
Universidade de Aveiro
2007
Departamento de Electrónica, Telecomunicações e Informática

Gaspar Manuel Sousa Dias

Sistema de Recuperação Automática de Informação Biomédica


Universidade de Aveiro
2007
Departamento de Electrónica, Telecomunicações e Informática

Gaspar Manuel Sousa Dias

Sistema de Recuperação Automática de Informação Biomédica

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

Dedico este trabalho à Filipa, minha mulher.

o júri

presidente

Doutor Joaquim Arnaldo Carvalho Martins
Professor Catedrático da Universidade de Aveiro

Doutor Orlando Manuel de Oliveira Belo
Professor Associado do Departamento de Informática da Escola de Engenharia da Universidade do Minho

Doutor José Luís Guimarães Oliveira
Professor Associado da Universidade de Aveiro

palavras-chave

Integração de dados, Bases de dados Biomédicas, Doenças Genéticas Raras, Genética, Genómica.

resumo

Os avanços mais recentes nas áreas da genómica e proteómica têm resultado num crescimento significativo em termos de informação disponibilizada ao público. Espera-se que tais quantidades de informação tragam grandes vantagens à prática clínica onde diagnósticos e tratamentos passarão a ser suportados ao nível molecular.

Contudo, a navegação através das bases de dados bioinfomáticas e genéticas revela-se actualmente uma tarefa complexa e improdutiva para grande parte dos profissionais de saúde. Além disso, no contexto das doenças genéticas raras, verifica-se que o conhecimento sobre determinadas doenças se encontra confinado a um pequeno grupo de utilizadores experientes. Novas perspectivas serão introduzidas ao nível da compreensão de varias doenças raras com a criação de interfaces intuitivas que permitam a extracção, a manutenção e a partilha desta informação por um maior número de utilizadores.

Nesta dissertação é apresentado um sistema de recuperação automática, disponibilizado através de um portal web, diseasecard.org, utilizado para reunir e integrar informação relativa a doenças raras, cobrindo conceitos que vão desde o fenótipo até ao genótipo.

keywords

Database integration Systems, Biomedical databases, Genetic rare diseases, Genetics, Genomics.

abstract

The recent advances on genomics and proteomics research bring up a significant grow on the information that is publicly available. This huge amount of data is expected to give rise to a new clinical practice, where diagnosis and treatments will be supported by information at the molecular level.

However, navigating through genetic and bioinformatics databases can be a too complex and unproductive task for a primary care physician. Moreover, considering the rare genetic diseases field, we verify that the knowledge about a specific disease is commonly disseminated over a small group of experts. The capture, maintenance and sharing of this knowledge over user-friendly interfaces will introduce new insights in the understanding of some rare genetic diseases.

In this thesis we present an information retrieval engine that is being used to gather and join information about rare diseases, from the phenotype to the genotype, in a public web portal – diseasecard.org.

Índice

Lista de Siglas e Acrónimos

ADN	Ácido Desoxirribonucleico
ARN	Ácido Ribonucleico
API	*Application Programming Interface*
DC	*DiseaseCard* (Sistema de Informação)
DIP	*Deletion Insertion Polymorphism*
DLL	*Dynamically Linked Library*
GUI	*Graphical User Interface*
HGP	*Human Genome Project*
HTTP	*HyperText Transfer Protocol*
HTML	*HyperText Markup Language*
IHGSC	*International Human Genome Sequencing Consortium*
IIS	*Internet Information Services*
JSP	*Java Server Pages*
NAR	*Nucleic Acids Research*
NCBI	*National Center for Biotechnology Information*
NIH	*National Institute of Health*
MVC	*Model-Viewer-Controller*
OMIM	*Online Mendelian Inheritance in Man*
OQL	*Object Query Language*
PDB	*Protein Data Bank*
SGBD	Sistema de Gestão de Bases de Dados
SNP	*Single Nucleotide Polymorphism*
SQL	*Structured Query Language*
SVG	*Scalable Vector Graphics*
XCD	*XML Card Descriptor*
XML	*Extensible Markup Language*
XPD	*XML Protocol Descriptor*

XPDE *XML Protocol Descriptor Engine*

1 Introdução

1.1 Contextualização

A descodificação do genoma humano e de outros seres vivos constitui um avanço importante no conhecimento da biologia nomeadamente contribuindo para o melhor entendimento do processo evolutivo das espécies bem como das relações entre genes e doenças. As expectativas da comunidade científica em torno destes avanços são bastante grandes. Como exemplo, espera-se que a integração de informação genética no meio clínico proporcione desenvolvimentos técnicos significativos ao ponto dos diagnósticos e tratamentos serem suportados com informação ao nível molecular [1, 2].

Contudo, a rápida e crescente expansão do conhecimento na área da biomedicina que daí advém, bem como a redução significativa dos custos das tecnologias computacionais e a vulgarização do acesso à Internet propiciam condições suficientes para se gerarem enormes quantidades de informação digital [3]. Com efeito, os últimos avanços tanto na biologia molecular como na genómica, estão a provocar um crescimento explosivo em termos de informação biológica (Figura 1.1), originada pela comunidade científica e fornecida por um vasto leque de bases de dados nas mais diversas implementações.

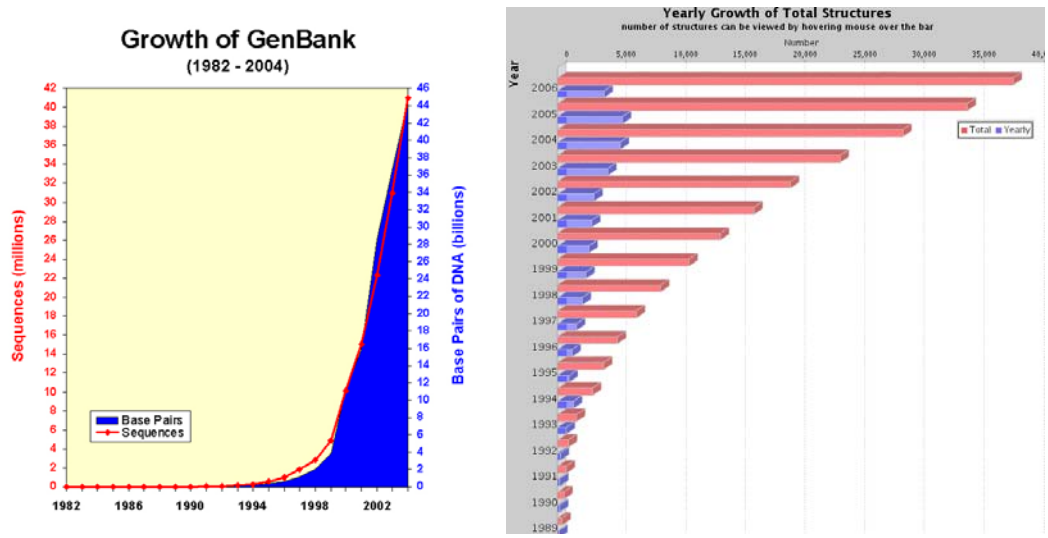


Figura 1.1 – Exemplos da rápida expansão do volume de informação gerado pela sequenciação do ADN. À esquerda: Crescimento da base de dados de sequências de ADN do GenBank [4]. À direita: Crescimento da base de dados PDB [5].

Hoje em dia, existem bases de dados dispersas por todo o mundo e acessíveis via Internet que contêm dados de natureza biomédica que vão desde informação clínica de pacientes individuais até à estrutura genética do ser humano. Assim, a informação genética do cidadão tem cada vez mais relevância no estudo do seu estado de saúde e predisposição para determinadas doenças [6], sendo que, quando disponibilizada, representa um conhecimento potencialmente valioso para a prática clínica.

Esta grande quantidade de informação disponível tem aumentado de um modo disperso na medida em que cada organização científica desenvolve modelos de dados próprios respondendo prioritariamente às suas necessidades, descurando a preocupação no sentido de coordenar e normalizar as suas implementações tanto a nível de modelo de dados como de conteúdos. Esta natureza descentralizada da comunidade científica resulta deste modo num aumento da diversidade das fontes de informação originando um vasto grupo de bases de dados heterogéneas o que dificulta a agregação, o cruzamento ou a integração das diversas fontes [3].

À medida que a complexidade e a dimensão dos recursos disponíveis na Internet aumentam, aumenta também o esforço dos utilizadores para encontrar informação específica dentro de um qualquer domínio de pesquisa. Por outro lado, tendo em conta a especificidade e heterogeneidade das bases de dados existentes, tanto a nível dos dados clínicos (pacientes) como a nível biológico (caracterização do genoma), não é de esperar que os profissionais médicos incluam este conhecimento nas suas investigações de rotina [3, 7].

Para obter um benefício real do conhecimento existente, os profissionais médicos necessitam de ferramentas ou plataformas que lhes proporcionem uma perspectiva integrada e global sobre esta grande quantidade de fontes de conhecimento, fornecendo-lhes funcionalidades de acesso, consulta e navegação.

1.2 Apresentação do Caso de Estudo

Uma sinergia eficaz entre a medicina e a biologia molecular exige uma grande cooperação e troca de informação entre os domínios clínico e biológico [8]. Actualmente existem mais de 800 [9] bases de dados públicas que cobrem as mais diversas áreas desde genes até às doenças.

Para além da grande quantidade de fontes de informação existente, outro grande obstáculo para a sua utilização prende-se com a natureza *ad hoc* de muitas delas relativamente à sua estrutura, ao modo de acesso e às terminologias utilizadas. A especificidade de tais fontes de dados bem como o conhecimento e destreza necessárias para as explorar reservam estas tarefas para utilizadores com grande experiência de navegação nas mesmas. Este problema tende a agravar-se quando esta exploração de informação se faz em torno das doenças raras, dificuldade que se deve à forte relação genótipo/fenótipo patente na sua natureza (80% destas têm origem genética).

Embora classificadas de “raras", por afectarem um número muito limitado de pessoas (menos de 1/2000), estão já diagnosticadas entre 5000 e 8000 doenças deste tipo, afectando entre 6 e 8% da população total, o que na Comunidade Europeia representa entre 24 e 36 milhões de cidadãos. Muitas destas doenças são difíceis de diagnosticar devido à ausência de sintomas ou à sua manifestação tardia. Além disso, é também comum o diagnóstico incorrecto devido à falta de conhecimento destas doenças por parte dos profissionais de saúde. Tendo em conta este cenário, torna-se urgente maximizar a disponibilização de informação correcta aos médicos, geneticistas, pacientes e famílias.

Não é portanto a ausência de informação a principal preocupação dos profissionais de saúde que lidam com as doenças genéticas raras. A sua principal dificuldade centra-se na ausência de um sistema que uniformize de certa forma todo o conhecimento genético e médico até agora obtido.

Neste sentido, tendo em vista um enquadramento global das doenças desde o sintoma até ao gene, é crucial uma selecção cuidadosa das fontes de informação biomédica, cada uma adequando-se a uma categoria ou conceito de interesse. É com este conjunto de categorias ou conceitos, através das bases de dados associadas, que a interligação entre as categorias do foro clínico com categorias dos níveis genéticos e moleculares se vai efectivar.

Tendo em conta estes factores, é aqui apresentado o *Diseasecard*, um portal virtual que integra informação genética e médica existente em múltiplas bases de dados públicas, mapeadas num protocolo de conceitos de potencial interesse. Recorrendo a este ambiente integrado, os clínicos podem alcançar e relacionar os dados disponíveis na Internet sobre cerca de duas mil doenças genéticas raras incluídas no sistema.

1.3 Motivação e Objectivos

Como já foi referido, não é a falta de informação a principal preocupação dos profissionais de saúde que normalmente lidam com as doenças genéticas raras. Uma das principais dificuldades é a ausência de sistemas que uniformizem todo o conhecimento genético e médico disponível na Internet. As diferentes fontes de dados encontram-se dispersas na rede global e a sua informação necessita de ser localizada, acedida e extraída [10]. Os principais obstáculos à integração do conhecimento biomédico e genético em torno das doenças raras podem-se resumir nos itens seguintes:

- A troca de informação entre as diversas fontes de dados torna-se complexa uma vez que as bases de dados são suportadas por modelos de dados distintos. As nomenclaturas e terminologias utilizadas não são unificadas o que dificulta o cruzamento de informação [10, 11].
- Muitos profissionais da saúde não estão familiarizados com as fontes de informação disponíveis na Internet [12]. Por outro lado, dado o grande número de fontes de informação disponíveis torna-se difícil para eles manterem-se ao corrente das "últimas novidades" na área.

Tendo estes factores em conta e partindo da constatação inicial de que as actuais infra-estruturas e sistemas de informação não fornecem ainda os meios necessários com vista à integração de conhecimento nas diversas áreas em estudo, principalmente na interligação entre a medicina geral e a genómica, surgiu a necessidade de desenvolver um protótipo de uma plataforma de integração que proporcione um conjunto de vistas integradas e globais da informação biomédica relacionada com as doenças genéticas raras desde o nível do sintoma da doença até ao gene ou genes envolvidos na mesma.

O objectivo geral desta dissertação é conceber e concretizar um sistema de informação *web*, público, que integre informação, com ênfase na área das doenças raras, proveniente de bases de dados heterogéneas médicas, biomédicas e genéticas. Mais especificamente pretende-se:

- Identificar bases de dados públicas de medicina e bioinformática nomeadamente com informação potencialmente relevante no contexto das doenças raras;

- Desenvolver um modelo de navegação sobre bases de dados públicas através de um protocolo que incluirá um conjunto de categorias tais como: sintomas, patologia, centros, mutações, fármacos, SNP, estrutura proteica, testes genéticos, etc.;
- Construir um sistema de informação para recuperação e visualização de dados sobre doenças raras que permita ir do sintoma da doença até ao gene associado à mesma;
- Desenvolver uma aplicação que permita integrar num cartão virtual por cada doença rara, informação disponível sobre os sintomas, os genes associados e as causas que as podem originar. Recorrendo a este ambiente integrado, pretende-se que os clínicos possam alcançar, entender e relacionar de uma forma intuitiva os dados disponíveis na Internet sobre as doenças genéticas raras já existentes.

1.4 Estrutura da Dissertação

Além da Introdução, esta dissertação divide-se em mais cinco capítulos cujos conteúdos são apresentados a seguir.

Capítulo 2 –Genética, Genómica e Medicina

Nesta secção são descritas algumas noções básicas da biologia molecular celular que serão úteis para entender e associar alguns dos conceitos abordados nos capítulos seguintes. Estas noções servem também de motivação para o trabalho descrito nesta dissertação. Começa-se por explorar os conceitos fundamentais associados à vida sendo eles as células, os genes e as proteínas bem como a sua constituição e propósito. Aborda ainda as diferenças entre a genética e genómica apresentando a necessidade de se associar o domínio da genética à medicina (associação genótipo/fenótipo). Termina com a alusão à motivação central deste trabalho – as doenças genéticas raras, apresentando também a problemática do acesso às enormes quantidades de informação útil e que se encontra disponível mas dispersa e fragmentada nas diversas fontes de dados na Internet.

Capítulo 3 – Integração de Informação Heterogénea

O cerne do desenvolvimento deste trabalho está fortemente associado ao contexto da integração de informação proveniente de bases de dados heterogéneas. O desafio de associar informação dispersa nas inúmeras bases de dados actuais é um problema

amplamente discutido e estudado desde o aparecimento, disponibilização e expansão das primeiras bases de dados. Neste capítulo são apresentadas algumas formas de avaliar e qualificar a heterogeneidade das fontes de informação.

Capítulo 4 – Fontes de Informação Biomédica

Tendo em conta o objectivo de reunir informação de várias áreas disciplinares à volta das doenças genéticas raras, cobrindo conceitos desde o fenótipo até ao genótipo, é apresentado neste capítulo um modelo ou protocolo que associa um conjunto de conceitos ou categorias cobrindo áreas como a medicina, a genética e farmacologia, consideradas relevantes para a pesquisa deste tipo de doenças. É também sugerido um conjunto de bases de dados cuja informação pode responder aos diversos conceitos do protocolo.

Capítulo 5 – Um modelo para a integração virtual de dados heterogéneos

Este capítulo apresenta a aplicação *DiseaseCard*, uma solução desenvolvida no âmbito deste mestrado cujo objectivo é fornecer uma vista que integre todos os conceitos do protocolo apresentado no capítulo 4, em torno das várias doenças raras. Aqui é descrito o modelo de integração do sistema bem como os blocos funcionais mais importantes da sua arquitectura.

Capítulo 6 – Conclusões e trabalho futuro

Neste capítulo são sintetizados os aspectos mais relevantes do trabalho realizado, e são apresentados alguns resultados obtidos através da aplicação. Finalmente são traçadas algumas orientações possíveis para a evolução do sistema.

