, para cada, vários métodos podem ser usados para quantificação dos valores de expressão. O Mind permite ainda o registo dos parâmetros usados na digitalização e no armazenamento dos ficheiros originais e dos processados com os valores de expressão.

3.4.4 Armazenamento da MGED Ontology

No caso dos microarrays, existem vantagens evidentes na separação entre a ontologia e o modelo de dados, sendo, no entanto, indispensável o seu uso para a completa e correcta anotação dos dados inseridos. No sistema Mind, a MGED Ontology é armazenada localmente, de forma a assegurar a existência de uma relação directa de cada elemento com o termo da ontologia que o descreve.

À data de início de desenvolvimento do Mind, não existia nenhuma solução normalizada disponível para armazenamento da ontologia MGED em base de dados. Foi, então, proposto e implementado um esquema relacional composto por quatro tabelas que respondia aos requisitos existentes. Este esquema, apresentado na Figura 3.10, encontra-se em uso no sistema Mind.

A ontologia é disponibilizada no formato OWL, tendo sido, depois, convertida para viabilizar a sua inserção nas quatro tabelas apresentadas. A tabela principal, *MAGEOntologyTerm*, contém todos os termos existentes na ontologia, independentemente do tipo. Para cada termo, existe uma relação com a tabela *OntologyTermType* que especifica o tipo do termo anterior. Os tipos mais comuns são as classes, as propriedades e os indivíduos. Sendo a ontologia uma estrutura hierárquica, a relação entre os termos é definida pela tabela *MGEDOntologyRelationship* e, também, pela tabela *OntologyRelationshipType,* que indica o tipo de relação existente. Por exemplo, a classe *Sex*, subclasse da classe *BioMaterialCharacteristics*, possui, entre outros, os elementos *male* e *female*. Neste exemplo, os termos *BioMaterialCharacteristics*, *Sex*, *male* e *female* são armazenados na tabela *MGEDOntologyTerm*, indicando a tabela *OntologyTermType* o tipo de termo.

Consciente da necessidade de armazenamento da ontologia em base de dados o grupo MGED promoveu, no ano de 2007, um *programming jamboree* em Seattle, em que foi, precisamente, endereçada esta questão. O esquema anterior foi apresentado nesta mesma reunião, tendo servido de base à solução actualmente proposta pelo grupo MGED.

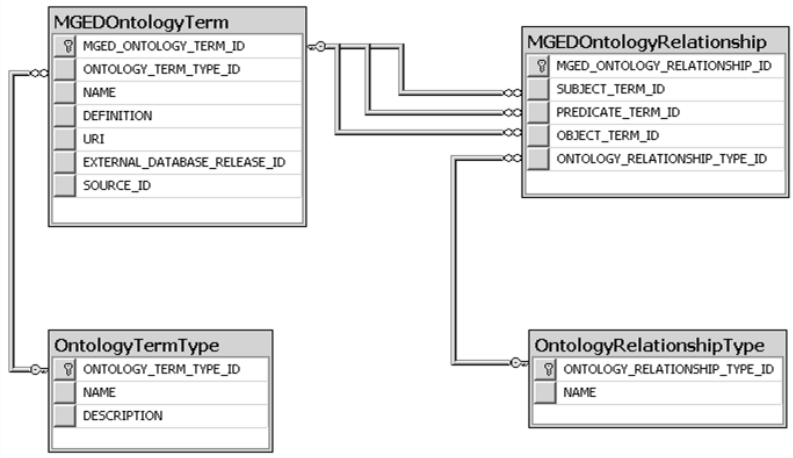


Figura 3.10: Modelo de dados de suporte à ontologia MGED.

3.4.5 Arquitectura

A arquitectura utilizada é baseada em *thin clients*, o que impõe poucos requisitos aos clientes (apenas um *web browser*). Essa arquitectura compreende três camadas fundamentais: camada de apresentação, camada de lógica e camada de dados. Os clientes comunicam apenas com a camada de apresentação, sendo a camada de lógica, que se encontra no servidor, a única que tem acesso à base de dados. A escolha desta arquitectura permitiu a concretização de alguns dos requisitos, tais como a facilidade de desenvolver um sistema distribuído, escalável, seguro e de alta disponibilidade, além de facilitar a sua manutenção.

A camada de apresentação (Figura 3.11.a) é responsável pela interacção com o utilizador. Esta camada foi desenvolvida utilizando a plataforma *.Net*, na linguagem C#, HTML e *JavaScript*. O conjunto de tecnologias denominado AJAX foi ainda usado em várias interfaces, para tratar os pedidos do servidor, melhorando a experiência de utilização. No desenvolvimento desta camada, foram tidos em consideração vários factores, como o controlo de acesso, o controlo de visibilidade e o fluxo de actividades. O controlo de acesso garante que apenas os utilizadores registados possam aceder ao sistema, enquanto o controlo de visibilidade restringe o acesso aos dados armazenados apenas aos utilizadores que possuem permissões. O fluxo de actividades evita o acesso a qualquer página do sistema apenas através do seu endereço. A questão é, que nos sistemas *web* convencionais, nada impede o utilizador de aceder às páginas numa ordem diferente da pré-definida, o que constitui uma potencial falha de segurança do sistema. O controlo de fluxo vem, justamente, determinar quais as sequências de acessos permitidas.

Os pedidos realizados na camada de apresentação, após validados, são transmitidos à camada lógica (Figura 3.11.b), responsável pelo acesso à base de dados e pelo processamento dos mesmos. Esta camada possui quatro componentes: LIMS, análise de dados, importação/exportação de dados e repositório público. Embora toda a lógica do

sistema esteja em C#, a linguagem R também foi utilizada no desenvolvimento dos módulos de análise de dados. A principal motivação para o uso da linguagem R foi a elevada quantidade de bibliotecas de análise de dados de *microarrays* disponibilizadas pelo projecto BioConductor [70], que vieram possibilitar a integração transparente de novas funcionalidade de análise.

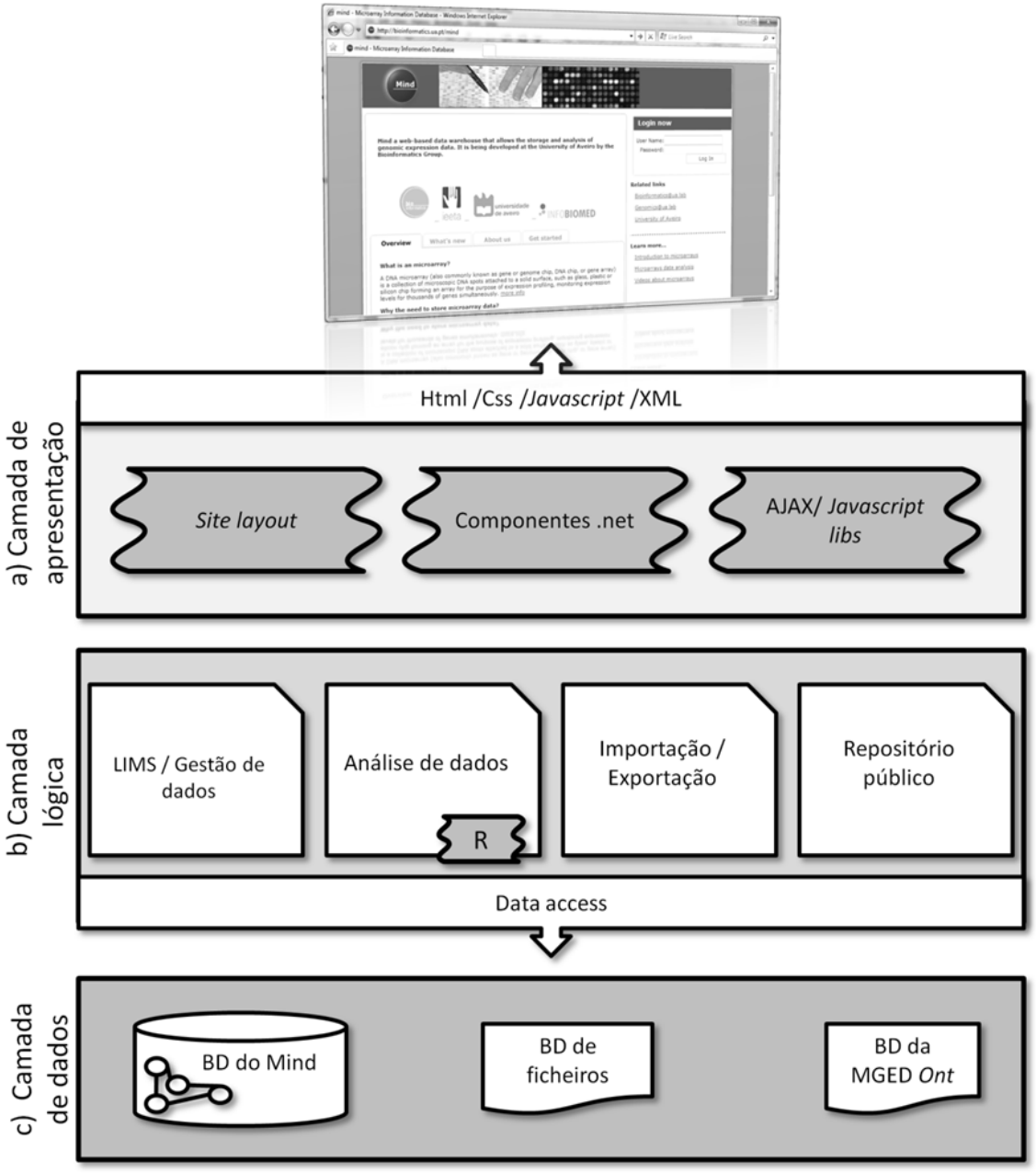


Figura 3.11:Arquitectura do sistema Mind dividida em três camadas: a) camada de apresentação b) camada lógica e a c) camada de dados.

A terceira camada é responsável pelo armazenamento dos dados (Figura 3.11.c). No Mind, existem múltiplos repositórios de dados, dos quais se destacam a base de dados de experiências, o armazenamento de ficheiros e o armazenamento de ontologias. Para armazenar os dados, foi criada uma base de dados em *SQL server*, que implementa o modelo de dados anteriormente apresentado. De modo, a poder relacionar directamente os dados, armazenados na base de dados com a sua respectiva anotação, e os termos da ontologia MGED, optou-se por juntá-los fisicamente numa única base de dados.

O programa Mind é composto por 37 classes que auxiliam no acesso à base de dados e que e que encapsulam a complexidade, por 56 páginas *aspx* e por 18.000 linhas de código C#, 14.000 linhas HTML/*aspx* e cerca de 500 linhas de *scripts* em R. Este encontra-se disponível em http://bioinformatics.ua.pt/mind.

3.5 Uso do sistema Mind

O desenvolvimento do sistema Mind iniciou-se em 2005, tendo sido concluído em Julho desse ano o primeiro protótipo funcional, que endereçava os principais requisitos. Esta versão teve como propósito testar diferentes metodologias de interacção, de maneira a verificar qual a mais adequada para cada um dos problemas apresentados. Nesta fase, foi identificada a dificuldade em balancear correctamente o modelo de navegação, entre o número de níveis e o número de entidades por nível.

O resultado dos testes realizados no protótipo inicial foi usado no desenvolvimento da versão final, colocada em produção em Maio de 2006, através do endereço http://bioinformatics.ua.pt/mind. Devido à especificidade da aplicação, nunca foi espectável um elevado número de utilizadores, no entanto, era de esperar que cada utilizador possuísse um conjunto elevado de dados.

Em Setembro de 2009, o sistema contava com cerca de 15 utilizadores, dos quais 5 possuíam, pelo menos, uma experiência inserida. Estes 5 utilizadores provêm de 3 unidades de investigação portuguesas. Estão armazenadas no Mind, 11 experiências, que totalizam 225 hibridações. No total, o espaço ocupado pela base de dados é de 12 GB.

Recentemente o Mind passou também a ser usado enquanto sistema de suporte à Rede Nacional de Genómica, projecto financiado pela Fundação para a Ciência e Tecnologia (REEQ/737/BIO/2005) que visa promover o desenvolvimento de redes de excelência em transcriptómica, proteómica, bioinformática e biomatemática. Assim, é espectável que o sistema continue a crescer, na medida em que a alocação do laboratório de *microarrays* já está completa até ao final de 2010.

3.6 Sumário

Neste capítulo foi apresentado o fluxo de dados típico de uma experiência de *microarrays,* assim como um levantamento das normas, das ontologias e dos sistemas de gestão de dados existentes. Verificou-se ainda a importância que o grupo MGED detém no processo de desenvolvimento de normas que possibilitam a consistência entre os sistemas LIMS e que permitem a efectiva partilha de dados. Aos vários sistemas LIMS disponíveis, foi feita uma análise que possibilitou a detecção de várias falhas, tais como a falta de usabilidade, de escalabilidade e de facilidade de instalação e de manutenção. Foi, então, proposto e desenvolvido o sistema Mind, que pretende colmatar as lacunas encontradas nos sistemas existentes.

As principais inovações do sistema apresentado consistem na sua interface intuitiva e fácil de usar, na possibilidade de trabalho colaborativo, no controlo de qualidade dos dados automatizada, na submissão directa para repositórios públicos, tais como ArrayExpress ou *Gene Expression Omnibus* (GEO) e numa integração transparente com aplicações externas.

O sistema Mind é uma alternativa às soluções actualmente disponíveis. Com o seu desenvolvimento pretendeu-se satisfazer uma necessidade do laboratório de *microarrays* da Universidade de Aveiro, assim como obter uma plataforma base sobre a qual pudessem ser testadas algumas das ideias propostas nesta tese. Apesar de não ter sido realizada uma avaliação formal da interface, foram recolhidas várias impressões entre os utilizadores do sistema. Por fim, a quantidade de dados armazenados e as estatísticas de uso do sistema comprovam a sua validade.

Capítulo 4

4 Partilha de dados de *microarrays*

Um dos princípios da investigação científica é a possibilidade das conclusões obtidas poderem ser verificáveis e reprodutíveis [71, 72]. Para que tal seja possível, é necessário que o resultado dos estudos realizados, assim como todos os dados associados, sejam registados e disponibilizados. No caso dos estudos de expressão génica em larga escala, é fundamental partilhar os ficheiros com os níveis de expressão e a informação que descreve a experiência, segundo a norma MIAME.

De modo a evitar constrangimentos na pesquisa e no acesso a estes dados, foram construídos vários repositórios centrais de armazenamento. Estes têm como principais objectivos evitar a dispersão dos resultados por vários servidores, assegurar a correcta formatação dos dados e garantir a disponibilidade e perpetuidade das experiências. Existem actualmente três repositórios (ArrayExpress [55], GEO: Gene Expression OmniBus [56] e Cibex [57]) que, apesar de utilizarem estratégias distintas e de terem como público-alvo diferentes comunidades geográficas, possuem um mesmo princípio comum: armazenar de forma centralizada e pública os dados de experiências de expressão génica. De facto, estes repositórios constituem uma valiosa fonte de dados para diferentes perfis de utilizadores, desde experimentalistas, interessados em estudar técnicas utilizadas por outros investigadores, a biólogos, motivados pelos estudos de genes específicos, ou a estatísticos, que pretendam obter dados de várias experiências para testar os seus próprios algoritmos.

Apesar da prevalência destes repositórios enquanto contentores globais de dados de *microarrays*, o seu interesse é limitado, quando a questão se coloca na partilha de dados antes da conclusão da experiência. Isto acontece porque estes foram planeados tendo em consideração que a experiência já se encontrava terminada. Sucede que, em várias, situações o trabalho que conduz a uma publicação é realizado por vários investigadores, que podem, ou não, pertencer ao mesmo laboratório. No caso destes utilizarem o mesmo sistema de gestão de informação e assumindo que este tem capacidades de trabalho

colaborativo, a solução é relativamente simples. No entanto, vários desafios se colocam quando esta situação não se verifica.

Neste capítulo são apresentados os principais repositórios de dados de *microarrays,* assim como os processos de normalização de dados. É proposto e implementado um modelo de partilha de dados em genómica que possibilita que diferentes grupos de investigação partilhem dados entre si, mesmo se usando sistemas LIMS distintos. É ainda exemplificado o uso deste modelo, através da implementação de dois cenários reais de utilização.

4.1 Repositórios de dados de *microarrays*

O desenvolvimento de repositórios de dados de *microarrays* constituiu um relevante passo para a valorização dos dados produzidos. Antes da existência destes repositórios os dados nem sempre eram armazenados e, quando o eram, estavam dispersos por vários servidores, o que dificultava a sua pesquisa e o seu acesso. Com base no levantamento prévio, realizado por Catherine A. Ball *et al,* foram encontradas as seguintes mais-valias no estabelecimento de repositórios de *microarrays* [72]:

- Assegurar, com elevado nível de segurança, a perpetuidade dos dados de expressão génica e facilitar o seu acesso através de um ponto único;
- Poder, durante o processo de submissão, validar os dados e garantir que estes se encontram concordantes com as normas existentes;
- Facilitar o processo de revisão, na medida em que os repositórios disponibilizam acesso condicionado aos dados durante o processo de publicação;
- Desenvolver, a partir dos dados armazenados, novas ferramentas de análise e de integração, com o objectivo de facilitar a tarefa de aceder, pesquisar e partilhar dados;
- Simplificar, com base no uso de normas para o armazenamento dos dados, a acessibilidade dos mesmos, bem como possibilitar a integração dos dados de expressão com outros tipos de dados, nomeadamente de sequência, SNPs, literatura, ou quaisquer recursos que possam ser usados na interpretação de padrões de expressão.

De seguida são apresentadas as características dos três repositórios de armazenamento de dados de expressão génica existentes: ArrayExpress, GEO e Cibex.

4.1.1 ArrayExpress

O repositório de dados ArrayExpress[1], desenvolvido e mantido pelo EBI (*European Bioinformatics Institute*), encontra-se online desde 2003 e possui, à data, a mais relevante colecção de estudos de expressão génica, contendo, aproximadamente, 9000 experiências e 270.000 hibridações [55, 73, 74].

Os principais objectivos para o seu desenvolvimento eram: servir como arquivo de dados de *microarrays,* associados com publicações científicas; fornecer o acesso facilitado aos dados num formato normalizado; e, por último, promover a partilha de desenhos experimentais e de protocolos entre a comunidade.

Para além do repositório, onde são armazenados todos os dados de experiências submetidas, existe uma segunda base de dados, o ArrayExpress Data Warehouse. Nesta apenas são importados estudos de elevada relevância, pois cada experiência necessita de ser manualmente validada e convertida para um novo esquema. A principal vantagem reside no facto de, sobre esta base de dados, poderem ser realizadas pesquisas mais complexas, que relacionam vários parâmetros, incluindo perfis de expressão, genes e eventos. Embora possua uma dimensão bastante inferior à do repositório principal, o ArrayExpress Data Warehouse, continha em Setembro de 2009, mais de 29.000 hibridações.

De acordo com os autores, uma das características que distingue o ArrayExpress dos restantes repositórios é a sua proactividade na adopção das normas existentes. Com efeito, o ArrayExpress foi o primeiro repositório a possibilitar a importação e a exportação em MAGE-ML e MAGE-TAB, tendo sido também a primeira base de dados concordante com a norma MIAME. Os autores salientam ainda a existência de várias ferramentas, tais como o *Expression Profiler*, que podem ser usadas para exploração e análise dos dados armazenados.

4.1.2 GEO

O NCBI *(National Center for Biotechnology Information*), através do desenvolvimento do GEO[2] (*Gene Expression Omnibus*), foi o primeiro a reconhecer a necessidade de um repositório central para o armazenamento de dados de expressão génica [56, 75]. O GEO encontra-se online desde 2001, contendo, em Setembro de 2009, mais de 13.000 experiências e 340.000 hibridações.

O GEO consiste numa plataforma flexível que facilita a submissão, a pesquisa e o acesso aos dados. Uma das vantagens deste sistema resulta na facilidade e na rapidez de

[1] http://www.ebi.ac.uk/microarray-as/ae
[2] http://www.ncbi.nlm.nih.gov/geo/

submissão de dados por meio da sua interface *web*. No entanto, possui outras alternativas, todas elas concordantes com a norma MIAME, entre as quais o formato SOFT (*Simple Omnibus Format*), MINiML (*MIAME Notation in Markup Language*) e o MAGE-ML. De salientar que os formatos SOFT e MINiML são proprietários e não reconhecidos pelo MGED.

Outra das mais-valias do sistema é a possibilidade de utilização destes dados no contexto mais alargado das restantes ferramentas do NCBI, gerando-se um benefício mútuo. Assim, sempre que possível, são fornecidos apontadores para recursos como o PubMed, o GenBank, o Entrez Gene, o UniGene, o MapViewer, o OMIM, entre outros. O GEO lucra ainda directamente com a utilização da anotação das sequências inseridas, assim como com o mapeamento e com os recursos bibliográficos.

Ao afirmar-se como o mais vasto recurso público de dados de expressão génica, o GEO apresenta como principais possibilidades a identificação de padrões de expressão comuns, a obtenção de redes reguladoras e a inferência da função de genes não caracterizados.

4.1.3 Cibex

O Cibex[1] (*Center for Information Biology gene EXpression database*) é, em termos de dados armazenados, o menos relevante dos repositórios apresentados [57]. Em Setembro de 2009, possuía 53 experiências e 1815 hibridações.

A submissão dos dados é realizada através de uma aplicação implementada em *Java*. Os dados descritivos da experiência são inseridos através do preenchimento de um formulário *web,* sendo possível a importação de dados em formato tabular, correspondendo à descrição do *microarray* e aos valores de expressão. De notar que esta base de dados possibilita o armazenamento de acordo com a norma MIAME.

Apesar de não ser importada a imagem resultante da digitalização do *microarray* é disponibilizada uma aplicação que possui, entre outras funcionalidades, a capacidade de reconstruir uma representação virtual da imagem.

4.2 Processos de normalização na partilha de dados

A submissão dos resultados de experiências de *microarrays* a repositórios públicos é essencial para assegurar a perpetuidade dos dados. Este processo constitui, no entanto, um relevante desafio, sobretudo devido à variação existente entre os formatos seleccionados para o armazenamento dos dados pelos diferentes laboratórios. De forma a endereçar esta

[1] http://cibex.nig.ac.jp

questão, foram desenvolvidas várias normas que possibilitam a submissão de dados a repositórios públicos.

De seguida, são apresentados os métodos existentes de submissão de dados ao ArrayExpress. Neste estudo, por uma questão de objectividade, focámo-nos no repositório ArrayExpress, eleito pelo facto de ser o mais proactivo no uso e na adopção de normas, assim como por possuir o mais relevante conjunto de experiências. Porém, muitos dos mecanismos apresentados são extensíveis aos restantes repositórios. Limitámos ainda o nosso estudo aos métodos propostos pelo grupo MGED tendo, não obstante, consciência da existência, se bem que marginal, de outros métodos proprietários.

4.2.1 MAGE-ML

A norma MAGE-ML *(Microarray and Gene Expression Markup Language*) foi desenhada com o objectivo de descrever e, consequentemente, promover a comunicação de experiências de *microarrays* [54]. O MAGE-ML é baseado em XML (*eXtensible Markup Language*) e descreve desenhos de *microarrays*, para além de informação de construção do *microarray* e de preparação e de execução da experiência. O MAGE-ML procede de um decalque para XML do modelo de objectos da norma MAGE-OM, apresentado no capítulo anterior.

No processo de conversão do MAGE-OM para MAGE-ML foram apenas usadas regras simples. Primeiro, cada classe, no modelo de objectos, passa a estar representada como um elemento XML com uma lista de atributos correspondentes aos elementos da classe. De seguida, para cada associação que essa classe possui, é criado um sub-elemento com o da associação concatenado com 'assn'. Caso a associação seja unívoca, é concatenado 'ref'. No entanto, caso a cardinalidade da associação seja superior a um, é concatenado 'list'. Foram ainda criados elementos especiais, correspondentes a cada um dos pacotes do modelo MAGE-OM.

Devido à elevada complexidade do modelo MAGE, não é espectável que se aceda directamente ao MAGE-ML. Em alternativa, foi desenvolvida uma ferramenta baseada no MAGE-OM, designada MAGE-STK (MAGE - *Software ToolKit*). Esta ferramenta define uma API de acesso facilitado ao MAGE-OM e possui, neste momento, três implementações: Perl, Java e C++. O propósito do MAGE-STK é fornecer uma camada intermédia de acesso ao modelo MAGE-OM que possibilite a exportação para MAGE-ML, a persistência dos dados numa base de dados relacional, ou a geração de entradas para ferramentas de análise de dados.

A Figura 4.1 apresenta um exemplo da descrição de uma sequência biológica em MAGE-ML retirado da experiência armazenada no ArrayExpress com a referência E-MANP-1.

4.2.2 MAGE-TAB

O MAGE-TAB é uma alternativa recente à norma MAGE-ML. O principal objectivo para o seu desenvolvimento era conseguir expressar a complexidade de uma experiência de *microarrays* de uma forma mais inteligível do que a apresentada pelo MAGE-ML [76]. Para tal, recorreu ao uso de ficheiros em formato tabular. Este formato é constituído por três ficheiros:

- IDF (*Investigation Design File*): trata-se do ficheiro empregue para fornecer uma visão geral da experiência, incluindo os factores experimentais utilizados, descrição dos protocolos usados, as estratégias de controlo de qualidade, os detalhes da publicação e os contactos. Inclui, ainda, uma referência para o ficheiro SDRF;
- SDRF (*Sample and Data Relationship Format*): é o ficheiro que abrange as relações entre as amostras (do *BioMaterial* ao *Labelled Extract*) aplicadas na hibridação com os *microarrays* e as referências para os ficheiros com os dados de expressão no formato bruto e normalizado. É também responsável pelo armazenamento da informação relativa a factores experimentais, protocolos usados e respectivos parâmetros. A Figura 4.2 contém um excerto de um ficheiro SDRF;
- ADF *(Array Design File*): este ficheiro inclui, para cada spot do *microarray*, a sequência da sonda ou a identificação do transcrito alvo. Ele é apenas é necessário no caso de *microarrays* construídos no laboratório, na medida em que nos comerciais basta o uso do código correspondente. Usualmente o desenho do *microarray* constitui uma submissão separada, podendo esta ser reutilizada entre experiências.

4.2.3 XML *vs* TAB

O MAGE-ML e o MAGE-TAB são duas aproximações distintas que possuem um mesmo objectivo: possibilitar a comunicação de dados de experiências de *microarrays*. Apesar de, a nível global, se verificar uma notória tendência do uso de XML enquanto contentor de transporte dados, neste particular, a norma baseada em XML precede a mais recente, baseada em ficheiros tabulares.

De acordo com os autores da norma MAGE-TAB [76], no caso concreto dos *microarrays,* o uso de ficheiros tabulares apresenta várias vantagens relativamente ao XML. O principal argumento reside no facto dos formatos baseados em XML serem mais apropriados para representar dados cuja estrutura se assemelhe a uma árvore. No entanto, a estrutura que melhor representa o desenho experimental de uma experiência de *microarrays* resulta num DAG (*Directed Acyclic Graph*) e a sua representação em XML, apesar de possível, torna-se bastante ineficiente. No que se refere ao *Array Design*, que contém para cada *spot* a localização e a descrição da sonda, mais uma vez a forma natural de armazenamento é tabular.

A favor do uso da versão tabular está ainda o facto desta se revelar uma proposta mais simples, permitindo a edição e a visualização dos dados. Na realidade, o facto de o MAGE-ML ter sido directamente derivado de um modelo de objectos já em si complexo apenas veio criar uma versão completa, mas também ela complexa, de armazenamento de dados. O elevado número de classes e principalmente as relações existentes entre estas são factores responsáveis pela complexidade do MAGE-ML.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE MAGE-ML SYSTEM "file://localhost/H:/MAGE-ML.dtd">
<MAGE-ML identifier="MAGE-ML:E-MANP-1">
  <BioSequence_package>
    <BioSequence_assnlist>
      <BioSequence identifier="BS:clone.1" length="822">
        <SequenceDatabases_assnlist>
          <DatabaseEntry accession="M11147">
            <Database_assnref>
              <Database_ref identifier="DB:embl"/>
            </Database_assnref>
          </DatabaseEntry>
        </SequenceDatabases_assnlist>
        <PolymerType_assn>
          <OntologyEntry category="polymertype" value="RNA"/>
        </PolymerType_assn>
        <Type_assn>
          <OntologyEntry category="biosequence:type" value="clone"/>
        </Type_assn>
      </BioSequence>
 [...]
</MAGE-ML>
```

Figura 4.1: Exemplo da descrição de uma sequência biológica em MAGE-ML. (Retirado da experiência E-MANP-1, armazenada no ArrayExpress).

Hybridization Name	ArrayDesign REF	Protocol REF	Array Data File	Protocol REF	Normalization Name	Derived Array Data File	FactorValue [strain_or_line]
EC1118_T1	A-MEXP-1185	SCANNING_30% Red/Green Laser PMT	ev_file_79.txt	TRANSFORMATION	Median Normalized data_MIND	Norm mediam all data.txt	Lalvin EC-1118
EC1118_T1	A-MEXP-1185	SCANNING_30% Red/Green Laser PMT	ev_file_79.txt	TRANSFORMATION	Median Normalized data_MIND	Norm mediam all data.txt	S288c (MAT alpha SUC2 mal mel gal2 CUP1 flo1 flo8-1)
EC1118_T1_ds	A-MEXP-1185	SCANNING_30% Red/Green Laser PMT	ev_file_85.txt	TRANSFORMATION	Median Normalized data_MIND	Norm mediam all data.txt	Lalvin EC-1118
EC1118_T1_ds	A-MEXP-1185	SCANNING_30% Red/Green Laser PMT	ev_file_85.txt	TRANSFORMATION	Median Normalized data_MIND	Norm mediam all data.txt	S288c (MAT alpha SUC2 mal mel gal2 CUP1 flo1 flo8-1)
EC1118_T2	A-MEXP-1185	SCANNING_30% Red/Green Laser PMT	ev_file_80.txt	TRANSFORMATION	Median Normalized data_MIND	Norm mediam all data.txt	Lalvin EC-1118
EC1118_T2	A-MEXP-1185	SCANNING_30% Red/Green Laser PMT	ev_file_80.txt	TRANSFORMATION	Median Normalized data_MIND	Norm mediam all data.txt	S288c (MAT alpha SUC2 mal mel gal2 CUP1 flo1 flo8-1)

Figura 4.2: Excerto de um ficheiro SDRF correspondente a uma experiência armazenada no Mind.

4.3 Proposta e implementação de um modelo de partilha de dados em genómica

O objectivo inicial das normas anteriormente apresentadas consistia em facilitar a tarefa de submissão dos resultados de estudos de *microarrays* a repositórios públicos. A anotação da experiência é tipicamente realizada ao longo da sua execução no sistema LIMS existente no laboratório. Após o término da experiência, é necessário exportar os dados num dos formatos normalizados (MAGE-TAB ou MAGE-ML), de forma a estes serem submetidos a um dos repositórios públicos. Este cenário base de partilha de dados é caracterizado por uma única operação de submissão e por várias operações de pesquisa e leitura dos dados (Figura 4.3).

Uma visão mais lata das mesmas normas compreende, não obstante, o seu uso na qualidade de meio de comunicação normalizado de dados de *microarrays*. Esta abordagem possibilita que dados correspondentes a uma experiência possam ser sucessivamente trocados entre laboratórios antes mesmo da sua conclusão. Na próxima secção é proposto um novo cenário que possibilita a partilha de dados entre laboratórios.

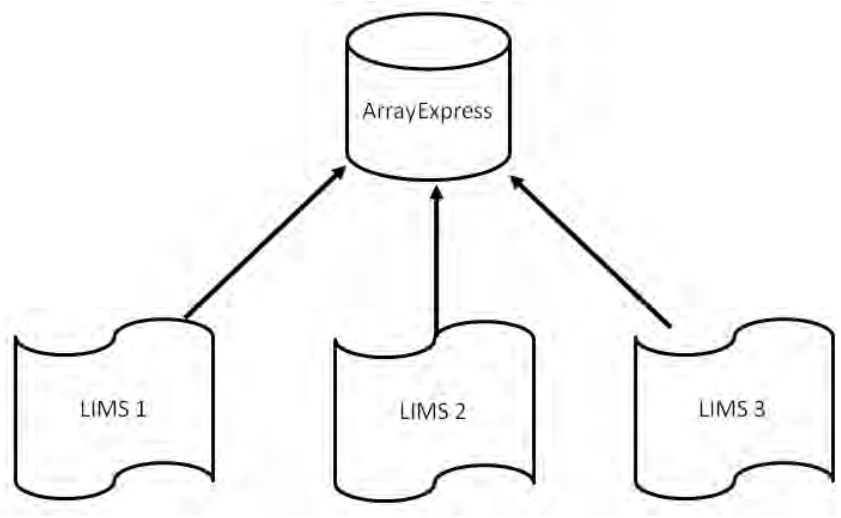


Figura 4.3: Cenário base de partilha de dados de expressão génica.

4.3.1 Cenário proposto

Surgiram, de facto, novos requisitos que suportam o uso destas normas para partilha de dados de experiências parciais. Por um lado, pese a generalização da tecnologia enquanto ferramenta laboratorial, não é viável a existência, em todos os laboratórios, de todo o equipamento necessário à construção e ao uso dos *microarrays*. Por outro lado, a premência de realizar estudos mais complexos leva à incapacidade de um único laboratório conduzir estudos isoladamente. Como consequência surge a necessidade de existirem cooperações entre laboratórios. A relevância na partilha de recursos entre laboratórios,

assim como o estabelecimento de parcerias para realização de experiências, conduz, obrigatoriamente, à necessidade de um meio de permuta de dados entre organizações.

Este problema é tipicamente resolvido através do uso de um mesmo sistema LIMS com funcionalidades de partilha de dados entre utilizadores. O próprio sistema Mind possibilita o trabalho colaborativo, permitindo a tranferência de elementos entre utilizadores. Porém, o problema a que se pretende dar resposta remete para um cenário em que vários investigadores colaboram num mesmo projecto mas pertencem a laboratórios com sistemas LIMS diferentes (Figura 4.4). Este facto cria novos desafios, visto cada LIMS possuir o seu modelo de dados interno, sendo, então necessário fazer uso de uma camada de abstracção partilhada por todos, sobre a qual possa ser realizada a comunicação dos dados referentes à experiência.

O uso de uma das normas MAGE enquadra-se, pois fornece um modo normalizado de comunicação de dados de experiências de *microarrays*. Uma vantagem adicional é ainda o reduzido número de alterações necessárias de modo a compatibilizar os dados com esta abordagem. Na realidade, é apenas necessário implementar o mecanismo de importação, visto o de exportação já ser comum.

No intuito de assegurar a sua congruência, em cada instante, apenas um laboratório pode deter a experiência e, consequentemente pode proceder a alterações na mesma. Neste cenário, a posse da experiência e a transferência de dados são realizadas por meio de canais informais de comunicação, nomeadamente, o correio electrónico.

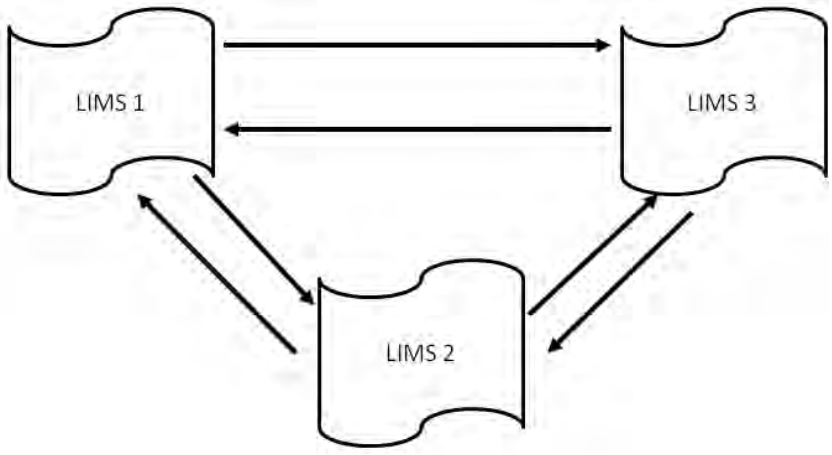


Figura 4.4: Representação esquemática do cenário proposto.

4.3.2 Desafios de implementação

Apesar da tarefa a realizar ser conceptualmente simples, a sua implementação apresentou uma série de desafios que necessitaram de ser identificados e endereçados. Os principais estão relacionados com a elevada complexidade do modelo MAGE e com a dificuldade em mapear correctamente conceitos entre vários esquemas. Existem também outras questões,

tais como garantir a segurança e a autenticação na comunicação dos dados. São, de seguida, examinados em detalhe os principais desafios encontrados, sendo, ainda, apresentadas as estratégias usadas para os superar:

- **Multiplicidade de normas de comunicação de dados.** Conquanto ambas as normas apresentadas para a partilha de dados de *microarrays* sejam suportadas pelo grupo MGED, estas consistem em implementações bastante distintas. E, embora existam algumas ferramentas de conversão estas apresentam algumas limitações, especialmente na tarefa de transformação do formato XML para o formato tabular. Assim, o uso simultâneo de ambas no modelo proposto pode causas situações de incongruências;
- **Mapeamento entre o modelo MAGE-TAB e o esquema da base de dados relacional.** Ainda que se assuma que todos os sistemas LIMS a integrar no modelo proposto são concordantes com a norma MIAME, é natural que diferentes estratégias sejam usadas ao nível do modelo de armazenamento. Esta situação pode criar problemas na tarefa de mapeamento entre esquemas podendo causar ambiguidades. Uma solução pode passar pela adição, ao ficheiro exportado, de campos com dados opcionais. Outra questão relacionada com o mapeamento resume-se ao facto de existirem dados que não são armazenados de forma explícita no sistema LIMS local, tipicamente por serem comuns a todas as experiências realizadas. Estes dados devem ser usados como padrão em todas as exportações, ou, em alternativa, serem inseridos pelos utilizadores no momento da exportação;
- **Violação da integridade dos dados durante a tarefa de importação.** De forma a evitar problemas de integridade dos dados, a tarefa de importação requer atenção redobrada. Soma-se a esta problemática, a possibilidade de diferentes sistemas atribuírem relevâncias distintas a diferentes necessidades díspares e, consequentemente, mesmo sendo importação bem sucedida, é necessário verificar que os dados e as relações no sistema de destino são criados. A estratégia usada na resolução deste eventual problema consiste na pré-importação dos dados para um modelo de dados temporário, sobre qual a integridade é testada;
- **Perda de informação com sucessivos mapeamentos entre esquemas.** Tendo em consideração que, em cada processo de exportação ou de importação existe a possibilidade de ocorrerem falhas no mapeamento e de, em consequência disso, ocorrer uma perda de dados, num cenário com múltiplos processos de exportação e importação esta situação torna-se complexa de gerir. Uma forma de superar este constrangimento pode passar pelo armazenamento local das sucessivas versões e pela comparação das alterações introduzidas.

4.3.3 Implementação do módulo de exportação/importação

No intuito de concretizar o cenário anteriormente apresentado, é fundamental que cada um dos sistemas LIMS possua capacidade de exportação e importação de documentos MAGE. O módulo de seguida apresentado tem por objectivo servir de interface entre a base de dados do sistema LIMS e o modelo MAGE. A sua inclusão num sistema LIMS requer uma completa compreensão da estrutura da base de dados de origem, sendo ainda estritamente necessário que esta seja compatível com a recomendação MIAME. Para evitar incongruências de importação, foi incluída a possibilidade de adicionar informação complementar relativa à experiência. Em caso de ambiguidade, pode mesmo ser necessária interacção por parte do utilizador. O módulo suporta, ainda, o uso das normas MAGE-ML e MAGE-TAB.

O processo de exportação inicia-se com o acesso ao sistema LIMS e com a obtenção de todos os dados relacionados com a experiência a exportar. Seguidamente é realizada uma análise sintáctica aos dados e são adicionados os elementos em falta, seja através de dados fornecidos pelo utilizador, seja pela adição de dados armazenados em ficheiros. O próximo passo consiste em realizar uma análise semântica dos dados e um pré-processamento de modo a formatá-los. Depois, dependendo se os dados vão ser exportados para MAGE-ML ou MAGE-TAB, uma das instâncias dos modelos locais é usada para carregar localmente os dados. Por fim, após a importação local dos dados, estes são exportados para MAGE-TAB/ML, podendo ser enviados para o ArrayExpress ou para importação noutro sistema LIMS (Figura 4.5).

No processo de importação, apesar dos elementos usados serem essencialmente os mesmos, a sua ordem de actuação é inversa. O primeiro passo compõe-se da leitura dos ficheiros no formato MAGE-TAB/ML para o modelo de dados local. Posteriormente, é realizada uma validação da existência dos dados mínimos para criar uma nova experiência. Esta tem o objectivo de evitar incongruências ao inserir a experiência na base de dados. Segue-se o pré-processamento, que possibilita a limpeza e reorganização dos dados, o analisador sintáctico e semântico e a adição de elementos contextuais em falta.

4.3.4 Submissão ao ArrayExpress

O módulo de exportação/importação implementado pode ser usado em vários cenários. O primeiro a ser testado a automatação da tarefa de submissão de uma experiência armazenada no Mind ao repositório público do ArrayExpress. O módulo foi implementado de forma a ser independente do sistema LIMS, tendo sido fundamental, antes de mais, proceder à sua configuração para o sistema Mind. Esta tarefa implicou o mapeamento do esquema da base de dados com as respectivas classes do modelo MAGE. Foi ainda necessário desenvolver a interface web para a interacção com o utilizador. Optou-se pelo uso de MAGE-TAB para submissão dos dados.

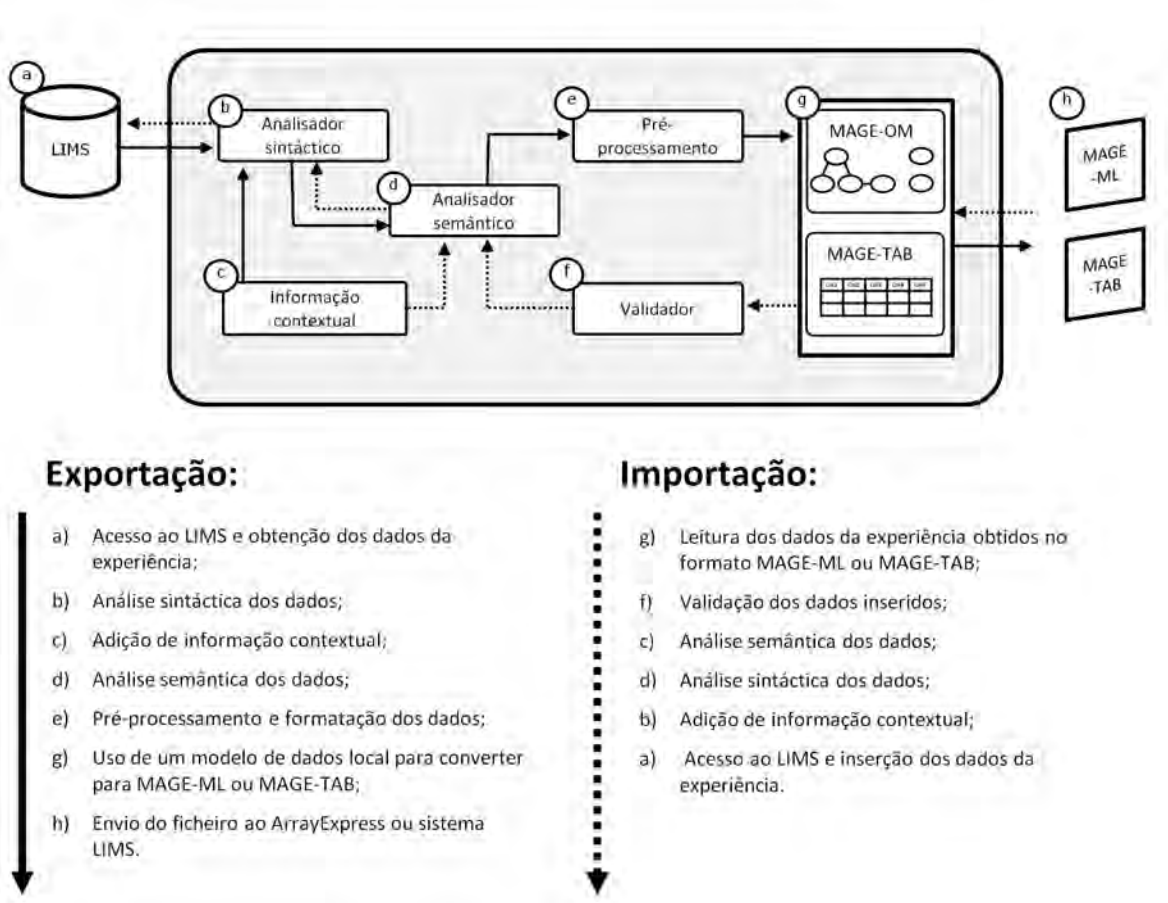


Figura 4.5: Esquema de funcionamento do módulo de exportação e de importação de dados.

Após a completa inserção da experiência no Mind o procedimento de submissão supõe os seguintes passos:

- **Selecção da experiência a exportar;**
- **Inserção de detalhes da experiência e da publicação:** os detalhes da experiência incluem o *ArrayDesign* usado, a data de disponibilização da experiência, assim como os procedimentos de normalização e de controlo de qualidade usados. Como detalhes da publicação entende-se o seu título, autores, referências e estado;
- **Obtenção dos ficheiros:** é obtido o ficheiro IDF, que contém a descrição geral da experiência, e um arquivo zip, que contém o ficheiro SDRF, todos os ficheiros com os valores expressão para cada um dos *microarrays,* e opcionalmente, os dados no formato normalizado (Figura 4.6);
- **Submissão dos ficheiros no MiameExpress:** a submissão da experiência ao ArrayExpress envolve o acesso ao MiameExpress, a selecção do formato MAGE-TAB, a autenticação no sistema, e, por fim, a submissão dos ficheiros anteriores.

Figura 4.6: Interface do Mind com o relatório da exportação da experiência *Heat Shock*.

4.3.5 Partilha de dados entre dois LIMS

Este cenário resulta do uso do modelo implementado para partilha de dados entre dois sistemas LIMS distintos. De forma a testá-lo com sucesso, foi necessário considerar, para além do Mind, outro sistema LIMS. Foi, deste modo, estabelecida uma colaboração com o ScGTI (*Scottish Centre for Genomic Technology and Informatics*), com o propósito de se poder ter acesso ao GPX.

O GPX é um sistema LIMS de *microarrays*, concordante com a recomendação MIAME [77]. Este sistema encontra-se em uso desde 2003, para fornecer acesso a dados, no formato bruto e processado, produzidos no laboratório local. Os principais estudos realizados têm-se centrado na análise dos níveis de expressão de macrófagos, através da indução de pró-inflamatório, de anti-inflamatório e de intrusos benignos e patogénicos. Em Setembro de 2009, o sistema armazenava cerca de 54 experiências, que contêm mais de 800 hibridações. O esquema da base de dados do GPX encontra-se no Anexo 1.

O teste realizado implicou o desenvolvimento de um *pipeline* do sistema GPX para o sistema Mind. De modo a avaliar o funcionamento do *pipeline,* foi seleccionado um estudo realizado no ScGTI e armazenado no GPX (*Transcription profiling time series of mouse bone marrow derived macrophages after interferon-gamma treatment*), que se exportou para MAGE-ML e se importou para o sistema Mind. Este mesmo estudo encontra-se armazenado no ArrayExpress com o identificador E-MEXP-1490.

Mapeamento dos dois modelos

A Figura 4.7 apresenta a fracção do modelo da base de dados responsável pelo armazenamento da experiência, e a Figura 4.8 representa a sua homóloga no modelo do Mind. O objectivo que esta tarefa se propõe realizar, a transferência de dados do GPX para o Mind, poderia ser facilmente concretizada através do mapeamento directo entre os

modelos de ambas as bases de dados. Esta solução, apesar de efectiva para o problema proposto, constituiria uma elevada perda de generalidade, na medida em que não suportaria a adição de novos sistemas.

O procedimento usado no estabelecimento do *pipeline* consistiu em acoplar o módulo de importação/exportação, anteriormente apresentado, a cada um dos sistemas LIMS. Deste modo, é possível usar experiências em MAGE-ML para importar e exportar dados de ambas as bases de dados. É necessário, para cada um dos pacotes do modelo MAGE, proceder à correspondência entre todo e qualquer elemento e o seu homólogo no esquema da base de dados.

A Tabela 4.1 resume este processo de mapeamento para o pacote *Experiment*. Para proceder à migração da experiência, foi imperativo estender este procedimento aos restantes pacotes do modelo MAGE. O código MAGE-ML gerado correspondente à experiência encontra-se no Anexo 2.

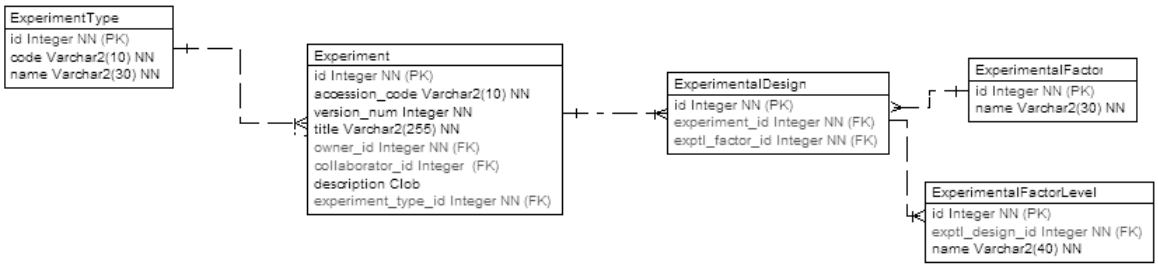


Figura 4.7: Secção do modelo de dados do GPX usado no armazenamento da experiência.

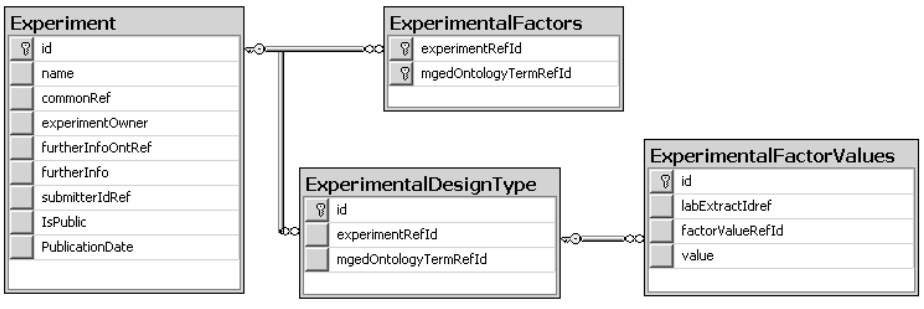


Figura 4.8: Secção do modelo de dados do Mind usado no armazenamento da experiência.

Tabela 4.1: Resumo do mapeamento entre o GPX e o Mind através do uso do modelo MAGE para o pacote *Experiment.*

GPX	MAGE-ML	Mind
Experiment:accession_code	Experiment_package/Experiment_assnlist/Experiment: identifier	Experiment:comonRef
Experiement:version_num	Experiment_package/Experiment_assnlist/Experiment/ PropertySets_assnlis/NameValueType:value	Experiment:PublicationDate
Experiment:title	Experiment_package/Experiment_assnlist/Experiment:name	Experiment:name
Experiment:description	Experiment_package/Experiment_assnlist/Experiment/Descripti ons_assnlist/Description:text	Experiment:furtherInfo
ExperimentType:name	Experiment_package/Experiment_assnlist/ExperimentDesign_as sn: Types_assnlist/OntologyEntry: value	ExperimentalFactorValues:valu es
ExperimentType:code	Experiment_package/Experiment_assnlist/ExperimentDesign_as sn: Types_assnlist/OntologyEntry: value	ExperimentalFactorValues:fact orValueRefId
ExperimentalFactor:name	Experiment_package/Experiment_assnlist/ExperimentDesign_as sn: Types_assnlist/OntologyEntry: value	ExperimentalFactors:mgedOnt ologyTermRefId
ExperimentalFactorLevel:na me	Experiment_package/Experiment_assnlist/ExperimentDesign_as sn: Types_assnlist/OntologyEntry: value	ExperimentalFactors:mgedOnt ologyTermRefId

4.4 Sumário

Neste capítulo foi exposto um levantamento dos repositórios de dados de *microarrays*, das normas existentes e dos cenários de partilha de dados possíveis. Foi, por fim, dado a conhecer a arquitectura de um modelo que possibilita que diferentes investigadores participem num mesmo projecto, ainda que utilizando sistemas LIMS distintos.

A submissão dos resultados de experiências de *microarrays* a repositórios públicos é essencial para assegurar a perpetuidade dos dados. O desenvolvimento destes repositórios beneficiou, em muito, com os avanços das normas e ontologias, na medida em que estas permitiram evitar os problemas de variação existente entre os formatos de armazenamento dos dados usados em vários laboratórios.

A par do que sucedia nas questões relacionadas com o armazenamento, também na partilha as normas MAGE desempenham um importante papel, ao facilitarem a submissão de dados a repositórios centrais. Existem, actualmente, duas alternativas: uma baseada em XML (MAGE-ML) e outra baseada no formato tabular (MAGE-TAB). Apesar de existir uma tendência global na adopção de formatos baseados em XML, no caso concreto dos *microarrays,* o uso de ficheiros tabulares apresenta vantagens claras.

Para além do cenário base de submissão a um repositório público, foi proposto um novo para a partilha de dados entre laboratório. A implementação deste modelo apresentou uma série de desafios, estando os principais relacionados com a dificuldade em correctamente mapear conceitos entre vários esquemas. Finalmente, de forma a testar com sucesso o

modelo que instancia os cenários propostos, foi necessário considerar outro sistema LIMS para além do Mind. Para tal, foi estabelecida uma colaboração com o ScGTI, de forma a incluir o sistema GPX no modelo. Como resultado, é apresentado o processo de migração de uma experiência do sistema GPX para o Mind.

Capítulo 5

5 Análise e interpretação biológica de dados de *microarrays*

Dos desafios que se colocam à execução de uma experiência de *microarrays* os mais complexos, relacionam-se com a avaliação e a interpretação dos resultados. A existência de vários factores que podem influir na qualidade dos dados, a intrínseca relação existente entre a questão biológica a endereçar e o desenho experimental, assim como as implicações que estes têm na escolha dos métodos de análise a utilizar são, precisamente as contingências apontadas. Nos últimos anos, vários algoritmos e ferramentas foram propostos no intuito de ultrapassar esta situação, no entanto, devido à especificidade de cada experiência, a escolha dos métodos mais adequados continua a recair sobre o investigador [78, 79].

Uma das dificuldades encontradas prende-se com o elevado número de ferramentas computacionais distintas que necessitam de ser usadas de forma a completar uma análise. Este processo requer que o investigador domine independentemente cada uma, assim como que consiga transferir os dados entre estas, o que, em vários casos, implica alterações no formato em que se encontram armazenados. Estas sucessivas conversões não só são morosas, como ainda são muito susceptíveis de gerar erros difíceis de identificar e corrigir.

Este capítulo encontra-se organizado em três partes. Na primeira é identificado o fluxo típico da análise de uma experiência de *microarrays,* sendo apresentado um levantamento das principais ferramentas e métodos de tratamento e análise de dados. Na segunda é descrita a implementação, sobre o sistema Mind, de um fluxo integrado de ferramentas de análise de dados. Através do uso do Mind, torna-se possível a realização da análise completa de uma experiência, eliminando os problemas associados com a manipulação manual dos ficheiros ou a necessidade de utilizar várias ferramentas. Por último, é apresentada uma demonstração do funcionamento do fluxo de análise do Mind por meio de dados obtidos no laboratório.

5.1 Fluxo da análise de dados

O fluxograma apresentado na Figura 5.1 resume os passos directamente associados com a análise de dados de uma experiência de *microarrays*. O processo, que se inicia com a digitalização e quantificação da imagem de cada *microarray,* prossegue com a correcção do *background*, o pré-processamento e a normalização dos dados. De seguida podem ser aplicados dois tipos de análise: identificação de genes diferencialmente expressos ou análise exploratória. Por fim, é realizada a análise funcional dos dados. No contexto deste capítulo é assumido que os procedimentos associados com a obtenção e quantificação dos valores de expressão da imagem já foram previamente realizados.

O primeiro passo, a correcção do *background,* consiste num ajuste dos valores de intensidade através da subtracção da intensidade associada com hibridação não específica. O pré-processamento dos dados, que se lhe segue, resulta na filtragem dos *spots* que não cumpram critérios mínimos de qualidade e na transformação dos dados de forma a facilitar a sua subsequente manipulação. Posteriormente, a normalização cumpre o propósito de evidenciar as diferenças biológicas existentes em detrimento das variações técnicas.

Existem duas metodologias de análise que dependendo do tipo de experiência e da questão biológica a endereçar, podem ser empregues. A primeira, e tipicamente a mais utilizada, assenta no recurso a testes de significância estatística para identificação dos genes diferencialmente expressos. Em alternativa, a análise exploratória apoia-se no uso de métodos de classificação de forma a encontrar padrões nos dados. Conquanto esta estratégia de análise não seja considerada neste documento, descrições detalhadas podem ser encontradas em [78, 80, 81]. Da execução dos métodos anteriores resultam, tipicamente, sub-listas de genes, cuja interpretação biológica é essencial. Com esse objectivo, a análise funcional serve-se de ontologias ou de informação biológica prévia para análise dos dados existentes. Esta análise será detalhadamente descrita no capítulo 6.

5.2 Obtenção e tratamento dos dados de *microarrays*

Esta secção resume os principais métodos de validação e transformação dos dados de *microarrays*. Aqui inclui-se, portanto, o controlo da qualidade dos dados obtidos, a correcção do *background,* o pré-processamento e a normalização dos dados.

5.2.1 Estratégias de controlo da qualidade

Grande parte do trabalho estatístico baseia-se na limpeza dos dados obtidos e em assegurar a sua qualidade. No entanto, o processo de controlo de qualidade dos dados inicia-se no laboratório, ainda antes da execução da experiência. Existe, deste modo, uma série de passos que devem ser realizados de forma a assegurar a qualidade dos dados.

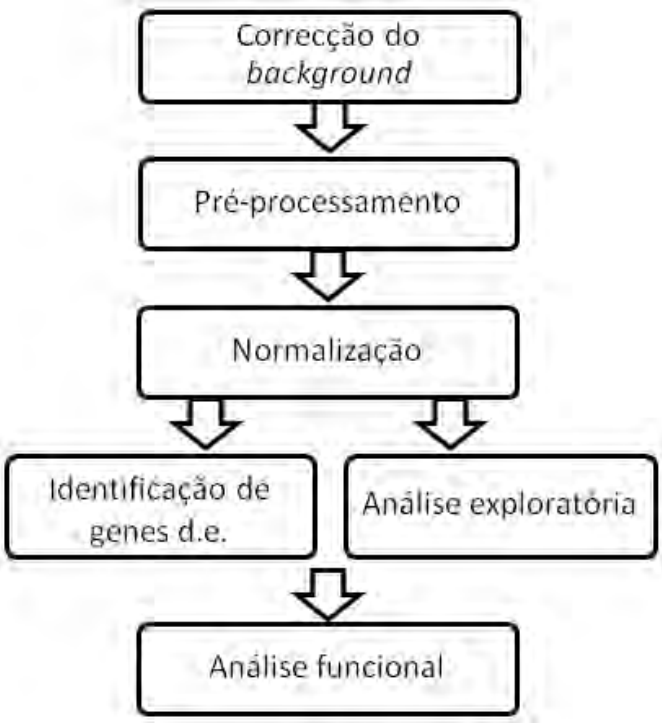


Figura 5.1: Fluxo de análise de dados resultantes de uma experiência de *microarrays.*

Controlo de qualidade no laboratório

Os primeiros passos de verificação da qualidade dos dados são efectuados no laboratório mesmo antes das hibridações. Uma vez que a detecção e correcção de problemas é mais complicada após a hibridação, esta verificação inicial permite a eliminação antecipada de *microarrays* problemáticos.

Duas verificações frequentemente usadas são a validação da qualidade do RNA e a eficiência de incorporação do marcador. Entre o momento em que a amostra é retirada e purificada, enzimas na célula iniciam um processo de degradação do mRNA através da sua separação em fragmentos menores. Estes fragmentos vão hibridar mais facilmente com outras sondas do *microarray,* o que pode levar a distorções nas medições realizadas. Existem actualmente vários procedimentos padrão que permitem a verificação destes dados. Também as medições efectuadas dependem directamente da quantidade de marcador presente na amostra. Deste modo, faz sentido verificar a quantidade que foi efectivamente incorporada pois esta pode variar entre amostras, especialmente para o marcador *Cy5*. Por exemplo, moderados níveis de ozono, comuns em verões húmidos, contribuem para uma degradação do marcador *Cy5*, isto apesar de não afectarem o marcador *Cy3*.

Uso de controlos

O uso de controlos no *microarray* é essencial para a correcta validação dos resultados. São três os tipos de controlos que podem ser utilizados: negativos, positivos e *spike-in*.

Controlos negativos compõem-se de sondas desenhadas para sequências de DNA que não encontrem correspondência nas amostras. Os controlos negativos devem possuir um baixo valor de intensidade, pelo que o seu valor médio de intensidade pode ser usado para se ter uma ideia do valor de *background* devido à hibridação não específica. Controlos positivos consistem em réplicas de sequências que devem ser abundantes na amostra. Estes possuem funções bem definidas, como, por exemplo, a obtenção do valor máximo esperado de intensidade. Finalmente, controlos *spike-in* correspondem a sondas para transcritos propositadamente adicionados à amostra em quantidades conhecidas. Estes podem ser usados para se ter uma ideia da validade e da linearidade das intensidades. Esta informação pode ser bastante útil durante a fase de normalização, no entanto, a sua interpretação deve ser cautelosa na medida em que ainda apresenta algumas limitações e nem sempre os resultados correspondem ao esperado.

Análise individual dos *spots*

Após a validação dos anteriores, este passo pretende verificar a qualidade dos dados para os *spots* individuais, isto é, tem como principal objectivo a detecção de problemas de impressão em vez de anomalias de hibridação. Uma vez que uma inspecção individual e manual dos *spots* não é praticável, esta deve ser realizada com base em procedimentos de filtragem automatizados, que consideram parâmetros como a área do *spot* ou a uniformidade do *background*. Os próprios programas de quantificação já anotam os *spots* que saem fora dos critérios de qualidade e raramente é boa ideia considerá-los na análise.

Outro problema comum reside na formação dos *spots,* que nem sempre possuem a forma desejada. Quando, por exemplo, o tamanho destes é demasiado elevado, dois *spots* podem conjugar-se, o que gera erros de leitura. Na prática, o uso de *spots* demasiado pequenos ou demasiado grandes é desaconselhável uma vez que pode provocar níveis de ruído excessivamente elevados, propagando-se na uniformidade dos rácios entre cores.

5.2.2 Correcção do *background*

Devido à existência de hibridação não específica, o valor da intensidade de um *spot* corresponde à soma do valor efectivo da sua intensidade com o valor correspondente ao *background*. A correcção do valor de *background* é, portanto, fundamental para a obtenção dos correctos valores de intensidade dos *spots*[82].

Os vários algoritmos disponíveis utilizam diferentes estratégias para obtenção dos valores corrigidos de *background*, assim como múltiplas variantes na correcção desses valores. O método base consta na subtracção directa dos valores de intensidade de *background* aos valores do *spot*. O valor de intensidade de *background* é obtido através do cálculo do valor médio da área circundante ao *spot*. Existem, no entanto, outros métodos tais como o uso do valor médio de *background* de n *spots* circundantes ou de *spots* vazios, isto é, que não possuem elementos de mRNA.

Porém, alguns estudos demonstram que em certos casos existe uma maior apetência para os alvos se ligarem ao substrato do *microarray* do que a um *spot* que possua DNA não específico com esse alvo. Deste modo, a utilização dos métodos de correcção anteriores pode eventualmente resultar na obtenção de um valor de intensidade corrigido negativo [38, 82].

Em opção, outras estratégias podem ser usadas, nomeadamente através do uso de controlos. Uma forma simples de evitar intensidades negativas é proposta por Edwards [83] e consiste em ajustar os valores de intensidade através da subtracção do *background* sempre que a diferença entre este dois valores seja superior a um valor pré-definido. Outros métodos disponíveis são o Normexp, que faz uso de um modelo de convolução da normal mais a exponencial; o Vsn que faz uso do método de estabilização da variância, inicialmente proposto por Huber; e o Morph que se traduz na obtenção da abertura morfológica através do cálculo do mínimo e do máximo local. A Tabela 5.1 resume os principais métodos de correcção de *background*. Uma análise extensiva do uso destes, assim como da sua aplicabilidade pode ser encontrada em [82].

Tabela 5.1: Resumo dos métodos de correcção de *background* mais comuns.

Método	Estimativa do *background*	Ajustamento
Standard	Mediana local	Subtracção
Kooperberg	Média local	Modelo de *Bayes* empírico
Edwards	Mediana local	Subtracção + função monotónica suave
Normexp	Mediana local	Modelo de convolução normal + exponencial
Normexp + offset	Mediana local	Igual à anterior a menos da adição de uma constante
Vsn	Mediana local	Método de estabilização da variância de Huber
Morph	Mediana local	Realização de uma abertura morfológica através do cálculo do mínimo e do máximo local

5.2.3 Pré-processamento dos dados

O pré-processamento tem por objectivo extrair ou evidenciar determinadas características dos dados. A normalização é apenas uma das várias transformações que podem ser aplicadas aos dados de *microarrays*, existindo, no entanto, outros tipos de pré-processamento.

Um deles resulta no uso de valores logarítmicos. A motivação prende-se com o facto dos efeitos na intensidade dos *microarrays* tenderem a ser multiplicativos. Por exemplo, a

duplicação da quantidade de RNA deve duplicar a intensidade do sinal para um intervalo abrangente de intensidades. O uso de um logaritmo converte estes efeitos multiplicativos (rácios) em efeitos aditivos (diferenças) que são mais fáceis de modular. Esta transformação não só simplifica a interpretação dos resultados como possui mais relevância biológica. Tipicamente é usada a base dois.

Outra consta na filtragem de *spots* tendo em consideração vários factores tais como o nível de intensidade ou de ruído. Uma das filtragens mais comuns baseia-se na eliminação de todos os *spots* cujo valor de intensidade é apenas ligeiramente acima do valor de *background*. A questão é que estes *spots* possuem valores de imprecisão elevados e, consequentemente, é provável que sejam de baixa qualidade. Alternativamente, podem ser filtrados todos os *spots* cuja relação sinal/ruído mínimo seja inferior a um valor pré-definido. De forma geral, a filtragem tem como finalidade desprezar os *spots* menos estáveis de forma a aumentar a qualidade global dos resultados finais.

5.2.4 Normalização

O objectivo da normalização é o de compensar diferenças técnicas sistemáticas entre *microarrays* de modo a que as diferenças biológicas entre amostras sobressaiam. Os principais desvios que motivam o uso de normalização são:

- Diferentes níveis de intensidade entre *microarrays*;
- Diferentes *backgrounds* típicos entre *microarrays*;
- Diferentes quantidades de mRNA utilizadas;
- Diferentes eficiências de incorporação dos marcadores usados;
- Diferentes velocidades de hibridação e, ou, diferentes condições experimentais usadas;
- Diferentes parâmetros de utilização no *scanner*.

A aproximação mais simples à normalização calcula-se através da divisão, ou subtracção no caso do logaritmo, pela média ou mediana dos valores. Este método apresenta, não obstante, algumas limitações no caso dos *microarrays* de cDNA, devido à distorção causada pela diferente eficiência de incorporação dos dois marcadores. São três os principais métodos que possibilitam lidar com este problema: *curve fitting and correction*, Lowess/Loess e *piece-wise linear normalization*. Neste documento focamo-nos no Lowess/Loess, todavia, uma descrição mais detalhada dos métodos disponíveis pode ser encontrada em [49, 84, 85].

O método Lowess (*LOcally WEighted linear regreSSion*), inicialmente proposto por Cleveland [86], foi aplicado aos *microarrays* com o principal objectivo de reduzir os efeitos de distorção causado pelos marcadores [87]. O seu funcionamento resulta da

divisão num conjunto de intervalos sobrepostos, sendo, então, encontrada uma função polinomial que se adapta a cada intervalo. Em alternativa a considerar todos os *spots*, uma variação do algoritmo resume-se à na sua aplicação a grupos de spots. Designado, de PrintTip Lowess esta aplica o algoritmo do Lowess a todos os *spots* associados a uma mesma agulha de impressão e tem como principal vantagem a possibilidade de considerar os efeitos individuais que cada agulha de impressão tem nos valores de intensidade.

5.3 Identificação de genes diferencialmente expressos

Apesar de um *microarray* poder conter dezenas de milhar de sondas, correspondentes a um número aproximado de genes, apenas um número reduzido destes se expressa de forma diferenciada para a condição experimental testada. A análise dos níveis de expressão permite a identificação destes genes. No entanto, visto tratar-se de dados experimentais susceptíveis de vários tipos de erros, é essencial proceder à quantificação do nível de confiança nos resultados obtidos.

Um dos métodos de análise disponíveis é o teste de significância estatística (também conhecido por teste de hipótese) inicialmente proposto por Ronald Fisher [88]. Este teste apoia-se no estabelecimento de duas hipóteses: uma hipótese nula H0 em que se assume a não existência de nenhuma diferença na expressão entre condições; e uma hipótese alternativa H1 em que se assume a existência de diferença na expressão. O nível de confiança nos resultados é representado por *p*, que traduz na probabilidade do gene ter sido seleccionado como diferencialmente expresso tendo em consideração que a hipótese nula se verifica, isto é, que o gene não se encontra diferencialmente expresso [89].

Dependendo do tipo de estudo, diferentes métodos podem ser usados no cálculo do valor de *p*, sendo que muitos destes métodos seguem funções de distribuição probabilística teórica, como por exemplo, a normal, *t* de *student*, qui-quadrado ou binomial. Contudo, nem todos os testes de significância são baseados no conhecimento de uma função de distribuição de probabilidade. No caso concreto dos *microarrays* o tamanho da amostra é tipicamente reduzido, o que torna difícil verificar a efectiva distribuição dos dados. Neste caso é aconselhável o uso de testes não paramétricos, ou seja, que não seguem nenhuma função de distribuição pré-estabelecida, substituindo-as pelo uso de *rankings*.

Em ambos os casos é, porém, necessário estabelecer uma métrica para decidir se devemos rejeitar a hipótese nula em detrimento da alternativa. De facto, quanto menor o valor de *p,* maior a evidência contra a hipótese nula. Por convenção, considera-se 5% como o valor máximo aceitável para rejeitar a hipótese nula. Deste modo, rejeita-se a hipótese nula e indica-se que os resultados são significantes, quando $p<=5\%$. Em contraste, sempre que $p>5\%$, conclui-se que não existe evidência suficiente para rejeitar a hipótese nula. Deve, nãoobstante, ter-se em conta que isto não indica que a hipótese nula seja verdadeira, simplesmente indica que não existe evidência suficiente para a sua rejeição.

5.3.1 Métodos de quantificação de genes diferencialmente expressos

A escolha do teste estatístico a utilizar está directamente relacionada com o desenho experimental, o tipo de variáveis e a distribuição dos dados. O esquema apresentado Figura 5.2 resume as decisões na escolha do método mais adequado para a quantificação de genes diferencialmente expressos através do uso do teste de significância em que os dados são numéricos [85, 90].

O primeiro factor determinante na escolha do teste estatístico a empregar fundamentar-se no número de condições em estudo. Com efeito comparar um tecido em estado normal com outro associado a uma determinada doença corresponde a um teste com duas condições, no entanto, se estiver a ser estudada a resposta de um micro-organismo a um fármaco ao longo de uma série temporal, trata-se claramente de um teste com mais de duas condições.

No caso de estarem a ser testadas duas condições, interessa saber se existe relação entre as amostras ou se estas são independentes. O exemplo anterior de comparação de um tecido em estado normal com um em estado patológico corresponde a uma situação em que existe independência entre as amostras. Um exemplo de uma experiência em que exista relação entre as amostras em teste será o de uma linha celular antes de depois de sofrer um tratamento químico. No caso de existirem mais de duas condições, é feita uma distinção entre a existência de independência entre as amostras e a realização de uma série temporal.

De seguida são apresentados alguns dos métodos mais usados para testar a expressão diferencial entre condições.

Fold Change

A experiência de *microarrays* mais simples é conduzida através da detecção de diferenças na expressão entre duas condições (por exemplo, normal *vs* doente). Nos *microarrays* de duas cores esta comparação de duas amostras pode ser realizada com único teste. Nestes casos o método mais simples de identificação de genes diferencialmente expressos consiste na avaliação da razão do valor da expressão das duas condições. Caso se utilize uma base logarítmica, são de interesse os genes cujo valor absoluto seja o mais afastado de zero. Um cenário comum é considerar como diferencialmente expressos os genes cujo valor de expressão de uma condição seja o dobro da outra, o que corresponde a um valor absoluto de 1 no logaritmo.

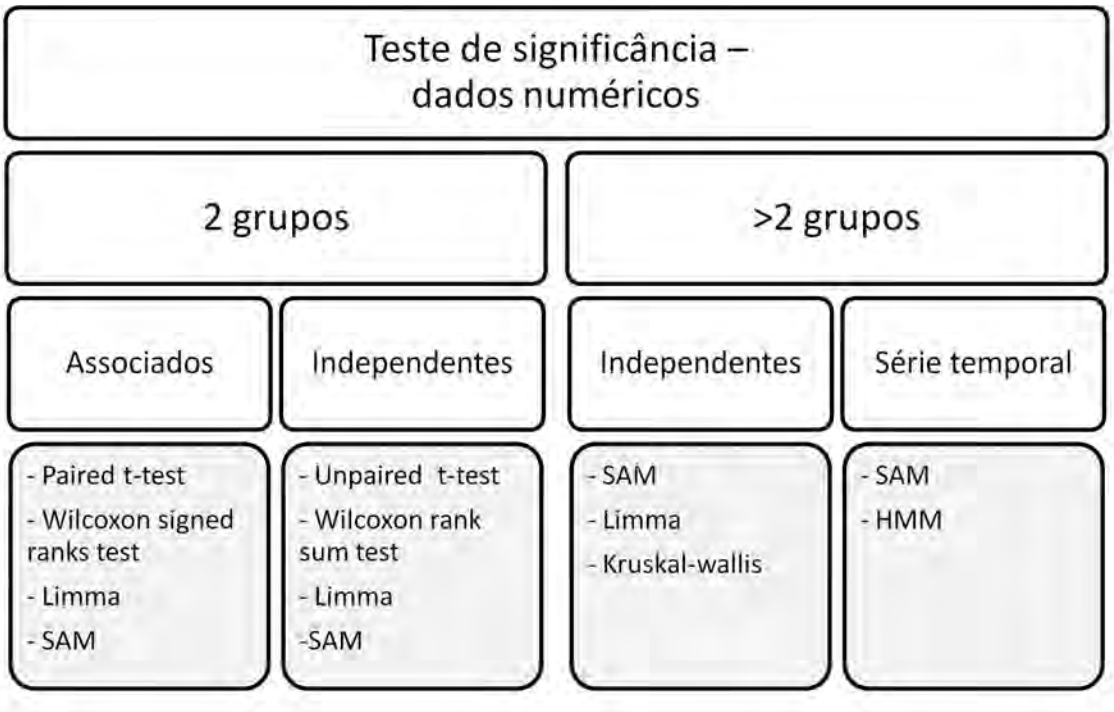


Figura 5.2: Métodos de quantificação da expressão diferencial aplicados a dados numéricos.

Este método apresenta, todavia, bastantes limitações. Por um lado, não pode ser considerado um teste estatístico, na medida em que não fornece nenhuma medida de confiança nos resultados. Por outro lado, os genes cujo nível de intensidade é bastante reduzido podem facilmente ser, erroneamente, identificados como diferencialmente expressos. Isto sucede porque os genes com baixos níveis de expressão possuem uma variância superior aos que possuem níveis de expressão mais elevados. Como solução para este problema foram propostos limiares de selecção dependentes da intensidade [91].

Teste *t* de *Student*

O teste *t* é um método estatístico que pode ser facilmente usado no cálculo de genes diferencialmente expressos. Ao tratar os genes individualmente, o teste *t* não é afectado pela heterogeneidade da variância existente entre genes. Este pode, ainda assim, ter uma capacidade reduzida devido ao facto da amostra - o número de *spots* para cada gene - ser pequena. Outro problema reside no facto das variâncias estimadas de cada gene não serem estáveis: por exemplo, para um gene com valor da variância estimado pequeno, o valor de *t* pode ser elevado, mesmo se o rácio das intensidades se aproximar de 1.

Esta dificuldade poe ser ultrapassada através do uso de uma estimativa de erro da variância comum a todos os genes, assumindo que a variância é homogénea entre genes distintos. Porém, assim, tratar-se-á, na verdade, de um teste de *fold change,* pois o valor de *t* global é ordenado da uma forma semelhante ao *fold change*, isto é, este não ajusta a variabilidade dos genes individualmente. E pode, então, sofrer do mesmo efeito de *bias* que o teste de *fold change,* caso a variância não seja constante para todos os genes.

Inspecção visual: *volcano plot*

O *volcano plot* resulta num método de inspecção visual que possibilita uma fácil identificação de genes de interesse. Este combina ambas as características do *fold change* com os critérios do teste *t*. O gráfico resultante relaciona o $-\log_{10}$ dos valores de transformados de *p* específicos para cada gene contra os valores de $\log_2$ *fold change* (Figura 5.3). Genes com significância na expressão encontram-se acima da linha horizontal. Genes com elevado *fold change* encontram-se na parte externa de ambas as linhas verticais. Os genes de interesse tendem, deste modo, a encontrar-se nos quadrantes superiores esquerdo e direito.

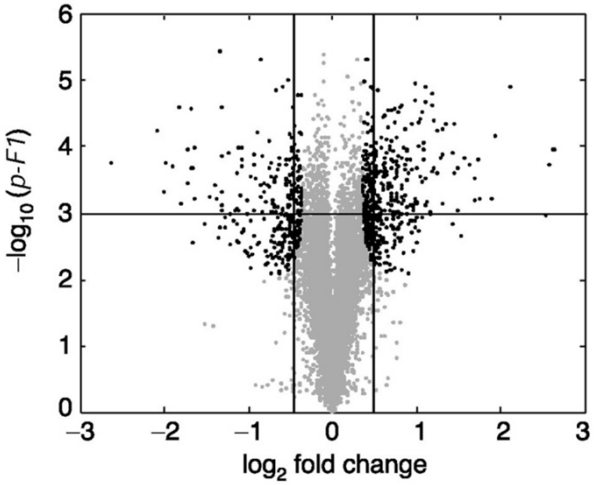


Figura 5.3: *Volcano plot* que relaciona os $-\log_{10}$ dos valores de transformados de *p* específicos para cada gene contra os valores de $\log_2$ *fold change*.

Outros métodos disponíveis

Existem, contudo, outros métodos que possibilitam a identificação de genes diferencialmente expressos. O método SAM (*Significance Analysis for Microarrays*), proposto por Tusher *et al* [92], foi desenvolvido para endereçar o problema dos reduzidos valores de variâncias através da utilização de um factor de expansão. Este método encontra-se disponível no pacote SAMR do BioConductor [93] ou como um Excel *add-in*. Outro método, proposto por Kerr et al [94], baseia-se na aplicação do modelo ANOVA para obter estimativas dos valores relativos de expressão para cada gene em cada amostra. Este encontra-se disponível através de pacote do Bioconductor designado de maanova. Smyth *et al* [95] propõem a combinação do método de *Baiseano* empírico como um teste *t* moderado e modelos lineares. O uso de um modelo linear oferece como principal vantagem a possibilidade de modelar desenhos experimentais mais complexos, incluindo *dye-swaps*. Este método é disponibilizado no pacote limma do BioConductor.

Para além dos métodos anteriores, existe ainda um conjunto de testes não paramétricos. Estes não assumem a existência de uma distribuição padrão nos dados, pelo que fazem uso de listas de ordenação para seleccionar os genes diferencialmente expressos. O teste de

Wilcoxon redunda numa aproximação muito mais robusta, quando comparado com os métodos anteriores, apresentando, no entanto, a desvantagem de perda de informação resultante da passagem dos dados numéricos para lista de resultados. Este teste possui diversas variantes, nomeadamente, a Wilcoxon *signed-rank test* que se aplica quando existe uma relação entre as amostras, e o Wilcoxon *rank-sum test,* que se aplica quando existe independência entre as amostras. Uma generalização do teste anterior para o uso de mais de duas condições resultou no teste de Kruskal-Wallis. A Tabela 5.2 resume os principais métodos disponíveis, e uma revisão detalhada do funcionamento destes encontra-se em [85, 96].