2 Genética, Genómica e Medicina

2.1 Breve Introdução aos Elementos da Biologia – Células, Genes e Proteínas

Ao longo da existência humana, várias hipóteses foram formuladas por sábios, filósofos e cientistas na tentativa de compreender a origem da vida no planeta. Até ao século XIX, imaginava-se que os seres vivos poderiam surgir não só a partir do cruzamento entre si mas também a partir de matéria bruta, de uma forma espontânea. Esta teoria, proposta há mais de 2000 anos, era conhecida por geração espontânea ou abiogénese[1] (do grego *a-bio-genesis*, "origem não biológica"). Os defensores desta hipótese na antiguidade, dos quais fazia parte Aristóteles, supunham a existência de um "princípio activo" nos materiais brutos, o qual se traduziria numa força capaz de comandar uma série de reacções que culminariam com a súbita transformação do material inanimado em seres vivos.

Só no século XVII é que vários estudiosos começaram a sugerir novas ideias que se opunham à abiogénese e procuravam refutá-la partindo de experiências baseadas no método científico. Com efeito, nesse século, o biólogo italiano Francesco Redi (1626-1697) elaborou experiências simples que na época abalaram profundamente a teoria da geração espontânea. Colocou pedaços de carne no interior de frascos, deixando alguns abertos e fechando outros com uma tela. Verificou então que nos frascos fechados, onde as moscas não tinham acesso à carne, as larvas não apareciam ao contrário dos frascos abertos. A experiência de Redi favoreceu a biogénese, teoria segundo a qual a vida é originada somente a partir de outra vida preexistente.

Contudo, mais tarde, a invenção e aperfeiçoamento do microscópio reavivou a polémica da geração espontânea, renovando a aceitação da abiogénese. Em 1683, Anton van Leeuwenhoek descobriu as bactérias, e logo foi notado que não importava o quão cuidadosamente a matéria orgânica fosse protegida por telas, ou fosse colocada em

[1]Além da definição aristotélica, o termo abiogénese é actualmente utilizado como referência à origem química da vida a partir de reacções de compostos orgânicos originados abioticamente.

recipientes tapados. Uma vez que a putrefacção ocorresse, era invariavelmente acompanhada de uma miríade de bactérias e outros organismos sendo a sua origem atribuída à geração espontânea.

Foi principalmente devido ao biólogo francês Louis Pasteur (1822-1895), por volta de 1860, que a ocorrência da abiogénese no mundo microscópico foi refutada. Pasteur preparou um caldo de carne, excelente meio de cultura para micróbios, e submeteu-o a uma cuidadosa técnica de esterilização, com aquecimento e resfriamento, técnica essa que passou a ser conhecida como "pasteurização". Uma vez esterilizado, o caldo de carne era conservado no interior de um balão "pescoço de cisne". Devido ao longo gargalo do balão de vidro, o ar entrava, mas as impurezas ficavam retidas na curva do gargalo. Nenhum microrganismo poderia chegar ao caldo de carne. Assim, apesar de estar em contacto com o ar, o caldo mantinha-se estéril, provando a inexistência da geração espontânea. Muitos meses depois, Pasteur exibiu estas amostras na Academia de Ciências de Paris. O caldo de carne estava perfeitamente estéril. A geração espontânea estava definitivamente desacreditada.

Apesar de refutar a abiogénese aristotélica, a experiência de Pasteur nada diz quanto à origem química da vida. Esta forma de abiogénese supostamente ocorreu sob condições totalmente diferentes, dentro de períodos de tempo muito maiores à escala geológica, não sendo algo que se suponha ocorrer a qualquer instante ou hoje em dia. Além disso o conceito actual não propõe a origem espontânea de formas de vida complexas mas antes uma origem mais singular da vida fruto de um complexo processo gradual composto por vários estágios. A propor esta hipótese surge em 1936 Aleksandr Ivanovitch Oparin (1894-1980), que demonstrou que moléculas orgânicas podem ser criadas numa atmosfera ausente de oxigénio, através da acção da luz solar, radiações ultravioletas ou de descargas eléctricas atmosféricas sobre moléculas simples existentes no planeta como o metano, amónia, hidrogénio e vapor de água. Das moléculas resultantes fariam parte os aminoácidos que sob aquecimento prolongado, se combinariam entre si para formar proteínas. Algumas destas combinações teriam a capacidade de se quebrar em duas réplicas da original. Esta característica, hoje observada na replicação das cadeias de ADN, pode ser encarada como uma forma primitiva de reprodução e de metabolismo.

Oparin não teve condições de provar esta hipótese. Contudo em 1953, Stanley Miller, na Universidade de Chicago, realizou uma experiência colocando num balão de vidro: metano, amónia, hidrogénio e vapor de água e submetendo-os a aquecimento prolongado. Um arco eléctrico de alta tensão cortava continuamente o ambiente onde estavam contidos os gases. Ao fim de certo tempo, Miller comprovou o aparecimento de moléculas de aminoácidos no interior do balão, que se acumulavam no tubo em U [13] (Figura 2.1).

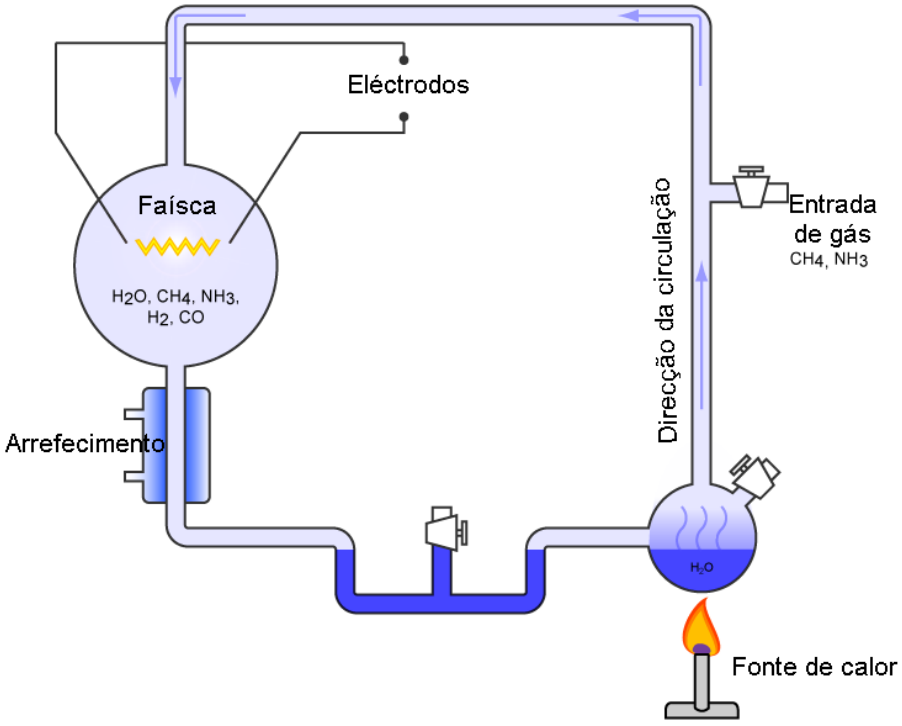


Figura 2.1 – Aparato concebido por Miller-Urey para recriar em laboratório as condições químicas da Terra primitiva e para sintetizar as moléculas primordiais da vida.

Pouco tempo depois, em 1957, Sidney Fox submeteu uma mistura de aminoácidos secos a aquecimento prolongado e demonstrou que eles reagiam entre si, formando cadeias peptídicas, com o aparecimento de moléculas proteicas pequenas.

Nos últimos 120 anos, constatou-se que não há diferença entre matéria viva e a "bruta" ou "inanimada". Os seres vivos não são compostos de algo fundamentalmente diferente de outros objectos, nem possuem um "princípio activo" que lhes dá a vida como outrora se pensava. Carbono, hidrogénio, oxigénio e nitrogénio são os elementos predominantes dos seres vivos, e também se encontram fora deles. Em última análise a vida consiste numa organização material de compostos formados por estes elementos. A abiogénese ocorre

então através de processos e etapas que cumulativamente dão origem à organização básica dos seres vivos [14].

Sendo a vida uma organização complexa de moléculas formadas por compostos básicos, os parágrafos seguintes descrevem alguns dos seus principais constituintes onde são descritas algumas noções básicas da biologia molecular celular que serão úteis para entender e associar alguns dos conceitos abordados nos capítulos seguintes.

2.1.1 Os domínios da vida

Comparações ao nível molecular mostram que a vida no planeta se divide em três grupos primários distintos chamados domínios, sendo eles *Bacteria*, *Archaea* e *Eucarya* [15] (ver Figura 2.2). As diferenças que os separam são mais profundas do que as diferenças que separam reinos típicos como os animais e as plantas. Cada um destes domínios contém dois ou mais reinos, pertencendo os reinos dos animais e das plantas ao domínio *Eucarya*. Este sistema de três domínios foi contraposto ao anterior sistema que dividia os organismos em procariótas e eucariótas. A introdução do sistema dos três domínios surgiu como consequência da descoberta de microrganismos que partilham características que eram tidas como exclusivas dos procariótas ou eucariótas. Apesar dos organismos do novo domínio (*Archaea*) serem semelhantes a outros procariótas em aspectos relativos à estrutura celular e ao metabolismo, apresentam também características típicas dos eucariótas no que respeita a determinados processos importantes da biologia molecular (transcrição e traslação genética). Em suma estes novos organismos estão mais próximos dos eucariótas do que das verdadeiras bactérias.

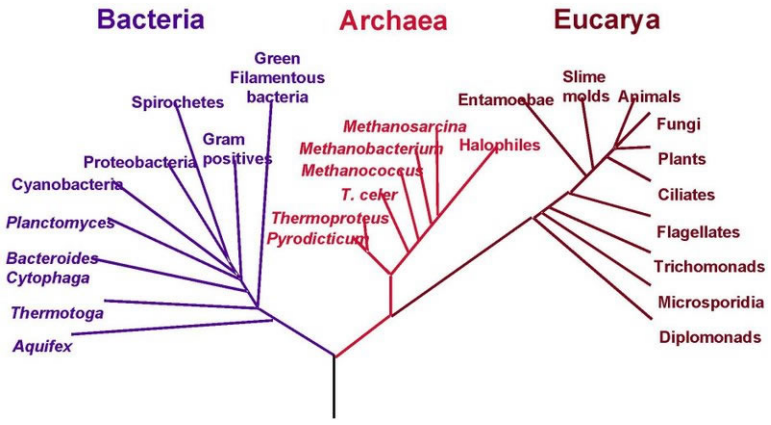


Figura 2.2 – Árvore filogenética dos organismos vivos, baseada na informação do ARN. Esta árvore ilustra a separação entre os três domínios *(eu)bacteria*, *archaea* e *eucarya*.

2.1.2 A célula

Comuns a todos os organismos de qualquer um dos domínios, as células são as unidades estruturais e funcionais dos organismos vivos, representando a menor porção de matéria viva dotada da capacidade de auto-duplicação independente. Tal como os seres humanos, as células individuais que formam todos os organismos vivos podem crescer, reproduzir-se, processar informação, responder a estímulos externos e suportar um vasto conjunto de reacções químicas. Estas competências definem a vida [16].

Em alguns organismos tais como as bactérias e os protozoários as células são os próprios indivíduos ou seja, cada célula é independente do ponto de vista funcional, suportando todas as actividades necessárias à vida. Estes são conhecidos como organismos unicelulares. Nos organismos mais complexos, os multicelulares, as actividades principais que garantem a vida são delegadas a grupos especializados de células. Contudo, cada célula de um organismo multicelular está ainda apta a executar tarefas independentes.

As células de todos os organismos dividem-se em dois grandes grupos de acordo com a quantidade e organização das membranas celulares bem como com a complexidade da região nuclear: células procarióticas e células eucarióticas [17]. A Figura 2.3 ilustra as principais características.

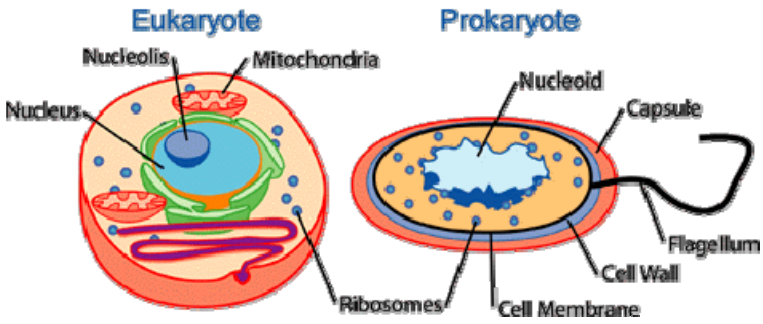


Figura 2.3 – Comparação das células eucariótas e procariótas representando respectivamente uma célula humana e uma bactéria típicas. A ilustração da esquerda realça as estruturas internas das células eucariótas, incluindo o núcleo, o nucléolo e as mitocôndrias [18].

As células procariótas são tipicamente mais pequenas e possuem uma estrutura mais simples que as eucariótas; por exemplo, não possuem membranas intra celulares e estão geralmente associadas a organismos unicelulares. Esta definição engloba todos os organismos dos domínios *bacteria* e *archaea*. São sempre organismos unicelulares, reproduzindo-se assexualmente por fissão binária (consultar glossário em anexo). Outras

formas de recombinação de ADN entre procariótas incluem a transformação e transdução. Estas últimas ocorrem entre organismos de diferentes géneros ou seja, um organismo de um género cede características próprias por meio de ADN a um outro organismo de outro género. Um exemplo deste processo é a aquisição da resistência a antibióticos através da transferência de plasmídeos contendo genes que conferem essa mesma resistência (Figura 2.4).

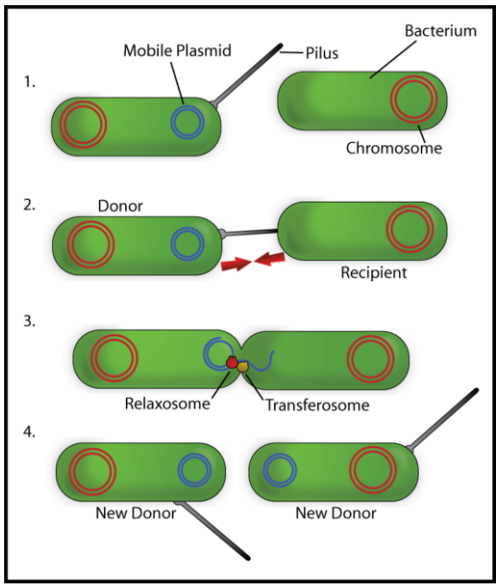


Figura 2.4 – Transferência de plasmídeos entre duas bactérias [19].

Constituindo todos os membros dos reinos vegetal e animal, incluindo também os fungos e protozoários, as células eucarióticas são bastante mais complexas que as primeiras, possuindo um núcleo que está separado do resto da célula por uma membrana. A região entre a membrana e o núcleo, o citoplasma, comporta uma miríade de estruturas em suspensão com funções especializadas, denominadas por organelos (ex.: mitocôndria, retículo endoplasmático, complexo de Golgi, vesículas, etc.) (Figura 2.5). Estima-se que o número de células do corpo humano é cerca de $6x10^{13}$, divididas em cerca de 320 tipos diferentes (células da pele, músculos, células cerebrais, etc.). É altamente provável que os eucariótas tenham surgido por um processo de aperfeiçoamento contínuo das células procarióticas.

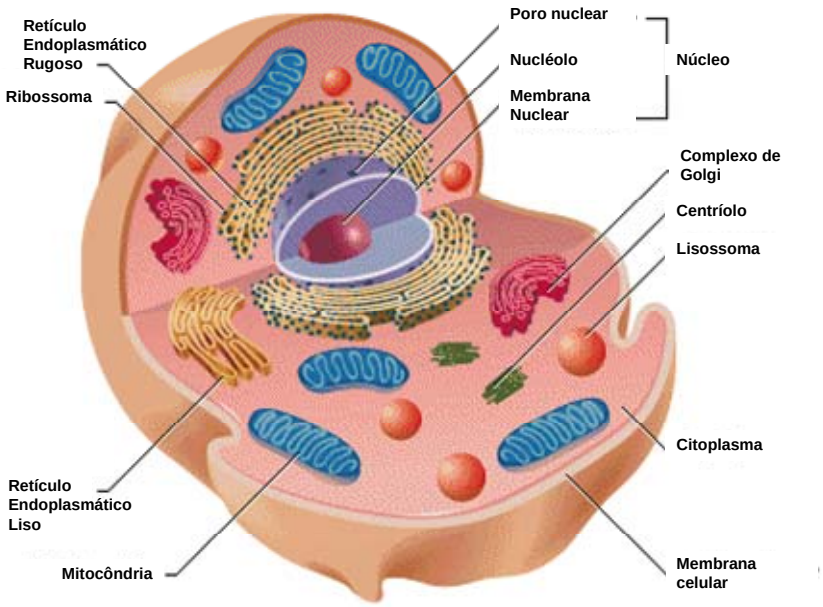


Figura 2.5 – Modelo de uma célula eucariótica animal com os principais organelos. Este tipo de células contém um núcleo confinado por uma membrana. A região entre a membrana do núcleo e a membrana exterior chama-se citoplasma, no qual se encontram dispersos os organelos (estruturas com funções especializadas) [20].