Tabela 5.2: Métodos disponíveis para identificação de genes diferencialmente expressos.

Método	Pacote	URLs
Fold change	R	http://cran.r-project.org/
Teste *t* de *Student*	R	http://cran.r-project.org/
Volcano plot	R/Limma	http://bioinf.wehi.edu.au/limma/
ANOVA	R/MaAnova	http://cran.r-project.org/web/packages/maanova
Wilcoxon's rank-sum	R/Multtest	http://cran.r-project.org/web/packages/multtest/
Limma	R/Limma	http://bioinf.wehi.edu.au/limma/
SAM	R/SAMR	http://www-stat.stanford.edu/~tibs/SAM/
Kruskal-Wallis	R/CMA	http://www.bioconductor.org/packages/release/bioc/html/CMA.html

5.3.2 Significância e testes múltiplos

Na realização de um teste de significância podem ser cometidos dois tipos de erros. Erro do tipo I, ou falso positivo, sempre que se afirma que um gene é diferencialmente expresso, quando na realidade não o é. Ou um erro do tipo II, ou falso negativo, quando se falha na identificação de um gene diferencialmente expresso [38, 89].

Numa experiência de *microarrays* são realizados milhares de testes estatísticos (um para cada gene) pelo que um número substancial de falsos positivos pode ser acumulado. De seguida são apresentados alguns dos métodos existentes para endereçar este problema, designado de problema de testes múltiplos.

Family-wise error-rate control

Uma das aproximações ao teste múltiplo consta no FWER (*Family-wise error-rate control*), que avalia a probabilidade de acumular um ou mais falsos positivos num conjunto de testes estatísticos.

O procedimento mais simples para efectuar a baseia-se na correcção de Bonferroni. Segundo esta, o nível de significância é dividido pelo número de testes. Existem, porém, outros métodos, tais como a correcção de um passo baseada em permutações e o ajustamento de Westfall and Young [90]. Conquanto de estes métodos garantam melhores resultados, também são mais exigentes em termos computacionais.

False-discovery-rate control

Uma aproximação alternativa ao problema dos testes múltiplos está na avaliação da FDR (*False Discovery Rate*). Esta cifra-se na proporção de falsos positivos no conjunto de todos os genes inicialmente considerados como diferencialmente expressos, ou seja, entre todas as hipóteses nulas rejeitadas.

Ao contrário do nível de significância, que é determinado antes de se olhar para os dados, o FDR é uma medida de confiança pós-dados. Esta usa informação disponível nos dados para estimar a proporção de falsos positivos que ocorreram. Numa lista de genes diferencialmente expressos que satisfaz o critério imposto pela FDR, pode esperar-se que uma proporção destes corresponda a falsos positivos. O critério FDR permite uma taxa superior de falsos positivos, conseguindo obter melhores resultados do que os procedimentos FWER.

5.4 Implementação da análise de dados no Mind

Na secção anterior foram apresentados os principais procedimentos de tratamento e de análise de uma experiência de *microarrays*. Na concretização da análise são tipicamente aplicadas várias ferramentas, sendo que cada uma possui as suas vantagens, tendo em consideração os métodos que disponibilizam. Por exemplo, a aplicação LimmaGUI, apesar de possuir vários métodos de correcção do *background* e de normalização, apenas usa o algoritmo Limma na identificação de genes diferencialmente expressos. Em opção, o MeV [66], embora possua um conjunto alargado de métodos de análise diferencial, apresenta limitações na filtragem e normalização.

O facto de ser necessário usar mais do que uma ferramenta implica a migração dos dados entre aplicações, o que constitui uma limitação não só pelo tempo dispendido, como pela possibilidade de introdução de erros com sucessivas transformações realizadas. Tendo em consideração que o Mind já armazena os dados relativos à expressão dos genes, a integração sobre esta plataforma das principais ferramentas de análise constitui uma mais-

valia para o investigador. De seguida são apresentados os detalhes da implementação sobre o sistema Mind do fluxo de análise de dados.

5.4.1 Modelo de navegação

O modelo de navegação apresentado na Figura 5.4 consiste no detalhe da análise de dados do modelo global do sistema Mind. Foram considerados dois elementos de topo que englobam as funcionalidades de análise apresentadas anteriormente: controlo de qualidade e quantificação de genes diferencialmente expressos. Foi ainda adicionada uma entrada, transversal aos conceitos anteriores, que consiste na gestão de *datasets*. A pertinência da existência de *dataset* advêm do facto de nem sempre ser de interesse considerar todos os *microarrays* de uma experiência para análise. A isto acresce a possibilidade de criar várias vistas sobre a mesma experiência, o que pode ser útil no teste de vários desenhos experimentais, ou de procedimentos de análise alternativos. O controlo de qualidade é constituído por duas fases: definição dos parâmetros de análise e visualização do relatório com os resultados. A quantificação de genes possibilita a aplicação de um conjunto de quatro algoritmos aos dados.

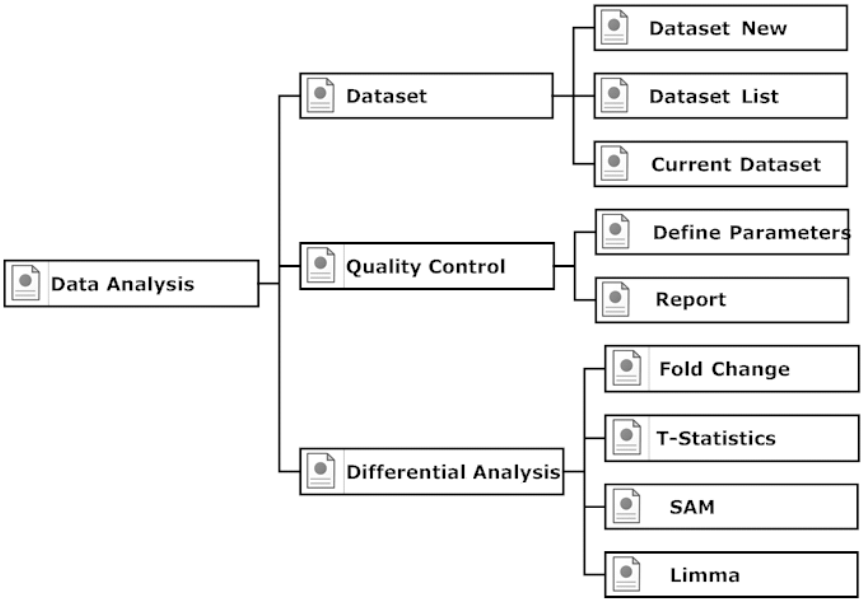


Figura 5.4: Modelo de navegação da análise de dados no Mind.

5.4.2 Levantamento de ferramentas

De forma a evitar a replicação de trabalho anterior, foi realizado um esforço na identificação de bibliotecas programáticas que disponibilizassem as funcionalidades pretendidas. Neste particular, foi de especial relevo o projecto do BioConductor [70, 97] que promove o desenvolvimento de software livre para a análise e compreensão de dados experimentais produzidos em biologia molecular. Este projecto assenta primariamente na

linguagem de programação R [98], e, entre a extensa lista de pacotes, está incluída uma quantidade considerável para análise de dados de *microarrays*.

Na Tabela 5.3 encontram-se, para cada fase da análise, os métodos seleccionados assim como a respectiva biblioteca e referência. A fase do controlo de qualidade foi essencialmente implementada com recurso à biblioteca Limma [49, 98], tendo, no entanto, a fase de filtragem sido realizada directamente na linguagem R. A justificação desta decisão reside no facto de, uma vez lidos os dados, ser mais fácil o uso de R do que C#. A fase de identificação de genes diferencialmente expressos foi também implementada em R, tendo sido usado um conjunto mais alargado de bibliotecas, tais como o SAMR [92] e a MaAnova [99]. Por fim, para a realização de testes múltiplos foi usada a biblioteca Multtest [99].

Tabela 5.3: Resumo dos métodos de análise usados e das respectivas estratégias de implementação.

Fase	Métodos usados	Biblioteca	
Controlo de qualidade	Filtragem	R	[98]
	Correcção do *background*	R/Limma	[49, 98]
	Normalização	R/Limma	[49, 98]
Identificação de genes diferencialmente expressos	*Fold change*	R	[98]
	Teste *t* de *Student*	Limma	[95]
	SAM	SAMR	[92]
	LIMMA	Limma	[95]
	ANOVA	MaAnova	[99]
	Wilcoxon	Multtest	[99]
	Kruskal-Wallis	CMA	[100]
Testes múltiplos	*False-discovery-rate*	Multtest	[99]

5.4.3 Integração do processamento em R

Apesar das vantagens evidentes no uso das bibliotecas em R para tratamento e análise dos dados, esta opção apresentou um conjunto de dificuldades técnicas relacionadas com o facto do Mind se encontrar desenvolvido em C#. Foi deste modo necessário estabelecer um mecanismo que possibilitasse a execução de código R em ambiente .Net.

O primeiro passo consistiu no teste e validação dos métodos existentes de comunicação entre estas duas plataformas. A solução originalmente testada, R-(D)COM[1], foi desenvolvida com o objectivo de usar o motor DCOM para facilitar o desenvolvimento de aplicações cliente R. Foi ainda testada a biblioteca ServeR [101] que possibilita, de um modo bastante intuitivo, a execução de métodos R sobre objectos C#.

Se bem que interessantes, os resultados dos testes realizados demonstraram falta de estabilidade com o aumento da quantidade de dados partilhados entre os dois ambientes. De facto, na análise de uma experiência com mais de 12 *microarrays,* o sistema demonstrou-se bastante instável. Ainda, no caso do ServeR, o encapsulamento de entidades R em objectos C#, que à partida se apresentava como uma mais-valia, acabou por se revelar confuso, dificultando a organização e a transparência do código obtido. Em alternativa às soluções anteriores, optou-se pelo uso de uma estratégia em que o código R é directamente invocado através de *shell scripts*.

Armazenamento e geração de scripts

Tal como o esquema da Figura 5.5 ilustra, para um determinado pedido de execução lançado pelo utilizador é gerado o *script* R correspondente. Este processo realiza-se através do acesso à base de dados onde são armazenados *templates* dos *scripts* disponíveis e à base de dados do Mind para obtenção dos ficheiros de expressão. Após instanciação dos *scripts* com os dados da análise em questão, procede-se à sua execução. Esta concretiza-se pela criação de um processo em C#, no qual o programa R é executado tendo como argumento de execução o *script* anteriormente gerado. Ao ser lançado num processo separado do Mind, a execução decorre de forma assíncrona. Após finalização do processamento é enviado um alerta para o Mind sendo os ficheiros resultantes da análise armazenados numa *cache* temporária.

Distribuição da carga de processamento

A execução dos processos é realizada de forma a rentabilizar as capacidades de processamento do computador. Ainda que, de acordo com as actuais taxas de utilização, não exista uma necessidade específica de processamento intensivo, no desenvolvimento foi tida em consideração a posterior inclusão da distribuição da carga de processamento. Estando o processo de análise fisicamente separado do código do Mind, é fácil, através da adição de novas máquinas, a execução remota de *scripts*.

[1] http://www.sciviews.org/_rgui/projects/RDcom.html

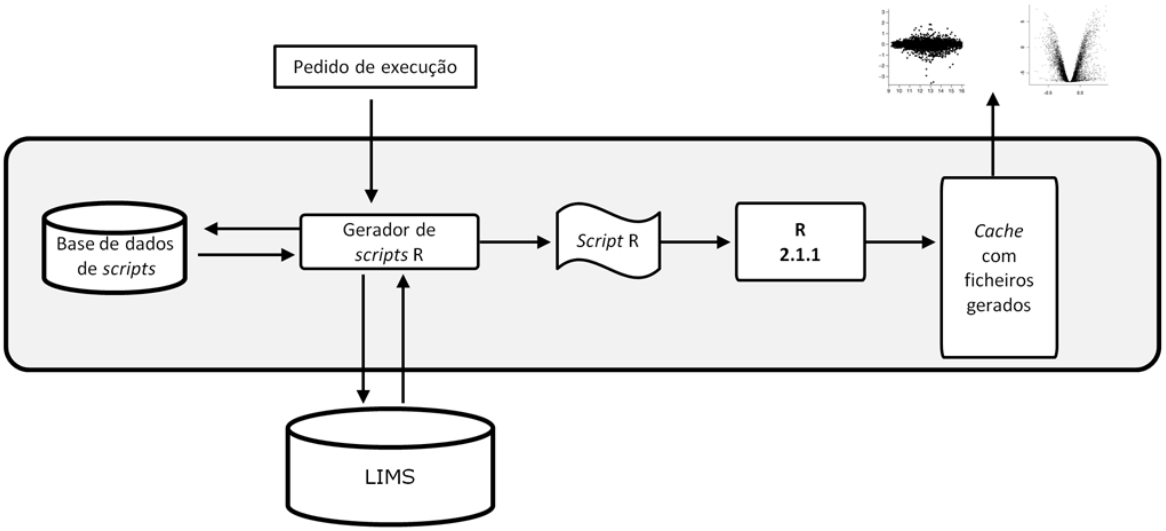


Figura 5.5: Execução de *scripts* R no Mind.

5.4.4 Modelo de dados

Adicionalmente ao esquema da base de dados do Mind, apresentado na Figura 3.9, existe um conjunto adicional de tabelas responsável pelo armazenamento da informação associada com a análise de dados (Figura 5.6). Este modelo permite o armazenamento dos *datasets* criados pelo utilizador, dos respectivos detalhes e dos resultados das análises realizadas.

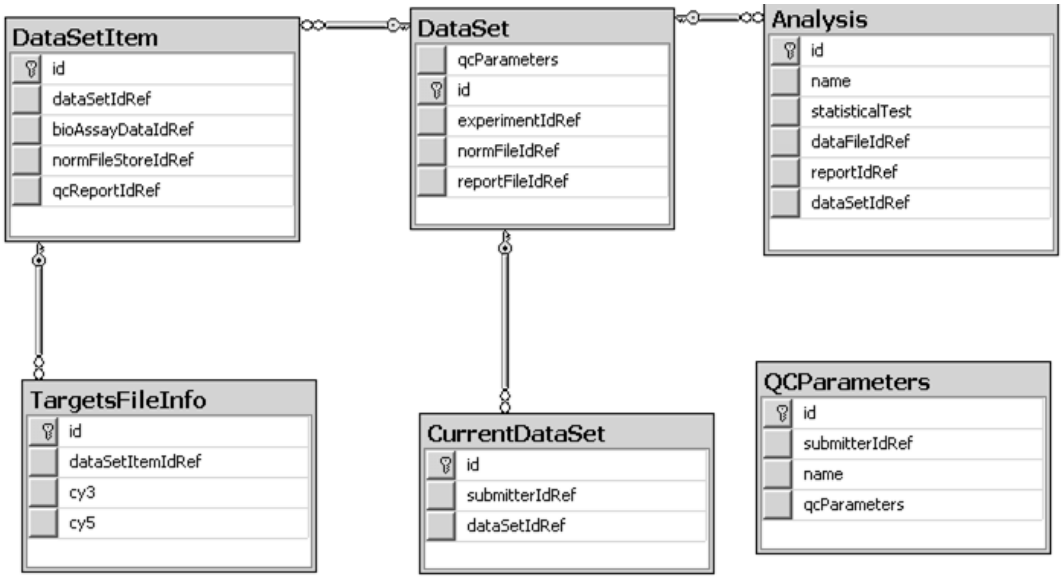


Figura 5.6: Modelo de dados da análise no Mind.

5.5 Exemplo de utilização

Esta secção visa exemplificar as diversas fases da análise de dados no Mind e como algumas das funcionalidades desenvolvidas podem ser exploradas. Para tal, é analisado um conjunto de dados resultantes de uma experiência executada no laboratório de *microarrays* da Universidade e previamente armazenada no sistema Mind.

5.5.1 Descrição geral da experiência

A experiência em questão assenta na comparação dos níveis de expressão de levedura (*Saccharomyces cerevisiae*) antes e depois de esta ser sujeita a um choque térmico. As células de controlo foram obtidas através de crescimento num meio de YPD (*Yeast Peptone Dextrose*), a uma temperatura de 30ºC, até um OD600 (*Optical Density at 600nm*) de 0,5. As células do choque térmico foram incubadas a 37ºC durante 20 minutos. A amostra correspondente ao choque térmico (HS) e a respectiva amostra de controlo (C) foram etiquetadas com diferentes fluoróforos (Cy3 e Cy5) e hibridado contra o mesmo *microarray*. Nesta experiência foram usados seis *microarrays*, um em cada uma de três culturas (réplicas biológicas) combinadas com os respectivos *dye-swap*. O desenho experimental usado encontra-se na Figura 5.7.

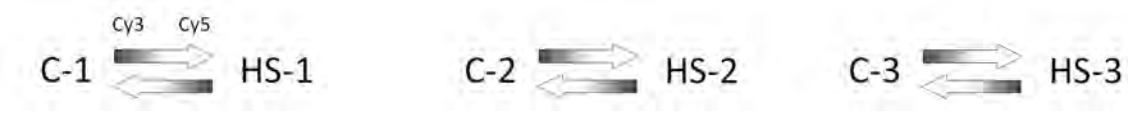


Figura 5.7: Desenho experimental da experiência de choque térmico. Três réplicas biológicas usadas com *dye-swap*.

5.5.2 Criação do *dataset*

O primeiro passo assenta na criação de um *dataset* que corresponde ao conjunto de dados a analisar. No Mind, o acesso às funcionalidades de análise de dados encontra-se no menu superior (Figura 5.8). De seguida, as opções do nível dois do modelo de navegação foram transpostas para o menu lateral esquerdo, estando as de terceiro nível no topo do painel de contexto. Caso exista algum *dataset* que tenha sido previamente seleccionado, são apresentados os seus detalhes, caso contrário, é apresentada a lista de todos os *datasets* disponíveis.

A criação de um novo *dataset* requer a selecção da experiência e dos ficheiros com os valores de expressão a considerar. Na existência de mais do que um ficheiro para o mesmo *bioassay,* é necessário resolver a ambiguidade, na medida em que é apenas possível uma única selecção.

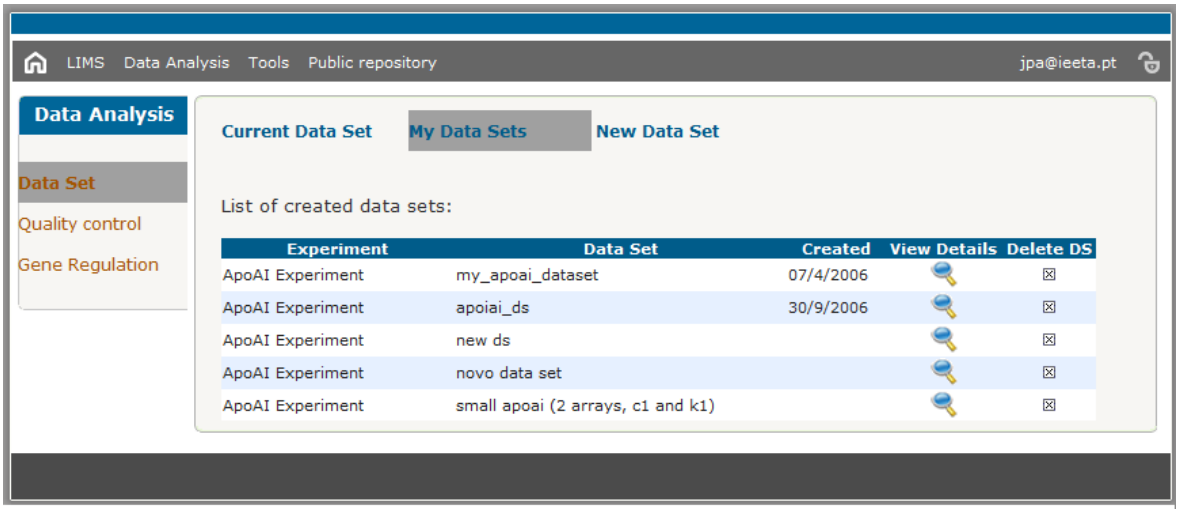


Figura 5.8: Gestão de *datasets* na análise de dados.

5.5.3 Controlo de qualidade

O passo seguinte resulta na validação da qualidade dos dados. Neste exemplo, apenas são apresentados os resultados para um dos *microarrays* (C-1 -> HS-1). No entanto, o mesmo procedimento deve ser aplicado aos restantes cinco.

Inserção dos parâmetros de controlo de qualidade

O procedimento de controlo de qualidade requer, antes de mais, a inserção dos parâmetros a serem usados na análise. É necessário indicar controlos, critérios de filtragem, correcção de *background*, normalização e gráficos a criar. Esta tarefa é, não obstante, facilitada através do uso de perfis de controlo de qualidade previamente usados e guardados (Figura 5.9.a). Após a finalização do processamento, fica disponível, para cada *microarray*, um relatório com o resultado do controlo de qualidade (Figura 5.9.b).

Visualização e interpretação dos resultados

Procede-se, então, a análise dos resultados do relatório de controlo de qualidade. O próximo passo é a correcção do *background*. Com este propósito recorre-se ao método *Standard* que se baseia na subtracção da mediana local aos valores de intensidade. Para facilitar a visualização e comparação dos resultados são usados os gráficos MA apresentados na Figura 5.10.

O gráfico MA relaciona os valores de M, no eixo vertical, com os valores de A, no eixo horizontal. M é definido por $M = log_2 R - log_2 G$ e representa a diferença de expressão entre as duas condições experimentais. A é definido por $A = \frac{1}{2} \times (log_2 R + log_2 G)$ e representa os valores médios de intensidade de cada *spot*. A análise deste gráfico assenta na premissa de que a maioria dos *spots* não possui alterações na sua expressão e consequentemente, se distribuem ao longo da linha horizontal (M=0). Permite ainda

detectar em que intensidades existem *spots* com diferenças de expressão, assim como, através do uso de cores, a localização de controlos ou *spots* específicos.

a) b)

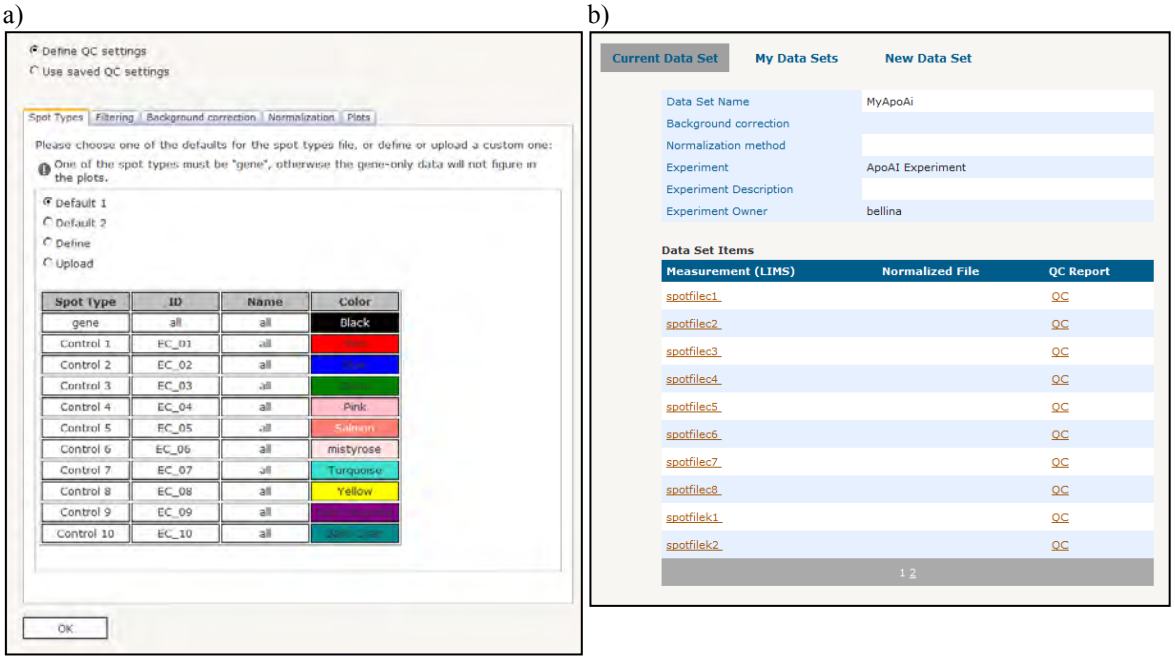


Figura 5.9: Controlo de qualidade do *microarray* (C-1 -> HS-1): a) definição dos parâmetros; b) relatório.

Da análise dos gráficos da Figura 5.10, verifica-se que a correcção do *background* fez com que mais *spots*, essencialmente na zona central de intensidade, ficassem com um valor absoluto de M superior.

a) b)

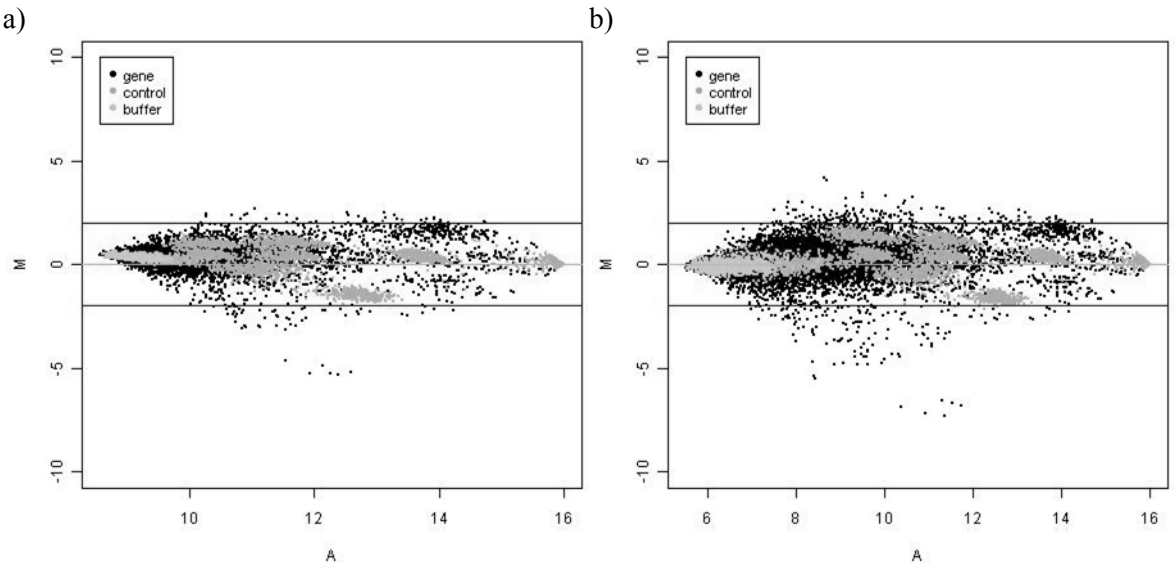


Figura 5.10: Gráficos MA: a) sem correcção do *background*; b) com o método *Standard* (subtracção). Os pontos a negro representam genes e os a cinzento os controlos.

Segue-se, então a normalização dos dados. A Figura 5.11 contém o gráfico MA para dois métodos de normalização: a) *lowess* e b) *print-tip lowess*. Para a análise foi considerado o uso do método *print-tip lowess* aplicado unicamente aos genes.

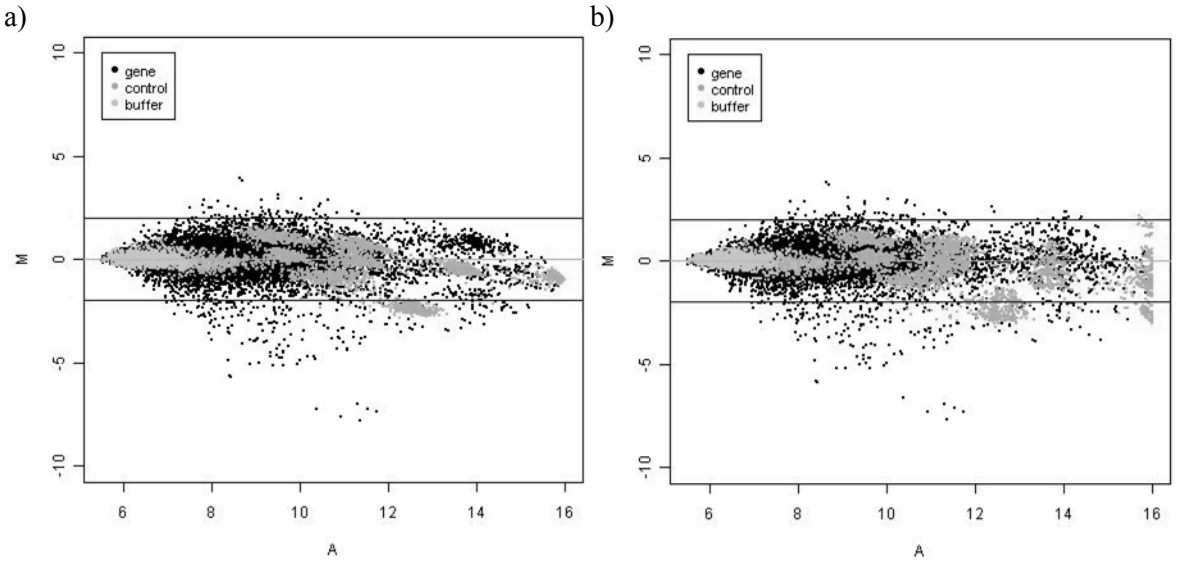


Figura 5.11: Gráfico MA com comparação dos métodos de normalização apenas aplicados aos genes: a) método *lowess*; b) método *print-tip lowess*.

A representação anterior não possibilita uma distinção clara entre os dois métodos de normalização. Esta tarefa pode ser facilitada através da visualização de *boxplots,* apresentadas na Figura 5.12, em que cada coluna representa uma zona impressa por uma determinada agulha. Existe uma notória uniformização da distribuição dos dados no *boxplot* com *print-tip lowess* (Figura 5.12.b).

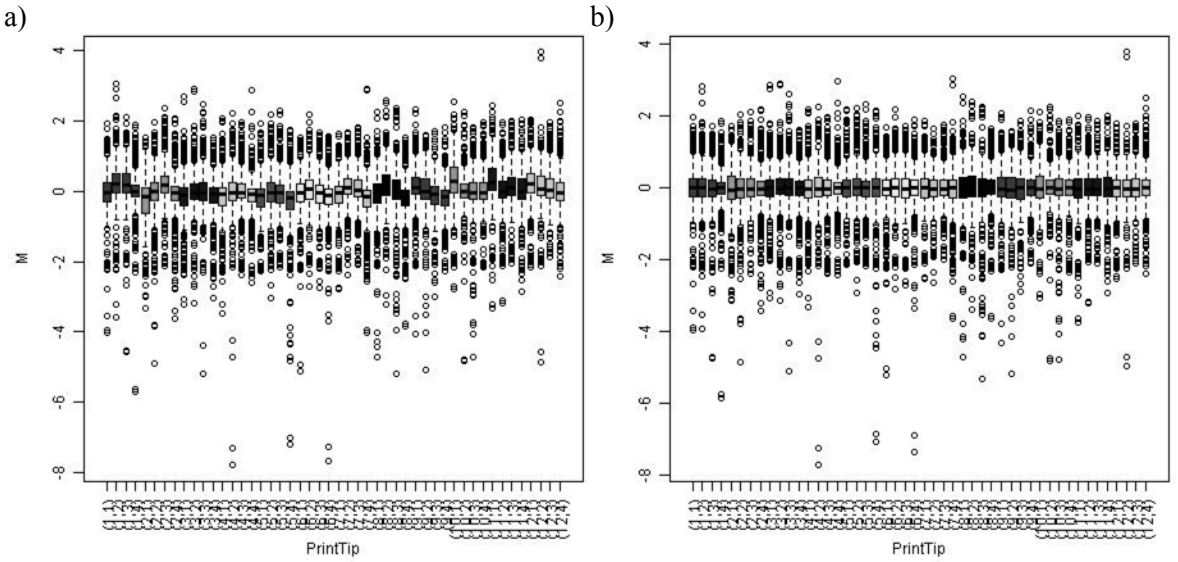


Figura 5.12: *Boxplot* com comparação dos valores de M com: a) *lowess*; b) *print-tip lowess*.

5.5.4 Identificação de genes diferencialmente expressos

Após a execução dos procedimentos de controlo de qualidade, apresentados na subsecção anterior para o *microarray* (C-1 -> HS-1), procede-se à identificação dos genes diferencialmente expressos. Deve seleccionar-se a opção *Gene Regulation* e confirmar os dados a analisar. Ficam assim acessíveis os métodos de análise disponíveis, bem como a aplicação de filtros (Figura 5.13).

A escolha do método de análise deve ter em consideração o desenho experimental. Neste caso existem dois grupos (HS e C) com três réplicas biológicas, assumindo-se a independência entre ambos. Deste modo, na identificação dos genes diferencialmente expressos são usados três métodos: teste *t*, Limma e SAM.

Os parâmetros utilizados no teste *t* envolvem uma distribuição com duas caudas e com 4 graus de liberdade. Tal como o *volcano plot* da Figura 5.14.a ilustra são seleccionados os genes com um valor M>5 e com um valor $p<0.01$. Com estas condições, obtém-se uma lista inicial de 256 genes, a qual, depois de eliminados controlos e genes repetidos, apresenta 49 genes. Na Figura 5.14.b encontra-se um *volcano plot* com os resultados do Limma. Através do uso da estatística de B usada pelo Limma, que, do mesmo modo do que o teste *t* consiste numa medida de confiança nos resultados, são seleccionados os 35 genes do topo. Na análise SAM são eleitas duas classes independentes, tomando-se 15 como valor de delta. Com estes parametros foi obtida uma lista de 150 genes da qual se eliminaram controlos e genes repetidos, obtendo-se a lista de 41 genes apresentados (Figura 5.15.a). O passo final consiste na comparação das listas devolvidas pelos vários métodos de análise. O resultado, apresentado na Figura 5.15.b, mostra os 21 genes que foram identificados em todos os métodos considerados.

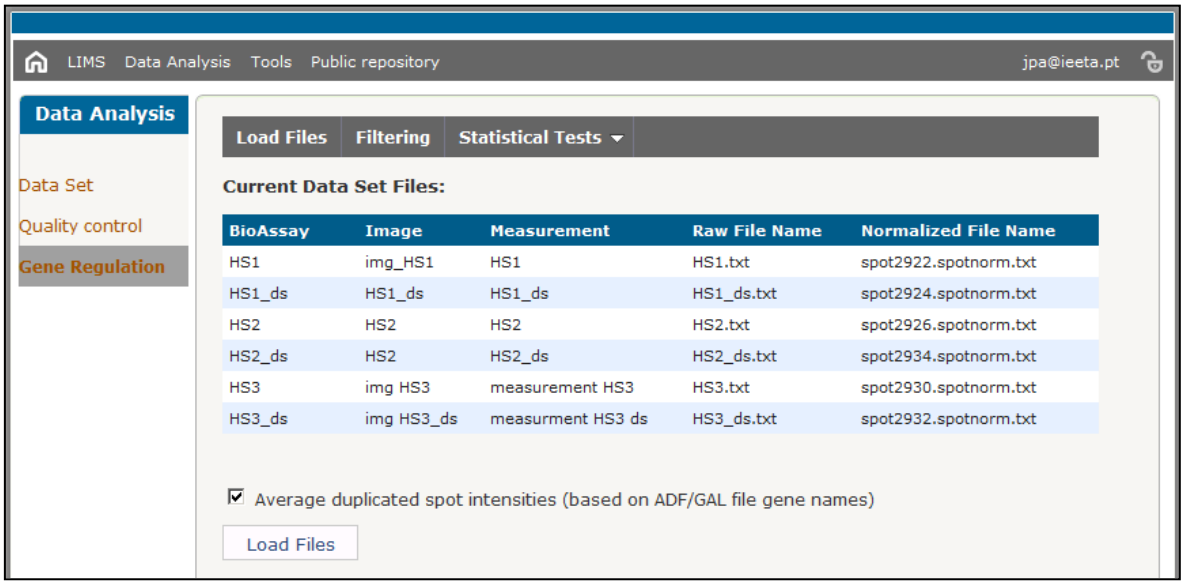
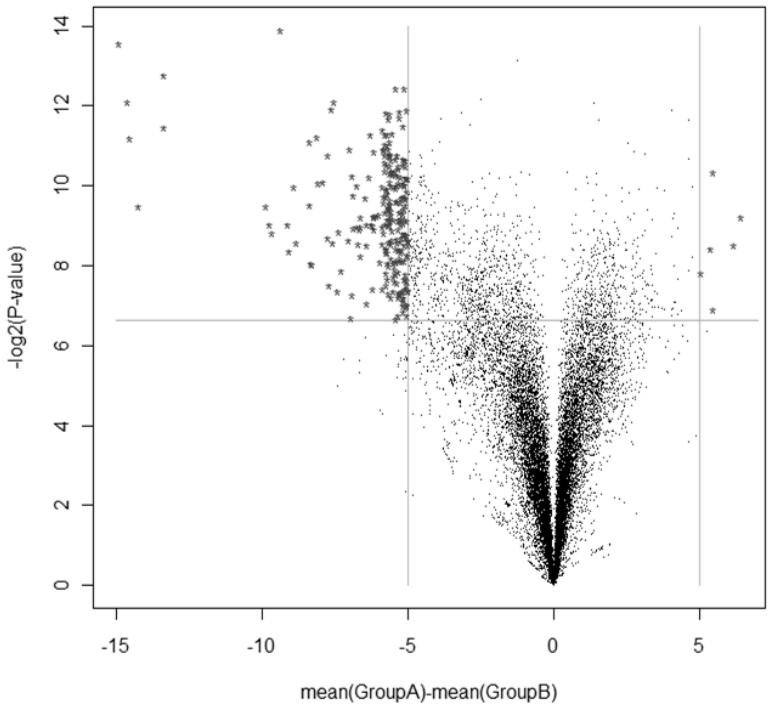


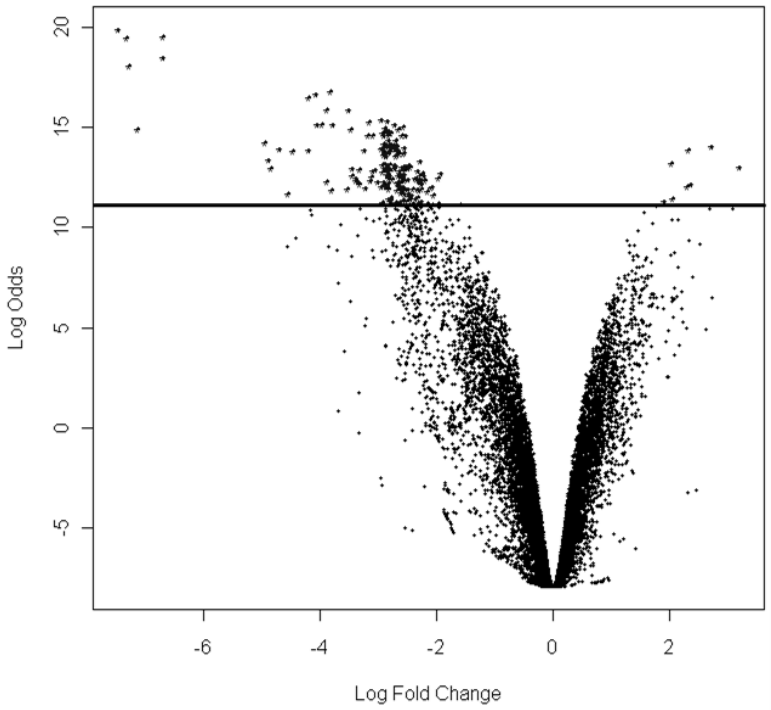
Figura 5.13: Leitura dos dados para realizar a análise diferencial.

a) teste *t* de *Student*



YBL064C	YHR055C
YBR054W	YHR087W
YBR072W	YHR104W
YBR169C	YKL096W
YBR233W-A	YKL103C
YBR299W	YKL151C
YCL040W	YLL026W
YCL064C	YLR178C
YDL130W-A	YLR216C
YDL159W-A	YLR258W
YDL204W	YLR259C
YDR070C	YLR327C
YDR171W	YML100W
YDR214W	YML128C
YDR258C	YMR105C
YER103W	YMR169C
YFL014W	YMR173W
YFR053C	YMR173W-A
YGL037C	YMR196W
YGR008C	YNL015W
YGR043C	YOR020C
YGR088W	YOR120W
YGR248W	YOR374W
YGR292W	YPR160W
YHR053C	

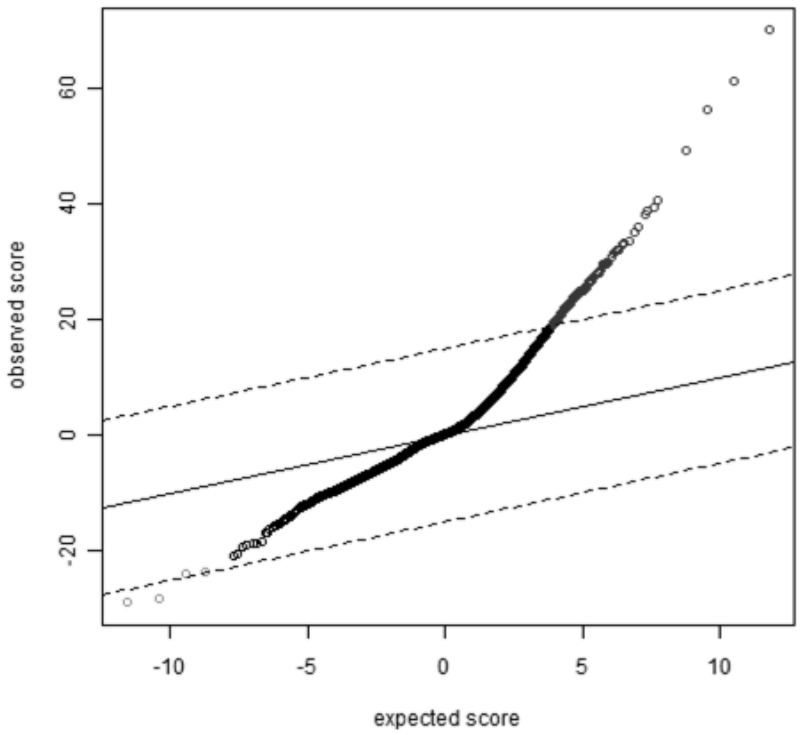
b) Limma



YBR054W	YGL037C
YBR072W	YGR008C
YBR149W	YGR061C
YBR169C	YGR292W
YBR233W-A	YHR053C
YCL030C	YHR104W
YCL035C	YKL096W
YCL040W	YLL026W
YDL204W	YLR106C
YDR001C	YLR178C
YDR214W	YLR216C
YDR258C	YML100W
YER103W	YMR058W
YFL014W	YMR120C
YFR053C	YMR169C
YNL015W	YMR173W
YOR374W	YMR173W-A
YPR160W	

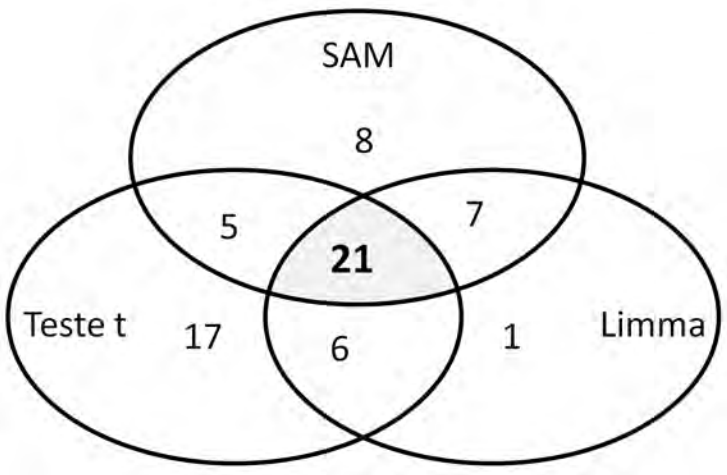
Figura 5.14: Resultado dos métodos de identificação de genes diferencialmente expressos: a) 49 genes identificados com teste *t* de *Student*; b) 35 genes identificados no Limma.

a) SAM



YBL064C	YGR287C
YBR053C	YGR292W
YBR054W	YKL024C
YBR067C	YKL096W
YBR072W	YLL026W
YBR126C	YLR106C
YBR149W	YLR178C
YBR169C	YLR216C
YCL030C	YLR438W
YCL035C	YML100W
YDL130W-A	YML128C
YDL204W	YMR120C
YDR001C	YMR169C
YDR214W	YMR170C
YDR258C	YMR173W
YER103W	YMR196W
YFL014W	YNL015W
YFR053C	YOR374W
YGL037C	YPL154C
YGR061C	YPR160W
YGR248W	

b) Diagrama de *Venn*



YBR054W	YKL096W
YBR072W	YLL026W
YBR169C	YLR178C
YDL204W	YLR216C
YDR214W	YML100W
YDR258C	YMR169C
YER103W	YMR173W
YFL014W	YNL015W
YFR053C	YOR374W
YGL037C	YPR160W
YGR292W	

Figura 5.15: Resultado dos métodos de identificação de genes diferencialmente expressos. a) 41 genes identificados com a análise SAM; b) 21 genes seleccionados com a intersecção dos três métodos.

5.6 Sumário

Neste capítulo foi apresentado o fluxo típico da análise de uma experiência de *microarrays,* assim como um levantamento das principais ferramentas e métodos disponíveis. Foi proposto e implementado, sobre o sistema Mind, um conjunto de ferramentas que possibilitam a análise completa de um estudo de *microarrays*. Por fim, foi exemplificada a sua utilização com recurso a uma experiência em que se testa o efeito de choque térmico em levedura.

Conquanto o conjunto de algoritmos apresentado não seja inovador, pois faz parte do que é o estado da arte neste domínio, a sua utilização integrada no ambiente LIMS consiste numa grande mais-valia, na medida em que facilita a tarefa. De facto, ao integrar, sobre uma interface única, algoritmos e mecanismos de visualização que possibilitam a análise exaustiva de uma experiência, consegue-se evitar os problemas associados com o uso de várias ferramentas. Com efeito, a necessidade de conhecer várias interfaces distintas, assim como a transferência de dados entre ferramentas tipicamente conduzem a erros de conversão. No global, com o uso integrado do sistema Mind para armazenamento e análise dos dados consegue-se agilizar os procedimentos de análise de uma experiência de *microarrays*.

Capítulo 6

6 Integração de dados biológicos

Nos últimos anos assistiu-se a um aumento exponencial da quantidade de dados disponível na área da biologia molecular. Este crescimento manifesta-se não só pelo contínuo incremento da informação armazenada por bases de dados estabelecidas, como o UniProt [102] e o GenBank [103], mas também pelo aparecimento de bases de dados focalizadas em novos domínios.

O principal motor de criação de todo este conjunto de dados consistiu no desenvolvimento de novas técnicas de sequenciação e de monitorização da expressão génica em larga escala. O facto dos resultados obtidos serem disponibilizados publicamente em bases de dados estruturadas veio criar novas oportunidades, permitindo a aplicação de métodos estatísticos e computacionais para inferir novas relações entre as entidades armazenadas, o que, consequentemente contribuiu para o aumento dos dados. Um exemplo recorrente é o conjunto de bases de dados de interacção entre proteínas, como o InterPro [104], obtidas através da aplicação de modelos computacionais a dados pré-existentes.

Na realidade, os dados disponíveis, sejam eles originais ou derivados, possuem um elevado interesse para responder a novas questões biológicas. Estas questões requerem que dados de várias fontes sejam relacionados de forma a extrair relações até então desconhecidas. Por exemplo, um geneticista interessado no estudo da obesidade em humanos restringe o seu objecto de estudo a uma região de 5Mb do cromossoma 1. No intuito de dar resposta a este problema, e em alternativa a realizar experiências que analisem de uma forma cega toda a região, recorre a uma base de dados sobre a qual formula a seguinte pergunta: Existem genes nesta região homólogos aos de alguns dos organismos modelo e que estejam envolvidos na regulação do metabolismo de lípidos? Encontrar uma resposta coloca um desafio técnico nos domínios de acesso e da relação entre dados proveniêntes de várias bases de dados geograficamente dispersas e que possuem diferentes modelos de acesso e de partilha (questão adaptada de [105]). A integração de dados caracterizados pela sua heterogeneidade constitui mesmo um dos maiores desafios ao desenvolvimento de

ferramentas capazes de extrair relações de dados biológicos. No caso concreto dos *microarrays*, após a obtenção dos dados de expressão e do *cluster* de genes de interesse, é imperativo perceber o contexto biológico adjacente, pelo que o acesso a estas bases de dados é essencial [106, 107].

Este capítulo visa propor ferramentas que façam uso dos dados biológicos disponíveis para auxiliar a interpretação de resultados de estudos de expressão génica. É realizado um levantamento das principais bases de dados biológicas, das políticas de disponibilização dos dados, assim como das condicionantes existentes. É apresentado o desenvolvimento de duas aplicações *web* (GeneBrowser e Quext) que permitem a integração das relações existentes entre genótipo e fenótipo. De forma a facilitar a tarefa de aceder a fontes dispersas e, consequentemente, o desenvolvimento destas aplicações, é ainda proposta uma plataforma de integração de dados biológicos (Gens).

6.1 Interpretação biológica de estudos de expressão génica

O estudo das relações entre o genótipo e o fenótipo é fundamental para compreender a complexidade dos sistemas biológicos. A principal dificuldade coloca-se no facto dos genes não operarem sozinhos nas células mas numa intrincada rede de interacções em que múltiplos genes podem, directa, ou indirectamente, influenciar o mesmo fenótipo. O desenvolvimento de tecnologias de análise do genoma em larga escala, tais como os *microarrays*, criou a possibilidade de ter uma vista holística de um sistema biológico. De facto, genes co-expressos tendem a estar envolvidos nos mesmos processos e alguns estudos mostram que, mesmo em eucariontes superiores, genes funcionalmente relacionados tendem a estar juntos no genoma.

A compreensão dos resultados destas experiências deve ser realizada no contexto da biologia de sistemas, em que múltiplos conceitos são simultaneamente relacionados de forma a obter-se o resultado final. Com a popularização do *microarray* de DNA, nos últimos anos, várias ferramentas informáticas têm sido propostas para endereçar esta questão. Uma aproximação usualmente aplicada baseia-se no enriquecimento de classes funcionais que agrupam genes presentes no *dataset* da experiência. Ferramentas como o OntoExpress [108], o FatiGO [106] ou o GOminer [109] são algumas das mais relevantes. No entanto, devido à complexidade do fenómeno, uma aproximação mais sistemática é necessária.

O trabalho apresentado neste capítulo divide-se em três secções. Primeiro é analisada a problemática de acesso aos dados biológicos, sendo exposto um levantamento das principais bases de dados, das condicionantes ao seu acesso e das estratégias de acesso integrado existentes. É ainda apresentado o desenvolvimento de uma plataforma de integração de dados biológicos, designada por Gens, usada como base na construção das restantes aplicações. De seguida, é explanada a elaboração da aplicação *web* GeneBrowser,

que combina várias fontes de dados com mecanismos de visualização, de forma a facilitar a tarefa de interpretar biologicamente o resultado de uma experiência de expressão génica. Por fim, propõe-se e detalha-se o desenvolvimento da ferramenta Quext que possibilita a análise da literatura associada com um conjunto de genes fornecidos.

6.2 Motivação à integração de dados biológicos

As primeiras bases de dados em biologia molecular eram constituídas por repositórios de reduzidas dimensões que continham informação relativa a um determinado tópico de estudo. A difusão de técnicas laboratoriais em larga escala, tais como *microarrays*, levaram ao crescimento, tanto em dimensão como em número, destes arquivos, o que originou uma enorme dispersão e fragmentação dos dados representativos do conhecimento biológico.

De forma a colmatar esta lacuna foram propostos vários repositórios a nível mundial com o objectivo de servirem de ponto integrador. São exemplos o Ensembl [110], o UniProt [102], o Entrez [111] e o KEGG [112]. Porém, a constante criação de novas áreas de estudo e o aumento de dados derivados, obtidos através do processamento dos dados originais, fez com que o número total de bases de dados biológicas continuasse em crescimento. Actualmente existem mais de 1170 bases de dados disponíveis [113].

São dois os propósitos que justificam a integração de dados biológicos. Primeiro, porque os dados relativos a uma entidade biológica podem-se encontrar dispersos em várias bases de dados. Por exemplo, para um gene a informação relativa à sequência do nucleótido encontra-se armazenada no GenBank, às vias metabólicas no KEGG Pathway e aos dados de expressão no ArrayExpress [114]. A capacidade de obter uma vista unificada destes dados é crucial para compreender o papel desempenhado pelos genes. Uma segunda razão reside no facto de bases de dados, sobre o mesmo tópico de estudo, conterem dados incompletos, redundantes ou sobrepostos. A integração destas bases de dados permite a complementaridade da informação e a detecção de incongruências através da comparação directa dos dados armazenados.

Se, de início, o facto das bases de dados se encontrarem dispersas não ter representado grande problema, visto cada investigador se centrar num determinado domínio de estudo, actualmente, cada vez mais se pretende uma visão holística dos dados e, consequentemente, maior é a necessidade de os integrar [115].

Microarrays

O resultado típico de uma experiência de *microarrays* é um conjunto de ficheiros tabulares com os valores de expressão de cada gene. Destes, e tendo em consideração o desenho experimental, é obtida uma ou mais listas com os genes de interesse. Estas podem ser provir da análise dos genes diferencialmente expressos, ou da detecção de *clusters* de

genes. Em ambos os casos, a obtenção destas listas não representa a finalização do estudo. É necessário, através do recurso a informação armazenada em bases de dados públicas, compreender como estes genes se relacionam, isto é, quais os processos em que estão envolvidos e qual a sua influência nesses mesmos processos.

Uma das técnicas utilizada apoia-se na realização de uma análise funcional dos genes através do uso de uma ou mais terminologias. Este procedimento, comummente aplicado através do uso da *Gene Ontology* [106], tem sido expandido a outros conceitos, tais como a KEGG Orthology, vias metabólicas [107] ou classes de homologias. Todavia, esta metodologia depende directamente do acesso às bases de dados que originalmente armazenam a informação e, dependendo da questão biológica a endereçar, outras fontes de dados podem necessitar de ser acedidas.

Biologia de sistemas

A biologia de sistemas tem por objectivo a compreensão do comportamento de um determinado sistema biológico como um todo. Para tal, é essencial o acesso a um vasto conjunto de dados heterogéneos e de modelos matemáticos que possibilitam a modelação e posterior simulação de sistemas biológicos complexos. Em alguns projectos em que se simulou o comportamento de organismos eucariontes foram usados dados de enzimas, factores de transcrição, vias metabólicas, redes de genes e dados de expressão génica [116]. Como suporte a toda esta área está então a necessidade de aceder de forma facilitada aos dados que originalmente se encontram armazenados em bases de dados públicas.

Medicina genómica

Outra área em que a integração de dados é fundamental é a medicina genómica. Nesta área trabalham geneticistas humanos que procuram genes responsáveis por doenças genéticas, fármaco-geneticistas interessados em perceber como genes, ou grupos de genes, estão envolvidos na resposta diferencial, após sujeição a um determinado fármaco, ou que procurem reunir dados biológicos, clínicos e conhecimento químico no desenvolvimento de novos fármacos; e ainda clínicos que pretendam, em tempo real, obter informação para tratamento dos pacientes.

Apesar da aparente multiplicidade de sub-disciplinas na área da medicina genómica, existe um desafio comum que consiste no estabelecimento de associações entre o genótipo e o fenótipo. O genótipo é definido como a variação genética individual determinada pela sequência de ADN, sendo o fenótipo as características visíveis de um organismo obtidas através da interacção entre o genótipo e o ambiente. Neste contexto, faz sentido estudar as ligações entre polimorfismos num genótipo e ou diferentes respostas de um genótipo a um tratamento de uma determinada doença. É então essencial integrar num esquema único a informação relativa a toda a cascata de relações que, sobre determinadas pressões ambientais, impelem um determinado genótipo, a manifestar-se como fenótipo [117, 118].

6.3 Levantamento e classificação de fontes de dados.

De acordo com a edição de 2009 da revisão de bases de dados realizada pela revista *Nucleic Acids Research,* existem 1170 bases de dados referenciadas na área da biologia molecular [113]. Se, por um lado, o potencial que oferecem ainda está subaproveitado, por outro lado, é também fácil reconhecer a dificuldade em explorar os dados disponíveis. Isto deve-se ao facto de abrangerem um vasto conjunto de domínios de estudo e de possuírem diferentes organizações internas, o que limita a capacidade dos investigadores usarem mais do que um pequeno conjunto de bases de dados. Sendo então impossível, ou, pelo menos, bastante complicado obrigar cada investigador a conhecer individualmente as capacidades de cada base de dados, pode, no entanto, o potencial oferecido por estas ter um melhor aproveitamento com a sua organização em classes. Conquanto existam vários métodos de classificação, são apresentadas aqui duas perspectivas distintas sobre as quais as bases de dados podem ser classificadas. Uma análise mais detalhada destes e doutros métodos de classificação está disponível em [113, 119].

A primeira proposta de classificação das fontes de dados consiste em avaliá-las de acordo com a sua cobertura, tendo em consideração os conceitos e espécies abrangidas. Atendendo a esta classificação foram identificadas as três classes de base de dados seguintes:

- **Horizontais:** estão focadas num determinado tópico de estudo, mas possuem essa informação para várias espécies. São exemplos desta classe as bases de dados GeneBank, Gene Ontology e ArrayExpress;
- **Verticais:** possuem informação relativa a vários domínios para uma determinada espécie. São exemplo os projecto SGD, CGD e ZFIN;
- **Mistas:** são geralmente *hubs* integradores de dados que possuem, para várias espécies, informação relativa a vários domínios de estudo. São exemplos Ensembl, Entrez e KEGG.

Olhar para as fontes de dados desta perspectiva, se bem que um pouco simplista, permite uma nítida separação em função da sua abrangência e a definição de utilizador tipo. Tal como a Figura 6.1 ilustra, algo que também fica evidente com esta aproximação é a colisão de termos entre dados armazenados nas diversas fontes, o que implica a existência de redundância nas bases de dados.

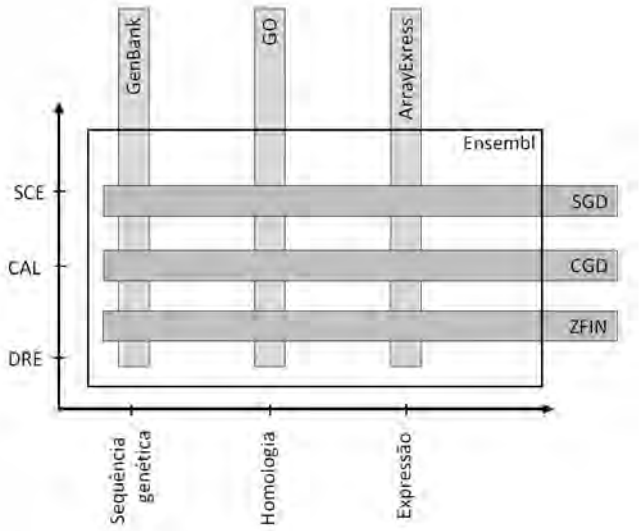


Figura 6.1: Esquema que exemplifica a diferente organização de bases de dados biológicas. No eixo vertical são considerados três organismos (SCE: *Saccharomyces cerevisiae*; CAL: *Candida albicans*; ZFIN: *Zebrafish*) e no horizontal três classes de dados.

Uma segunda aproximação mais detalhada possibilita a classificação das fontes de acordo com o tipo de dados que armazenam. Deste modo, de forma a melhor compreender estas bases de dados, foram definidas as seguintes categorias, que, apesar de não serem estanques, nem únicas, permitem facilitar a posterior análise:

- Sequência de nucleótidos
  - DNA codificante e não codificante;
  - Estrutura de genes, intrões, exões e *splicing sites;*
  - Regulação transcricional e factores de transcrição.
- Sequência de RNA
  - Propriedades de proteínas;
  - Localização de proteínas e *targeting*;
  - Sequências de *motifs* e *active sites*;
  - Domínios proteicos.
- Sequência de proteínas
  - Pequenas moléculas;
  - Carbohidratos;
  - Estrutura de ácidos nucleicos;
  - Estrutura proteica.
- Anotação do genoma, ontologias e nomenclaturas
- Taxonomia e identificação
- Vias de sinalização
  - Genómica comparativa e organismos modelo.
- Doenças e fármacos

- Expressão génica
- Recursos proteómicos
- Literatura científica

6.4 Políticas de disponibilização de dados

As primeiras bases de dados biológicas consistiam em simples páginas *web*, cujo conteúdo era adicionado através da edição directa do código HTML. Estas tinham como propósito a partilha dos dados de um estudo, ou projecto, com a restante comunidade científica. No entanto, com o aumento da complexidade das questões biológicas a endereçar a necessidade de aceder programaticamente aos dados foi-se tornando mais relevante. Para ultrapassar este problema, foram inicialmente desenvolvidos *parsers* que, através da análise do código HTML, extraíam a informação pretendida. Contudo, rapidamente se tornaram evidentes as limitações destes métodos. Primeiro, porque a mais pequena alteração na formatação da página poderia invalidar o *parser* desenvolvido. Segundo, porque nem todas as páginas possuíam URL explícito, invalidando a abordagem. A acrescentar a estes problemas, os intensivos acessos programáticos causavam degradação do desempenho para os utilizadores do sistema.

Uma primeira solução consistiu na disponibilização de todo o conteúdo da base de dados em ficheiros XML ou tabulares num servidor FTP. Apesar da disponibilização dos dados para processamento local ter representado um avanço significativo, estes métodos tornam-se limitados quando apenas se pretende uma fracção dos dados totais.

De forma a colmatar esta lacuna foram propostas várias soluções das quais as mais interessantes se baseiam no uso de *web services*. Um *web service* é uma interface de programação distribuída sobre a *web* que possibilita a execução de operações em servidores remotos. Esta tecnologia faz uso do protocolo http, para troca de mensagens, e da linguagem XML para descrever o formato da transmissão. Uma vantagem desta abordagem resulta do facto de não ser dependente de nenhum sistema operativo ou linguagem de programação.

No desenvolvimento de *web services* podem identificar-se duas metodologias distintas: uma tradicional, designada de SOAP (S*imple Object Access Protocol*), e outra, conceptualmente mais simples mas que tem tido bastante receptividade, denominada por REST (*Representative State Transfer*). Na área da biologia molecular, ambas as aproximações têm sido usadas podendo-se, todavia, afirmar que a baseada em SOAP continua a ser a mais representativa.

Um dos exemplos mais conhecidos do uso da tecnologia SOAP é o sistema BioMoby [120]. O BioMoby define um padrão de troca de mensagens baseado numa ontologia que possibilita aos clientes a descoberta e a interacção com os fornecedores do serviço sem a

necessidade de uma manipulação directa dos formatos ou dos fluxos dos dados. Já se encontram disponíveis várias aplicações cliente do BioMoby, sendo a mais carismática o Taverna [121], que possibilita a criação e execução de fluxos de tarefas sobre as fontes de informação disponibilizadas.

Relativamente à tecnologia REST o sistema DAS (*Distributed Annotation System*) [122] é o mais relevante. Este sistema possibilita a disponibilização de dados de anotação genómica. As anotações fornecidas possuem para cada sequência, notas, observações e predições tais como a identificação de exões (zonas codificantes dos genes), intrões (zonas não codificantes dos genes) e a categorização de repetições no genoma.

Uma comparação entre as duas tecnologias permite inferir que a interface REST é mais indicada para realizar pedidos de dados já pré-calculados enquanto os SOAP são mais indicados para o processamento remoto de dados. Uma revisão da evolução dos serviços biológicos actualmente disponibilizados através de *web services* pode ser encontrada em [123] e uma listagem completa no BioCatalogue[1].

Não obstante a importância dos *web services,* existem outras estratégias de disponibilização programática de dados. O Ensembl permite o acesso aos seus dados através de uma API implementada em Perl e Java. Esta consiste na instanciação de classes locais que mascaram uma interface de acesso à base de dados remota do Ensembl. Apesar de funcionalmente semelhante aos *web services*, do ponto de vista da arquitectura, a estratégia adoptada pelo Ensembl diverge visto utilizar o acesso directo à base de dados. Porém, esta estratégia pode apresentar problemas quando se acede através de uma *firewall*.

6.5 Limitações no acesso aos dados

Se bem que essencial, a integração de dados apresenta ainda vários desafios. A maioria dos dados armazenados está publicamente disponível como ficheiros de texto semi-estruturado ou através de interfaces *web,* e para os obter tem que se aceder a cada base de dados, fazer *download*, *parsing* e, finalmente, agregá-los num repositório único e não redundante. Esta tarefa, para além de morosa, pode ser surpreendentemente difícil devido a várias limitações encontradas.

Mapeamento de identificadores

Um dos maiores problemas na área da biologia molecular decorre do facto de cada uma das bases de dados existentes utilizar referências através de "*accession numbers*" ou através do uso de terminologias próprias. A principal questão é que cada base de dados possui a sua própria estratégia de identificação. Apesar de várias bases de dados já possuírem informação relativa a referências cruzadas, esta continua incompleta.

[1] http://beta.biocatalogue.org

Dimensão dos *datasets* a integrar

Com a actual dimensão dos conjuntos de dados biológicos é inevitável surgirem problemas de consistência, sincronização e actualização. No entanto, os maiores desafios apresentam-se não pelo efectivo tamanho dos dados mas mais pela quantidade de elementos individuais a integrar e pelo número de inter-relações existentes.

Heterogeneidade das bases de dados

Um dos obstáculos à integração advám da heterogeneidade existente entre as bases de dados. Com isto pretende-se referir que cada base de dados possui uma interface distinta para acesso e obtenção dos dados dificultando, deste modo, a tarefa de explorar e adquirir os dados pretendidos. Entretanto, um maior impedimento reside na heterogeneidade ao nível da semântica pois, apesar dos esforços de unificação já realizados, diferentes terminologias continuam a ser empregues por bases de dados distintas.

Gestão de versões

O acesso a fontes de dados distintas é ainda frequentemente dificultado pelo desconhecimento da versão da base de dados em utilização. Na realidade, devido à constante mutação da informação armazenada, o resultado de um estudo realizado hoje pode diferir do resultado alcançado alguns meses depois. Esta dificuldade pode ser ultrapassada através do registo de todos os dados intermédios obtidos da análise. Isto é especialmente relevante quando se realizam pesquisas complexas que abrangem múltiplos tipos de dados e bases de dados distintas.

6.5.1 Abordagens à integração de dados

Pese o consenso existente quanto à necessidade de integrar dados biológicos, o mesmo não acontece no que se refere ao método para proceder à integração. Nesta secção são revistos os três principais métodos de integração: hiperligação, mediador e *warehouse.*

A integração baseada em hiperligações foi a primeira e continua a ser a mais bem sucedida aproximação à integração de dados biológicos [105]. O motivo deste sucesso radica na sua semelhança com a própria *web*. No contexto da biologia molecular a questão coloca-se no facto de existir um incremento no número de bases de dados com informação de interesse. Para obter esta informação o investigador necessita de individualmente aceder a cada base de dados e de manualmente pesquisar a informação pretendida. Para além disso, uma vez que cada base de dados possui a sua própria interface, o utilizador tem que "aprender" a navegar em cada uma das bases de dados. De forma a resolver este problema e simplificar a tarefa do investigador, foram desenvolvidos sistemas que agregam hiperligações de acesso directo às bases de dados biológicas. Deste modo, o utilizador apenas necessita de aceder a um sítio *web* e de fornecer o critério de pesquisa uma única vez, devolvendo o sistema a informação disponível em todas as bases de dados integradas. Algumas implementações não armazenam quaisquer dados localmente, sendo os identificadores,

necessários à construção do url, obtidos, aquando da execução, através de um mecanismo de *crawling*. No entanto, noutras implementações desta técnica, uma base de dados local é usada para o armazenamento dos identificadores previamente obtidos. Exemplos desta metodologia são DiscoveryLink [124] e DiseaseCard [125].

Na integração baseada em mediadores é criada sobre as fontes de dados uma vista unificada que é fornecida ao utilizador. Através do uso deste método, o motor do mediador reformula durante a execução, a questão solicitada numa ou em várias questões que são colocadas sobre as respectivas bases de dados. Os resultados são então agregados e processados de forma a alcançar um resultado final que é devolvido ao utilizador. Exemplos da utilização deste método são o BioMediator [126] e o SEMEDA [127].

Por fim, a integração de dados baseada na metodologia *warehouse* consiste na integração física de múltiplas fontes numa base de dados local, de forma a possibilitar a execução de questões directamente nesta base de dados e não nas fontes originais. A utilização de *warehouses* implica o desenvolvimento de um modelo unificador que possibilite acomodar toda a informação já armazenada nas bases de dados originais. Adicionalmente, é também necessária a existência de *scripts* que acedam às fontes de dados no intuito de obter e processar os dados, fazendo com que estes correspondam ao esquema unificador local. Após este passo inicial de configuração, a *warehouse* pode ser usada para responder a questões suportadas pelas fontes de dados originais, assim como a questões que necessitem da relação de conceitos armazenados em múltiplas bases de dados.

6.6 Gens: plataforma de integração de dados biológicos

O acesso integrado a fontes de dados dispersas constitui um elemento basilar no desenvolvimento de aplicações bioinformáticas. O modelo proposto, designado de Gens (*Genomic Name Server*), tem por objectivo facilitar o desenvolvimento de novas aplicações que necessitem de aceder a dados, sem terem que se preocupar com a camada de acesso, obtenção e processamento.

As escolhas efectuadas ao nível do modelo de armazenamento tiveram como principal propósito responder a uma série de requisitos, tais como eficiência, flexibilidade e escalabilidade do sistema. Como resultado, não se chegou a um modelo único mas sim a um conjunto de dois. Um de base, designado modelo físico, responsável pelo armazenamento efectivo dos dados independentemente do seu formato ou da sua origem, e um de topo, conceptual designado de meta-modelo, responsável pela definição dos dados a integrar e a armazenar. Sobre este último assentam os mecanismos de obtenção de dados e a ele acedem as aplicações externas, tais como GeneBrowser e Quext, que usam como fonte integradora de dados uma implementação do modelo proposto (Figura 6.2).

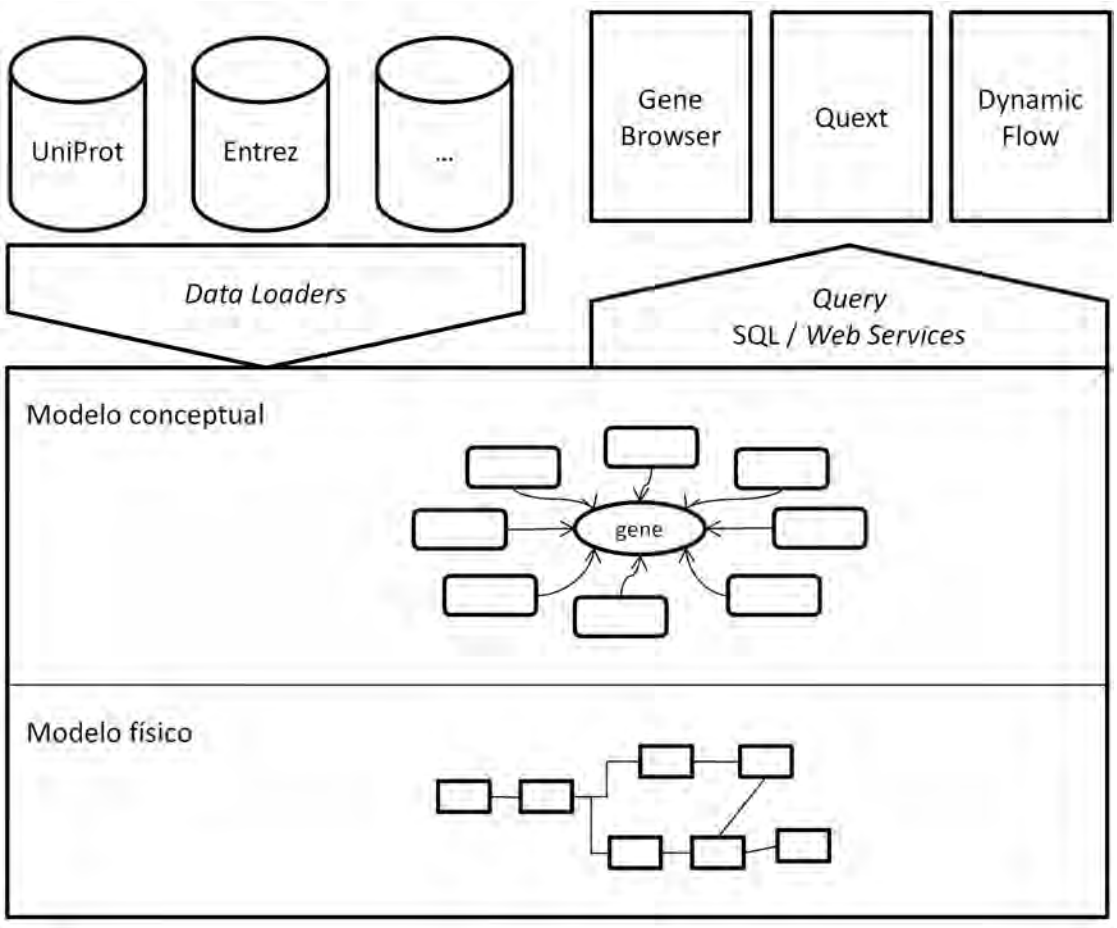


Figura 6.2: Modelo de integração constituído pelo modelo físico e pelo modelo conceptual.

Nas subsecções seguintes é explanado com maior detalhe o funcionamento deste modelo de integração, nomeadamente quais as fontes de dados seleccionadas, os métodos utilizados na obtenção dos dados e as ferramentas de *parsing* desenvolvidas (6.6.1); uma descrição pormenorizada do meta-modelo de integração no que refere às classes de dados usadas e às relações entre estas (6.6.2); o modelo físico da aplicação (6.6.3); o mapeamento entre o modelo conceptual e o modelo físico (6.6.4); um exemplo de utilização do sistema (6.6.5); e, por fim, a informação relativa ao acesso aos dados (6.6.6).

6.6.1 Bases de dados integradas

O primeiro passo na construção do sistema de integração de dados visa a selecção das bases de dados e na identificação dos métodos mais adequados para os extrair. O esquema da Figura 6.3 ilustra as bases de dados e os respectivos métodos de acesso empregues. Sistematizando, as bases de dados usadas foram:

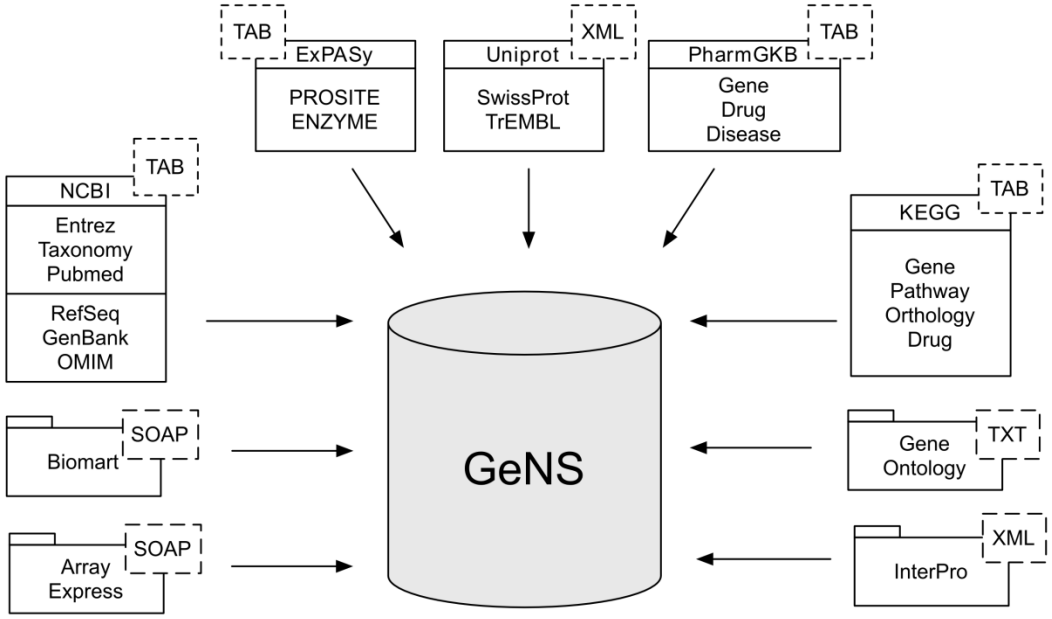


Figura 6.3: Bases de dados integradas no Gens.

- **ArrayExpress[1]:** Esta base de dados possui dados de experiências de expressão genética e foi acedida através do uso da interface SOAP [114];
- **BioMart[2]:** Esta base de dados, que corresponde a uma interface alternativa aos ao Ensembl, foi seleccionada pelas facilidades oferecidas na selecção personalizada dos dados pretendidos [110, 128]. Os dados são obtidos por meio da interface SOAP;
- **NCBI[3]**: O NCBI (*National Center for Biotechnology Information*) [129] possui uma vasta colecção de bases de dados (Entrez Gene, Taxonomy, Pubmed, ReSeq, GenBank e OMIM), que foram integradas com recurso à importação dos dados disponibilizados em formato tabular;
- **ExPASy[4]**: A base de dados ExPASy (*Expert Protein Analysis System*) [130] do SIB (*Swiss Institute of Bioinformatics*), focalizada na análise da sequência e estrutura de proteínas, foi integrada através de ficheiros tabulares;
- **UniProt[5]:** A base de dados UniProt [102], constituída pela versão curada (SwissProt) e pela versão não curada (TrEMBL), foi integrada pela obtenção dos dados no formato XML;

[1] http://www.ebi.ac.uk/microarray-as/ae/
[2] http://www.biomart.org/
[3] http://www.ncbi.nlm.nih.gov/
[4] http://www.expasy.ch
[5] http://www.uniprot.org

- **PharmGKB[1]:** Esta base de dados armazena informação relativa ao impacto que variações genéticas possuem na resposta a fármacos e foi integrada recorrendo-se ao *download* de ficheiros tabulares [131];
- **KEGG[2]:** O KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [112] possui uma representação completa da célula, do organismo e da envolvência que permite a predição computacional de processos celulares e de informação molecular. Apesar de o KEGG possibilitar a obtenção de dados por meio de vários métodos foi seleccionado o uso de ficheiros tabulares;
- ***Gene Ontology[3]:*** O GO (*Gene Ontology*) [132] consta de um vocabulário controlado que possibilita a descrição de genes e de produtos de genes em qualquer organismo. Os dados foram obtidos através do *parsing* do ficheiro de texto disponibilizado no formato OBO (*The Open Biomedical Ontologies*) [58];
- **InterPro[4]:** A base de dados InterPro [104] consiste num repositório de famílias de proteínas, domínios, regiões e repetições já identificadas em proteínas conhecidas que possam ser usadas no processo de descoberta de novas. A obtenção dos dados desta base de dados foi realizada pela análise dos mesmos no formato XML.

Não obstante as bases de dados apresentadas apenas representarem uma pequena porção das existentes, algumas delas são, já por si, *hubs* integradores, pelo que o resultado final redunda num conjunto alargado de conceito e relações biológicas. Deste modo, ao integrar estes dados numa instância única, foram obtidos mais de 500.000 genes únicos, cerca de 50 milhões de identificadores e 50 milhões de relações entre entidades biológicas, o que perfaz um total de 100 milhões de entidades biológicas sobre um único esquema.

6.6.2 Meta-modelo de integração de dados

As bases de dados apresentadas cobrem um vasto conjunto de áreas de estudo, podendo ser usadas na resolução de questões biológicas complexas que necessitem de relacionar vários conceitos [133]. No entanto, é necessário construir um modelo que integre os conceitos biológicos de cada base de dados, assim como as relações existentes entre estes.

O meta-modelo obtido, apresentado na Figura 6.4, teve como especial foco a análise de estudos de expressão génica, pelo que a selecção das classes de dados a integrar e as relações apresentadas reflectem isso mesmo. Este apresenta-se como uma rede de conceitos centrados na proteína visto a maioria das relações derivarem desta unidade. Optou-se pelo uso da proteína e não do gene, enquanto conceito central, pelo facto de um gene poder dar origem a mais do que uma proteína.

[1] http://www.pharmgkb.org
[2] http://www.genome.jp/kegg
[3] http://www.geneontology.org
[4] http://www.ebi.ac.uk/interpro

Directamente associados com cada proteína estão outros conceitos: o organismo a que a proteína pertence, a sua sequência, a sua localização genómica e os seus nomes alternativos. Estão também relacionadas diversas entidades, tais como doenças genéticas, informação de fármacos e seus respectivos genes alvo. As vias de sinalização possuem uma relação múltipla com o gene, visto um gene poder encontrar-se em várias vias de sinalização e, de igual modo, uma via de sinalização possuir vários genes.

6.6.3 Modelo físico

O meta-modelo apresentado contém informação relativa às classes de dados a considerar e às respectivas relações. Numa aproximação comum, este modelo poderia mesmo ser convertido em modelo relacional e ser usado na base de dados para o armazenamento dos dados a integrar. Esta estratégia, embora mais directa, apresenta várias limitações, na medida em que a adição de uma classe implica alterações no modelo físico de armazenamento. A opção seguida passou pelo estabelecimento de um modelo físico de armazenamento, suficientemente genérico, de forma a suportar o meta-modelo anterior, assim como futuras evoluções do mesmo.

Nesta subsecção são, primeiro, descritos os requisitos à implementação do modelo, de seguida, são apresentados os principais argumentos para o uso de *warehouse,* sendo, por fim, apresentada uma descrição do modelo físico.

Requisitos do modelo físico

Foram estabelecidos os seguintes requisitos na implementação do modelo físico:

- De modo a que a implementação do modelo seja usável, o seu esquema tem que ser fácil de compreender e de manter. Para tal, vários esforços foram colocados no desenvolvimento de um esquema com um limitado número de tabelas;
- É fundamental que o sistema seja escalável em dimensão, de forma a conter vários *gigabytes* de dados e centenas de milhões de entidades biológicas;
- O sistema deve também ser escalável em termos do número de bases de dados armazenadas, o que deve ser obtido sem necessitar de alterar o esquema da base de dados;
- Quanto aos dados armazenados, é necessário que sejam acessíveis através do uso de vários métodos. Para tal, é possível pesquisar no sistema através do uso de *web services* e uso directo de sql;

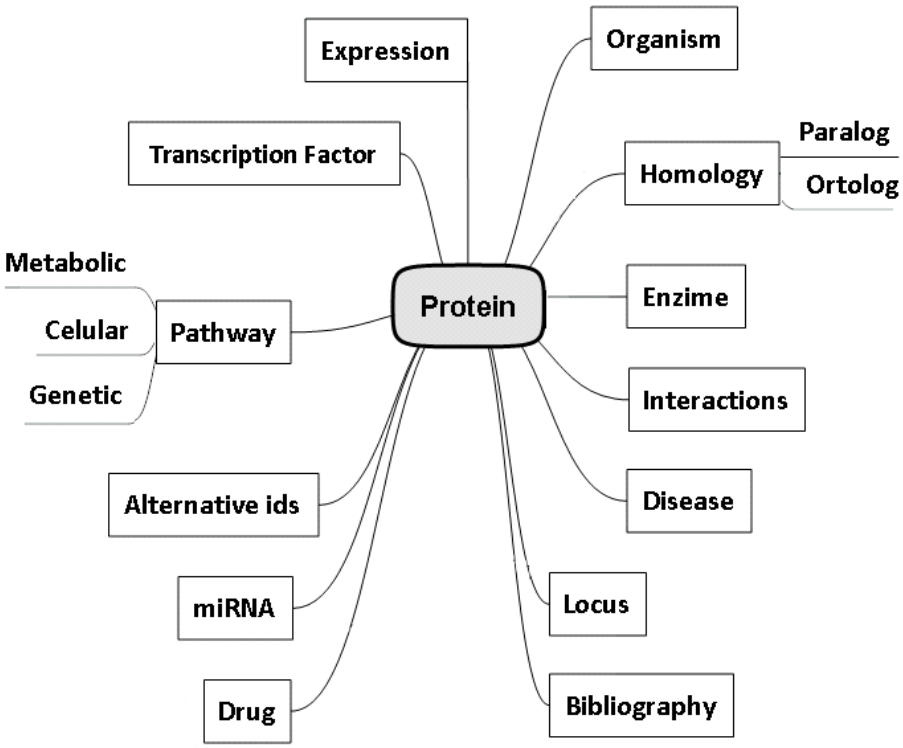


Figura 6.4: Meta-modelo de dados de integração do Gens.

- Mesmo contendo uma elevada quantidade de dados, o sistema está obrigado a ser eficiente de modo a possuir tempos de resposta baixos para as pesquisas mais frequentes. Este aspecto é de especial importância, pois esta ferramenta serve para responder a perguntas definidas pelo utilizador, assim como de plataforma ser usada como fonte de dados por outras ferramentas. De modo a atingir este objectivo, os identificadores dos genes são armazenados numa tabela distinta dos dados referentes a entidades biológicas, assim como são adicionados índices nas colunas associadas com os critérios de pesquisa mais comuns;
- O sistema deve ainda incluir a possibilidade de guardar várias versões da mesma base de dados.

Justificação para o uso de *warehouse*

Tendo em consideração os requisitos impostos e atendendo às características específicas de cada um dos métodos de integração, optou-se pelo uso de *warehouse.* Uma descrição mais detalhada dos argumentos, que justificam o uso de *warehouse* e que são apresentados de seguida, pode ser encontrada em [134].

- **Desempenho:** A *warehouse* é o único método em que o tempo de obtenção dos resultados depende unicamente de factores locais (processador, memória, disco), ao contrário dos restantes que estão dependentes do atraso da comunicação dos dados e do tempo de processamento de servidores remotos. Este aspecto é especialmente

relevante para a execução de questões complexas, que requerem a decomposição da questão original em várias sub-questões;

- **Restrições de acesso:** Ao contrário da *warehouse*, algumas das fontes de dados não disponibilizam mecanismos de pesquisa sobre as suas bases de dados (*web services*, url directo), e outras incluem mecanismos nas suas interfaces (tais como variáveis de sessão, *cookies*, etc) que dificultam o uso de métodos de acesso remoto.
- **Disponibilidade:** As restantes metodologias, por oposição à *warehouse*, não permitem garantir a qualidade do serviço devido à incontornável dependência da disponibilidade de fontes de dados externas. Assim, o serviço prestado pode ficar comprometido no decurso de uma avaria no servidor; porque os métodos de acesso aos dados foram alterados; ou, por exemplo, se a construção de um url se alterar, um procedimento não invulgar numa *web* sempre em constante evolução. Em qualquer dos casos o acesso à base de dados torna-se indisponível, e com agravamento destes serem usualmente resolvidos com intervenção humana;
- **Processamento dos dados:** Outro problema apresentado pelas restantes metodologias que não a *warehouse* consiste na dificuldade em manipular e processar o conjunto de dados como um todo. No caso da integração baseada em *links*, a unidade atómica é a própria página *web*, pelo que não é possível distinguir elementos. Mesmo os mediadores que oferecem uma maior granularidade, ao permitirem a manipulação directa de elementos, têm dificuldade em aceder a todos os detalhes.
- **Gestão de versões:** a *warehouse* é a única abordagem que possibilita a monitorização da versão de cada uma das bases de dados usadas, funcionalidade não muito relevante em pequenos projectos mas essencial em projectos de maior dimensão.

Descrição do modelo físico

O modelo implementado no Gens é baseado num modelo hierárquico centrado na proteína, no qual todas as relações são construídas tendo por base os identificadores de proteínas associados. A motivação para centrar o modelo na proteína, e não no gene, reside no facto deste último se apresentar como um conceito mais geral, na medida em que cada gene, devido a vários mecanismos biológicos, poder dar origem a várias proteínas. Deste modo, se o modelo estivesse centrado no gene, estar-se-ia a perder informação relativa a associações únicas entre proteínas e outras entidades biológicas.

O esquema da base de dados do Gens encontra-se na Figura 6.5.

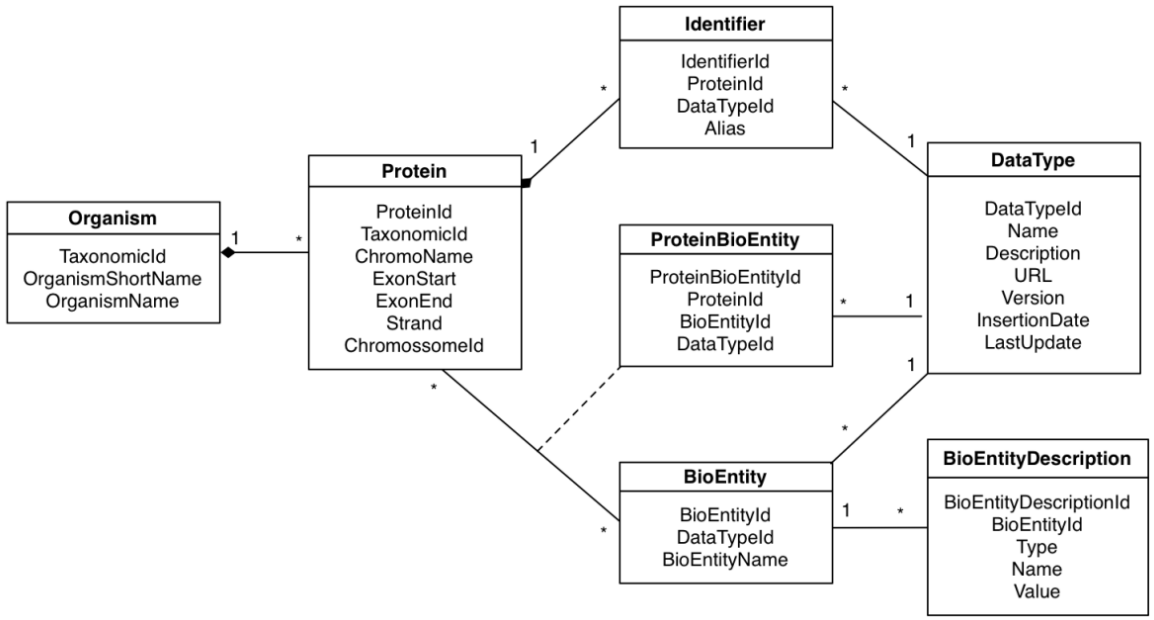


Figura 6.5: Esquema da base de dados do Gens.

De seguida são descritos em pormenor os conceitos aplicados no desenho da base de dados:

- ***Organism*:** Armazena a informação taxonómica, sendo que cada entrada corresponde ao organismo. Esta tabela pode ser considerada como a raiz do modelo, pois dela dependem todas as proteínas. É registado para cada organismo informação detalhada como o seu nome científico e a sequência de referência;
- ***Protein***: Guarda a informação de cada proteína. Esta informação inclui a localização no cromossoma, a sequência do gene e da proteína, assim como as relações para a tabela *Identifier* e *BioEntity;*
- ***Identifier**:* Contém informação relativa a nomes e identificadores alternativos para a proteína em questão;
- ***BioEntity:*** Armazena informação relativa a todas as entidades biológicas associadas com a proteína. Apesar de não existir uma lista fechada de elementos que podem constituir esta lista, esta inclui entradas em ontologias, vias metabólicas e doenças;
- ***DataType*:** Define, cada entrada uma classe à qual um conjunto de identificadores de entidades biológicas ou de genes pertence. Este conceito é distinto do de base de dados, na medida em que tanto uma base de dados, pode possuir mais de um tipo de dados (por exemplo, o KEGG possui o KO, para ortologias, e o KEGG Pathway, para vias metabólicas), como o mesmo tipo de dados pode ser usado em várias bases de dados (por exemplo, o *geneid* é usado para identificar um gene humano no

Entrez e no KEGG). É ainda mantida, pelo *DataType* a informação relativa à última actualização;

- ***BioEntityDescription:*** Recolhe-se nesta tabela a informação relativa à descrição dos elementos armazenados na *BioEntity*. Exemplos de utilização incluem a descrição detalhada da via metabólica ou da mutação de uma determinada doença genética.

A organização hierárquica não se limita a simplificar o esquema da base de dados (tornando o sistema mais fácil de perceber e de manter), como também permite um melhoramento no desempenho do sistema. O sistema obtido é bastante flexível devido à forma como os elementos são mapeados. A adição de um novo tipo de elemento biológico, como uma via metabólica, por exemplo, apenas requer a criação de um novo tipo de dados, o estabelecimento da correspondente descrição por *BioEntityDescription* e, finalmente, a inserção de relações entre vias metabólicas e proteínas por *ProteinBioEntity*. Este esquema está ainda preparado para acomodar dados derivados.

6.6.4 Mapeamento entre o modelo conceptual e o modelo físico

A estratificação do modelo em dois níveis, físico e conceptual, foi concretizada tendo com principal argumento a possibilidade de a adição de fontes de dados sem necessidade de alterar o esquema físico de armazenamento dos dados. Exemplefica-se de seguida como o conjunto de base de dados apresentado é mapeado no modelo conceptual e como estes dados são armazenados no modelo físico.

O primeiro passo consiste em, para cada base de dados através do uso do método de *parsing* adequado, analisar e extrair as relações pretendidas. A título de exemplo, para o UniProt é necessário analisar o ficheiro XML, sendo depois, para cada elemento, correspondente a uma proteína extraídas as relações de acordo com o modelo conceptual. Deste modo, para cada proteína, elemento central do modelo, o *parser* extrai as relações associadas, tais como os identificadores de genes alternativos, a sequência, as vias de sinalização ou os fármacos associados. É desta forma que o modelo conceptual é iterativamente construído.

Após a análise e extracção das relações de todas as fontes de dados, procede-se à persistência dos dados através do modelo físico. Cada entrada de proteína no modelo conceptual é registada na tabela *Protein,* que armazena ainda a sequência genética e genómica. A cada entrada de proteína encontra-se associada uma entrada com o organismo correspondente. No modelo conceptual, as associações relativas a *alternative ids* são guardado na tabela *Identifier*, sendo estabelecida uma associação com a tabela *DataType,* que define univocamente qual o tipo de identificador usado. De seguida, as restantes associações são recolhidas na tabela *BioEntity*. Por exemplo, o elemento *Pathway,* no modelo conceptual, é armazenado na tabela *BioEntity,* sendo a associação com a proteína estabelecida pela tabela *ProteinBioEntity*. A indicação de que uma entrada na tabela

BioEntity corresponde a um determinado elemento do modelo conceptual é definida pela tabela *DataType*. Ainda, caso o elemento em questão possua uma ou mais descrições, estas podem ser conservadas na tabela *BioEntityDescritpion*.

A adição de uma classe biológica ao meta-modelo, situação bastante comum, não implica alterações no esquema da base de dados. São, portanto, evidentes as vantagens que esta aproximação apresenta quando comparada com a tradicional.

6.6.5 Exemplo de utilização

O seguinte exemplo demonstra um de vários cenários de utilização do Gens: um investigador pretende obter a rede de conceitos relacionada com o gene 'sce:Q0085'.

Começa por se determinar o identificador interno da proteína através da tabela *Identifier*. Com este identificador pode então, chegar-se à lista dos identificadores de genes alternativos. Seguidamente e através do acesso à tabela *Protein,* é possível conseguir informação genérica tal como a sua sequência ou a sua localização cromossomal. Dando sequência a este processo e por meio da consulta da tabela *BioEntity* são encontradas todas as entidades biológicas directamente relacionadas como o gene em questão, tais como homologia, bibliografia, dados de expressão, ontologias, vias de sinalização ou enzimas. Por fim, informação mais detalhada sobre cada uma destas entidades pode ser encontrada na tabela *BioEntityDescription.* Seguindo este procedimento de navegação, é agora possível saber, por exemplo, para a via metabólica 'sce00190', identificada como contendo o gene 'sce:Q0085', todos os genes directamente associados. Para tal é necessário perceber todas as relações inversas entre a tabela *BioEntity* e *Protein*. É, assim, obtida a rede de conceitos representada na Figura 6.6.

6.6.6 Utilização pública

O acesso ao Gens encontra-se disponível através de dois métodos: *queries* SQL efectuadas directamente sobre a base de dados e uma interface de *web services*.

O recurso a *queries* SQL é aconselhado para acessos intensivos e para questões que necessitem de relacionar simultaneamente um número elevado de entidades biológicas. Neste caso, é necessária a instalação de uma instância local com uma dimensão aproximada de 10 GB. O *download* da base de dados, assim como informação detalhada sobre a instalação encontra-se disponível em http://bioinformatics.ua.pt/applications/gens.

Em alternativa, é possível pesquisar na base de dados através da interface de *web services,* disponível em http://bioinformatics.ua.pt/GeNS/WS/. Os métodos implementados fazem uso da flexibilidade do esquema do Gens para, através de um número limitado de métodos, conseguir expressar um conjunto abrangente de questões. Estes permitem a listagem dos tipos de dados e proteínas armazenados, a pesquisa por organismos e proteínas e a dupla

conversão entre entidades biológicas e identificadores. A Tabela 6.1 resume os métodos actualmente disponibilizados.

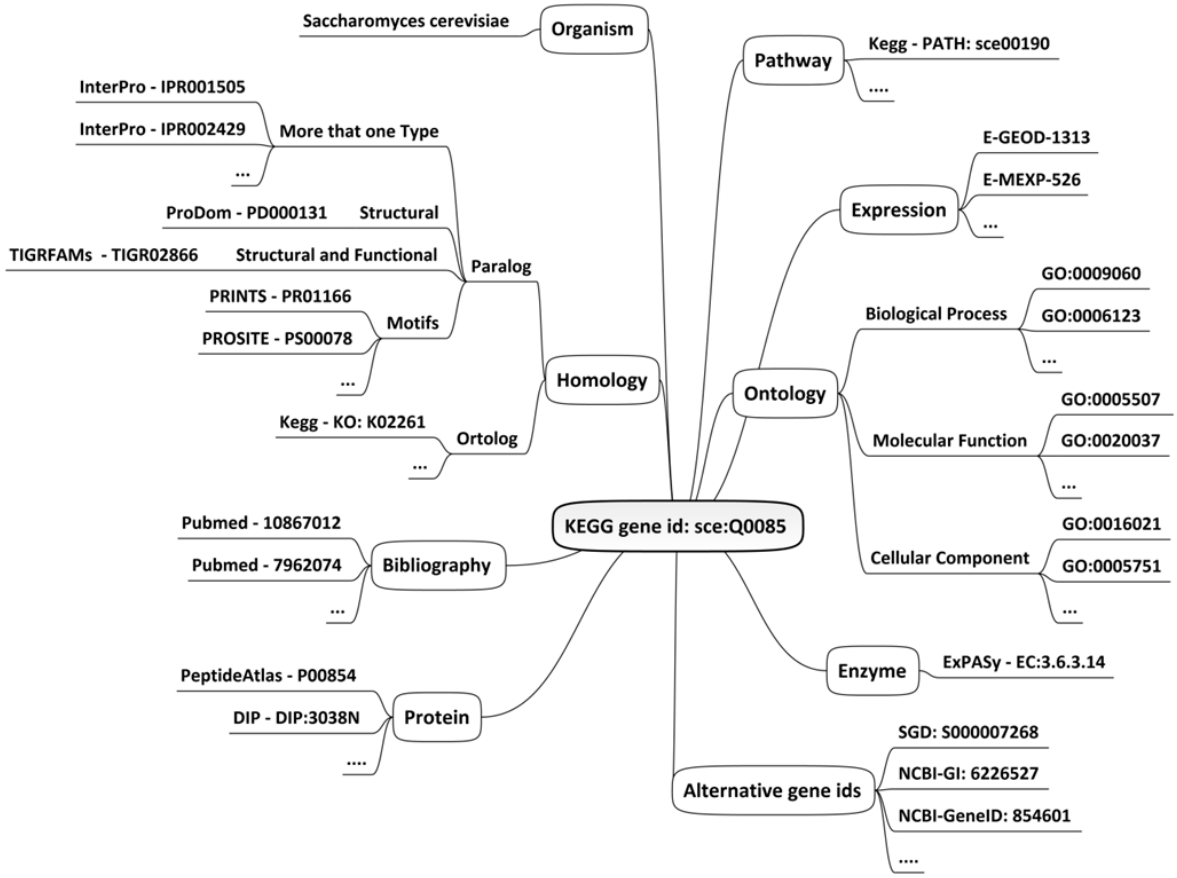


Figura 6.6: Exemplo que ilustra o uso do Gens para obter a rede de conceitos associados com o gene 'sce:Q0085'.

6.7 GeneBrowser

O GeneBrowser é uma aplicação *web,* disponível em http://bioinformatics.ua.pt/gb2, que possibilita explorar e identificar relacções dentro de um conjunto de genes. O GeneBrowser reúne características comuns a portais *web*, tais como o Entrez Gene [111] ou o GeneCards [135], com características presentes em ferramentas de extracção de conhecimento de experiências de *microarrays,* tais como o OntoExpress [108]. De facto, quando comparado com o primeiro grupo de ferramentas, que apenas combinam dados de um único gene, no GeneBrowser múltiplos genes são simultaneamente analisados. Por outro lado, quando em comparação com as ferramentas *desktop* de análise funcional, que apenas possibilitam a análise de uma fonte de dados, no GeneBrowser várias fontes de dados estão disponíveis para análise. Por fim, como o sistema é *web,* são evitados os comuns problemas associados com instalação e manutenção.

Tabela 6.1: Resumo dos métodos disponibilizados pelos *Web Services* do Gens.

Método	Parâmetros	Descrição
ListDataTypes	*()*	Devolve todos os *DataType* disponíveis.
	(search_name_string)	Devolve todos os *DataType* que corrrespondem ao critério de pesquisa *search_name_string.*
	("Identifier")	Devolve todos os *DataType* associados a identificadores.
	("BioEntity")	Devolve todos os *DataType* associados a entidades biológicas.
SearchOrganism	*()*	Devolve todos os organismos.
	(search_name)	Devolve todos os organismos cujos nome científico ou abreviatura correspondam ao critério de pesquisa *search_name*.
	(search_taxonomicid)	Devolve todos os organismos cujo identificador taxonómico corresponda ao critério de pesquisa *search_taxonomicid*.
ListProteins	*(taxonomic_id, datatype)*	Devolve todas as proteínas que correspondem ao organismo fornecido por *taxonomic_id* no formato fornecido por *datatype*.
	(taxonomic_id, datatype, lower_limit, upper_limit):	Igual ao anterior, mas apenas devolve o sub-conjunto compreendido entre o intervalo *lower_limit* e *upper_limit*.
SearchProteinMatches	*(alias)*	Devolve todas as proteínas que correspondem ao critério de pesquisa fornecido por *alias.*
	(taxonomic_id, alias)	Igual ao anterior, mas apenas devolve as pertencentes ao organismo dado por *taxonomic_id.*
ConvertIdentifier	*(taxonomic_id, alias, datatype)*	Procede ao mapeamento de um identificador (*alias*) para outro tipo de dados (*datatype*).
GetBioEntitiesByProtein	*(taxonomic_id,alias)*	Devolve todas as entidades biológicas associadas com a proteína *alias*
GetProteinsInBioEntity	*(taxonomic_id,alias)*	Devolve todas a proteínas associadas com a entidade biológica *alias.*

6.7.1 Metodologia estatística

Apesar do GeneBrowser conter várias vistas da organização dos dados, um dos objectivos consiste no uso de múltiplas estratégias para agrupar o conjunto de genes inserido.

O procedimento seguido traduz-se em tirar partido das várias terminologias biológicas disponíveis para classificar genes e identificar as classes biológicas que se evidenciam. Um exemplo recorrente de uma destas metodologias de classificação é a *Gene Ontology,* apresentada no capítulo anterior. A metodologia usada na detecção das classes que podem ser consideradas como relevantes baseia-se no uso da significância estatística através do cálculo do valor de *p*.

Motivação ao uso do valor de *p*

A disponibilidade do GO, assim como de outras bases de dados biológicas, parece resolver o problema de interpretação dos resultados de experiências de *microarrays* do ponto de vista biológico. Com efeito, a maioria das bases de dados disponíveis possibilita a pesquisa por gene, devolvendo todas a categorias onde este se encontra. Porém, é necessário um processamento adicional desta lista de forma a encontrar quais os processos biológicos mais frequentes para a condição em estudo. Por exemplo, é relevante que, para 200 genes marcados como diferencialmente expressos, 160 estejam associados ao processo da mitose, 80 ao da oncogénese, 40 ao do transporte da glucose e 60 ao da proliferação celular (Tabela 6.2). O que intuitivamente se pode concluir é que a mitose é o processo mais significativo, na medida em que é o que apresenta um número absoluto de genes superior. Todavia, isto não se verifica, pois a mitose possui, à partida, uma elevada quantidade de genes. Deste modo, é necessário fazer uso de medidas estatísticas adicionais para avaliação das classes que efectivamente são mais expressivas.

À semelhança do que sucedia na identificação de genes diferencialmente expressos, também aqui se recorre ao teste de significância. Neste caso, porém os dados não são numéricos mas, antes, categóricos. O procedimento de análise fundamenta-se em, para cada gene, obter todos os termos em que este se encontra anotado na GO. De seguida, a informação é compilada de maneira a pôr em evidência os termos que possuem anotações de genes. Com uma primeira aproximação, poder-se-ia pensar que os termos com maior número de referência a genes seriam os mais relevantes. Contudo, como nem todos os termos possuem, à partida, um número igual de genes anotados, esta dedução não é correcta. É, então, necessário comparar a relação entre os números do *dataset* anotados com o número de genes inicialmente anotado.

Tabela 6.2: Exemplo da análise funcional de 200 genes considerados diferencialmente expressos. Adaptado de [38].

Processo biológico	Genes encontrados	Genes esperados	Avaliação
Mitose	**160**	160	Não significativo
Oncogénese	**80**	80	Não significativo
Transporte da glucose	**40**	10	**Bastante significativo**
Controlo positivo da proliferação celular	**60**	20	Significativo

Exemplo do uso do cálculo

Existem vários métodos de cálculo do valor de *p* nomeadamente binomial, qui-quadrado e hipergeométrica. Neste trabalho foi desenvolvida uma biblioteca que inclui os métodos

anteriores, assim como a possibilidade de correcção do valor de *p*. No entanto, por omissão, é usado o método binomial de seguida enunciado.

De um conjunto de *N* genes foram identificados *K* genes como sendo de interesse ao estudo. O conjunto de *N* genes tipicamente representa todos os genes da espécie, podendo também equivaler aos genes usados no *microarray*. Dos *N* genes foram identificados *M* genes como pertencentes à classe *F*. O que se pretende determinar é a probabilidade P de nos *K* genes existirem *x* genes pertencentes à classe *F*:

$$P\left(X = x \middle| K, \frac{M}{N}\right) = \binom{K}{x}\binom{M}{N}^{x}\left(1 - \frac{M}{N}\right)^{K-x}$$

O valor de *p* corresponde à probabilidade de existirem *x*, ou menos, genes na classe *F*, pelo que é dado por:

$$p = \sum_{i=0}^{x}\binom{K}{i}\binom{M}{N}^{i}\left(1 - \frac{M}{N}\right)^{K-i}$$

6.7.2 Acesso integrado aos dados

O funcionamento do GeneBrowser assenta no acesso intensivo a fontes de dados externas. Efectivamente, para cada conjunto de genes inseridos esta ferramenta, necessita de aceder a mais de 10 bases de dados. A capacidade de realizar este processo, mantendo a usabilidade, apenas foi possível através do uso da plataforma Gens. De facto, para um *dataset* típico, de aproximadamente 100 genes, a tarefa de validação de identificadores e de obtenção da rede de conceitos associados é realizada em menos de um segundo. Em contraste, um sistema semelhante, o FatiGO, apresenta tempos de espera na ordem dos minutos, desde a inserção do *dataset* até à possibilidade de visualizar os resultados.

Não obstante os dados serem, sempre que possível, obtidos através do sistema Gens, tal nem sempre é viável, sobretudo quando estes são demasiado específicos ou com elevada taxa de actualização. Nestes casos, os dados são obtidos aquando da execução, através do acesso directo à fonte, como, por exemplo, a informação relativa à estrutura terciária das proteínas apresentada no explorador de genes, extraída da base de dados PDB.

6.7.3 Funcionalidades disponíveis

São descritas, de seguida, as principais funcionalidades do GeneBrowser, organizadas em sete grupos que se encontram resumidos na Tabela 6.3.

Tabela 6.3: Resumo das funcionalidades disponíveis no GeneBrowser.

Grupo	Descrição
Explorador de genes	Agrega toda a informação disponível relativa ao conjunto de genes inseridos. Aquela encontra-se agrupada em sete secções: sumário do gene; classes de ontologia; vias metabólicas; estrutura de proteínas; sequência genética e proteica; referências bibliográficas e referências para bases de dados externas.
Gene Ontology	Permite a classificação de genes através dos termos da GO, possuindo dois modos de visualização: numa tabela e em árvore.
Homologias	Aplica o princípio de acumulação de genes em categorias por meio do uso de várias bases de dados de homologias.
Explorador de vias de sinalização	Engloba um conjunto de funcionalidades que possibilita a detecção das vias metabólicas mais relevantes e a visualização das mesmas.
Localização cromossomal	Permite identificar a localização cromossomal do conjunto de genes inseridos.
Expressão génica	Faz uso de dados de expressão génica armazenados no ArrayExpress para identificar o perfil típico dos genes inseridos.
Literatura científica	Contém os artigos presentes no Pubmed que melhor reflectem estudos semelhantes ao associado com o conjunto de genes inserido.

Inserção dos dados

O primeiro passo para a utilização do sistema consiste na criação de um novo *dataset* pelo preenchimento do formulário da página inicial (Figura 6.7). É necessário fornecer a lista de genes que se pretende analisar, a lista de genes usados como referência, a espécie do organismo em questão e o nome do *dataset*. Após submissão do formulário, o GeneBrowser usa o Gens para validar os identificadores inseridos. Como entrada, podem ser usados mais de 20 tipos de identificadores distintos de genes e proteínas. Após remoção dos identificadores repetidos e dos não válidos, os restantes são mapeados para um formato interno. É ainda possibilitada a submissão de um dataset com vários tipos de identificadores usados simultaneamente.

Caso o utilizador se encontre autenticado, o *dataset* fica disponível para futuros acessos, caso contrário é eliminado após 7 dias (sendo este, um valor configurável).

Explorador de genes

O explorador de genes fornece ao utilizador, para cada gene, acesso instantâneo à informação disponível relativamente à lista de genes inserida. Estes dados recolhidos através do uso do sistema Gens encontram-se estruturados em sete secções: sumário do

gene; classes de ontologia; vias metabólicas; estrutura de proteínas; sequência genética e proteica; referências bibliográficas e referências para bases de dados externas (Figura 6.8.a).

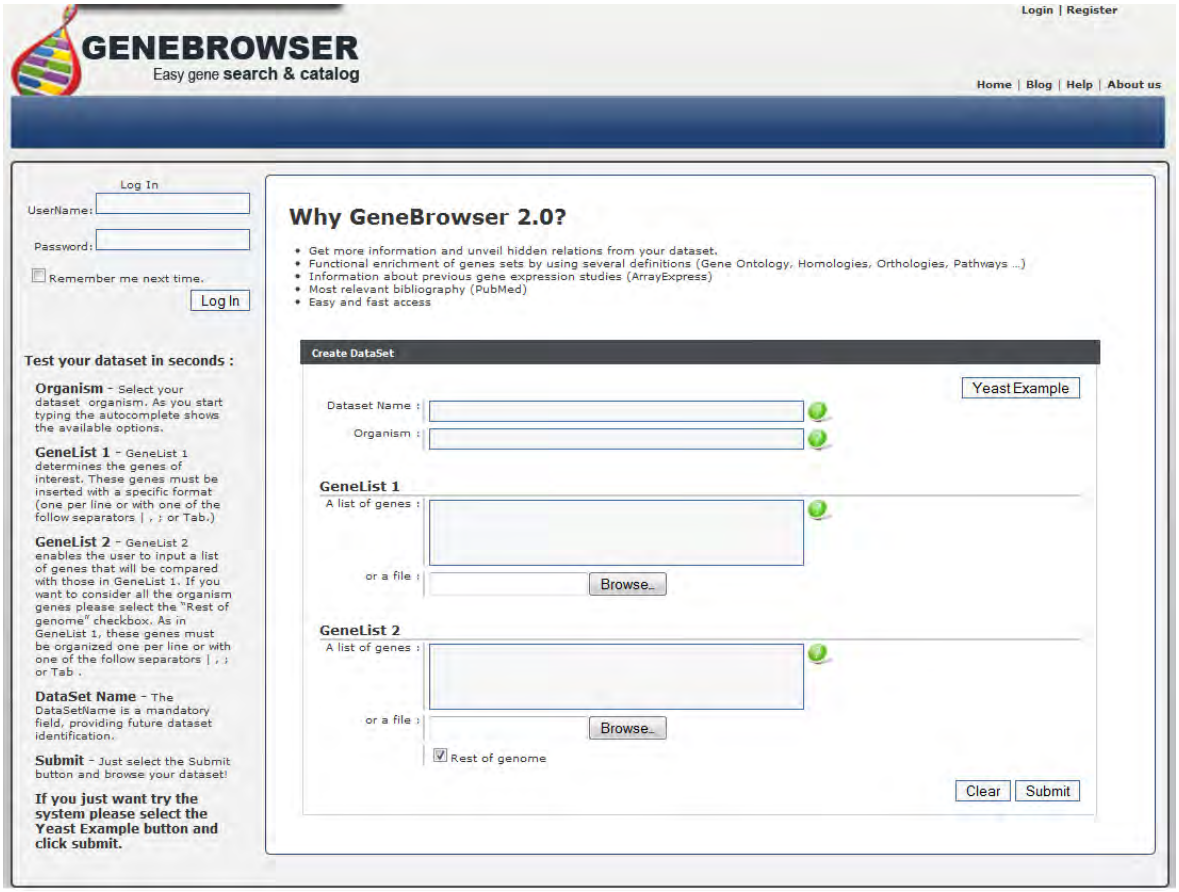


Figura 6.7: Interface inicial do GeneBrowser.

A navegação no conjunto total de genes é ainda auxiliada por dois mecanismos de filtragem. O primeiro permite seleccionar quais as classes de dados que se pretendem tornar disponíveis. O segundo possibilita filtrar genes de acordo com critérios que estes apresentem. Os critérios actualmente disponíveis são: vias metabólicas; classes de homologias e localização cromossomal. Um exemplo de utilização resulta da selecção de todos os genes que se situam num determinado cromossoma ou que se encontram envolvidos numa determinada classe funcional.

Análise funcional através da *Gene Ontology*

Um dos recursos mais comuns na construção desta análise é o GO (*Gene Ontology*) [132], que consta de um vocabulário controlado e estruturado, utilizado para descrever genes e produtos de genes, independentemente do organismo em questão.

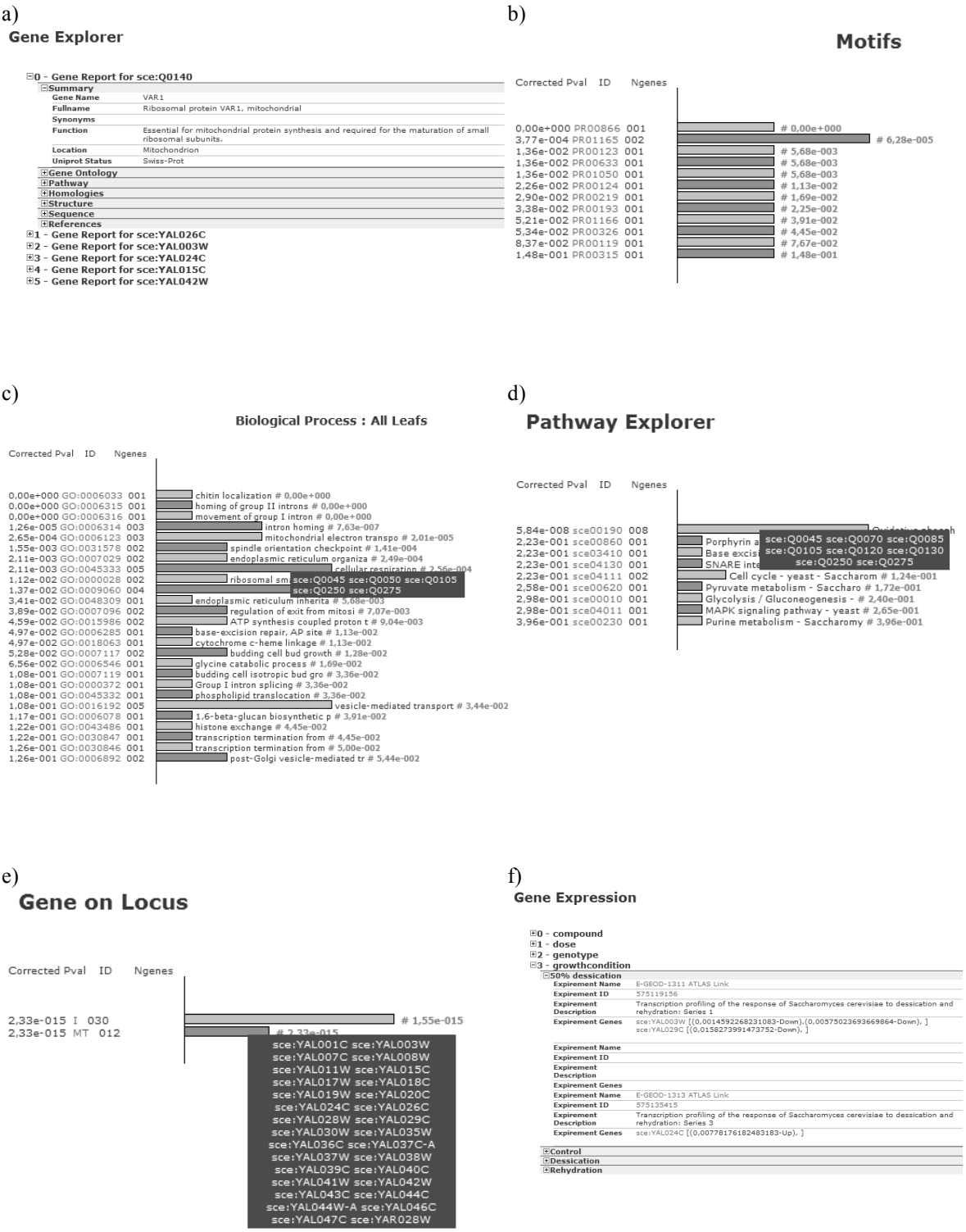


Figura 6.8: Principais funcionalidades do GeneBrowser: a) explorador com informação geral organizada por gene; b) análise funcional de homologias, neste caso motivos; c) análise funcional do processo biológico da *Gene Ontology*; d) análise funcional de vias metabólicas; e) análise funcional da localização cromossomal; f) dados de expressão génica de estudos anteriores.

O GO é constituído por três ontologias independentes: função molecular, processo biológico e componente celular. A função molecular descreve a actividade bioquímica do gene. O processo biológico representa o objectivo biológico do gene. Por fim, o componente celular corresponde ao local da célula onde o produto celular se encontra activo.

De forma a assegurar flexibilidade e extensibilidade na anotação de termos da ontologia, esta é constituída por um grafo acíclico unidireccional, em que cada termo do GO se encontra ligado ao termo superior através de uma relação do tipo "*is_a*" ou "*part_of*". Os novos termos adicionados possibilitam a adição de maior detalhe sem qualquer perda dos dados já existentes. Cada gene encontra-se associado com um ou mais termos da ontologia.

As potencialidades oferecidas pelo GO na análise funcional de genes já são reconhecidas de longa data, pelo que existe um conjunto de ferramentas que possibilita esta mesma análise: GOSurfer [136], OntoExpress [137] e FatiGO [106, 138]. Apesar de já existirem várias implementações que usam esta metodologia para analisar um conjunto de genes, no GeneBrowser a análise dos dados é facilitada através do uso de duas vistas alternativas: em árvore e num gráfico de barras (Figura 6.8.c). A vista em gráfico garante uma análise imediata relativamente às classes funcionais com maior significância. A vista em árvore permite explorar as classes funcionais disponíveis. Em ambos os casos a significância é avaliada através do cálculo dos valores de *p*.

Uso de homologias para extrair conhecimento dos *datasets*

Do mesmo modo que a informação armazenada nas ontologias pode ser usada para analisar conjuntos de dados, o mesmo procedimento pode também ser aplicado com classes de homologias (Figura 6.8.b). Por definição, dois genes são considerados homólogos se possuírem a mesma sequência, no entanto, numa visão mais abrangente várias classes de ontologias podem ser identificadas. A organização apresentada tem por base o trabalho prévio de Turchin [139]. O GeneBrowser tem implementado um procedimento genérico em que, para cada uma das bases de dados de homologias disponível, são identificadas as classes mais relevantes. Actualmente, os tipos de bases de dados utilizadas são: motivos, domínios funcionais e estruturais e ortólogos.

Explorador de vias de sinalização

Vias de sinalização consistem em séries de reacções químicas que se encadeiam de forma a produzirem um determinado processo biológico. Os genes e os seus produtos desempenham um papel essencial nestas vias, na medida em que depende da sua activação, ou não, que a cascata de reacções se realize e que o resultado final seja alcançado.

A possibilidade de conhecer as vias que potencialmente se encontram activas num dado conjunto de genes é essencial para compreender o papel desempenhado por estes. Por isso, no GeneBrowser é disponibilizada uma interface que visualiza a verificação das vias

metabólicas mais significativas e da forma como estão os genes envolvidos nesses mesmos processos (Figura 6.8.d).

Localização cromossomal

Estudos recentes mostram que, nos próprios eucariontes, genes que participam nos mesmos processos tendem a encontrar-se juntos no genoma [140]. Assim, a possibilidade de no GeneBrowser se verificar qual a distribuição de genes no genoma constitui uma mais-valia (Figura 6.8.e). A aproximação usada consiste em fornecer uma primeira vista em que são identificados os cromossomas aos quais os genes pertencem. É ainda fornecido um *link* para o navegador de genomas do NCBI, no qual é possível ver os genes do *dataset* no seu contexto genómico.

Expressão génica

Bases de dados como o ArrayExpress possuem uma valiosa colecção de estudos de expressão génica, em que, para cada par, gene e condição experimental, apresentam o perfil de expressão obtido. Estas informações possuem um elevado nível de qualidade, não só pelo facto de corresponderem a publicações científicas, como ainda por os dados terem sido manualmente validados.

O procedimento usado no GeneBrowser passa pela obtenção da lista de todas as condições experimentais sob as quais os genes já foram testados e dos valores experimentais associados a cada gene (Figura 6.8.f). Os dados são agrupados de acordo com as condições experimentais testadas. É além disso, possibilitado o acesso directo aos dados originais no ArrayExpress, de acordo com os critérios: perfil do gene, experiência/ publicação e condição experimental testada.

6.7.4 Implementação

Processamento síncrono e assíncrono

Das funcionalidades anteriormente apresentadas duas, a análise do GO e a Literatura cientíica, são intensivas ao nível do processamento, sendo o tempo típico de execução aproximadamente um minuto. Pela análise dos tempos de execução das principais etapas do processo, verifica-se que o problema não se encontra no mapeamento de identificadores no sistema Gens (tarefa executada em menos de um segundo) mas, antes, no cálculo dos valores de significância estatística, e, no caso do GO, na reconstrução do grafo de termos.

Na medida em que apenas duas das sete opções disponíveis possuem latência, optou-se por, após a inserção do conjunto de genes, disponibilizar de imediato as restantes funcionalidades. Em paralelo, é lançado no servidor um processo que se encarrega do processamento daquelas duas tarefas. Quando este termina, os resultados processados são armazenados na base de dados, ficando, por fim, estas funcionalidades tembém disponíveis. Apesar de não ser obrigatório, caso o utilizador se encontre autenticado no

sistema, este pode, inclusivamente, inserir um conjunto de genes, abandonar o sistema e voltar mais tarde já com acesso a todas as funcionalidades.

Com a metodologia seguida não só se consegue melhorar a usabilidade, uma vez que a maioria das funcionalidades fica disponível de forma automática, como ainda se torna possível minimizar a carga de processamento no servidor, pois que a execução de uma experiência é apenas realizada uma única vez.

Módulos usados

De forma a facilitar a tarefa de desenvolvimento, foram inicialmente identificadas as interfaces típicas e mais frequentemente utilizadas. Fazendo uso de *WebUser Controls* da plataforma .Net, foram estabelecidos três módulos que, através de uma interface bem definida, vieram concretizar a simplificação da tarefa de desenvolvimento: gráfico de barras, vista em árvore e vista em grelha.

Conquanto já existam várias bibliotecas que permitem a implementação destas funcionalidades, elas são maioritariamente comerciais, fechadas e, sobretudo, muito pesadas em termos de processamento e de tráfego gerado. Já os módulos desenvolvidos possuem a vantagem de estar totalmente enquadrados com a restante arquitectura da aplicação.

A estrutura típica destes módulos divide-se em três componentes: estrutura de dados em JSON, estilo definido em CSS e geração da interface em *JavaScript*. Com o propósito de minorar a carga de processamento no servidor, a geração do gráfico é realizada por meio de *JavaScript* no cliente. A Figura 6.9 ilustra uma vista em grelha e a respectiva estrutura de dados JSON, a Figura 6.10 a construção da árvore e a Figura 6.11 um gráfico típico.

Esta abordagem pretende diminuir a quantidade de informação supérflua a ser transmitida ao cliente, assim como racionalizar o processamento a efectuar no servidor.

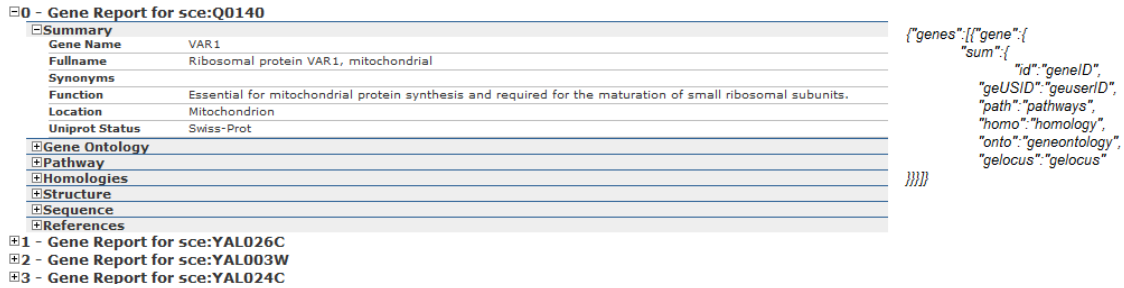


Figura 6.9: Exemplo de vista em grelha e respectiva estrutura de dados em JSON.

Figura 6.10: Exemplo de árvore e respectiva estrutura de dados em JSON.

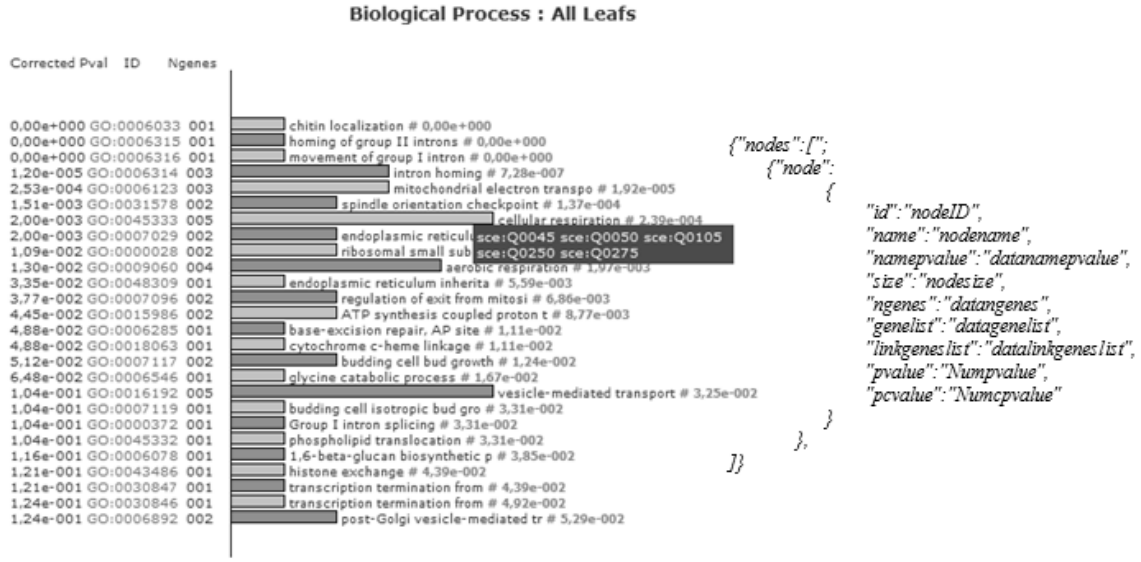


Figura 6.11: Exemplo de um gráfico e respectiva estrutura de dados em JSON.

Modelo de dados

Ainda que a informação armazenada localmente seja mínima, visto os dados serem extraídos do Gens, ou directamente das fontes durante a execução, foi, ainda assim, necessário estabelecer uma base de dados para o armazenamento dos dados referentes ao *dataset* e às contas de utilizadores. Relativamente à gestão das contas de utilizador, foi empregue o mecanismo disponibilizado pelo .Net, essencialmente devido à sua facilidade de utilização e funcionalidades disponibilizadas. Deste modo, foi possível chegar-se a um esquema constituído por três tabelas (Figura 6.12). Para além da tabela *Dataset*, que armazena o dataset inserido, encontra-se a tabela *Genelist_relation,* que armazena as relações múltiplas entre genes e *datasets,* e a tabela *DataSetGORelation*, que armazena a informação processada da *Gene Ontology*.

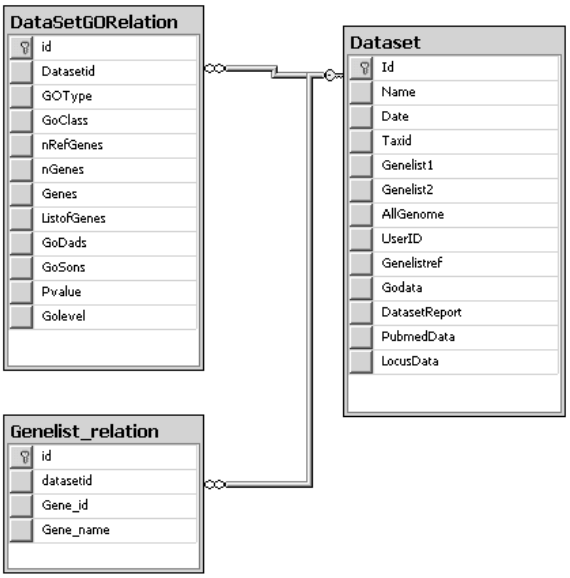


Figura 6.12: Modelo de dados do GeneBrowser.

6.8 Quext (Query Expansion Tool)

Apesar das funcionalidades oferecidas pelas bases de dados e pelas ferramentas de extracção de conhecimento de um conjunto de genes, muita da informação relevante continua a estar presente na literatura sob a forma de texto. Isto deve-se ao facto de muitas bases de dados biológicas, possuirem capacidade de armazenamento dos dados numa forma estruturada e até possibilitarem a leitura e processamento por computadores, mas continuarem a ser populadas por meio de pesquisas manuais ou semi-manuais da literatura. Deste modo, a capacidade de se manterem actualizadas não consegue acompanhar o número de evidências biológicas diariamente publicadas. Embora este processo esteja a mudar com o facto dos editores começarem a obrigar os autores a enviar os resultados das experiências para bases de dados conhecidas (tal como o ArrayExpress, para estudos de expressão génica), existe ainda uma lacuna considerável entre a informação contida nestas bases de dados e a armazenada na literatura, fazendo desta uma preciosa fonte de informação actualizada.

Na área biomédica e das ciências da vida, a base de dados MEDLINE da *United States National Library of Medicine* é a mais relevante fonte de informação. A sua vasta colecção contém mais de 17 milhões de artigos, de um total de 5000 jornais que datam desde 1960[1]. Diariamente, são adicionados aproximadamente 3000 novos artigos. Para pesquisar e aceder aos artigos desta base de dados está disponível uma interface *web* designada de PubMed[2].