2.1.3 A estrutura química das células

A maioria das actividades celulares deve-se fundamentalmente ao envolvimento de quatro tipos básicos de moléculas nos organismos:

- Pequenas moléculas;
- Proteínas;
- ADN;
- ARN.

As pequenas moléculas são utilizadas na construção de macromoléculas ou então têm papéis específicos como a transmissão de sinais, servir de fonte de energia ou material estrutural da célula. Para além das moléculas de água, pertencem a este tipo de moléculas, os açúcares, os ácidos gordos, aminoácidos e nucleótidos. Por exemplo, as membranas celulares são construídas com ácidos gordos, formando uma estrutura que comporta vários tipos de macro moléculas.

Uma das pequenas moléculas mais conhecidas é a adenosina trifosfato (ATP) cujo propósito consiste no armazenamento de energia nas suas ligações químicas. Quando a

célula quebra essas ligações a energia libertada é utilizada para suportar processos tais como a contracção muscular ou a síntese de proteínas.

Dentro deste tipo de moléculas existem também os aminoácidos, blocos elementares importantes a partir dos quais são construídas as proteínas (Figura 2.6). Existem 20 aminoácidos que diferem entre si na sua estrutura química. Todos eles apresentam a fórmula geral $R-CH(NH_2)-COOH$, sendo R um radical orgânico que varia em todos os aminoácidos. Quando combinados em cadeias lineares, devido às suas propriedades distintas e devido à ordem pela qual estão organizados, originam proteínas com estruturas tridimensionais específicas que lhes conferem funções distintas. Tipicamente as proteínas têm um comprimento entre 100 e 1000 aminoácidos.

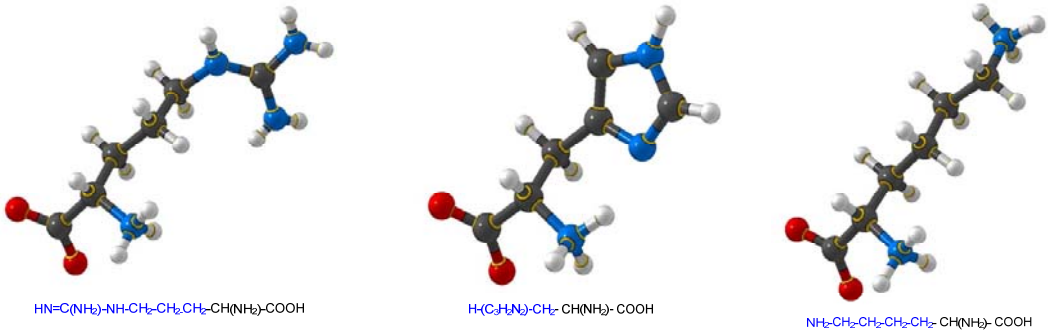


Figura 2.6 – Exemplos de moléculas de três aminoácidos, Arginina, Histidina e Lisina respectivamente [21].

As proteínas (Figura 2.7) são outro tipo importante de moléculas nos organismos. A grande diversidade de estruturas permite que estas desempenhem numerosas funções na célula. Algumas servem como componentes estruturais da célula formando por exemplo a estrutura óssea. Outras agem como sensores que mudam de forma quando a temperatura, concentrações de iões ou outras propriedades da célula sofrem alterações. Outras proteínas podem também importar ou exportar substâncias através da membrana celular. Outras são enzimas que catalizam (aceleram ou provocam) reacções químicas dentro da célula; podem ligar-se a determinados genes, activando ou desactivando a sua expressão; podem agir como sinais extra celulares enviados por uma célula para comunicar com outras células ou então ser sinais intracelulares para transportar informação dentro da própria célula. Há outras ainda que também transportam moléculas entre diferentes localizações da célula.

Ao conjunto destas reacções, tendo em conta a sua sequência, dá-se o nome de metabolismo, o qual se divide em vias metabólicas. Uma vez que é a estrutura da proteína

que define a sua função, quando a sua estrutura é semelhante à de outras proteínas, diz-se que elas pertencem à mesma família de proteínas.

As proteínas são construídas na célula com base em combinações de um conjunto de vinte tipos diferentes de aminoácidos. Os aminoácidos necessários à sua construção são obtidos a partir de outras moléculas ou a partir de proteínas provenientes dos alimentos. Dos vinte aminoácidos, apenas oito não são sintetizáveis pelo organismo humano pelo que é necessário obtê-los a partir dos alimentos ingeridos.

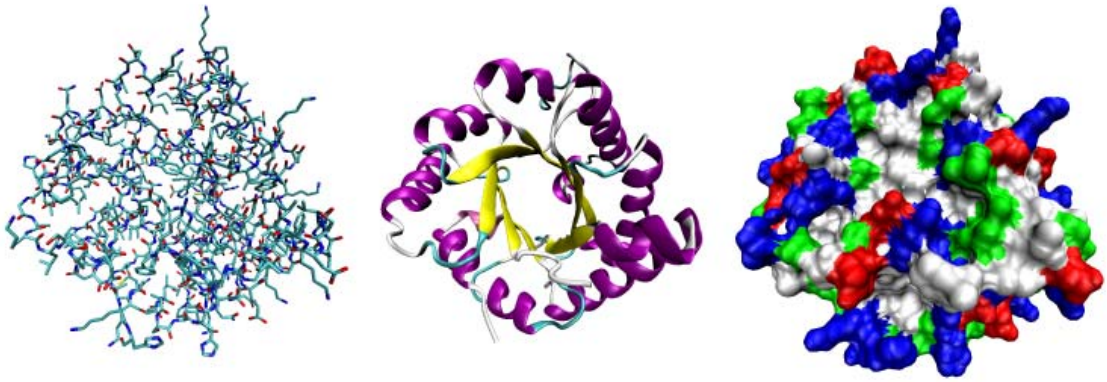


Figura 2.7 – Três representações possíveis da estrutura tridimensional da proteína triose-fosfato isomerase. Da esquerda para a direita, a primeira imagem representa a estrutura atómica da proteína; a segunda é uma ilustração do tipo *cartoon*, que representa apenas a estrutura secundária da proteína; a terceira representa a superfície da proteína [22].

2.1.4 O código genético

Apesar da grande variedade de tamanhos, formas e tipos de actividade, todas as células se dividem em duas grandes regiões internas, a região nuclear e o citoplasma, que reflectem uma divisão fundamental em termos de competências. O citoplasma suporta a maioria das funções vitais, englobando a criação de proteínas e outras estruturas moleculares de acordo com as directivas copiadas do código genético residente no núcleo. Por outro lado a região nuclear contém as moléculas de ADN (Ácido Desoxirribonucleico) organizadas em cromossomas e genes que são responsáveis pelo armazenamento de informação hereditária necessária para o crescimento, reprodução e controlo da célula.

A estrutura tridimensional do ADN consiste em duas longas cadeias de moléculas enroladas sobre um eixo comum, formando uma dupla hélice conforme a Figura 2.8. Esta

estrutura foi proposta há cerca de 50 anos atrás por James Watson e Francis Crick em Cambridge, Inglaterra [23].

Estas cadeias de ADN são compostas por monómeros (pequenas moléculas que têm a capacidade de se ligarem a outras moléculas similares) denominados por nucleótidos ou bases. Os nucleótidos estão agrupados três a três, sendo cada um destes grupos denominados de codões. Cada codão na cadeia de ADN codifica um aminoácido, a unidade constituinte das proteínas.

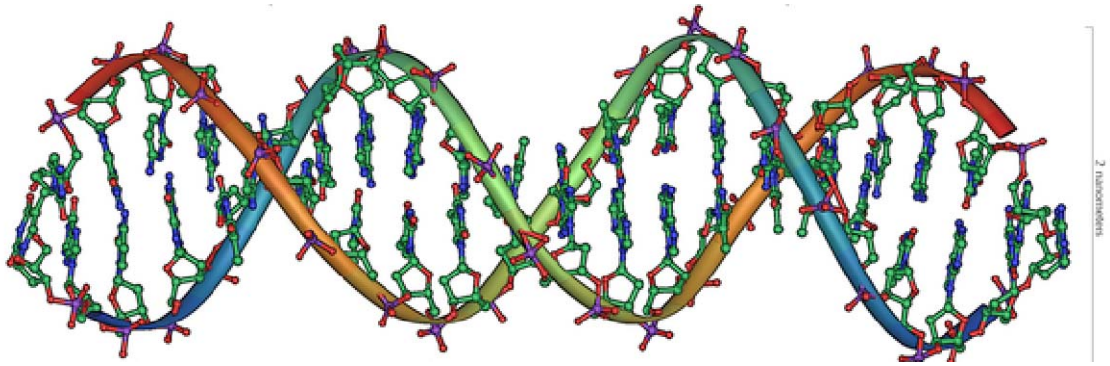


Figura 2.8 – Estrutura molecular da proteína de ADN. Consiste em duas cadeias complementares formando uma dupla hélice [24].

A cadeia de ADN é portanto formada por combinações de quatro nucleótidos diferentes, sendo eles abreviados pelas letras A, C, T, G (Adenina, Citosina, Timina e Guanina respectivamente).

As hélices da dupla cadeia de ADN têm uma construção simples: Onde existe um A numa cadeia, existe um T na mesma posição da outra cadeia e a cada C corresponde um G na outra cadeia. Esta complementaridade na correspondência das duas cadeias é tão forte que se forem separadas uma da outra, elas voltam a juntar-se quase espontaneamente.

A informação que o ADN transporta reside na sua sequência, ou seja, na ordem linear dos nucleótidos que a compõem ao longo da cadeia. Esta informação está dividida em unidades funcionais discretas, os genes, cujo comprimento está tipicamente compreendido entre 5000 e 100000 nucleótidos. Na maioria das bactérias, a quantidade de genes é da ordem dos poucos milhares enquanto que no ser humano estimam-se entre 20000 e 25000 genes, números muito abaixo daqueles que se esperavam no início da sequenciação do genoma humano [25]. Os genes que transportam informação que codifica proteínas estão geralmente divididos em duas partes: a região codificante que especifica a sequência de

aminoácidos da proteína e a região reguladora que controla quando e em que célula a respectiva proteína é produzida.

2.1.5 A síntese de proteínas

A maioria das actividades biológicas na célula é suportada por proteínas cuja informação para a sua síntese está armazenada no ADN. O encadeamento correcto dos aminoácidos pela ordem especificada nas instruções do ADN é um processo crítico na medida que dele depende a produção de proteínas funcionais e, por conseguinte, o funcionamento correcto das células e organismos.

A síntese de proteínas nas células eucarióticas ocorre no citoplasma onde três tipos de moléculas de ARN contribuem em diferentes funções:

- ARN mensageiro (ARNm);
- ARN de transferência (ARNt);
- ARN ribossomal (ARNr).

O ARN mensageiro (ARNm) transporta para fora do núcleo a informação genética transcrita da cadeia de ADN sob a forma de sequências de nucleótidos agrupados três a três, os codões, em que cada um especifica um aminoácido particular. A Figura 2.9 ilustra um pequeno exemplo de uma sequência de ARNm.

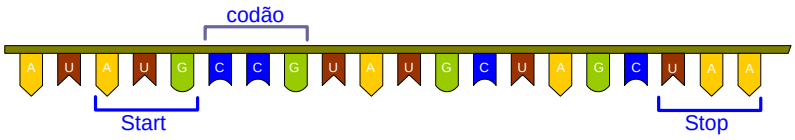


Figura 2.9 – Exemplo de uma pequena sequência de ARN mensageiro (ARNm) contendo os codões de iniciação (AUG) e de finalização (UAA neste caso).

O ARN de transferência (ARNt) é a chave da descodificação dos codões no ARNm. Trata-se de uma pequena molécula (Figura 2.10) que, durante o fabrico da proteína, transporta um aminoácido específico até ao extremo da cadeia de aminoácidos da proteína em crescimento, numa zona especial do ribossoma. Cada molécula de ARNt possui um local onde se liga um aminoácido e uma região de três bases chamada de anticodão a qual identifica o correspondente codão na cadeia de ARN mensageiro via emparelhamento complementar de bases.

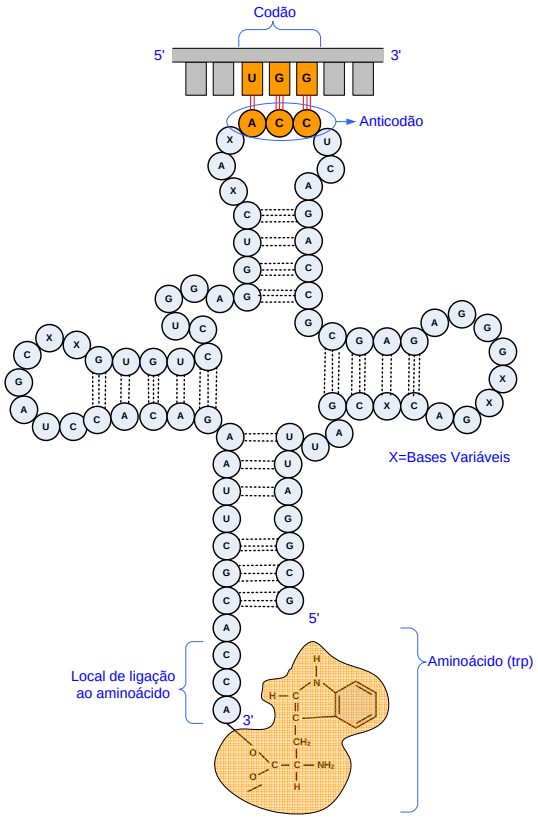


Figura 2.10 – Estrutura de uma molécula de ARN de transferência (ARNt) contendo o local de anexação do aminoácido e a zona do anticodão.

Cada tipo de molécula de ARNt pode apenas agregar um único tipo de aminoácido mas devido ao facto do código genético ser degenerado (ver glossário ou explicação mais à frente), vários tipos de ARNt possuindo diferentes anticodões podem transportar o mesmo aminoácido.

Um outro tipo de molécula é o ARN ribossomal (ARNr). Estas moléculas associam-se a um conjunto de proteínas para formar o ribossoma que consiste numa complexa máquina molecular que se pode mover fisicamente ao longo das cadeias de ARNm lendo-as, permitindo a agregação de aminoácidos numa cadeia polipeptídica dando assim origem a uma proteína. Neste processo o ribossoma permite que, para um dado codão lido, o respectivo ARNt se ligue ao ARNm e o aminoácido por ele transportado seja agregado à sequência polipeptídica em crescimento. Este processo, a síntese de proteínas, está ilustrado na Figura 2.11.