[1] http://www.nlm.nih.gov/pubs/factsheets/medline.html
[2] http://www.pubmed.com

A principal questão relacionada com o MEDLINE/PubMed deve-se ao facto dos dados serem armazenados sem respeitarem nenhuma estrutura ou organização o que dificulta a extracção de dados relevantes. No intuito de endereçar esta questão, nos últimos anos várias técnicas de extracção de conhecimento e de mineração de texto têm sido aplicadas [141-143].

O objectivo destas ferramentas visa implementar uma nova e melhorada metodologia para obter e ordenar os mais relevantes artigos partindo de uma lista de genes. As metodologias propostas podem ser comparadas a uma compilação das tarefas realizadas pelos investigadores quando estes manualmente processam os dados. De facto, o procedimento usual consiste em inserir esta lista, ou alguns dos seus elementos, na interface do PubMed. Simultaneamente, ferramentas, como o FatiGO+ [106] ou o GeneBrowser, podem ser usadas para detectar quais as categorias funcionais dentro da lista de genes fornecida. Estas categorias funcionais são, então, usadas como critério de pesquisa, para filtrar a lista de artigos resultante da busca. A realização deste procedimento, iterativamente, leva a uma restrita lista de artigos, propensa a conter apenas os artigos mais relevantes, mas tendo, como principal senão, o elevado tempo dispendido. A morosidade desta metodologia pode ser contornada através de uso de aplicações específicas, tendo sido precisamente esta a principal motivação para o desenvolvimento da ferramenta Quext, acessível em http://bioinformatics.ua.pt/Quext (Figura 6.13)

A metodologia aqui implementada segue um fluxo dividido em quatro etapas, que tem, como entrada, uma lista de genes devolvendo uma lista de identificadores de artigos no PubMed. O primeiro passo passa pela aceitação e validação da lista de identificadores de genes para uma convenção consistente. De seguida, através do uso da base de dados GENS, os termos originais - genes - são expandidos para outros conceitos relacionados. Esta lista de conceitos é, então, submetida ao mecanismo de pesquisa que itera sobre a base de dados PubMed localmente armazenada, devolvendo todas as correspondências positivas. O passo final compõe-sa da assemblagem de todos os resultados individuais, do estabelecimento de rankings e da disponibilização dos dados ao utilizador. A Figura 6.14 contém o *workflow* proposto para a descoberta de artigos através de uma lista de genes.

Uso de técnicas de expansão de termos

A expansão de termos pode ser definida como o processo de reformular uma *query* base, de forma a melhorar a performance na obtenção de informação. Assim, existem várias técnicas de realizar este processo, sendo que cada uma possui as suas vantagens. Dentro das mais comuns, encontra-se a expansão de uma *query* de pesquisa baseada em sinónimos, a pesquisa de termos relacionados e a adição destes à *query* original. A principal motivação para recurso à expansão de termos reside no incremento da qualidade dos resultados de pesquisa, através da consequção de artigos que, com a *query* original não seriam encontrados.

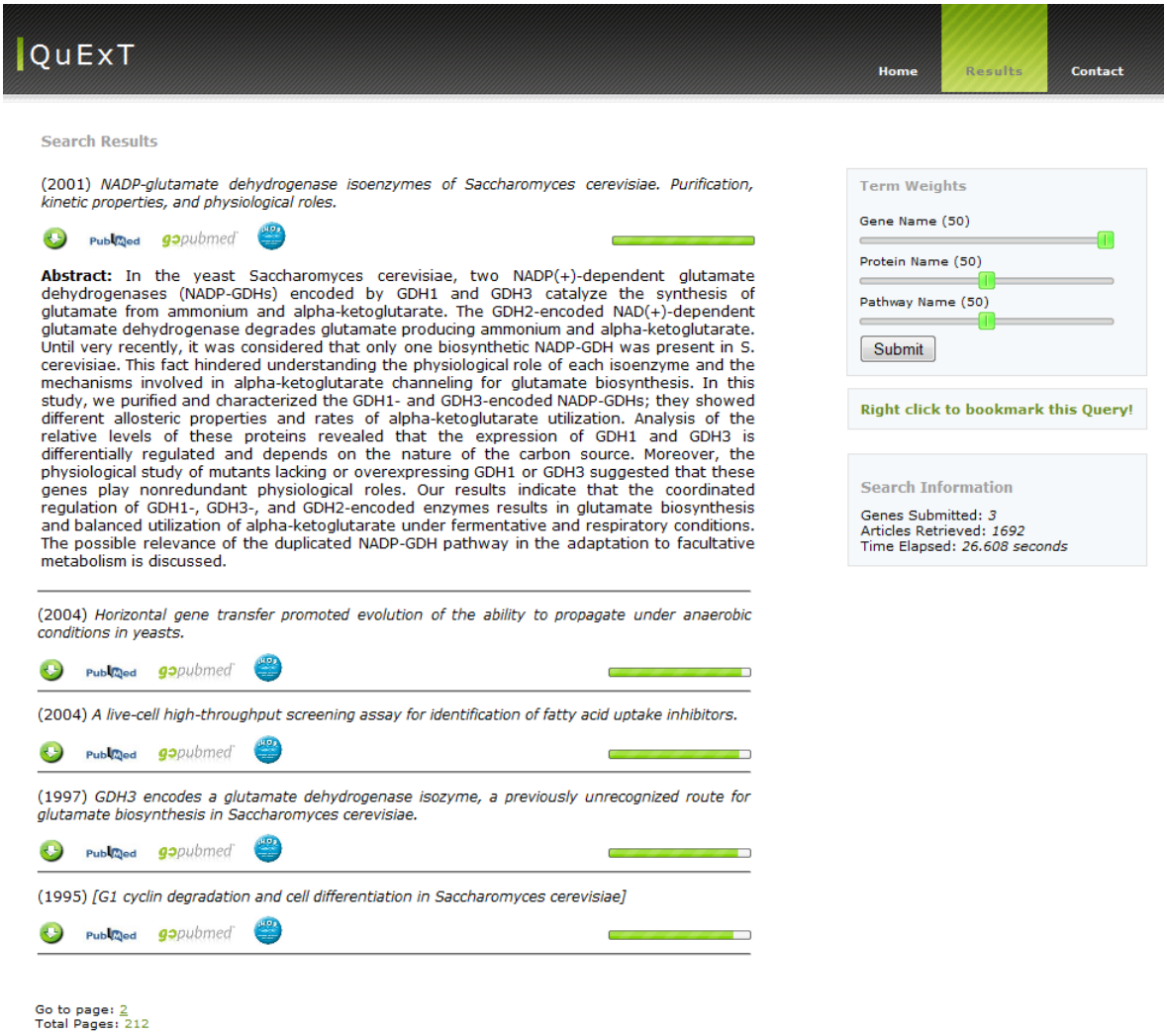


Figura 6.13: Resultado de uma pesquisa no Quext.

Subjacentes a todas as técnicas de obtenção de dados estão dois parâmetros que definem a qualidade desta: *recall* e precisão. O *recall* mede a fracção dos documentos relevantes devolvidos, considerando o número total de documentos. Por seu lado, a precisão mede a facção de documentos devolvidos ponderando o número total de artigos presentes. O ideal, se bem que difícil, é conseguir em simultâneo, valores elevados de precisão e de *recall*. O uso de técnicas de expansão de termos tem como principal objectivo o aumento do *recall*, promovendo, no entanto, como consequência, a diminuição dos valores de precisão. Uma forma de conseguir incrementos consideráveis nos valores de *recall,* sem grande prejuízo nos valores de precisão, implica uma escolha adequada das classes dos termos sobre os quais a expansão vai ser realizada. O sistema desenvolvido utiliza, como termos de expansão sobre a *query* original, as proteínas correspondentes às vias metabólicas e os nomes alternativos dos genes.

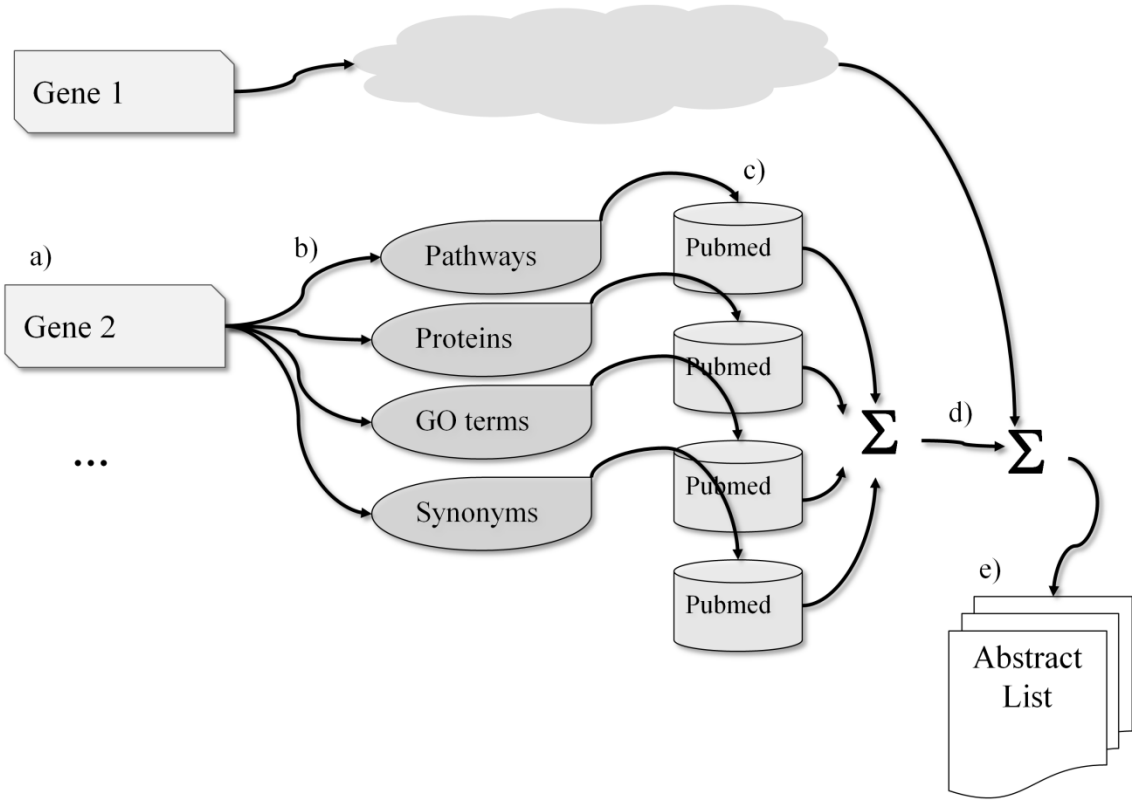


Figura 6.14: *Workflow* proposto para obtenção de artigos através de expansão de termos: a) lista inicial de genes b) expansão dos termos iniciais; c) pesquisa na base de dados Pubmed; d) assemblagem e ordenação dos resultados; e) devolução da lista final de artigos.

O exemplo da Figura 6.15 ilustra este procedimento. Para o gene YCR005C do organismo *Saccharomyces cerevisiae* é, em primeiro lugar, obtida a sua correspondência com o identificador Entrez, sendo, de seguida, expandido para os seguintes termos: nome do gene (*CIT2*), via metabólica (sce00020, sce00630), proteína (P08679) e Gene Ontology (0006101, 0006537, 0006097, etc).

Pesquisa dos termos no PubMed

O acesso a uma base de dados remota apresenta sempre como inconveniente atrasos de comunicação. No caso desta aplicação em que o número de termos iniciais é expandido aumentando o número efectivo de questões a realizar, o acesso, em tempo real, ao PubMed é impraticável.

De forma a contornar este problema, optou-se pela criação de um índice local da base de dados do PubMed, fazendo com que o factor limitativo passasse a ser definido pela capacidade dos recursos locais. O índice foi construído tendo em consideração que apenas o título e o resumo do artigo constituem informação relevante.

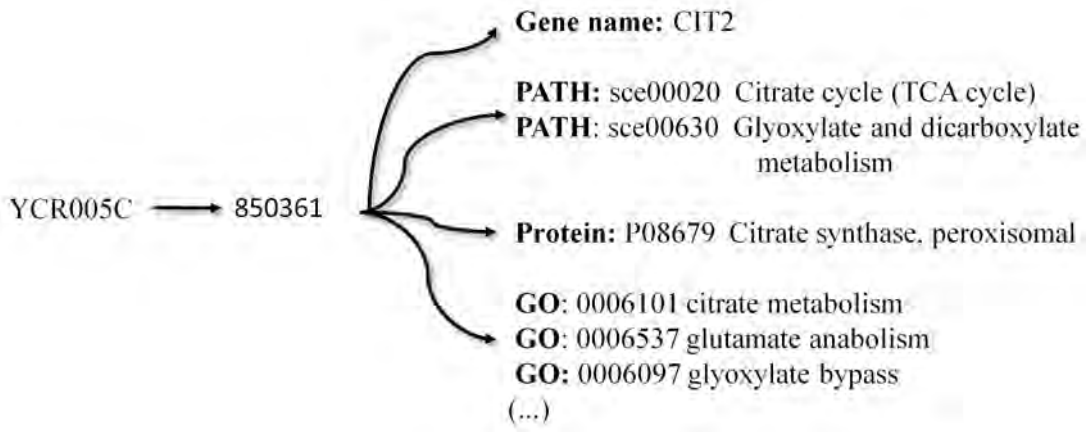


Figura 6.15: Exemplo da expansão de termos. O idenficador de gene é mapeado para o identificador do Entrez, sendo obtido o nome do gene, as vias metabólicas, proteínas e termos do GO.

O mecanismo de pesquisa usado sobre o índice é fornecido através do Lucene [144]. Cada *query* é, então, construída em tempo real através do uso do sistema Gens e submetida à função de pesquisa do Lucene, que devolve os identificadores do PubMed positivamente correspondentes aos critérios de pesquisa.

Assemblagem dos resultados

A pesquisa por um determinado termo no índice local devolve uma lista de identificadores de artigos. Sendo realizada uma *query* por cada gene, várias pesquisas são efectivamente realizadas, significando que é necessário assemblar os resultados. Esta assemblagem requer um algoritmo de ordenação, que deve reflectir diferentes pesos para os vários termos (por exemplo, proteínas, vias metabólicas). A ordenação da lista final de artigos deve espelhar a relevância de cada artigo face à lista de entradas e ao esquema de pesagem, de modo a que os mais relevantes estejam nas posições de topo. A fórmula que reflecte estes requisitos é definida por:

$$Rank_i = \sum_{j=1}^{Ntc} W_j \times n_{ij} \qquad (1)$$

i artigo específico; *j* classe de dados; N_{tc} número total de classes; W_j peso atribuído à classe *j*; n_{ij} número de vezes que a expansão da classe *j* produz resultados positivos para o artigo *i*.

6.9 Sumário

Neste capítulo foi apresentado o desenvolvimento de duas aplicações *web,* GeneBrowser e Quext, que permitem facilitar a tarefa de interpretação biológica de estudos de expressão

génica. Como suporte a estas aplicações, é também proposta e desenvolvida uma plataforma de integração de dados biológicos, que possibilita o acesso centralizado a fontes de dados dispersas. O modelo proposto, designado de Gens (*Genomic Name Server*), tem por objectivo facilitar o desenvolvimento de novas aplicações que necessitem de aceder a dados, sem terem que se preocupar com a camada de acesso, obtenção e processamento.

O problema da integração de dados biológicos apresenta-se como um dos mais relevantes problemas da bioinformática. O principal factor diferenciador da plataforma proposta consiste no estabelecimento de dois níveis de armazenamento: um de base, designado modelo físico, responsável pelo armazenamento efectivo dos dados independentemente do seu formato ou da sua origem; e um modelo de topo, designado de meta-modelo, responsável pela definição dos dados a integrar e a armazenar. A solução proposta, para além de simples, permite garantir a eficiência, flexibilidade e escalabilidade do sistema. A instância actual do Gens integra 9 bases de dados que totalizam 500.000 de genes únicos com cerca de 100 milhões de entidades biológicas.

O GeneBrowser integra as vantagens dos portais *web*, tais como o Entrez Gene ou o GeneCards, com características presentes nas ferramentas de extracção de conhecimento de experiências de *microarrays*. O GeneBrowser reúne a informação biológica contida nas principais bases de dados, juntamente com várias ferramentas de análise que possibilitam identificação de vários elementos biológicos, tais como homologias, ontologias, vias de sinalização, localização cromossomal e estudos de expressão relacionados.

A segunda aplicação, Quext, implementa uma nova e melhorada metodologia, baseada na expansão de termos, para obter e ordenar os mais relevantes artigos, partindo de uma lista de genes. A pertinência deste trabalho prende-se com o facto de muita da informação relevante continuar a ser unicamente armazenada nas bases de dados de literatura sob a forma de texto, fazendo destas uma valiosa fonte de informação actualizada.

Capítulo 7

7 Conclusões e trabalho futuro

No início do documento foi definida, como principal questão a endereçar, a forma como as actuais tecnologias de sistemas de informação podiam ser usadas para agilizar os procedimentos de armazenamento, partilha e análise de dados de experiências de *microarrays*. O trabalho realizado para responder à questão anterior revelou-se multi-disciplinar, na medida em que necessitou de conhecimentos de ciências da computação, de biologia molecular e de estatística. Deste modo, o primeiro passo consistiu no estudo dos princípios de biologia molecular necessários à compreensão do funcionamento dos *microarrays*. Esta fase mostrou-se essencial para abranger o domínio do problema, tendo facilitado as tarefas de levantamento de requisitos e de modelação das ferramentas propostas. De forma geral, durante todo o trabalho, se bem que mais especificamente nesta fase, revelou-se vital a possibilidade de colaborar directamente com os membros do laboratório de *microarrays* da Universidade de Aveiro.

A primeira tarefa visou a análise do fluxo de informação gerado durante uma experiência de *microarrays* e o levantamento das normas, ontologias e sistemas de gestão de dados existentes. Esta possibilitou a identificação de limitações nos procedimentos existentes, resultando na proposta de um sistema LIMS para gestão de dados de laboratoriais. Este sistema, o Mind, veio facilitar a gestão dos mesmos dados no laboratório.

O problema que de seguida foi colocado radicava na capacidade de partilhar dados de experiências. Apesar da dificuldade inicial ter consistido na exportação de dados para repositórios públicos, rapidamente surgiram novos requisitos que deram origem a diferentes cenários de utilização. Foi, então, proposto e implementado um modelo que possibilita que diversos investigadores participem num mesmo projecto, mesmo utilizando sistemas distintos. Aqui os desafios lançaram-se em duas direcções. Por um lado, a complexide do modelo MAGE, cuja compreensão implicava atender a mais de 200 classes com um elevado número de relações, não raro, na depedência de ontologias externas. Por outro e superando ainda a complexidade própria do modelo, a dificuldade na execução da

tarefa de mapeamento entre este e os dois sistemas LIMS usados. Conquanto este objectivo tenha sido alcançado com sucesso, surgiu, entretanto, a norma MAGE-TAB, cuja implementação se revelou muito mais simples.

A necessidade seguinte assentou no estudo das diferentes fases da análise do resultado de um estudo de *microarrays*. Foram examinados os procedimentos de pré-processamento e de normalização de dados, de selecção de genes diferencialmente expressos e de detecção de padrões, assim como de análise funcional. Por fim, foi proposto um conjunto de ferramentas implementadas sobre o sistema Mind, que inclui os procedimentos de análise mais comuns.

O trabalho apresentado no Capítulo 6 foi, sem dúvida, o que maior número de desafios gerou. A necessidade de um modelo que reunisse os requisitos pretendidos (flexibilidade, extensibilidade e simplicidade) não se traduziu numa tarefa óbvia. Na realidade, foram inicialmente testados vários modelos, e mesmo o modelo escolhido foi alvo de várias afinações até se ter alcançado a versão apresentada. O principal objectivo não era o da obtenção do modelo em si, mas sim o da consecução de uma plataforma sobre a qual pudessem ser implementadas outras aplicações.

Das aplicações subsequentes, o GeneBrowser foi a primeira a ser planeada, surgindo no seguimento de uma contribuição menor para um artigo [145], cujo objectivo era realizar uma análise funcional de um conjunto de miRNAs. A ideia original, segundo a qual esta análise incluía vários conceitos biológicos, rapidamente se estendeu, chegando-se a um primeiro protótipo. Este foi posteriormente redesenhado de forma a tornar-se mais genérico, culminando na ferramenta apresentada. A outra aplicação proposta, o Quext, permite obter, para uma lista de genes relacionados entre si, o conjunto de publicações de maior relevo. Para tal, usa, enquanto critério de pesquisa, não só o nome dos genes e respectivos sinónimos, mas também vários conceitos biológicos associados. Em ambas as aplicações a possibilidade de utilização da plataforma Gens constituiu uma mais-valia.

7.1 Contribuições

As contribuições deste trabalho expandiram-se em duas vertentes: publicações e ferramentas desenvolvidas (Figura 7.1). No total, foram publicados 11 artigos em jornais e actas de conferências e foram desenvolvidas 5 ferramentas actualmente disponíveis para toda a comunidade científica.

No âmbito das ferramentas concebidas, um dos principais resultados deste trabalho consiste no sistema Mind, que se encontra descrito nos Capítulos 3 a 5. Este encontra-se em funcionamento desde 2006, possuindo mais de 200 hibridações correspondentes a mais de 10 experiências, das quais duas já se encontram publicadas e três estão em fase final de publicação. Também do trabalho associado ao sistema Mind resultaram quatro publicações: uma com o levantamento das lacunas existentes nos sistemas LIMS e os

requisitos do sistema Mind [67]; outra explorando os problemas associados com a partilha de dados [60]; a terceira sobre os desafios na interligação dos sistemas LIMS com sistemas clínicos [146]; e um artigo com todas as funcionalidades do sistema Mind [68].

Foi ainda obtida a plataforma de integração de dados Gens e duas aplicações que dela fazem uso: GeneBrowser e Quext, como ficou exposto no Capítulo 6. Sobre esta plataforma e as respectivas aplicações foram realizadas cinco publicações: uma a respeito da base de dados Gens [147]; uma segunda acerca do sistema GeneBrowser [148]; outra relativa ao mecanismo de expansão de termos usado no Quext [149]; e, outrossim, duas publicações que associam o uso do Gens e de *workflows* para responder a questões biomédicas [150, 151].

Por fim, da colaboração existente com o grupo de genómica funcional, e pela aplicação directa das ferramentas desenvolvidas, resultaram ainda duas publicações [145, 152].

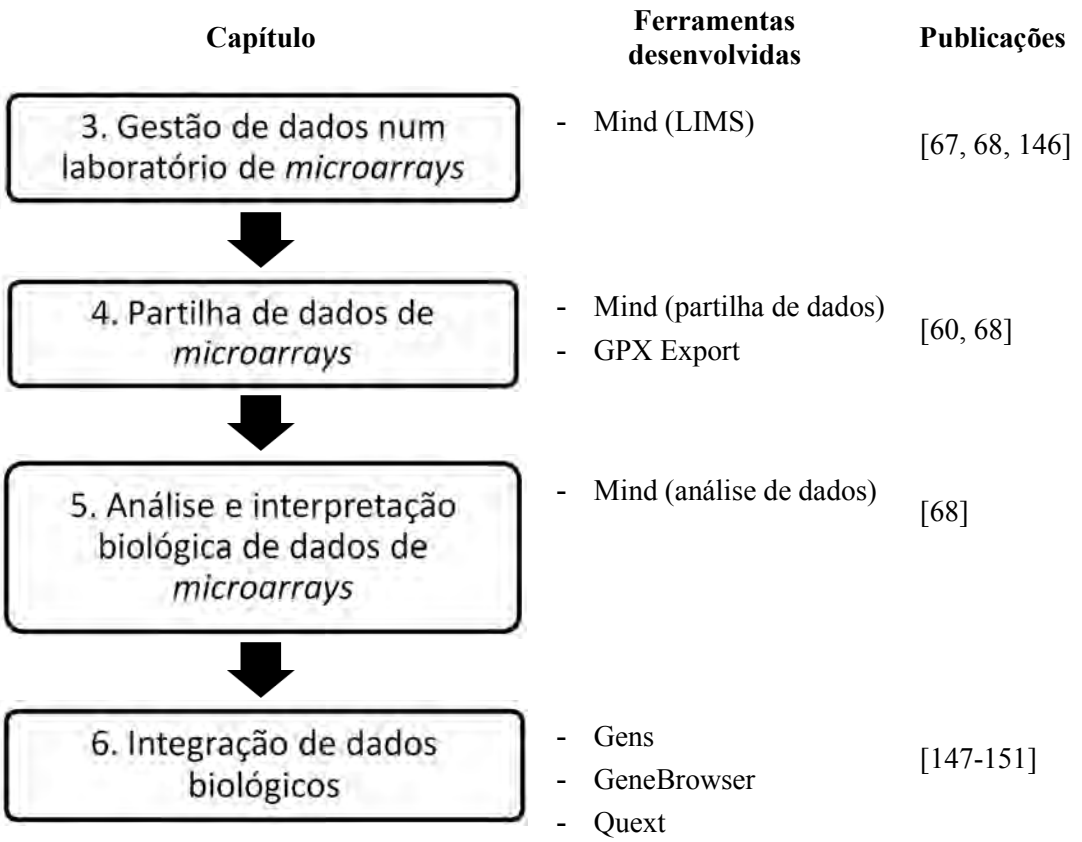


Figura 7.1: Resumo das contribuições.

7.2 Perspectivas de trabalho futuro

Por si só, várias das questões aqui abordadas podem constituir objecto de estudo mais detalhado. Optou-se, no entanto, por apresentar uma visão mais geral que pode, não obstante, ser utilizada como âncora para vários trabalhos no futuro.

Todas as quatro áreas em que o trabalho se focou possuem, com efeito, elevadas perspectivas de desenvolvimento de trabalho futuro. Nos sistemas LIMS, podem ser exploradas novas metáforas de interacção; ao nível da partilha de dados, há espaço para o estudo de estratégias de mapeamento automatizado entre modelos de dados através do uso de ontologias; relativamente à análise de dados, a inclusão de vários métodos de análise baseados em análise exploratória constituirá uma mais-valia para o sistema Mind. Outra área que se apresenta extremamente promissora é a que está relacionada com as estratégias de integração de dados biomédicos, apresentadas no Capítulo 6.

Na realidade, muitas das matérias relacionadas com a interpretação biológica dos dados necessitam de uma reflexão mais cuidada. A análise funcional das vias de sinalização realizada pelo GeneBrowser carece de precisão, na medida em que apenas identifica as vias que possuem um número significativo de genes, não considerando a efectiva relação existente entre estes. Esta análise pode, pois, ser estendida através do uso de muita da informação já existente no Gens combinada com dados provenientes de bases de dados de redes de genes (*gene networks*). Outro desafio resulta da possibilidade de compreender a cadeia de relações existentes entre genes e fenótipos, especialmente no caso de doenças genéticas.

Independentemente da direcção seguida, as ferramentas propostas neste trabalho, quer ao nível de integração de dados quer ao nível das metodologias de visualização e de análise, consistem numa plataforma essencial na resposta a questões na área da biologia molecular e da biomedicina.

Referências

[1] A. von Bubnoff, "Next-generation sequencing: the race is on", *Cell,* vol. 132, pp. 721-3, Mar 7 2008.

[2] E. Pettersson, J. Lundeberg, and A. Ahmadian, "Generations of sequencing technologies", *Genomics,* vol. 93, pp. 105-11, Feb 2009.

[3] J. Shendure and H. Ji, "Next-generation DNA sequencing", *Nat Biotechnol,* vol. 26, pp. 1135-45, Oct 2008.

[4] J. J. Keurentjes, M. Koornneef, and D. Vreugdenhil, "Quantitative genetics in the age of omics", *Curr Opin Plant Biol,* vol. 11, pp. 123-8, Apr 2008.

[5] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA microarrays", *Nat Genet,* vol. 21, pp. 33-7, Jan 1999.

[6] N. Sevenet and O. Cussenot, "DNA microarrays in clinical practice: past, present, and future", *Clin Exp Med,* vol. 3, pp. 1-3, May 2003.

[7] A. E. Frolov, A. K. Godwin, and O. O. Favorova, "Differential gene expression analysis by DNA microarrays technology and its application in molecular oncology", *Mol Biol (Mosk),* vol. 37, pp. 573-84, Jul-Aug 2003.

[8] G. R. Grant, E. Manduchi, A. Pizarro, and C. J. Stoeckert, Jr., "Maintaining data integrity in microarray data management", *Biotechnol Bioeng,* vol. 84, pp. 795-800, Dec 30 2003.

[9] H. F. Lodish, *Molecular cell biology*. New York: W.H. Freeman, 2007.

[10] R. M. Simon, *Design and analysis of DNA microarray investigations*. New York: Springer, 2003.

[11] D. R. Rhodes and A. M. Chinnaiyan, "DNA microarrays: implications for clinical medicine", *J Invest Surg,* vol. 15, pp. 275-9, Sep-Oct 2002.

[12] G. S. Cojocaru, G. Rechavi, and N. Kaminski, "The use of microarrays in medicine", *Isr Med Assoc J,* vol. 3, pp. 292-6, Apr 2001.

[13] B. Alberts, *Molecular biology of the cell*. New York: Garland Science, 2002.

[14] N. J. Trun and J. E. Trempy, *Fundamental bacterial genetics*. Malden, MA: Blackwell, 2004.

[15] J. M. Bartlett and D. Stirling, "A short history of the polymerase chain reaction", *Methods in molecular biology (Clifton, N.J.),* vol. 226, pp. 3-6, 2003.

[16] H. A. Erlich, *PCR technology : principles and applications for DNA amplification*. New York: Oxford University Press, 1994.

[17] E. Southern, "Southern blotting", *Nature protocols,* vol. 1, pp. 518-25, 2006.

[18] C. G. Kevil, L. Walsh, F. S. Laroux, T. Kalogeris, M. B. Grisham, and J. S. Alexander, "An Improved, Rapid Northern Protocol", *Biochemical and Biophysical Research Communications,* vol. 238, pp. 277-279, 1997.

[19] H. Towbin, T. Staehelin, and J. Gordon, "Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 76, pp. 4350-4, 1979.

[20] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray", *Science.,* vol. 270, p. 467, 1995.

[21] S. Choudhuri, "Microarrays in biology and medicine", *Journal of biochemical and molecular toxicology,* vol. 18, pp. 171-9, 2004.

[22] D. B. Searls, "Using bioinformatics in gene and drug discovery", *Drug Discovery Today,* vol. 5, pp. 135-143, 2000.

[23] D. Stekel, *Microarray bioinformatics*. Cambridge; New York: Cambridge University Press, 2003.

[24] M. Dufva, "Fabrication of DNA Microarray", *Methods in Molecular Biology,* vol. 529, pp. 63-80, 2009.

[25] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. Garcia, J. Geoghegan, G. Germino*, et al.*, "Multiple-laboratory comparison of microarray platforms", *Nat Methods,* vol. 2, pp. 345-50, May 2005.

[26] C. L. Yauk, M. L. Berndt, A. Williams, and G. R. Douglas, "Comprehensive comparison of six microarray technologies", *Nucleic Acids Res,* vol. 32, p. e124, 2004.

[27] S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas, "Light-directed, spatially addressable parallel chemical synthesis", *Science,* vol. 251, pp. 767-73, Feb 15 1991.

[28] T. R. Hughes, M. Mao, A. R. Jones, J. Burchard, M. J. Marton, K. W. Shannon, S. M. Lefkowitz, M. Ziman, J. M. Schelter, M. R. Meyer*, et al.*, "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer", *Nature biotechnology,* vol. 19, pp. 342-7, 2001.

[29] C. Lausted, T. Dahl, C. Warren, K. King, K. Smith, M. Johnson, R. Saleem, J. Aitchison, L. Hood, and S. R. Lasky, "POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer", *Genome biology,* vol. 5, 2004.

[30] A. L. Ghindilis, M. W. Smith, K. R. Schwarzkopf, K. M. Roth, K. Peyvan, S. B. Munro, M. J. Lodes, A. G. Stover, K. Bernards, K. Dill*, et al.*, "CombiMatrix

oligonucleotide arrays: genotyping and gene expression assays employing electrochemical detection", *Biosens Bioelectron,* vol. 22, pp. 1853-60, Apr 15 2007.

[31] F. N. Ahmed Fadiel, "Microarray applications and challenges: a vast array of possibilities", *Int Arch Biosci,* pp. 1111-1121, 2003.

[32] T. Moroy, "DNA Microarrays in Medicine: Can the Promises Be Kept?", *Journal of Biomedicine and Biotechnology,* vol. 2:1, 2002.

[33] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent, "Use of a cDNA microarray to analyse gene expression patterns in human cancer", *Nature genetics,* vol. 14, pp. 457-60, 1996.

[34] R. Mei, P. C. Galipeau, C. Prass, A. Berno, G. Ghandour, N. Patil, R. K. Wolff, M. S. Chee, B. J. Reid, and D. J. Lockhart, "Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays", *Genome Res,* vol. 10, pp. 1126-37, Aug 2000.

[35] C. Eftychi, J. M. Howson, B. J. Barratt, A. Vella, F. Payne, D. J. Smyth, R. C. Twells, N. M. Walker, H. E. Rance, E. Tuomilehto-Wolf*, et al.*, "Analysis of the type 2 diabetes-associated single nucleotide polymorphisms in the genes IRS1, KCNJ11, and PPARG2 in type 1 diabetes", *Diabetes,* vol. 53, pp. 870-3, 2004.

[36] X. S. Liu, "Getting started in tiling microarray analysis", *PLoS computational biology,* vol. 3, pp. 1842-4, 2007.

[37] D. Kampa, J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt*, et al.*, "Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22", *Genome research,* vol. 14, pp. 331-42, 2004.

[38] S. Draghici, *Data analysis tools for DNA microarrays*. Boca Raton: Chapman & Hall/CRC, 2003.

[39] Z. G. Goldsmith and N. Dhanasekaran, "The microrevolution: applications and impacts of microarray technology on molecular biology and medicine (review)", *Int J Mol Med,* vol. 13, pp. 483-95, Apr 2004.

[40] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes", *Proc Natl Acad Sci U S A,* vol. 93, pp. 10614-9, Oct 1 1996.

[41] M. Gardiner-Garden and T. G. Littlejohn, "A comparison of microarray databases", *Brief Bioinform,* vol. 2, pp. 143-58, May 2001.

[42] C. J. Stoeckert, Jr., H. C. Causton, and C. A. Ball, "Microarray databases: standards and ontologies", *Nat Genet,* vol. 32 Suppl, pp. 469-73, Dec 2002.

[43] C. A. Ball, G. Sherlock, H. Parkinson, P. Rocca-Sera, C. Brooksbank, H. C. Causton, D. Cavalieri, T. Gaasterland, P. Hingamp, F. Holstege*, et al.*, "Standards for microarray data", *Science,* vol. 298, p. 539, Oct 18 2002.

[44] P. L. Whetzel, H. Parkinson, H. C. Causton, L. Fan, J. Fostel, G. Fragoso, L. Game, M. Heiskanen, N. Morrison, P. Rocca-Serra*, et al.*, "The MGED Ontology: a resource for semantics-based description of microarray experiments", *Bioinformatics,* vol. 22, pp. 866-73, Apr 1 2006.

[45] K. Dobbin, J. H. Shih, and R. Simon, "Statistical design of reverse dye microarrays", *Bioinformatics,* vol. 19, pp. 803-10, May 1 2003.

[46] Y. H. Yang and T. Speed, "Design issues for cDNA microarray experiments", *Nat Rev Genet,* vol. 3, pp. 579-88, Aug 2002.

[47] T. Yuen, E. Wurmbach, R. L. Pfeffer, B. J. Ebersole, and S. C. Sealfon, "Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays", *Nucleic Acids Res,* vol. 30, p. e48, May 15 2002.

[48] Y. H. Yang, N. P. Thorne, Ims, and D. R. Goldstein, *Normalization for two-color cDNA microarray data*: Institute of Mathematical Statistics, 2003.

[49] G. K. Smyth and T. P. Speed, "Normalization of cDNA Microarray Data", *methods,* vol. 85, pp. 265-273, 2003.

[50] J. B. Bard and S. Y. Rhee, "Ontologies in biology: design, applications and future challenges", *Nat Rev Genet,* vol. 5, pp. 213-22, Mar 2004.

[51] C. A. Ball and A. Brazma, "MGED standards: work in progress", *OMICS,* vol. 10, pp. 138-44, Summer 2006.

[52] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton*, et al.*, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data", *Nat Genet,* vol. 29, pp. 365-71, Dec 2001.

[53] J. A. Blake and M. A. Harris, "The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis", *Curr Protoc Bioinformatics,* vol. Chapter 7, p. Unit 7 2, Sep 2008.

[54] P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage*, et al.*, "Design and implementation of microarray gene expression markup language (MAGE-ML)", *Genome Biol,* vol. 3, p. RESEARCH0046, Aug 23 2002.

[55] P. Rocca-Serra, A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, S. Contrino, J. Vilo, N. Abeygunawardena, G. Mukherjee, E. Holloway*, et al.*, "ArrayExpress: a public database of gene expression data at EBI", *C R Biol,* vol. 326, pp. 1075-8, Oct-Nov 2003.

[56] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", *Nucleic Acids Res,* vol. 30, pp. 207-10, Jan 1 2002.

[57] K. Ikeo, J. Ishi-i, T. Tamura, T. Gojobori, and Y. Tateno, "CIBEX: center for information biology gene expression database", *C R Biol,* vol. 326, pp. 1079-82, Oct-Nov 2003.

[58] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall*, et al.*, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration", *Nat Biotechnol,* vol. 25, pp. 1251-5, Nov 2007.

[59] D. A. Lindberg, B. L. Humphreys, and A. T. McCray, "The Unified Medical Language System", *Methods Inf Med,* vol. 32, pp. 281-91, Aug 1993.

[60] J. Arrais, J. L. Oliveira, G. Grimes, S. Moodie, K. Robertson, and P. Ghazal, "Microarray data sharing in BioMedicine", presented at the MIE 2006, Maastricht, Netherlands, 2006.

[61] A. T. McCray and R. A. Miller, "Making the conceptual connections: the Unified Medical Language System (UMLS) after a decade of research and development", *J Am Med Inform Assoc,* vol. 5, pp. 129-30, Jan-Feb 1998.

[62] L. H. Saal, C. Troein, J. Vallon-Christersson, S. Gruvberger, A. Borg, and C. Peterson, "BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data", *Genome Biol,* vol. 3, p. SOFTWARE0003, Jul 15 2002.

[63] P. J. Killion, G. Sherlock, and V. R. Iyer, "The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD)", *BMC Bioinformatics,* vol. 4, p. 32, Aug 20 2003.

[64] M. Maurer, R. Molidor, A. Sturn, J. Hartler, H. Hackl, G. Stocker, A. Prokesch, M. Scheideler, and Z. Trajanoski, "MARS: microarray analysis, retrieval, and storage system", *BMC Bioinformatics,* vol. 6, p. 101, 2005.

[65] D. Hancock, M. Wilson, G. Velarde, N. Morrison, A. Hayes, H. Hulme, A. J. Wood, K. Nashar, D. B. Kell, and A. Brass, "maxdLoad2 and maxdBrowse: standards-compliant tools for microarray experimental annotation, data management and dissemination", *BMC Bioinformatics,* vol. 6, p. 264, 2005.

[66] A. I. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan*, et al.*, "TM4: a free, open-source system for microarray data management and analysis", *Biotechniques,* vol. 34, pp. 374-8, Feb 2003.

[67] J. Arrais, L. Silva, M. Rodrigues, L. Carreto, J. L. Oliveira, and M. A. S. Santos, "Why another microarray LIMS", presented at the 3th European Medical and Biological Engineering Conference, Prage, Czech Republic, 2005.

[68] J. P. Arrais, L. Carreto, M. A. Santos, and J. L. Oliveira, "A Microarray Information Database", presented at the Biocomputation, Bioinformatics, and Biomedical Technologies, 2008. BIOTECHNO '08. International Conference on, Bucharest, 2008.

[69] J. Nielsen, *Usability Engineering*: Morgan Kaufmann, 1994.

[70] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry*, et al.*, "Bioconductor: open software development for computational biology and bioinformatics", *Genome Biol,* vol. 5, p. R80, 2004.

[71] B. Ventura, "Mandatory submission of microarray data to public repositories: how is it working?", *Physiol Genomics,* vol. 20, pp. 153-6, Jan 20 2005.

[72] C. A. Ball, A. Brazma, H. Causton, S. Chervitz, R. Edgar, P. Hingamp, J. C. Matese, H. Parkinson, J. Quackenbush, M. Ringwald*, et al.*, "Submission of microarray data to public repositories", *PLoS Biol,* vol. 2, p. E317, Sep 2004.

[73] U. Sarkans, H. Parkinson, G. G. Lara, A. Oezcimen, A. Sharma, N. Abeygunawardena, S. Contrino, E. Holloway, P. Rocca-Serra, G. Mukherjee*, et al.*,

"The ArrayExpress gene expression database: a software engineering and implementation perspective", *Bioinformatics,* vol. 21, pp. 1495-501, Apr 15 2005.

[74] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk*, et al.*, "ArrayExpress - a public database of microarray experiments and gene expression profiles", *Nucleic Acids Res,* vol. 35, pp. D747-50, Jan 2007.

[75] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall*, et al.*, "NCBI GEO: archive for high-throughput functional genomic data", *Nucleic Acids Res,* vol. 37, pp. D885-90, Jan 2009.

[76] T. F. Rayner, P. Rocca-Serra, P. T. Spellman, H. C. Causton, A. Farne, E. Holloway, R. A. Irizarry, J. Liu, D. S. Maier, M. Miller*, et al.*, "A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB", *BMC Bioinformatics,* vol. 7, p. 489, 2006.

[77] G. R. Grimes, S. Moodie, J. S. Beattie, M. Craigon, P. Dickinson, T. Forster, A. D. Livingston, M. Mewissen, K. A. Robertson, A. J. Ross*, et al.*, "GPX-Macrophage Expression Atlas: a database for expression profiles of macrophages challenged with a variety of pro-inflammatory, anti-inflammatory, benign and pathogen insults", *BMC Genomics,* vol. 6, p. 178, 2005.

[78] J. Quackenbush, "Computational analysis of microarray data", *Nat Rev Genet,* vol. 2, pp. 418-27, Jun 2001.

[79] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus", *Nat Rev Genet,* vol. 7, pp. 55-65, Jan 2006.

[80] S. I. Lee and S. Batzoglou, "Application of independent component analysis to microarrays", *Genome biology,* vol. 4, 2003.

[81] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On Clustering Validation Techniques", *Journal of Intelligent Information Systems,* vol. 17, pp. 107-145, 2001.

[82] M. E. Ritchie, J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G. K. Smyth, "A comparison of background correction methods for two-colour microarrays", *Bioinformatics,* vol. 23, pp. 2700-7, Oct 15 2007.

[83] D. Edwards, "Non-linear normalization and background correction in one-channel cDNA microarray studies", *Bioinformatics,* vol. 19, pp. 825-33, May 1 2003.

[84] J. Quackenbush, "Microarray data normalization and transformation", *Nat Genet,* vol. 32 Suppl, pp. 496-501, Dec 2002.

[85] C. Steinhoff and M. Vingron, "Normalization and quantification of differential expression in gene expression microarrays", *Brief Bioinform,* vol. 7, pp. 166-77, Jun 2006.

[86] W. S. Cleveland, E. Grosse, and W. M. Shyu. (1992) Local regression models. *Statistical Models in S.*

[87] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation", *Nucleic Acids Res,* vol. 30, p. e15, Feb 15 2002.

[88] R. A. Fisher, *Statistical methods for research workers : by Ronald A. Fisher*. New York: Hafner, 1970.

[89] A. Petrie and C. Sabin, *Medical statistics at a glance*. Malden, Mass. [u.a.]: Blackwell, 2007.

[90] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments", *Genome Biol,* vol. 4, p. 210, 2003.

[91] I. V. Yang, E. Chen, J. P. Hasseman, W. Liang, B. C. Frank, S. Wang, V. Sharov, A. I. Saeed, J. White, J. Li*, et al.*, "Within the fold: assessing differential expression measures and reproducibility in microarray assays", *Genome biology,* vol. 3, 2002.

[92] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response", *Proc Natl Acad Sci U S A,* vol. 98, pp. 5116-21, Apr 24 2001.

[93] R. Gentleman, *Bioinformatics and computational biology solutions using R and Bioconductor*. New York: Springer, 2005.

[94] M. K. Kerr, M. Martin, and G. A. Churchill, "Analysis of variance for gene expression microarray data", *J Comput Biol,* vol. 7, pp. 819-37, 2000.

[95] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments", *Stat Appl Genet Mol Biol,* vol. 3, p. Article3, 2004.

[96] Y. F. Leung and D. Cavalieri, "Fundamentals of cDNA microarray data analysis", *Trends Genet,* vol. 19, pp. 649-59, Nov 2003.

[97] M. Reimers and V. J. Carey, "Bioconductor: an open source framework for bioinformatics and computational biology", *Methods Enzymol,* vol. 411, pp. 119-34, 2006.

[98] R. Ihaka and R. Gentleman, "R: A Language for Data Analysis and Graphics", *Journal of Computational and Graphical Statistics,* vol. 5, pp. 299-314, 1996.

[99] G. Parmigiani, *The analysis of gene expression data: methods and software*. New York: Springer, 2003.

[100] M. Slawski, M. Daumer, and A. L. Boulesteix, "CMA: a comprehensive Bioconductor package for supervised classification with high dimensional data", *BMC Bioinformatics,* vol. 9, p. 439, 2008.

[101] D. R. Augustyn and L. Warchal, "ServeR: .NET-based infrastructure for remote services of statistical computing with R-project", *Commun. Comput. Info. Sci. Communications in Computer and Information Science,* vol. 39, pp. 192-199, 2009.

[102] "The Universal Protein Resource (UniProt) 2009", *Nucleic Acids Res,* vol. 37, pp. D169-74, Jan 2009.

[103] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank: update", *Nucleic Acids Research,* vol. 32, p. 23, 2004.

[104] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne*, et al.*, "InterPro: the integrative protein signature database", *Nucleic Acids Res,* vol. 37, pp. D211-5, Jan 2009.

[105] L. D. Stein, "Integrating biological databases", *Nat Rev Genet,* vol. 4, pp. 337-45, May 2003.

[106] F. Al-Shahrour, P. Minguez, J. Tarraga, I. Medina, E. Alloza, D. Montaner, and J. Dopazo, "FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments", *Nucleic Acids Res,* vol. 35, pp. W91-6, Jul 2007.

[107] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, and R. Romero, "A systems biology approach for pathway level analysis", *Genome Res,* vol. 17, pp. 1537-45, Oct 2007.

[108] S. Draghici, P. Khatri, P. Bhavsar, A. Shah, S. A. Krawetz, and M. A. Tainsky, "Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate", *Nucleic Acids Res,* vol. 31, pp. 3775-81, Jul 1 2003.

[109] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi*, et al.*, "GoMiner: a resource for biological interpretation of genomic and proteomic data", *Genome Biol,* vol. 4, p. R28, 2003.

[110] T. J. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke*, et al.*, "Ensembl 2009", *Nucleic Acids Res,* vol. 37, pp. D690-7, Jan 2009.

[111] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI", *Nucleic Acids Res,* vol. 35, pp. D26-31, Jan 2007.

[112] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes", *Nucleic Acids Res,* vol. 28, pp. 27-30, Jan 1 2000.

[113] M. Y. Galperin and G. R. Cochrane, "Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009", *Nucleic Acids Res,* vol. 37, pp. D1-4, Jan 2009.

[114] H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G. G. Lara, E. Holloway, M. Kapushesky*, et al.*, "ArrayExpress - a public repository for microarray gene expression data at the EBI", *Nucleic Acids Res,* vol. 33, pp. D553-5, Jan 1 2005.

[115] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, and P. Tarczy-Hornoch, "Data integration and genomic medicine", *J Biomed Inform,* vol. 40, pp. 5-16, Feb 2007.

[116] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood, "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network", *Science,* vol. 292, pp. 929-34, May 4 2001.

[117] T. E. Klein, J. T. Chang, M. K. Cho, K. L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D. E. Oliver*, et al.*, "Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base", *Pharmacogenomics J,* vol. 1, pp. 167-70, 2001.

[118] E. E. Schadt, S. A. Monks, and S. H. Friend, "A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets", *Biochem Soc Trans,* vol. 31, pp. 437-43, Apr 2003.

[119] O. Brazhnik and J. F. Jones, "Anatomy of data integration", *J Biomed Inform,* vol. 40, pp. 252-69, Jun 2007.

[120] M. D. Wilkinson and M. Links, "BioMOBY: an open source biological web services proposal", *Brief Bioinform,* vol. 3, pp. 331-41, Dec 2002.

[121] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat*, et al.*, "Taverna: a tool for the composition and enactment of bioinformatics workflows", *Bioinformatics,* vol. 20, pp. 3045-54, Nov 22 2004.

[122] R. D. Dowell, R. M. Jokerst, A. Day, S. R. Eddy, and L. Stein, "The distributed annotation system", *BMC Bioinformatics,* vol. 2, p. 7, 2001.

[123] P. B. Neerincx and J. A. Leunissen, "Evolution of web services in bioinformatics", *Brief Bioinform,* vol. 6, pp. 178-88, Jun 2005.

[124] L. Haas, P. Schwarz, P. Kodali, E. Kotlar, J. Rice, and W. Swope, "DiscoveryLink: A system for integrated access to life sciences data sources", *IBM Systems Journal,* pp. 489-511, 2001.

[125] J. L. Oliveira, G. Dias, I. Oliveira, P. Rocha, I. Hermosilla, J. Vicente, I. Spiteri, F. Martin-Sánchez, and A. S. Pereira, "DiseaseCard: A Web-Based Tool for the Collaborative Integration of Genetic and Medical Information", in *Biological And Medical Data Analysis: 5th International Symposium*, Springer, Ed., ed, 2004, pp. 409-417.

[126] E. Cadag, B. Louie, P. J. Myler, and P. Tarczy-Hornoch, "Biomediator data integration and inference for functional annotation of anonymous sequences", *Pac Symp Biocomput,* pp. 343-54, 2007.

[127] J. Kohler, S. Philippi, and M. Lange, "SEMEDA: ontology based semantic integration of biological databases", *Bioinformatics,* vol. 19, pp. 2420-7, Dec 12 2003.

[128] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk, "BioMart - biological queries made easy", *BMC Genomics,* vol. 10, p. 22, 2009.

[129] D. L. Wheeler, D. M. Church, R. Edgar, S. Federhen, W. Helmberg, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira*, et al.*, "Database resources of the National Center for Biotechnology Information: update", *Nucleic Acids Res,* vol. 32, pp. D35-40, Jan 1 2004.

[130] M. R. Wilkins, E. Gasteiger, A. Bairoch, J. C. Sanchez, K. L. Williams, R. D. Appel, and D. F. Hochstrasser, "Protein identification and analysis tools in the ExPASy server", *Methods Mol Biol,* vol. 112, pp. 531-52, 1999.

[131] C. F. Thorn, T. E. Klein, and R. B. Altman, "PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base", *Methods Mol Biol,* vol. 311, pp. 179-91, 2005.

[132] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig*, et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium", *Nat Genet,* vol. 25, pp. 25-9, May 2000.

[133] R. Stevens, C. A. Goble, and S. Bechhofer, "Ontology-based knowledge representation for bioinformatics", *Brief Bioinform,* vol. 1, pp. 398-414, Nov 2000.

[134] L. Wong, "Technologies for integrating biological data", *Brief Bioinform,* vol. 3, pp. 389-404, Dec 2002.

[135] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, "GeneCards: integrating information about genes, proteins and diseases", *Trends Genet,* vol. 13, p. 163, Apr 1997.

[136] S. Zhong, K. F. Storch, O. Lipan, M. C. Kao, C. J. Weitz, and W. H. Wong, "GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space", *Appl Bioinformatics,* vol. 3, pp. 261-4, 2004.

[137] S. Draghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz, "Global functional profiling of gene expression", *Genomics,* vol. 81, pp. 98-104, Feb 2003.

[138] F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo, "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes", *Bioinformatics,* vol. 20, pp. 578-80, Mar 1 2004.

[139] A. Turchin and I. S. Kohane, "Gene homology resources on the World Wide Web", *Physiol Genomics,* vol. 11, pp. 165-77, Dec 3 2002.

[140] H. Caron, B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M. C. Hermus, R. van Asperen, K. Boon, P. A. Voute*, et al.*, "The human transcriptome map: clustering of highly expressed genes in chromosomal domains", *Science,* vol. 291, pp. 1289-92, Feb 16 2001.

[141] C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, and U. Leser, "AliBaba: PubMed as a graph", *Bioinformatics,* vol. 22, pp. 2444-5, Oct 1 2006.

[142] A. Doms and M. Schroeder, "GoPubMed: exploring PubMed with the Gene Ontology", *Nucleic Acids Res,* vol. 33, pp. W783-6, Jul 1 2005.

[143] R. Hoffmann and A. Valencia, "Implementing the iHOP concept for navigation of biomedical literature", *Bioinformatics,* vol. 21 Suppl 2, pp. ii252-8, Sep 1 2005.

[144] E. Hatcher, O. Gospodnetic, and M. McCandless, *Lucene in action*. Greenwich, Conn.; London: Manning ; Pearson Education [distributor], 2009.

[145] A. R. Soares, P. M. Pereira, B. Santos, C. Egas, A. C. Gomes, J. Arrais, J. L. Oliveira, G. R. Moura, and M. A. Santos, "Parallel DNA pyrosequencing unveils new zebrafish microRNAs", *BMC Genomics,* vol. 10, p. 195, 2009.

[146] D. F. Polonia, J. Arrais, and J. L. Oliveira, "A Prospective Study on the Integration of Microarray Data in HIS/EPR", *Lecture notes in computer science.,* pp. 231-239, 2006.

[147] J. Arrais, J. E. Pereira, J. Fernandes, and J. L. Oliveira, "GeNS: a biological data integration platform", presented at the International Conference on Bioinformatics and Biomedicine, Venice, Italy, 2009.

[148] J. Arrais, B. Santos, J. Fernandes, L. Carreto, M. A. S. Santos, and J. L. Oliveira, "GeneBrowser: an approach for integration and functional classification of genomic data", *Journal of Integrative Bioinformatics,* vol. 4, 2007.

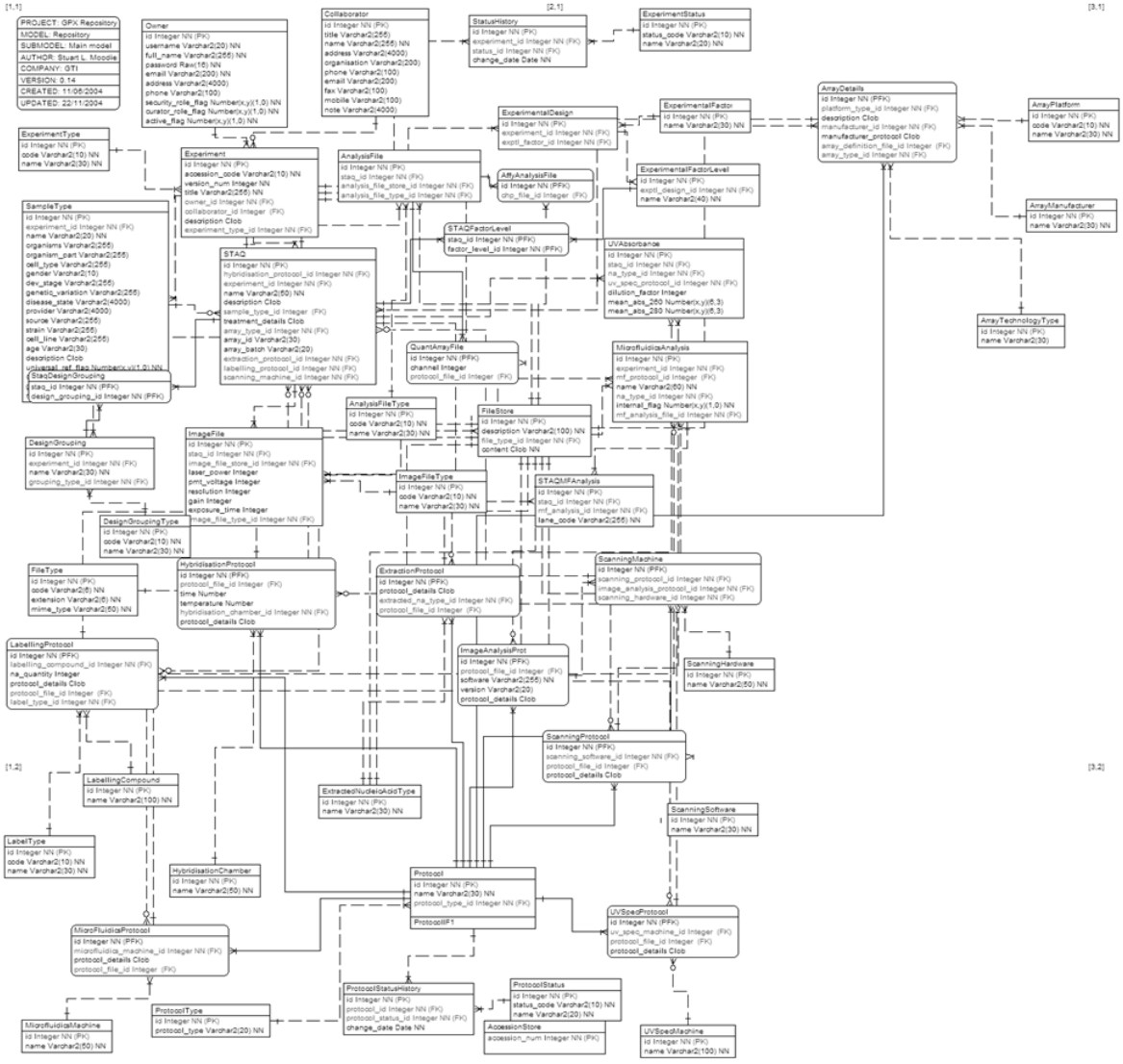
[149] J. Arrais, J. G. L. M. Rodrigues, and J. L. Oliveira, "Improving Literature Searches in Gene Expression Studies", in *Advances in Intelligent and Soft Computing : 2nd International Workshop on Practical Applications of Computational Biology and Bioinformatics*, J. M. Corchado*, et al.*, Eds., ed Berlin, DE: Springer Berlin / Heidelberg, 2009, pp. Capt. 10, p. 74 - 82.

[150] P. Lopes, J. Arrais, and J. L. Oliveira, "Dynamic service integration using web-based workflows", in *iiWAS '08: Proceedings of the 10th International Conference on Information Integration and Web-based Applications \& Services*, Linz, Austria, 2008, pp. 622-625.

[151] P. Lopes, J. Arrais, and J. L. Oliveira, "DynamicFlow: A Client-Side Workflow Management System", *Lecture notes in computer science.,* pp. 1101-1108, 2009.

[152] G. Moura, M. Pinheiro, J. Arrais, A. C. Gomes, L. Carreto, A. Freitas, J. L. Oliveira, and M. A. Santos, "Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure", *PLoS One,* vol. 2, p. e847, 2007.

Anexo 1: Esquema da base de dados do sistema GPX



Anexo 2: MAGE-ML da experiência E-MEXP-1490