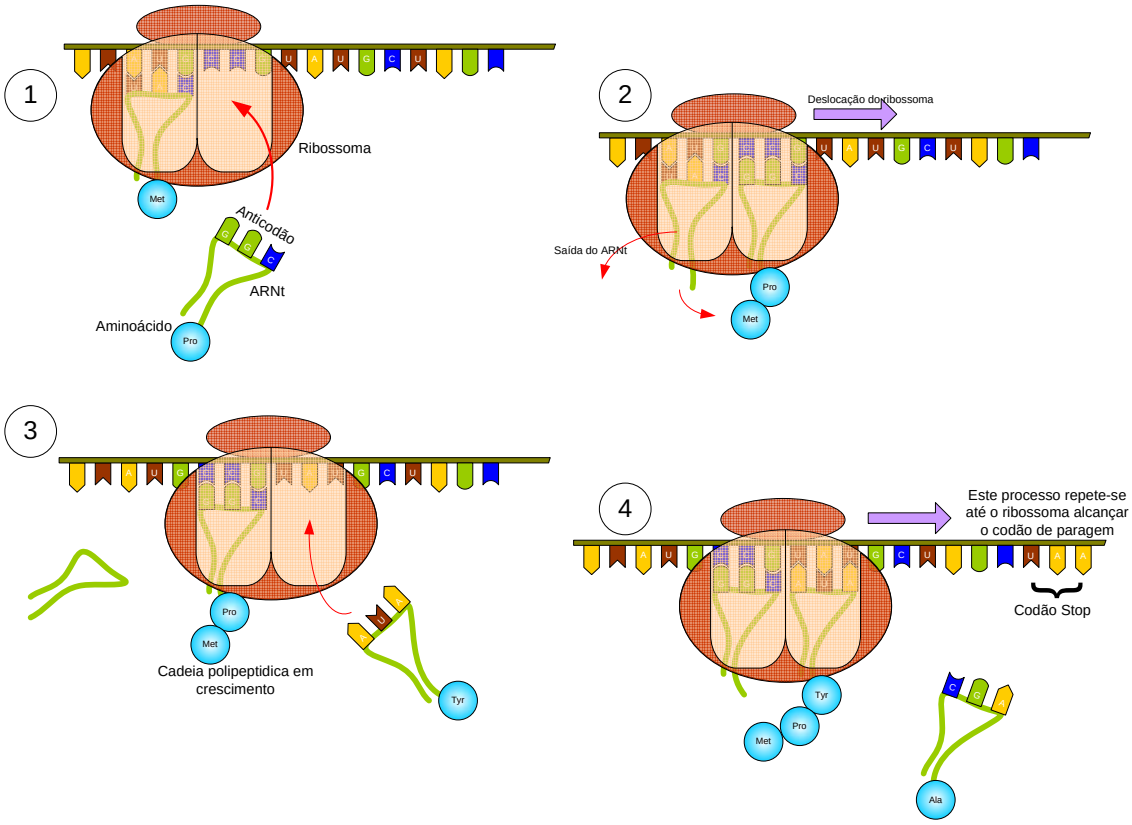


Figura 2.11 – Processo de síntese de proteínas a partir de uma cadeia de ARNm, de um ribossoma e das moléculas de ARNt. O ARN mensageiro (ARNm) é traduzido numa proteína através da acção conjunta do ARN de transferência (ARNt) e do ribossoma que é um conjunto de várias proteínas em que as duas maiores são as moléculas de ARN ribossomal. Cada codão (grupo de três nucleótidos) codifica um aminoácido específico.

Uma cadeia de ARNm tem sempre como codão inicial o AUG o que implica que a respectiva cadeia polipeptídica gerada no ribossoma (a cadeia de aminoácidos) tenha como aminoácido inicial a metionina (Met). De igual modo, a finalizar a cadeia de ARNm existe sempre um de três codões de finalização (UAA, UAG e UGA), contudo estes não codificam qualquer aminoácido.

Como existem quatro bases (A, C, T[2], G), existem 64 combinações de três bases diferentes a formar os codões onde 61 codificam aminoácidos específicos e os restantes 3 são codões de paragem (ver Tabela 2.1). No entanto, o número de aminoácidos codificados é apenas 20. Esta discrepância entre o número de codões e o de aminoácidos é devida à chamada

[2] Como o processo de tradução envolve apenas ARN, é comum apresentar neste tipo de sequências a base uracilo (U) em vez de timina (T).

degeneração do código genético ou seja, um aminoácido pode ser codificado por mais do que um codão. Contudo, como se pode observar na Tabela 2.1, o recíproco não se verifica.

Tabela 2.1 – Tabela de conversão de codões em aminoácidos. Cada codão de ARNm lido pelo ribossoma é convertido num aminoácido específico com base num ARNt. Por exemplo o codão CAA (1ª base = C, 2ª base = A, 3ª base = A) corresponde ao aminoácido Glutamina.

1ª Base	2ª Base: U	2ª Base: C	2ª Base: A	2ª Base: G	3ª Base
U	(Phe/F) Fenilalanina	(Ser/S) Serina	(Tyr/Y) Tirosina	(Cys/C) Cisteína	U
	(Phe/F) Fenilalanina	(Ser/S) Serina	(Tyr/Y) Tirosina	(Cys/C) Cisteína	C
	(Leu/L) Leucina	(Ser/S) Serina	"Ocre" (Stop)	"Opala" (Stop)	A
	(Leu/L) Leucina	(Ser/S) Serina	"Âmbar" (Stop)	(Trp/W) Triptofano	G
C	(Leu/L) Leucina	(Pro/P) Prolina	(His/H) Histidina	(Arg/R) Arginina	U
	(Leu/L) Leucina	(Pro/P) Prolina	(His/H) Histidina	(Arg/R) Arginina	C
	(Leu/L) Leucina	(Pro/P) Prolina	(Gln/Q) Glutamina	(Arg/R) Arginina	A
	(Leu/L) Leucina	(Pro/P) Prolina	(Gln/Q) Glutamina	(Arg/R) Arginina	G
A	(Ile/I) Isoleucina	(Thr/T)Treonina	(Asn/N) Asparagina	(Ser/S) Serina	U
	(Ile/I) Isoleucina	(Thr/T)Treonina	(Asn/N) Asparagina	(Ser/S) Serina	C
	(Ile/I) Isoleucina	(Thr/T)Treonina	(Lys/K) Lisina	(Arg/R) Arginina	A
	(Met/M) Metionina, Start	(Thr/T)Treonina	(Lys/K) Lisina	(Arg/R) Arginina	G
G	(Val/V) Valina	(Ala/A) Alanina	(Glu/E) Ácido glutâmico	(Gly/G) Glicina	U
	(Val/V) Valina	(Ala/A) Alanina	(Glu/E) Ácido glutâmico	(Gly/G) Glicina	C
	(Val/V) Valina	(Ala/A) Alanina	(Glu/E) Ácido glutâmico	(Gly/G) Glicina	A
	(Val/V) Valina	(Ala/A) Alanina	(Glu/E) Ácido glutâmico	(Gly/G) Glicina	G

As células recorrem a dois processos para converter a informação codificada no ADN em proteínas. No primeiro, denominado de transcrição, é criada uma versão do gene original sob a forma de uma cadeia única denominada de ácido ribonucleico (ARN). Nas células eucariótas, esta cadeia é processada e transformada em ARNm (ARN mensageiro) e transportada para fora do núcleo celular, para o citoplasma. Aí, os ribossomas executam o segundo processo chamado de tradução, descrito nos parágrafos anteriores (ver Figura 2.11), no qual montam uma cadeia de aminoácidos segundo uma ordem precisa que é ditada pela sequência de nucleótidos da cadeia de ARNm formando assim uma proteína. O processo completo da síntese de proteínas desde a transcrição do ADN em ARNm até à formação da cadeia de aminoácidos está ilustrado de uma forma resumida na Figura 2.12.

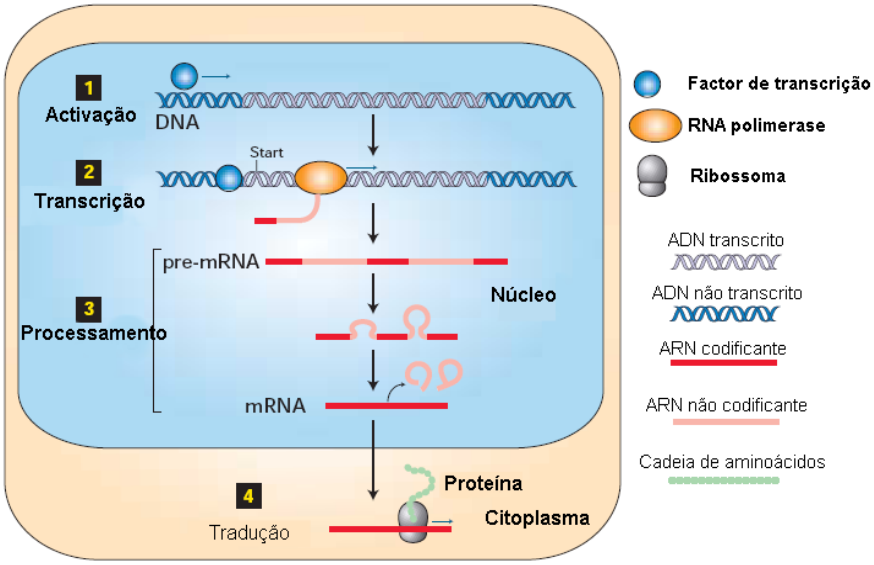


Figura 2.12 – O processo de conversão da informação codificada no ADN em sequências de aminoácidos – as proteínas. 1: Os factores de transcrição ligam-se à região reguladora do gene para o activar. Uma proteína especial, a ARN polimerase inicia a transcrição do gene activado a partir de um local específico – codão de iniciação. 2: A polimerase varre a cadeia de ADN produzindo uma versão do gene numa cadeia única de ARN. 3: A esta nova cadeia de nucleótidos são retiradas as subsequências não codificantes originando assim o ARN mensageiro (ARNm). 4: Nas células eucariótas o ARNm é deslocado para fora do núcleo, para o citoplasma onde se ligam os ribossomas que lêem a sua sequência, produzindo uma proteína ao ligar quimicamente aminoácidos numa cadeia linear [16].

Todos os organismos possuem métodos para controlar onde e quando determinados genes são transcritos. Quase todas as células do corpo humano possuem o mesmo conjunto de genes mas em cada tipo de célula só alguns destes estão activos e logo utilizados para criar proteínas. Por esta razão, por exemplo, as células do fígado produzem proteínas diferentes das células do rim, apesar de ambos os órgãos possuírem o mesmo código genético. Além disso, as células podem responder a sinais externos ou a alterações a condições exteriores com a activação ou desactivação de determinados genes, produzindo novas proteínas cuja funcionalidade vai colmatar as novas necessidades. Este tipo de controlo da actividade dos genes é feito através de proteínas especiais que se ligam ao ADN e se comportam como interruptores que, mediante sinais internos e externos, ora activam ora reprimem a transcrição de determinados genes. Estas proteínas chamam-se factores de transcrição. O processo de activação está ilustrado na Figura 2.12.

No núcleo das células eucarióticas está localizada a maior parte do ADN, enrolado em estruturas conhecidas por cromossomas. Cada cromossoma contém uma só molécula de ADN (Figura 2.13). O genoma de um organismo comporta todo o seu ADN. À excepção do

ovo e do espermatozóide, qualquer célula humana normal contém 46 cromossomas. Metade destes provêm da mãe e a outra metade do pai [16, 26].

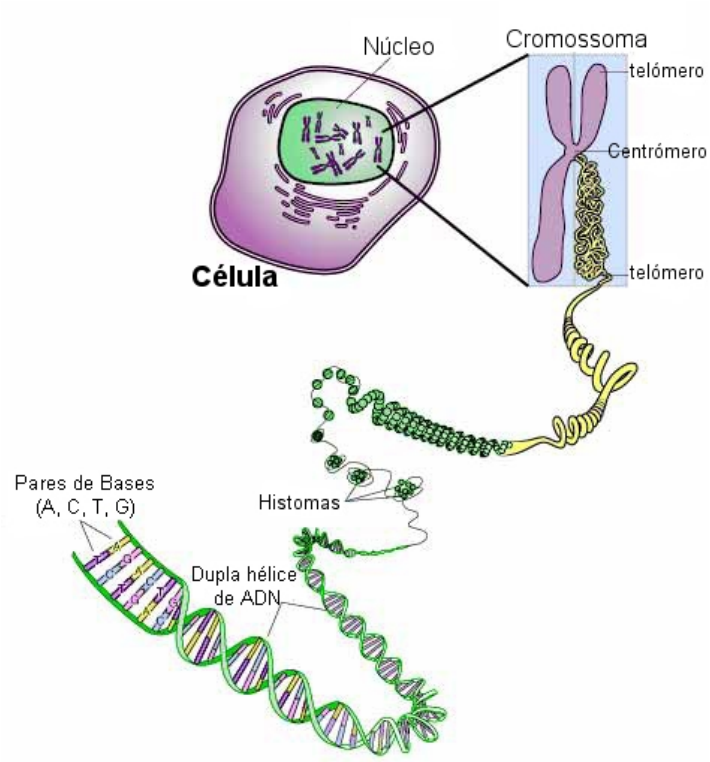


Figura 2.13 – Ilustração da organização dos genes e cromossomas dentro do núcleo de uma célula [27].

Sempre que uma célula se divide, uma máquina molecular replicadora constituída por proteínas complexas separa em cada cromossoma a dupla hélice do ADN em duas semi-cadeias, utilizando-as como modelo para agregar nucleótidos em novas cadeias complementares. O resultado final é um par de cadeias de ADN idênticas à original (Figura 2.14). A estrutura molecular do ADN e as propriedades notáveis destas máquinas moleculares garantem a execução da replicação do ADN em todos os cromossomas de uma forma rápida e fiável. Como exemplo disso, o genoma da mosca da fruta, que contém cerca de 1.2×10^8 nucleótidos, pode ser copiado em apenas três minutos.

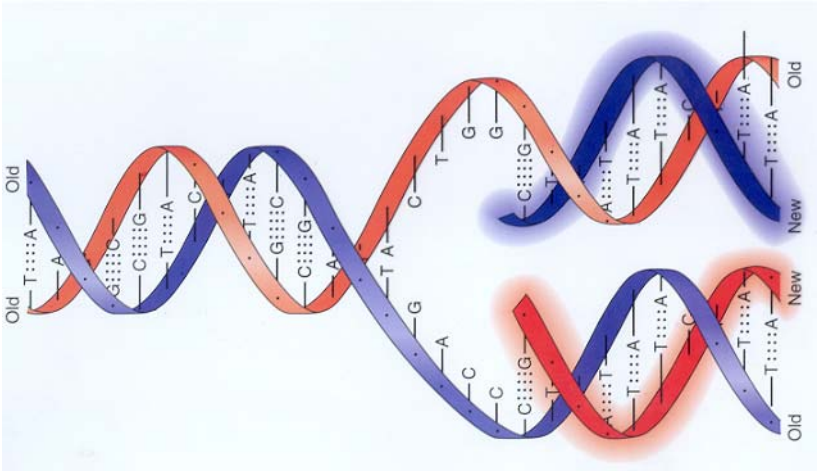


Figura 2.14 – Ilustração estilizada da replicação do ADN. A partir de uma cadeia velha são criadas duas novas [28].

Contudo, apesar da elevada fiabilidade no processo de replicação, ocasionalmente ocorrem erros nesta operação, daí resultando alterações na sequência de nucleótidos das cadeias filhas. Tais alterações, denominadas de mutações, podem ser induzidas por factores externos como as radiações ou por exposição a produtos químicos.

As mutações genéticas podem ocorrer devido à troca de um nucleótido por outro, pela remoção, inserção ou inversão de um nucleótido em milhões no cromossoma. A partir daqui, os genes com mutações podem codificar proteínas alteradas ou podem até deixar de ser devidamente controlados causando assim diversas doenças.

Um exemplo deste tipo de doenças é a anemia de células falciformes cuja causa se atribui à substituição de um simples nucleótido no gene da hemoglobina, gene esse responsável pela codificação da proteína que transporta o oxigénio nos glóbulos vermelhos. A consequente alteração de um aminoácido na proteína resultante da codificação do gene em causa reduz a capacidade de transporte de oxigénio dos referidos glóbulos. Este exemplo mostra como o correcto entendimento do funcionamento das células e os avanços na detecção deste tipo de mutações causadores de doenças pode oferecer grandes possibilidades na redução dos seus efeitos devastadores.

A descoberta da estrutura do ADN em 1953 por J. D. Watson e por F. H. C. Crick desencadeou a revolução molecular na biologia. Esta descoberta não só revelou a estrutura do ADN como também contribuiu para saber como é que a informação genética é codificada, utilizada, replicada e transmitida durante a divisão celular [17]. Além disso, esta descoberta despoletou todo um processo de investigação que levou nos últimos anos à

descodificação do genoma humano. Neste contexto, a informação genética do cidadão alcança uma relevância crescente no estudo do seu estado de saúde e predisposição para determinadas doenças [6].

Mais do que nunca, ao longo dos últimos anos a comunidade científica tem sido confrontada com uma explosão de informação acerca dos componentes da célula, as estruturas que contém e as interacções entre elas. Mesmo assim, muitos detalhes permanecem ainda obscuros, nomeadamente no que respeita ao fluxo de informação entre as células, bem como os critérios de decisão sobre os estímulos que recebe.

2.2 A Genética e a Genómica *na Medicina*

O progresso científico no campo da genética divide-se nas seguintes fases históricas. A primeira consistiu na atribuição da base celular da hereditariedade aos cromossomas. *Walter Flemming* comunica a descoberta dos cromossomas e da mitose em 1882, contudo é *August Weismann* que os relaciona com a hereditariedade um ano mais tarde. A segunda fase foi a atribuição da base molecular da hereditariedade à dupla hélice de ADN. Na terceira fase descobriram-se os mecanismos celulares responsáveis pela leitura, descodificação e tradução da informação contida nos genes e inventaram-se tecnologias de clonagem e sequenciação de genomas [29].

Em Fevereiro de 2001 as organizações IHGSC (*International Human Genome Sequencing Consortium*) e *Celera Genomics* revelam ao mundo, pela primeira vez uma versão geral, não completa, das sequências (*Draft Sequence*) do genoma humano. Além de outros benefícios, estas sequências conduziram à sistematização do conhecimento no genoma humano que por sua vez permitiu o desenvolvimento de novas ferramentas e abordagens que acentuaram a investigação na área biomédica. O desafio de obter a sequência completa levou a que passados 3 anos, as mesmas organizações apresentassem o genoma humano praticamente sequenciado, cobrindo cerca de 99% do genoma, com uma taxa de erros não superior a 1 em cada 10^4 bases. A actual sequência do genoma humano contém 2,85 biliões de nucleótidos, interrompidos apenas por 341 lacunas. Muitas destas lacunas estão associadas a zonas especiais do genoma que requerem novos métodos de sequenciação. A actual sequência permite consideráveis melhoramentos na precisão da análise biológica do genoma humano, incluindo estudos como o número total de genes, o nascimento e a morte.

Apesar da complexidade do genoma humano, as análises resultantes desta fase da sequenciação prevêem um número total de genes situado entre 20000 e 25000, quantidade consideravelmente inferior à esperada inicialmente. As primeiras estimativas apontavam para o número de 100000 genes, valor que foi refutado pelo artigo da primeira versão da sequenciação em Fevereiro de 2001 [29], que propôs um valor entre 30000 e 40000. Esta tendência de redução do número de genes no genoma humano tem sido mal recebida por vários elementos da comunidade científica uma vez que a contagem de genes era tida como um modo eficaz de quantificar a complexidade entre organismos. O choque deste resultado deve-se ao facto da contagem dos genes no ser humano ser pouco maior do que em alguns organismos como o verme *C. Elegans* [30].

Neste início do século XXI deparamo-nos com uma explosão de novos dados acerca dos componentes da célula, das estruturas que contém e de como elas se interrelacionam. Mesmo assim, continua por desvendar uma quantidade imensa de detalhes, nomeadamente como flúi a informação através das células e como estas decidem a melhor forma de resposta a estímulos. Os avanços mais recentes neste campo têm vindo a criar grandes expectativas no que respeita à sua utilização na medicina tendo o Projecto Genoma Humano (HGP) contribuído para a compreensão a nível genético da saúde humana [29, 31].

A integração da informação genética no plano clínico vem dar origem a novas abordagens onde os diagnósticos e tratamentos serão também suportados ao nível molecular [32]. Desta nova sinergia entre a medicina e a biologia molecular surge a Medicina Molecular de onde se esperam benefícios como diagnósticos mais precisos onde se incluem testes genéticos, fármacos personalizados com um espectro de actuação mais estreito e uma acção mais específica e métodos terapêuticos que actuam directamente no código genético [6]. Torna-se cada vez mais claro que é necessário analisar simultaneamente os resultados provenientes de diferentes experiências de genómica funcional (análise de mutações e perfis de expressão genética) com os dados clínicos e patológicos [11].

Ao longo deste texto surgem por diversas vezes os termos *genética* e *genómica* e convém referir que definem áreas diferentes. Genética é a ciência que estuda os genes isolados bem como os seus efeitos no organismo. A Genómica não só estuda os genes de uma forma

isolada mas também os estuda numa perspectiva global, à escala do genoma, investigando as suas funções e interacções com outros genes [33].

2.2.1 Genótipo e fenótipo

A necessidade de ligar o domínio da medicina à genética e genómica no contexto das doenças raras advém a forte relação entre fenótipo e genótipo. Antes de mais convém introduzir estes dois termos.

O genótipo é a descrição da informação genética do indivíduo ou seja, é a classe à qual um organismo pertence e é definida como a descrição do material físico de que é constituído o seu ADN. Esta informação é transmitida ao indivíduo pelos seus pais na altura da sua concepção. O fenótipo de um organismo é a classe à qual ele pertence e define-se como a descrição das suas características físicas e comportamentais como por exemplo o seu peso, cor dos olhos, altura, actividades metabólicas, o modo de andar, etc. [34]. O fenótipo de um indivíduo é determinado não só pelo seu genótipo como também pelo meio em que vive e pelo seu estilo de vida. Daqui se constata que dois indivíduos hipotéticos com iguais genótipos se viverem em ambientes diferentes, poderão apresentar também fenótipos diferentes. A expressão seguinte ilustra a relação entre fenótipo e genótipo. A sequência completa de ADN de um organismo ou seja, o seu genótipo, não contém informação suficiente para o especificar sendo que os resultados dos processos de desenvolvimento dependem tanto do genótipo como da sequência temporal dos meios em que o organismo se encontra.

$$Fenótipo = f(Genótipo, Meio)$$

Se os mecanismos de desenvolvimento do organismo fossem de tal forma que qualquer alteração no genótipo resultasse num fenótipo diferente e cada fenótipo diferente fosse a consequência de um genótipo específico o estudo do organismo seria bastante simples. Contudo esta relação não é unívoca ou seja, a correspondência entre o genótipo e fenótipo é uma relação do tipo “muitos–para–muitos” onde um dado genótipo corresponde a muitos fenótipos e existem diferentes genótipos correspondentes a um dado fenótipo [34]. A relação genótipo–fenótipo deve ser tida em conta no caso específico das doenças genéticas raras uma vez que estas são maioritariamente consequência de deficiências genéticas do indivíduo.

2.2.2 Doenças raras

A designação de “doença rara” ou “órfã” é geralmente aplicada a patologias de qualquer etiologia que constituam risco de vida ou de invalidez crónica e cuja prevalência é muito reduzida (na definição da UE inferior a 5:10000 indivíduos) [35]. Contudo, esta definição pode variar com o tempo e depende da localização geográfica a considerar (ver Tabela 2.2). Como exemplo, durante anos que a SIDA foi considerada uma doença extremamente rara, contudo, nos nossos dias tornou-se uma doença muito frequente em certas populações. Outro exemplo é a lepra, uma doença rara na Europa mas que não o é na África central. Até agora, estão descritas mais de sete mil doenças raras e em cada semana surgem em média cinco novas doenças na literatura médica. Não existe ainda grande consenso em relação a estes números pois dependem da exactidão da definição de doenças raras. A avaliação das doenças depende do estado do conhecimento actual, da exactidão das investigações e análises clínicas e da metodologia escolhida para as classificar.

Tabela 2.2 – Critérios de prevalência das doenças raras por continente [36]

Austrália	Japão	EUA	Europa
1/10 000	4/10 000	7.5/10 000	5/10 000

Dentro das doenças raras existe o grupo das doenças genéticas raras. Apesar de a grande maioria das doenças genéticas serem raras, o recíproco não é necessariamente verdadeiro, ou seja, nem todas as doenças raras têm origem em deficiências genéticas, contudo 80% destas são de origem genética. Outras doenças raras resultam de infecções (bacterianas ou virais) e alergias ou são devidas a causas degenerativas e que proliferam. Os sintomas de algumas doenças raras podem aparecer à nascença ou na infância, como no caso da atrofia muscular espinal infantil, da neurofibromatose, da osteogénese imperfeita, das doenças do armazenamento lisossomal, da condrodisplásia e do síndroma de Rett. Muitas outras, como a doença de Huntington, a doença de Chron, a doença de Charcot-Marie-Tooth, a esclerose amiotrófica lateral, o sarcoma de Kaposi e o cancro da tiróide, só aparecem na idade adulta. As doenças raras caracterizam-se pela ampla diversidade de distúrbios e sintomas que apresentam e variam não só de doença para doença, mas também de doente para doente que sofra da mesma doença [37].

Apenas as patologias severas foram distinguidas como doenças raras. Estas doenças podem ser caracterizadas quase sempre como:

- Doenças crónicas sérias, degenerativas e que normalmente colocam a vida em risco;
- Doenças incapacitantes, em que a qualidade de vida é comprometida devido à falta de autonomia;
- Doenças em que o nível de dor e de sofrimento do indivíduo e da sua família é elevado;
- Doenças para as quais não existe uma cura efectiva, mas os sintomas podem ser tratados para melhorar a qualidade de vida e a esperança de vida.

Embora individualmente tenham uma prevalência menor que outras patologias, globalmente afectam uma percentagem significativa da população – cerca de 20-25 milhões de pessoas na Europa, constituindo um complexo problema de saúde pública. No nosso país verifica-se uma quase total inexistência de informação disponível em português, o que se reflecte negativamente no seu impacto na sociedade em geral. Considerando o número de doentes existente por patologia e a sua diversidade, lidar com este problema exige cada vez mais acções concertadas europeias envolvendo todos os elementos-chave do processo – doentes e famílias, profissionais de saúde, indústria farmacêutica, laboratórios de diagnóstico e/ou investigação. Esta necessidade foi entendida pela UE que estabeleceu o tema das doenças raras como uma prioridade na área da saúde pública e criou um programa de acção comunitário específico (Decisão N.º 1295/1999/CE) [35]. Torna-se portanto cada vez mais evidente que, para compreender os mecanismos associados a uma doença, é necessário não só compreender a sua base genética (gerindo e explorando grandes quantidades de dados) como também é necessário associar e integrar o conhecimento normalmente gerido e processado no sector médico.

A falta de tratamento para este tipo de doenças deve-se essencialmente a quatro razões fulcrais [36].

- Limitações do conhecimento médico e científico associadas às causas de grande parte das doenças raras. Foram feitos poucos estudos destas patologias severas o

que, muitas vezes, põe a vida dos doentes em risco. Por se saber tão pouco acerca da maioria das doenças raras, o diagnóstico preciso, se feito, é feito tardiamente.

- Limitações de mercado: Devido ao reduzido grau de incidência deste tipo de doenças na população, não existe interesse significativo por parte do mercado na investigação de medicamentos;
- Custos de pesquisa e produção elevados: A origem genética da maioria das doenças raras (cerca de 80%) implica que o seu possível tratamento apresente custos avultados, associados à necessidade da utilização de equipamentos e produtos de elevada tecnologia.
- Dificuldade de patentear produtos resultantes da investigação de tratamentos como compostos biotecnológicos, moléculas etc.

Todos estes factores levam a que as indústrias biotecnológicas e farmacêuticas restrinjam os seus investimentos nesta área frustrando assim a eficácia dos tratamentos.

Em condições normais de mercado, o interesse da indústria farmacêutica pelo desenvolvimento de medicamentos que tenham em vista a prevenção, diagnóstico ou terapia de doenças raras é limitado. O reduzido número de pessoas afectadas por patologia leva a que o volume de vendas não amortize os elevados custos inerentes ao processo de investigação e desenvolvimento de um novo medicamento. Neste contexto surgem os medicamentos órfãos que podem ser definidos como aqueles que não são desenvolvidos pela indústria farmacêutica por razões económicas mas que respondem a uma necessidade de saúde pública. Estes medicamentos são portanto inerentemente não lucrativos devido ao número insuficiente de possíveis utilizadores tendo em conta os custos envolvidos no seu desenvolvimento e comercialização.

Para que os doentes tenham acesso às terapias que necessitam, são necessárias políticas de incentivo à investigação, desenvolvimento, produção e comercialização de medicamentos órfãos.

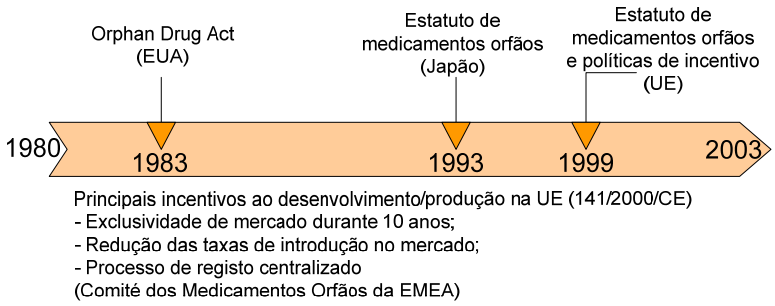


Figura 2.15 – Linha temporal das principais políticas de apoio ao desenvolvimento de medicamentos órfãos a nível mundial e principais medidas de incentivo na União Europeia [37].

O número de medicamentos órfãos europeus é ainda extremamente reduzido (20 em Janeiro de 2004) quando comparado com o dos EUA (249 em Dez. 2003) e um estudo recente efectuado pela Organização Europeia para as Doenças Raras (EURORDIS) mostra que existem iniquidades entre os estados-membros da comunidade em termos de número de produtos disponíveis, tempo de disponibilização no mercado, e preço [35].

Facilmente descuradas por médicos, investigadores e políticos, apenas as doenças raras que atraem a atenção do público beneficiam de uma política de investigação pública e/ou assistência médica. Normalmente são as associações e os grupos profissionais que fazem a consciencialização do público. O progresso feito no tratamento destas doenças permite àqueles que sofrem delas viver melhor e durante mais tempo, tendo como resultado maior sensibilização da opinião pública acerca da doença. Quase todas as pessoas com uma doença rara encontram os mesmos problemas: atraso e falha no diagnóstico, falta de informação acerca da doença, falta de referências para profissionais qualificados, falta de disponibilidade de cuidados com qualidade e de benefícios sociais, fraca coordenação dos cuidados de internamento e de consulta externa, autonomia reduzida, e dificuldade na reintegração no mundo do trabalho e ambientes social e familiar.

Dado que muitas doenças raras envolvem insuficiências sensoriais, motoras, mentais ou físicas, as pessoas afectadas por estas doenças são vulneráveis psicológica, social, cultural e economicamente. Em muitos casos as doenças raras não são diagnosticadas devido à escassez de conhecimento científico e médico ficando assim os pacientes excluídos do sistema de cuidados de saúde. Na melhor das hipóteses, alguns dos sintomas são reconhecidos e tratados. O grau de conhecimento de uma doença rara determina tanto a rapidez com que é diagnosticada como a qualidade das coberturas médica e social. A

percepção do doente da sua qualidade de vida está mais ligada à qualidade dos cuidados do que à gravidade da doença ou ao grau das deficiências associadas.

Apesar do progresso feito ao longo dos últimos dez anos, é frequente o diagnóstico de uma doença rara ser deficientemente comunicado. Muitos doentes e respectivas famílias descrevem a forma insensível e pouco informativa como o diagnóstico inicial é dado. Este problema é comum entre os médicos, que não estão organizados nem treinados em boas práticas de comunicação de diagnósticos. Após o diagnóstico, os doentes e respectivas famílias referem casos de cuidados seriamente desadequados. Não existe qualquer protocolo para a boa prática clínica para a vasta maioria das doenças raras. Nos casos em que tal protocolo existe, este conhecimento permanece isolado quando devia ser partilhado. Para além deste facto, a segmentação das especialidades médicas é uma barreira para o cuidado global de uma pessoa com uma doença rara. As famílias e os profissionais de saúde queixam-se frequentemente da dificuldade extrema em dar os passos administrativos necessários para receber benefícios sociais. Existem disparidades grandes e arbitrárias na atribuição da ajuda financeira e do reembolso de custos médicos de país para país e mesmo regionalmente dentro de alguns países. O custo dos tratamentos é muitas vezes mais elevado que o dos tratamentos das outras doenças devido à raridade da doença e ao número limitado de centros especializados sendo que uma parte significativa destas despesas é suportada pelas famílias. Para algumas doenças raras, como a febre mediterrânea familiar, a síndroma do X frágil e a fibrose quística, já existem protocolos de tratamento e programas médicos, sociais e educacionais definidos nalguns países, assim como programas de rastreio mais ou menos bem dirigidos. Estes novos métodos pré-natal e rastreio em fase assintomática para as doenças raras permitem que seja feita uma cobertura médica efectiva mais cedo, melhorando significativamente a qualidade e o tempo de vida. Outros programas de rastreio devem ser introduzidos mal existam testes fiáveis e tratamentos eficazes. O progresso qualitativo e quantitativo no prognóstico e no tratamento clínico está a levantar novas questões de saúde pública acerca das políticas de rastreio generalizado e direccionado de algumas doenças.

O conhecimento médico e científico acerca de doenças raras é escasso. O número de publicações científicas sobre doenças raras continua a aumentar, em particular aquelas que identificam novos sindromas. No entanto apenas menos de 1000 doenças, essencialmente aquelas que ocorrem mais frequentemente, beneficiam de um conhecimento mínimo. A

aquisição e a difusão do conhecimento científico são a base vital para a identificação das doenças e, ainda mais importante, para a investigação de novos procedimentos de diagnóstico e terapêuticos. Claramente é impossível desenvolver uma política de saúde pública específica para cada doença rara. Assim, em detrimento de uma abordagem fragmentada das doenças raras, uma abordagem global pode trazer soluções na medida em que permite que uma doença individual saia do anonimato e sejam estabelecidas políticas de saúde pública nas áreas de investigação científica e biomédica, investigação e desenvolvimento de medicamentos, política da indústria, informação e formação, benefícios sociais, hospitalização e consultas externas [37].