```
<Experiment_package>
    <Experiment_assnlist>
      <Experiment identifier="E-MEXP-1490"
                  name="Transcriptional Response of Mouse Bone Marrow Derived Macrophages
to Interferon-gamma Treatment">
        <PropertySets_assnlist>
          <NameValueType value="2008-03-17"
                         name="ArrayExpressReleaseDate">
          </NameValueType>
          <NameValueType value="2008-02-27 17:12:53"
                         name="ArrayExpressSubmissionDate">
          </NameValueType>
          <NameValueType value="Transcriptional Response of Mouse Bone Marrow Derived
Macrophages to Interferon-gamma Treatment"
                         name="AEExperimentDisplayName">
          </NameValueType>
        </PropertySets_assnlist>
        <Descriptions_assnlist>
          <Description text="We have examined the transcriptional events in mouse bone
marrow derived macrophages (MBDM) with interferon-gamma (Ifng)at 2, 4 &amp;amp; 8 h following
treatment or pre-treatment (0 h). ">
          </Description>
          <Description>
            <BibliographicReferences_assnlist>
              <BibliographicReference authors="Sobia Raza, Kevin A. Robertson, Paul A.
Lacaze, David Page, Anton J. Enright, Peter Ghazal, Tom C. Freeman"
                                      title="A logic-based diagram of signalling pathways
central to macrophage activation"
                                      publication="BMC Syst Biol"
                                      year="2008"
                                      pages="">
                <Parameters_assnlist>
                  <OntologyEntry value="journal_article"
                                 category="PublicationType">
                  </OntologyEntry>
                </Parameters_assnlist>
              </BibliographicReference>
            </BibliographicReferences_assnlist>
          </Description>
        </Descriptions_assnlist>
        <Providers_assnreflist>
          <Person_ref identifier="ebi.ac.uk:MIAMExpress:Person:Sobia_Raza.arrayexpress"/>
        </Providers_assnreflist>
        <BioAssayData_assnreflist>
          <DerivedBioAssayData_ref
identifier="ebi.ac.uk:MIAMExpress:DerivedBioAssayData:4256"/>
          <MeasuredBioAssayData_ref
identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssayData:37059"/>
          <MeasuredBioAssayData_ref
identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssayData:37060"/>
          <MeasuredBioAssayData_ref
identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssayData:37061"/>
          <MeasuredBioAssayData_ref
identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssayData:37062"/>
          <MeasuredBioAssayData_ref
identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssayData:37063"/>
          <MeasuredBioAssayData_ref
identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssayData:37064"/>
          <MeasuredBioAssayData_ref
identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssayData:37065"/>
          <MeasuredBioAssayData_ref
identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssayData:37066"/>
          <MeasuredBioAssayData_ref
identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssayData:37067"/>
        </BioAssayData_assnreflist>
        <BioAssays_assnreflist>
          <DerivedBioAssay_ref
identifier="ebi.ac.uk:MIAMExpress:DerivedBioAssay:4256.GEM_FGEM"/>
          <PhysicalBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:PhysicalBioAssay:37059"/>
          <MeasuredBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssay:37059"/>
          <PhysicalBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:PhysicalBioAssay:37060"/>
          <MeasuredBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssay:37060"/>
          <PhysicalBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:PhysicalBioAssay:37061"/>
          <MeasuredBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssay:37061"/>
          <PhysicalBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:PhysicalBioAssay:37062"/>
          <MeasuredBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssay:37062"/>
```