2.2.3 A problemática do acesso à informação

A especificidade dos recursos e o conhecimento necessário para os explorar restringe o seu acesso a um grupo muito limitado de utilizadores. Este problema torna-se ainda mais marcante no contexto das doenças raras onde a relação fenótipo – genótipo é bastante forte (80% das doenças raras têm origens genéticas).

Para entender as doenças genéticas raras é necessário ter em conta não só a componente genética mas também uma componente associada ao meio ambiente no qual está inserido o paciente. Como já foi dito anteriormente, além do genótipo é necessário conhecer também o fenótipo do paciente.

Para melhorar e aumentar o conhecimento dos processos associados às doenças genéticas raras, para desenvolver mais e melhores terapias para combater as doenças e para desenvolver estratégias para as prevenir, é necessário estabelecer uma sinergia num conjunto mais vasto de disciplinas sendo elas a medicina, biologia, informática, informática médica, bioinformática, bioquímica, farmacologia etc. [8]. A Figura 2.16 ilustra a interdisciplinaridade entre a genómica, a medicina e a informática, bem como as disciplina emergentes resultantes das diversas sinergias entre elas.

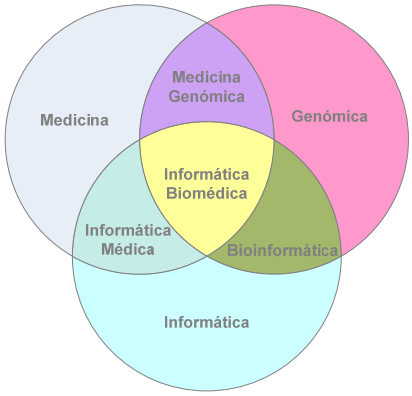


Figura 2.16 – A interdisciplinaridade entre a medicina, genómica e a informática leva ao aparecimento de novas disciplinas emergentes [8].

A crescente quantidade de informação proveniente de projectos associados ao genoma exige bases de dados que suportem a rápida assimilação de dados, modelos de dados e algoritmos computacionais eficientes para os gerir. Devido à natureza diversa da informação e ao seu crescimento disperso não existe um modelo único de dados e uma base de dados única onde se possa aceder a toda esta informação. Contudo existe um grande número de bases de dados contendo informação útil para os profissionais da saúde e investigadores disponível na Internet [32].

Além disso, esta informação tem vindo a crescer de um modo descentralizado em que cada organização científica desenvolve modelos de dados próprios respondendo apenas às suas necessidades sem qualquer preocupação no sentido de coordenar e estandardizar as suas implementações tanto a nível de dados como de conteúdos [3].

2.3 Sumário

Neste capítulo foram apresentadas de uma forma muito breve algumas noções básicas de biologia e genética, necessárias para compreender o contexto desta dissertação. Foi abordada a constituição das células desde as moléculas mais simples até à constituição do código genético.

De seguida foram introduzidas as doenças genéticas raras como motivação central deste trabalho, apresentando também a problemática do acesso a informação útil, necessária para cobrir todo o espectro de categorias que vão desde o sintoma até ao gene, e que se encontra disponível mas dispersa e fragmentada nas diversas fontes de dados na Internet. Esta

problemática serve de mote ao capítulo seguinte, onde são apresentados vários aspectos associados à integração de informação de bases de dados heterogéneas.

3 Integração de Informação Heterogénea

A integração de dados compreende o acesso uniforme e transparente a informação proveniente de bases de dados múltiplas [38]. Esta operação consiste na combinação de dados residentes em fontes diferentes, proporcionando ao utilizador uma vista unificada dessa mesma informação através de um esquema global [39]. Este esquema global (ou esquema mediador) é uma interface a partir da qual o utilizador edita e executa consultas ao sistema que por sua vez responde acedendo às fontes de dados apropriadas. O objectivo central é libertar o utilizador do conhecimento acerca da localização dos dados e como estes estão estruturados em cada uma das fontes de informação. O interesse neste tipo de sistemas tem aumentado ao longo dos últimos anos [40].

3.1 Heterogeneidade das Bases de Dados

A interoperabilidade entre sistemas heterogéneos tem sido um problema que toma dimensões mais importantes motivado pelo crescente número de sistemas computacionais e de informação e também pela necessidade de cruzamento de informação entre os vários sistemas [41]. Todos os indicadores sugerem que este crescimento se irá manter nos próximos anos. Contudo, a extracção de informação útil das bases de dados é muitas vezes complicada devido a algumas características. A primeira tem a ver com a distribuição da informação na medida que um utilizador não encontra resposta à sua consulta nos dados de uma base de dados única. A informação torna-se verdadeiramente útil quando se determina uma rede de relacionamentos entre várias bases de dados de tal modo que a informação que se encontra dispersa torna-se desfragmentada para o utilizador. Outro obstáculo importante na integração de bases de dados é a sua heterogeneidade. A heterogeneidade de bases de dados divide-se em vários tipos que variam sensivelmente de acordo com os autores.

Segundo Duschka e Genesereth [42], esta heterogeneidade pode ser notacional ou conceptual. Heterogeneidade notacional tem a ver com a linguagem ou protocolo de acesso à base de dados. Uma dada base de dados, por exemplo, utiliza SQL para executar consultas enquanto que outra utiliza OQL ou alguma notação *ad hoc*. Na realidade este tipo de heterogeneidade é de um modo geral resolvido através de diversas soluções

comerciais. Contudo, mesmo que se assuma que todas as bases de dados disponíveis utilizam linguagens e protocolos standard existe ainda o problema da heterogeneidade conceptual que tem a ver com as diferenças ao nível do esquema relacional e do vocabulário entre as bases de dados. Diferentes bases de dados podem utilizar diferentes palavras e nomenclaturas para representar o mesmo conceito ou podem utilizar a mesma palavra para representar conceitos distintos.

Outros autores, como Hull [43] apresentam a heterogeneidade tendo em conta um conjunto de problemas típicos. Alguns destes problemas devem-se a aspectos físicos dos sistemas como os componentes de *hardware*, arquitecturas, sistemas operativos, sistemas de gestão de bases de dados etc. Outros problemas devem-se a aspectos da camada lógica dos sistemas como o modelo de representação dos dados. Assim sendo, as diferentes barreiras que têm de ser tidas em consideração devido à heterogeneidade dos sistemas de informação vão desde as plataformas de hardware e de software que suportam o sistema, passando pelos esquemas e modelos de dados que providenciam a estrutura lógica de suporte aos dados até aos diferentes tipos e formatos em que os dados são armazenados.

Seth [44] identifica quatro tipos de heterogeneidade sendo eles heterogeneidade de sistema, sintáctica, semântica e estrutural. A primeira refere incompatibilidades de *hardware* e dos sistemas operativos; o nível sintáctico refere as diferentes linguagens e representações de dados; o nível estrutural inclui os diferentes modelos de dados e o nível semântico representa o significado dos termos utilizados em cada base de dados.

Wache [45] distingue a heterogeneidade em apenas dois grupos, heterogeneidade estrutural (esquemática) e heterogeneidade semântica. Heterogeneidade estrutural significa que diferentes sistemas de informação armazenam os seus dados em estruturas diferentes. Heterogeneidade semântica considera os conteúdos de cada item em termos das diferenças de significado entre fontes de dados.

A Figura 3.1 representa graficamente os limites das várias definições de heterogeneidade segundo os autores referidos anteriormente. Como se pode observar apenas a definição de Seth tem em conta as diferenças em termos de hardware e de sistema operativo.

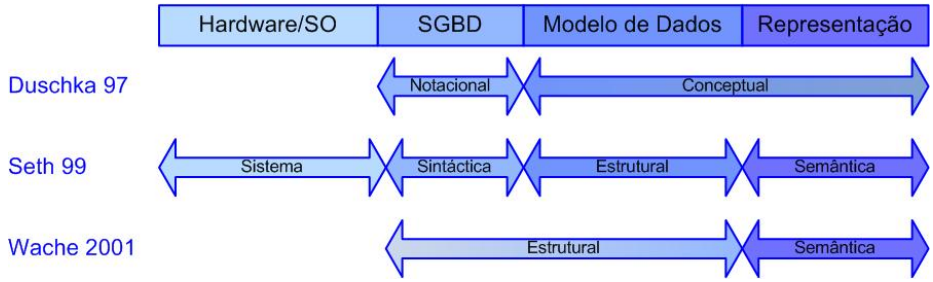


Figura 3.1 - Alcance das definições das categorias da heterogeneidade em bases de dados segundo Duschka [42], Seth [44] e Wache [45].

No contexto da biomedicina e na perspectiva do conteúdo das bases de dados, é importante salientar os diferentes tipos de integração de bases de dados heterogéneas segundo a definição de Sujansky [3].

Integração vertical: Agregação de dados semanticamente similares de múltiplas fontes heterogéneas como por exemplo um repositório virtual de acesso que providencia o acesso centralizado a dados de mamografias recolhidas e armazenadas em todas as bases de dados de um país.

Integração horizontal: Agregação de dados semanticamente complementares de múltiplas fontes heterogéneas como por exemplo um sistema centralizado de pesquisa a fontes de informação genómica, proteómica e clínica.

Integração para portabilidade de aplicações: Estandardização do acesso a fontes de dados semanticamente similares a bases de dados heterogéneas.

Em integração de bases de dados heterogéneas existem dois tipos de heterogeneidade:

- as bases de dados estão localizadas em plataformas distintas, espalhadas na Internet, com diferentes arquitecturas, sistemas operativos e SGDBs;
- as bases de dados apresentam diferentes modelos conceptuais de dados e diferentes arquitecturas de bases de dados.

No contexto da bioinformática e biomedicina, os desafios que têm de ser ultrapassados para integrar fontes de dados heterogéneas são inúmeros. A seguir são apresentadas as principais características destas fontes e os desafios que daí advêm na perspectiva da integração [46].

3.1.1 Variedade de dados

Os dados fornecidos pelas bases de dados disponíveis cobrem campos de investigação muito distintos dentro das áreas da genómica e da biologia. Tipicamente estas fontes de dados incluem informação sobre expressões e sequências genéticas, características de doenças, estruturas moleculares de proteínas, dados de *microarrays*, interacções de proteínas, etc. Dependendo da dimensão e da especificidade dos domínios de investigação e das suas fontes de dados, estas bases de dados têm de armazenar vários tipos de dados. Além disso, os dados bioinformáticos podem ser caracterizados pelos inúmeros relacionamentos entre objectos e conceitos, sendo difíceis de identificar formalmente ou porque se tratam de conceitos abstractos ou porque englobam vários tópicos de investigação. Além da grande quantidade de dados disponíveis numa fonte há também que ter em consideração a dimensão de cada registo que em muitas situações pode ser extremamente grande como é o caso das bases de dados de genomas e de estruturas 3D de proteínas.

3.1.2 Heterogeneidade na representação

Em geral, no caso das fontes de dados bioinformáticas, informação similar encontra-se em várias fontes de dados mas representada de modos diferentes dependendo da fonte de informação. De acordo com a definição de Sujansky [3], este tipo de heterogeneidade engloba as diferenças de representação entre fontes de dados as quais são estruturais, de nomenclatura, semânticas e diferenças de conteúdo. Por outras palavras, estas bases de dados não só são muito extensas como também possuem esquemas distintos e complexos (diferenças estruturais). Além disso, cada fonte de dados pode referir os mesmos conceitos com termos ou identificadores próprios o que, consequentemente origina discrepâncias semânticas entre as várias fontes de informação (diferenças de nomenclatura e semânticas). O oposto também pode acontecer onde várias fontes usam o mesmo termo ou identificador para representar diferentes objectos do ponto de vista semântico. As diferenças de conteúdo estão associadas a fontes de informação que para o mesmo objecto semântico contêm dados diferentes ou incompletos o que produz possíveis inconsistências entre as bases de dados (diferenças no conteúdo).

3.1.3 Autonomia das fontes de informação via *web*

A maioria destas fontes de dados opera autonomamente o que significa que têm liberdade para modificar esquemas de dados ou remover dados sem necessidade de qualquer notificação pública. Esta situação permite até que ocasionalmente o acesso a estas bases de dados seja bloqueado por razões de manutenção. Esta instabilidade é ainda agravada pelo facto destas fontes de dados serem *web-based* o que pode condicionar o seu acesso devido ao tráfego na rede.

3.1.4 Diferentes capacidades de consulta

Cada fonte de dados fornece ao utilizador interfaces de consulta próprias e cada uma delas exige uma aprendizagem por parte do utilizador de modo a extrair informação transversal a todas elas. Além disso, muitas destas fontes permitem o acesso e resposta a tipos de consulta muito específicos protegendo e prevenindo assim o acesso directo a todos os seus dados. Estas restrições intencionais no acesso forçam os utilizadores finais bem como os sistemas externos a adaptarem e a limitarem as suas consultas às condições fornecidas. Assim, devido a estas restrições, informação potencialmente útil para o utilizador pode não ser extraída e consultas potencialmente pertinentes podem não ser submetidas quando em muitos casos os dados necessários para a sua resposta estão efectivamente presentes na base de dados.

3.2 Requisitos para a integração de bases de dados heterogéneas

Partindo dos desafios anteriores é apresentada a seguir uma lista dos principais requisitos e pressupostos a ter em consideração no âmbito da concepção de um sistema de integração de bases de dados heterogéneas biomédicas [3, 47].

1. Apesar dos esforços e avanços no que respeita à criação de normas, a criação de um modelo único para a integração de bases de dados biomédicas é uma hipótese muito remota. Daqui se pode concluir que a heterogeneidade nas bases de dados está para durar.

2. Tanto os utilizadores como as aplicações devem ser capazes de editar consultas complexas que envolvam várias bases de dados. Consultas deste tipo combinam informação de várias fontes permitindo ao utilizador seleccionar informação à medida sem ter para isso de especificar os procedimentos técnicos para extrair essa mesma informação.
3. Aspectos como a existência de bases de dados, a sua localização física, mecanismos de acesso e estrutura dos dados devem ser transparentes para utilizadores e aplicações.
4. Não é de esperar que a maioria dos utilizadores tenha acesso com permissões de escrita sobre as bases de dados constituintes do sistema. Os conteúdos das bases de dados locais são geridos autonomamente e localmente.
5. As actualizações às bases de dados consideradas no sistema ocorrem frequentemente e os utilizadores depositam grande atenção sobre as novidades sendo o tempo de disponibilização desses dados prioritário para eles.
6. Os esquemas das bases de dados consideradas são alterados frequentemente. A justificação deste pressuposto está associada à grande complexidade da informação biológica. Este nível de complexidade exige várias iterações no desenvolvimento dos modelos conceptuais do sistema, o que pode durar vários anos.
7. Os utilizadores não devem ser forçados a restringirem à partida as suas consultas a um número limitado de bases de dados.
8. O sistema de integração deve disponibilizar ferramentas de alto nível que permitam, com relativa facilidade, alterar os esquemas de integração das bases de dados constituintes do sistema.
9. Todas as bases de dados relevantes para o sistemas de integração devem poder responder a consultas complexas via Internet.

3.3 Diferentes abordagens na integração de bases de dados

A heterogeneidade das fontes de dados, incluindo as implementações não convencionais bastante frequentes, torna o acesso à informação bastante difícil, nomeadamente nas áreas genómica e biomédica. Neste cenário, os investigadores são confrontados com diversos problemas para aceder à informação entre bases de dados heterogéneas de uma forma integrada.

As estratégias de integração actuais podem ser, numa primeira análise, classificadas em termos de modelo de dados utilizado – texto, dados estruturados ou registos ligados [46]. Para os sistemas cuja informação está organizada em formato de texto, a sua integração envolve procedimentos de pesquisa sobre texto baseada em palavras-chave. Quanto às fontes baseadas num modelo de dados estruturados, existem dois tipos de integração. O primeiro baseia-se na implementação de um repositório global (*warehouse*) no qual a informação das bases de dados locais é processada e armazenada numa base de dados central para posterior consulta. O segundo baseia-se na consulta directa às bases de dados locais através de um sistema mediador que depois de obter a informação a combina e devolve ao utilizador. Por último, para sistemas cuja informação é acedida através de *links*, a sua integração envolve processos de suporte à navegação e pesquisa desses *links* através das várias fontes de dados.

Nesta secção serão apresentadas diferentes abordagens para a implementação de sistemas de integração de bases de dados heterogéneas. Da análise de cinco autores compreende-se que, apesar das múltiplas definições e abordagens em relação às estratégias de integração, existem aspectos comuns entre elas.

A Tabela 3.1 mostra um agrupamento de todas as abordagens referidas nos parágrafos seguintes. Apesar de se observarem algumas diferenças entre as várias definições, pode-se constatar na sua análise que elas se agrupam de acordo com a tabela. Assim sendo, para facilitar esta leitura, as abordagens referidas passam a ser divididas nas categorias Integração por *Links*, Tradução de Consultas, Tradução de Dados e Bases de Dados Federadas.