```
          <PhysicalBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:PhysicalBioAssay:37063"/>
          <MeasuredBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssay:37063"/>
          <PhysicalBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:PhysicalBioAssay:37064"/>
          <MeasuredBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssay:37064"/>
          <PhysicalBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:PhysicalBioAssay:37065"/>
          <MeasuredBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssay:37065"/>
          <PhysicalBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:PhysicalBioAssay:37066"/>
          <MeasuredBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssay:37066"/>
          <PhysicalBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:PhysicalBioAssay:37067"/>
          <MeasuredBioAssay_ref identifier="ebi.ac.uk:MIAMExpress:MeasuredBioAssay:37067"/>
        </BioAssays_assnreflist>
        <ExperimentDesigns_assnlist>
          <ExperimentDesign>
            <Types_assnlist>
              <OntologyEntry value="time_series_design"
                             category="ExperimentDesignType">
              </OntologyEntry>
              <OntologyEntry value="co-expression_design"
                             category="ExperimentDesignType">
              </OntologyEntry>
              <OntologyEntry value="compound_treatment_design"
                             category="ExperimentDesignType">
              </OntologyEntry>
              <OntologyEntry value="in_vitro_design"
                             category="ExperimentDesignType">
              </OntologyEntry>
            </Types_assnlist>
            <ExperimentalFactors_assnlist>
              <ExperimentalFactor
identifier="ebi.ac.uk:MIAMExpress:ExperimentalFactor:4256.428"
                                  name="compound">
                <Category_assn>
                  <OntologyEntry value="compound"
                                 category="ExperimentalFactorCategory">
                  </OntologyEntry>
                </Category_assn>
                <FactorValues_assnlist>
                  <FactorValue
identifier="ebi.ac.uk:MIAMExpress:FactorValue:4256.209269.compound"
                               name="compound">
                    <Value_assn>
                      <OntologyEntry value="interferon gamma"
                                     category="Compound">
                      </OntologyEntry>
                    </Value_assn>
                  </FactorValue>
                </FactorValues_assnlist>
              </ExperimentalFactor>
              <ExperimentalFactor
identifier="ebi.ac.uk:MIAMExpress:ExperimentalFactor:4256.433"
                                  name="dose">
                <Category_assn>
                  <OntologyEntry value="dose"
                                 category="ExperimentalFactorCategory">
                  </OntologyEntry>
                </Category_assn>
                <FactorValues_assnlist>
                  <FactorValue
identifier="ebi.ac.uk:MIAMExpress:FactorValue:4256.209269.dose"
                               name="dose">
                    <Measurement_assn>
                      <Measurement value="10"
                                   otherKind="u/ml"
                                   type="absolute"
                                   kindCV="other">
                      </Measurement>
                    </Measurement_assn>
                  </FactorValue>
                </FactorValues_assnlist>
              </ExperimentalFactor>
              <ExperimentalFactor
identifier="ebi.ac.uk:MIAMExpress:ExperimentalFactor:4256.441"
                                  name="time">
                <Category_assn>
                  <OntologyEntry value="time"
                                 category="ExperimentalFactorCategory">
                  </OntologyEntry>
                </Category_assn>
                <FactorValues_assnlist>
                  <FactorValue
identifier="ebi.ac.uk:MIAMExpress:FactorValue:4256.209269.time"
                               name="time">
                    <Measurement_assn>
```

```
                    <Measurement value="0"
                                 type="absolute"
                                 kindCV="time">
                      <Unit_assn>
                        <TimeUnit unitNameCV="h">
                        </TimeUnit>
                      </Unit_assn>
                    </Measurement>
                  </Measurement_assn>
                </FactorValue>
                <FactorValue
identifier="ebi.ac.uk:MIAMExpress:FactorValue:4256.209271.time"
                             name="time">
                  <Measurement_assn>
                    <Measurement value="1"
                                 type="absolute"
                                 kindCV="time">
                      <Unit_assn>
                        <TimeUnit unitNameCV="h">
                        </TimeUnit>
                      </Unit_assn>
                    </Measurement>
                  </Measurement_assn>
                </FactorValue>
                <FactorValue
identifier="ebi.ac.uk:MIAMExpress:FactorValue:4256.209273.time"
                             name="time">
                  <Measurement_assn>
                    <Measurement value="2"
                                 type="absolute"
                                 kindCV="time">
                      <Unit_assn>
                        <TimeUnit unitNameCV="h">
                        </TimeUnit>
                      </Unit_assn>
                    </Measurement>
                  </Measurement_assn>
                </FactorValue>
                <FactorValue
identifier="ebi.ac.uk:MIAMExpress:FactorValue:4256.209274.time"
                             name="time">
                  <Measurement_assn>
                    <Measurement value="4"
                                 type="absolute"
                                 kindCV="time">
                      <Unit_assn>
                        <TimeUnit unitNameCV="h">
                        </TimeUnit>
                      </Unit_assn>
                    </Measurement>
                  </Measurement_assn>
                </FactorValue>
                <FactorValue
identifier="ebi.ac.uk:MIAMExpress:FactorValue:4256.209276.time"
                             name="time">
                  <Measurement_assn>
                    <Measurement value="8"
                                 type="absolute"
                                 kindCV="time">
                      <Unit_assn>
                        <TimeUnit unitNameCV="h">
                        </TimeUnit>
                      </Unit_assn>
                    </Measurement>
                  </Measurement_assn>
                </FactorValue>
              </FactorValues_assnlist>
            </ExperimentalFactor>
          </ExperimentalFactors_assnlist>
          <QualityControlDescription_assn>
            <Description>
              <Annotations_assnlist>
                <OntologyEntry value="biological_replicate"
                               category="QualityControlDescriptionType">
                </OntologyEntry>
              </Annotations_assnlist>
            </Description>
          </QualityControlDescription_assn>
        </ExperimentDesign>
      </ExperimentDesigns_assnlist>
    </Experiment>
  </Experiment_assnlist>
</Experiment_package>
```