Tabela 3.1 – Divisão das diferentes abordagens de integração de informação definidas por *Sujansky* [3], *Hammer* [48], *Davidson* [49] e [50], *Hernandez* [46], *Ouzzani* [51] e *Karp* [47]. Estas estratégias estão agrupadas em quatro categorias de acordo com as principais características.

Artigo	Integração por *links*	Tradução de consultas	Tradução de dados	Bds Federadas
Sujansky 2001		*Query Translation*	*Data Translation*	
Hammer 2003		*Query Driven Approach*	*Warehousing*	
Davidson 2001	*Linked Driven Federations*	*View Integration*	*Warehouse*	
Hernandez 2004	*Navigational Integration*	*Mediator Based Integration*	*Warehouse Integration*	
Ouzzani 2004	*Multidatabases Language Approach*	*Global Schema Integration*		*Federated Databases*
Karp 96	*Hypertext Navigation*	*UnMediated MultDB Queries*	*Data Warehouse*	*Federated Databases*

Sujansky [3] e *Hammer* e *Schneider* [48] dividem os sistemas de integração de bases de dados heterogéneas em duas categorias semelhantes. Enquanto que o primeiro distingue as implementações em Tradução de Dados (*Data Translation*) e Tradução de Consultas (*Query Translation*), os segundos apresentam as mesmas definições contudo com nomes diferentes: *Warehousing* e *Query Driven Integration* respectivamente. Em ambas as abordagens, os utilizadores acedem à informação das bases de dados locais através de um esquema global integrado.

Outros autores como *Davidson et al* [50] introduzem uma nova categoria às anteriores considerando como estratégia de integração os sistemas que referenciam bases de dados por intermédio de *links* nas páginas *web*. Assim, além da Integração por Vistas (*View Integration*) e *Warehouse*, é introduzida a abordagem que é denominada de *Linked Driven Federations*. Igualmente, *Hernandez* [46] define três tipos de abordagens *Mediator-based Integration*, *Warehouse Integration* e *Navigational Integration* (*Link-based Integration*) semelhantes às anteriores.

Outros autores como *Ouzzani* [51] e *Karp* [47] referem uma quarta abordagem mantendo as definições anteriores. Com efeito, ambos referem as bases de dados federadas como sendo uma colecção de fontes de dados autónomas mas cooperantes.

Ouzzani [38, 51] classifica as estratégias de integração de bases de dados heterogéneas em três categorias (*Multidatabase Language Approach*, *Global Schema Integration* e *Federated Databases*). Por último, *Karp* [47] inclui as quatro abordagens referidas anteriormente, denominando-as de *Hypertext Navigation*, *Unmediated MultiDB Queries*, *Data Warehouse* e *Federated Databases*, não existindo diferenças significativas relativamente às classificações anteriores.

3.3.1 Tradução de dados

A tradução de dados [3] é uma estratégia que envolve a transformação de dados nos vários formatos nativos das fontes locais de tal forma que são extraídos e armazenados num formato partilhado ou repositório comum com base no qual podem ser uniformemente acedidos. Este formato partilhado implementa todos os elementos do modelo de dados de modo a que tanto os utilizadores como as aplicações possam ignorar as especificações das bases de dados locais. Esta é uma definição semelhante à *Data Warehousing* de *Hammer* e *Schneider* [48] que introduz um repositório entre a aplicação de integração e as bases de dados locais o qual serve para armazenar vistas integradas dos dados locais. A Figura 3.2 ilustra este tipo de integração de dados.

Numa definição semelhante, *Davidson* [50] refere a estratégia *Warehouse*, a qual se baseia num esquema global único através de um modelo comum onde os esquemas das bases de dados constituintes do sistema são incorporados. Assim, os utilizadores consultam este esquema global utilizando uma linguagem de alto nível como por exemplo SLQ ou OQL. Uma vez instanciado, a consulta global é executada utilizando informação do repositório em vez de se delegarem as consultas locais às respectivas bases de dados.

Esta estratégia aumenta o desempenho do sistema de integração mas exige uma gestão dos procedimentos de actualização do repositório bastante complexa. De facto, este tipo de sistemas exige que os dados descarregados das fontes locais sejam convertidos num formato único através de um esquema de mapeamento antes de serem armazenados no repositório [46] [47].

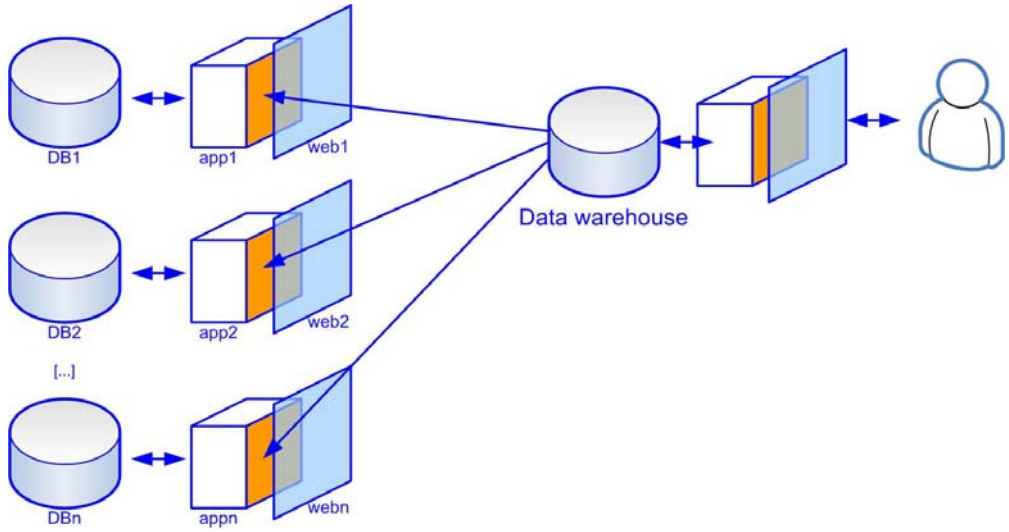


Figura 3.2 – Integração de informação por tradução de dados. Os dados das fontes locais são extraídos, processados e armazenados num repositório comum e partilhado no qual todas as consultas solicitadas pelo utilizador são executadas.

Existem dois modos de gestão de integração por tradução de dados: Tradução de dados manual e automática. A tradução manual é frequentemente usada na agregação de dados clínicos em bases de dados epidemiológicas em que os registos médicos são manualmente submetidos de acordo com normas e esquemas standard. Embora funcional, este método é muito dispendioso em termos de custos e de tempo o que inibe a criação deste tipo de sistemas. Mais comum que o caso anterior, a tradução de dados automática transforma mecanicamente os dados locais num formato comum. Comparativamente ao anterior, este método reduz os custos e o tempo necessário para traduzir os dados a integrar contudo não resolve dois problemas. Primeiro, a tradução de dados implica que a informação esteja armazenada em duplicado ou seja, os dados que residiam nas fontes locais passam também a fazer parte do repositório partilhado o que exige uma gestão cautelosa da integridade dos dados. Segundo, a correspondência entre o formato local e o formato partilhado está presente apenas nos algoritmos dos programas de tradução o que dificulta a inspecção e validação do modelo de integração do sistema.

3.3.2 Tradução de consultas

A estratégia alternativa à Tradução de Dados é a Tradução de Consultas [3] onde, o modelo de consultas é virtual, uma vez que a informação permanece armazenada somente nas bases de dados intervenientes. Assim sendo, uma consulta que é submetida ao modelo virtual é decomposta e traduzida num conjunto de consultas às bases de dados locais de

acordo com a linguagem e modelos de dados próprios. Estas consultas são executadas directamente sobre as fontes locais e os resultados são transmitidos, transformados e combinados para serem apresentados ao utilizador ou aplicação que solicitou a consulta. Contudo a implementação deste modelo implica uma perda adicional de desempenho devido à transformação e execução de consultas remotas. Outras desvantagens são a grande dificuldade de implementação e a dificuldade de acesso à informação em bases de dados que carecem da disponibilização *ad hoc* de interfaces de consulta.

Davidson [50] define uma estratégia semelhante a que dá o nome de Integração por Vistas (*View Integration*) na qual, tal como a estratégia *warehouse*, os esquemas das bases de dados constituintes do sistema são incorporados num esquema global único através de um modelo comum. Contudo, ao invés de utilizar um repositório adicional, o sistema determina sub conjuntos da consulta global aos quais pode responder, delega as consultas locais para as bases de dados apropriadas e combina as respostas retornadas pelas fontes de dados produzindo assim uma resposta à consulta global. A operação de combinação de informação pode ser expressa numa única consulta de alto nível, sendo delegada para o sistema a tarefa de exploração das bases de dados locais. Em geral as linguagens de consulta neste tipo de abordagens são bastante poderosas permitindo até a reestruturação arbitrária dos dados retornados pelas consultas.

Numa outra definição similar, *Hernandez* [46] introduz o conceito de mediador o qual, no contexto da integração de informação, é um subsistema responsável pela reformulação em tempo real de uma consulta colocada pelo utilizador de acordo com um esquema de dados único numa consulta segundo o esquema da base de dados local. Deste modo, recorre-se a um mapeamento das relações entre as fontes de dados e o mediador que permite que as consultas colocadas na camada do mediador sejam convertidas em consultas dedicadas às bases de dados locais. A Figura 3.3 ilustra a abordagem de integração de informação por tradução de consultas.

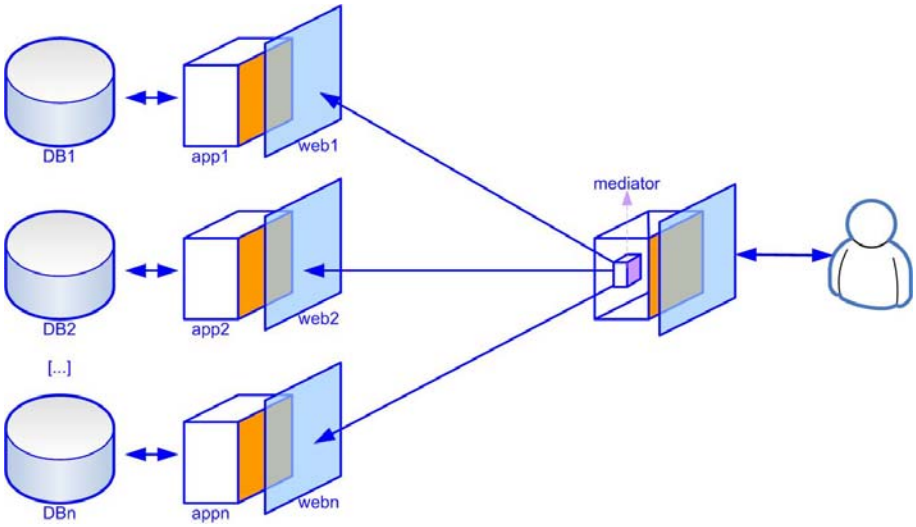


Figura 3.3 – Integração de informação por tradução de consultas. Ao contrário da tradução de dados, a informação permanece armazenada somente nas bases de dados nativas. Uma consulta submetida ao sistema de integração é decomposta num conjunto de consultas às bases de dados locais de acordo com as linguagens e modelos de dados nativos. Os resultados de cada consulta são depois combinados para serem apresentados ao utilizador.

Outras estratégias semelhantes são referidas por *Seth* e *Ouzzani* [38, 51] e *Karp* [47] apresentando-as como Integração por Esquema Global (*Global Schema Integration*) e *Unmediated MultiDb Queries* respectivamente.

3.3.3 Integração por links

Um grande número de fontes disponíveis na Internet permite apenas que os utilizadores pesquisem a informação manualmente através de várias páginas *web* de modo a obterem informação relevante [52].

Nesta abordagem, denominada por *Davidson* de *Linked Driven Federations* [50] (também referida como *Navigational Integration (Link-Based Integration)* por *Hernandez* [46]), os utilizadores começam por extrair informação de interesse a partir de uma dada base de dados explorando depois outras bases de dados associadas via *web links* disponibilizados propositadamente pelas equipas de desenvolvimento.

Esta abordagem é útil para os utilizadores com poucos conhecimentos em programação e formulação de consultas uma vez que a interface deste tipo de sistemas se limita a fornecer *links*. O motivo principal deste tipo de abordagem reside na dificuldade de, em determinadas fontes de informação, se aceder à informação com outro método que não a típica navegação por *browser*. O caminho seguido pelo utilizador através das diferentes

páginas constitui um *workflow* onde a saída de uma fonte de dados é redireccionada para a entrada da próxima fonte até que a informação desejada seja alcançada [53].

Nesta perspectiva, as consultas são transformadas em conjuntos de caminhos através das várias fontes em que, à medida que se avança, se alcançam maiores níveis de detalhe. Este modelo de integração afasta-se dos modelos relacionais de dados e aplica um modelo onde as fontes de informação são definidas como sendo conjuntos de páginas contendo interligações e pontos de entrada bem como o conteúdo em si, restrições nos percursos, parâmetros de entrada opcionais ou obrigatórios. Lacroix [54] explora a abordagem baseada em percursos analisando os caminhos entre fontes de dados biológicas, propondo-se a determinar várias propriedades das ligações entre as fontes de dados com o intuito de optimizar os resultados das consultas determinando os melhores caminhos.

A maioria das abordagens deste tipo, apesar de oferecer mecanismos de consulta muito restritivos, pode ser facilmente apreendida por utilizadores não técnicos. Contudo esta abordagem não é escalável já que, quando uma nova fonte de dados é adicionada à federação, é necessário adicionar novas entradas de ligação às antigas bases de dados, o que é uma tarefa bastante complexa. Bases de dados locais heterogéneas pré-existentes são geralmente integradas sem qualquer adaptação prévia. Assim, a informação armazenada nas diferentes bases de dados pode muitas vezes ser redundante, heterogénea e inconsistente [38, 51]. Além disso é habitual um utilizador estar interessado em combinar informação de duas fontes de dados federadas. Para tal tem de percorrer uma série de *links* até à segunda fonte de dados.

A estratégia equivalente sugerida por Karp [47] denomina-se de Navegação por Hiper texto (*Hypertext Navigation*), referindo que permite que o utilizador navegue interactivamente de uma dada base de dados para outra através de *links* em páginas de Internet. Geralmente os sistemas desta categoria suportam apenas dois tipos de operações: a pesquisa de uma dada entrada dentro de uma fonte de dados e o pedido de entradas ligadas a outras bases de dados.

A Figura 3.4 ilustra este tipo de estratégia.

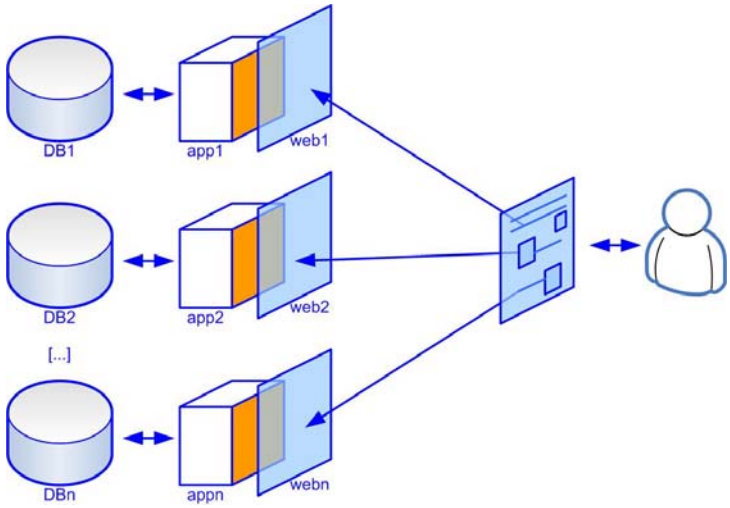


Figura 3.4 – Integração de informação por intermédio de *links*. Os utilizadores extraem informação de interesse a partir de uma dada base de dados explorando depois outras bases de dados associadas via *web links* disponibilizados localmente.

3.3.4 Bases de dados federadas

Outros autores como Ouzzani [51] e Karp [47] referem uma quarta abordagem mantendo as definições anteriores. Ambos referem como estratégia as bases de dados federadas, consistindo numa colecção de fontes de dados autónomas mas cooperantes. Estas bases de dados participam na federação de modo a permitirem uma partilha parcial e controlada dos seus dados [38]. Neste tipo de abordagem é mantida alguma autonomia nas bases de dados individuais, sendo a partilha de informação implementada através de esquemas de importação e exportação. Todas as bases de dados intervenientes estão registadas num dicionário federal. A Figura 3.5 representa graficamente a estratégia das bases de dados federadas.

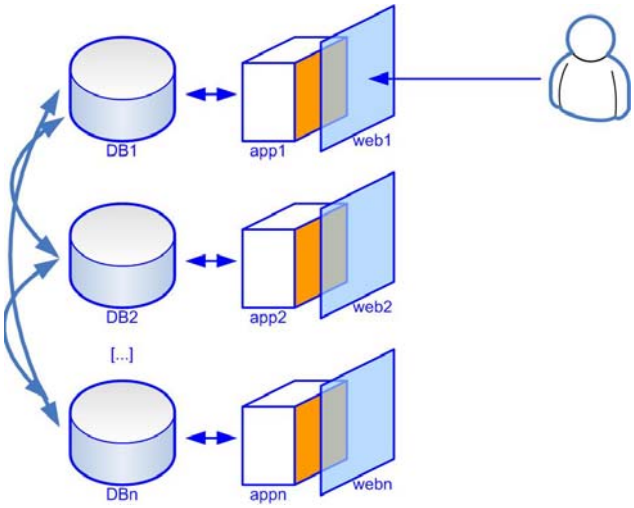


Figura 3.5 – Integração por bases de dados federadas. A partilha de informação é implementada localmente, sendo da responsabilidade de cada base de dados membro da federação.

3.4 Sumário

Neste capítulo foram apresentados os vários aspectos associados à problemática da heterogeneidade da informação bem como as principais estratégias de integração.

4 Fontes de Informação Biomédica

A especificação de um sistema de integração de bases de dados biomédicas e genéticas, contemplando a associação de conhecimento entre o fenótipo e genótipo tem latentes os seguintes desafios [31]:

- **A complexidade dos dados:** A complexidade e volume de dados gerados pelo *Human Genome Project* é de tal ordem que o seu manuseamento é considerado problemático até para a comunidade científica. Os termos de identificação dos elementos são complexos na medida em que existem muitos sinónimos para cada um, impossibilitando muitas vezes a interligação entre diferentes sistemas de informação. Como exemplo, um dado gene, pode ter nomes diferentes e, em alguns casos, o mesmo nome pode corresponder a genes diferentes.
- **A Natureza dinâmica dos dados:** Os sistemas de informação biomédica estão em constante evolução tanto a nível de conteúdo como a nível da sua estrutura e arquitectura. Isto deve-se essencialmente a esforço global que se tem vindo a sentir na pesquisa e obtenção de informação associada ao projecto *Human Genome Project*.
- **Quantidade crescente de bases de dados e de sistemas de informação:** Com mais de 800 fontes de dados disponíveis [9], torna-se muito complicado identificar e reunir a sua informação, bem como navegar entre os diferentes sistemas. Além disso, a informação de cada um está centrada em sub–áreas temáticas que pressupõem um conhecimento científico prévio de modo a interagir correctamente com as suas ferramentas.
- **A ausência ou insuficiência de normas de representação de conhecimento:** A ausência de consenso à volta das normas de representação da informação obtida no âmbito do HGP complica ou impossibilita a sua navegação e manipulação do ponto de vista computacional.

4.1 Fontes de Dados Consideradas

Tendo em conta o objectivo de reunir informação de várias áreas disciplinares à volta das doenças genéticas raras e cobrindo conceitos desde o fenótipo até ao genótipo, foi necessário encontrar uma lista de bases de dados de referência. A selecção de bases de dados biomédicas foi obtida a partir da publicação anual "*The Molecular Biology Database Collection*", publicada pela *Nucleic Acids Research* (NAR). Este documento contém uma lista das mais importantes bases de dados, classificadas por áreas de interesse no âmbito da biologia molecular, biomedicina, química, biologia computacional, genómica, etc. Na actualização de 2006 [9] estavam registadas 858 bases de dados, mais 139 do que no ano anterior [78]. A selecção de bases de dados fornecida pelo NAR salvaguarda aspectos muito importantes como a fiabilidade da informação armazenada, periodicidade das actualizações dos conteúdos, acesso gratuito etc.

Cada base de dados possui um domínio, arquitectura, interface e métodos de acesso próprios o que dificulta a obtenção de informação por parte do utilizador comum. São estes factores que caracterizam as bases de dados como sendo heterogéneas e que legitimam o propósito do trabalho apresentado nesta dissertação.

Foram seleccionadas 19 propriedades associadas às doenças genéticas raras, as quais foram organizadas em conceitos (Patologia, Fármacos, Polimorfismos, Gene, etc.). Seguidamente construiu-se um mapa de navegação que relaciona todos estes conceitos de uma forma lógica [6]. A Figura 4.1 ilustra o mapa obtido com todos os seus conceitos.

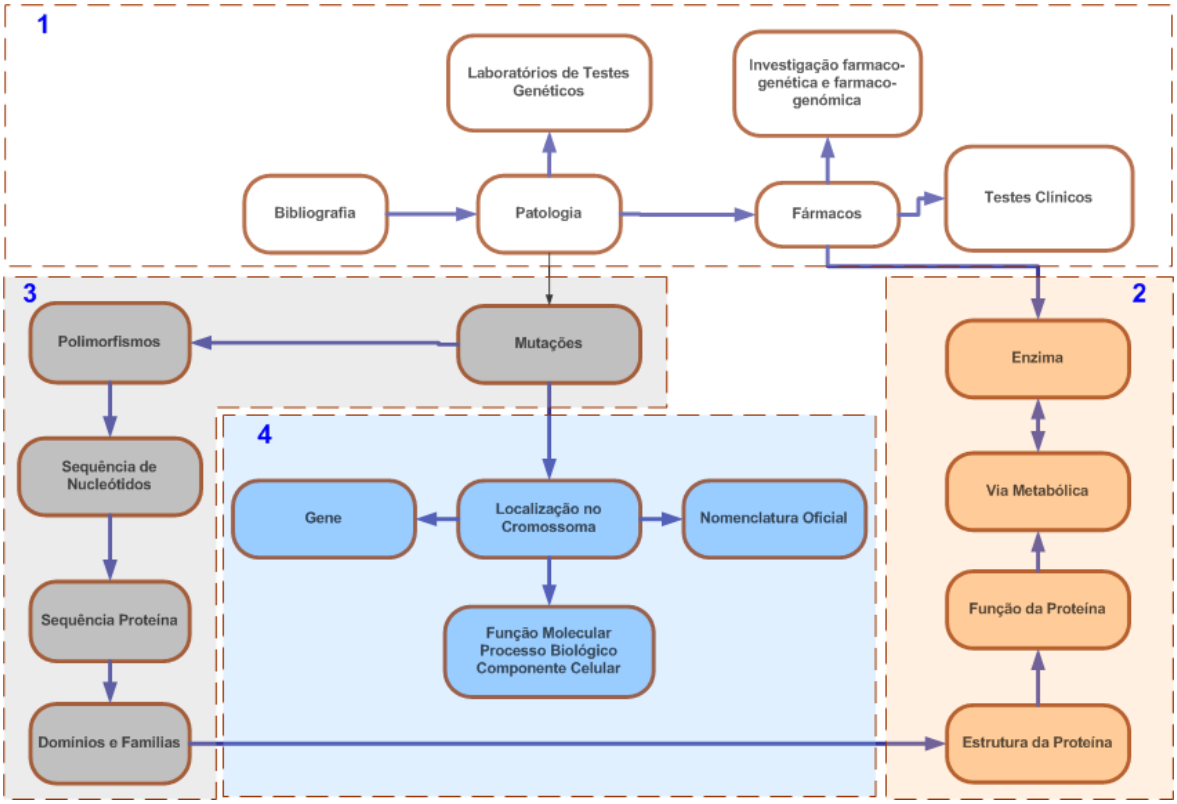


Figura 4.1 – Proposta de um mapa conceptual para caracterizar as doenças genéticas raras.

Obtido o mapa fica definido o protocolo que permite associar os conceitos do foro clínico a conceitos do foro genético e da biologia molecular. A Patologia é o ponto de entrada do protocolo. Este conceito descreve os aspectos que caracterizam a doença na perspectiva clínica. Uma doença genética rara deve-se a uma mutação (Mutações) numa combinação genética do paciente como por exemplo um Polimorfismo. Um polimorfismo deve-se à alteração de uma ou mais Sequências de Nucleótidos o que causa uma alteração na proteína correspondente (Sequência da Proteína). Cada proteína está classificada dentro de um dado domínio estrutural ou funcional (Domínios e Famílias). Como resultado da alteração da sequência, a estrutura 3D da proteína é também alterada (Estrutura da Proteína) e consequentemente a sua função (Função da Proteína). Esta proteína entra num processo bioquímico algures numa Via Metabólica participando em reacções enzimáticas (Enzima). É a este nível que os Fármacos, caso existam, actuam na doença. Associados a este tipo de fármacos estão os Testes Clínicos e informação acerca da investigação fármaco-genética e fármaco-genómica. Associado à doença existe ainda um conjunto de referências bibliográficas (Bibliografia) e existem centros onde são efectuados diagnósticos genéticos (Laboratórios de Testes Genéticos). Ao nível da investigação é

também importante a informação pormenorizada acerca do gene ou genes associados à doença como a sua Localização no Cromossoma, nome oficial (Nomenclatura Oficial), Função Molecular, Processo Biológico, Componente Celular e finalmente informação adicional sobre o gene (Gene).

Este protocolo está dividido em várias perspectivas de acordo com os diferentes perfis de utilizadores e com as suas necessidades específicas. No topo superior do protocolo (zona 1) estão os recursos que proporcionam informação pertinente ao médico de clínica geral tal como bibliografia sobre a doença, centros de diagnóstico existentes e tratamentos disponíveis. O especialista hospitalar, a quem o paciente é delegado pelo médico, irá focar a sua atenção, além dos conceitos anteriores, em informação mais detalhada (zonas 1 e 2). O outro perfil de utilização pertence ao geneticista que utiliza a informação contida nos blocos das zonas 3 e 4 (ver Figura 4.1).

O passo seguinte à identificação dos conceitos fundamentais para o estudo das doenças genéticas raras centra-se na associação desses conceitos a fontes de dados existentes e disponíveis através da Internet. Como já foi referido neste capítulo, a selecção foi efectuada a partir do documento "*The Molecular Biology Database Collection*".

A obtenção do protocolo de conceitos bem como a selecção de bases de dados para o preencher foi baseada em estudos prévios realizados por investigadores do *Instituto de Salud Carlos III* numa parceria no âmbito do projecto Infogenmed [6]. A versão actual do protocolo apresenta algumas diferenças relativamente à primeira proposta. Esta evolução deve-se não só à discussão mantida entre os parceiros do projecto com o intuito de melhorar optimizar o protocolo como também se deve à evolução dos recursos disponíveis na Internet. Com efeito, novas sugestões são feitas mediante novas ofertas de bases de dados e interfaces de consulta e também de novas ligações entre os recursos já existentes.

Nas secções seguintes são apresentadas algumas das bases de dados do conjunto seleccionado.

4.1.1 Orphanet

A ORPHANET [35] é uma base de dados relacional com informação sobre doenças raras e medicamentos órfãos, que pretende contribuir para melhorar o diagnóstico, seguimento e tratamento das doenças raras. É constituída por uma enciclopédia *online* escrita por peritos

europeus de várias especialidades médicas e proporciona uma lista de serviços adaptada às necessidades de doentes e famílias, profissionais de saúde e investigadores. Contém informação sobre consultas especializadas, laboratórios de diagnóstico, projectos de investigação, ensaios clínicos e grupos de apoio relacionados com este tipo de doenças. Está actualmente disponível em 6 línguas – Francês, Inglês, Italiano, Alemão, Espanhol e Português.

Em cada um dos países membros, a base de dados *Orphanet* é gerida por uma equipa autónoma composta por uma responsável proveniente do mundo académico, por um documentalista científico permanente e por uma comissão permanente de peritos de todas as especialidades, da endocrinologia à pediatria. A *Orphanet France* coordena a rede europeia e fornece o *software* e a formação inicial para além de ser responsável pela supervisão regular e pelo controlo de informação de modo a garantir homogeneidade e consistência da informação. Os visitantes desta base de dados dividem-se em idêntica proporção entre profissionais e particulares, a uma média de 11000 visitas diárias ao *site*. Com efeito, metade dos visitantes são profissionais de saúde, os doentes e familiares são responsáveis por um terço das visitas, sendo os restantes professores, estudantes, jornalistas, gestores de indústrias e outros interessados [35].

A versão actual da ORPHANET inclui mais de 3.600 doenças, com informação científica detalhada sobre cerca de 1.100 através de um texto descritivo, lista de sintomas (com frequências relativas), *links* para a base de dados OMIM e outros *sites*. O sistema de consulta da base de dados é simples e extremamente completo, permitindo a pesquisa de uma determinada doença por nome, nº OMIM, ICD (OMS), anomalias ou directamente a partir da listagem alfabética. Relativamente aos serviços, a pesquisa pode ser efectuada utilizando o nome (ou sigla) da entidade ou a localização.

Para além da descrição da doença, cada página resultante de uma pesquisa no *Orphanet*, fornece também ligações à base de dados OMIM, bibliografia relativa à doença (artigos científicos), sinais clínicos e sintomas, outros *sites* com informação sobre a doença, consultas, laboratórios de diagnóstico, projectos de investigação associados, redes profissionais, grupos de apoio, ensaios clínicos, registos e observatórios. A Figura 4.2 apresenta alguns detalhes de duas páginas da *Orphanet*.

Figura 4.2 – Detalhe de uma página *Orphanet* para a doença Anemia de Fanconi. Além de uma descrição textual da doença, esta página contém também um conjunto de informações adicionais. Seguindo a ligação "Ensaios Clínicos" obtém-se uma nova página que contém duas referências para institutos que realizam ensaios clínicos para esta doença.

A figura seguinte representa de um modo gráfico a informação disponível na base de dados *orphanet*, bem como as categorias de acesso às consultas.

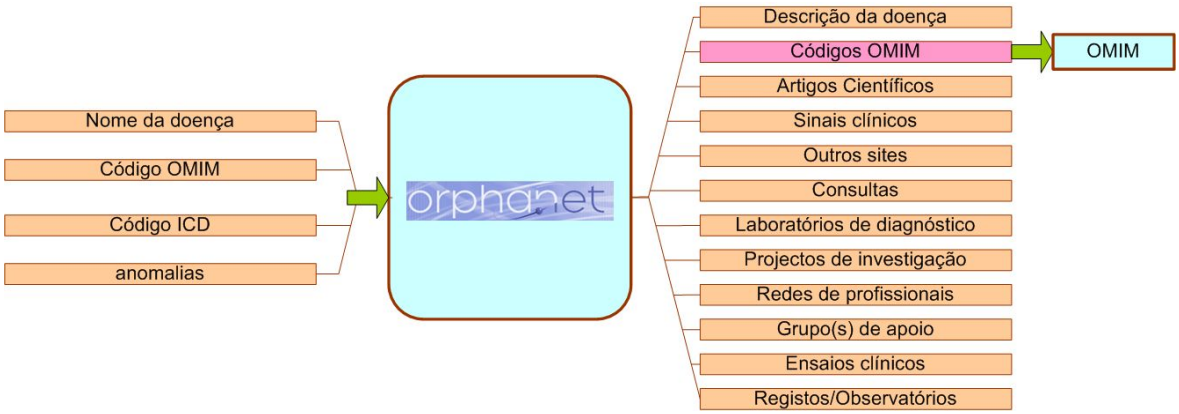


Figura 4.3 – Fluxo de informação na base de dados *Orphanet*. As caixas à esquerda representam as chaves de acesso nas consultas. As caixas da direita representam as diversas categorias de informação disponíveis. "Códigos Omim" é uma ligação à base de dados OMIM.

4.1.2 ClinicalTrials

Um teste clínico (*clinical trial*) consiste num estudo clínico em voluntários humanos cujo objectivo é responder a questões específicas da saúde sendo um método rápido de

encontrar e testar novos tratamentos e terapias. Com este tipo de testes é possível determinar se determinados tratamentos experimentais ou novas formas de usar terapias conhecidas são seguros e eficazes.

Neste contexto, o *site* norte-americano ClinicalTrials.gov [79] contém informação de descrição de testes clínicos para uma gama extensa de doenças. Actualmente a sua base de dados contém cerca de 29500 estudos clínicos patrocinados pelo NIH (*National Institute of Health*), várias agências federais e indústria privada. Os estudos presentes na base de dados são efectuados em todos os estados e em mais de 130 países [80].

Esta base de dados permite obter informação sobre testes clínicos a partir de vários tipos de palavras-chave como o nome da doença, nome do fármaco envolvido, localização do teste, patrocinador, etc. Para cada teste clínico retornado pela consulta existe no *site* informação imediata sobre o estado de desenvolvimento do teste, a sua entidade patrocinadora, o propósito do teste, etc. (Figura 4.4).

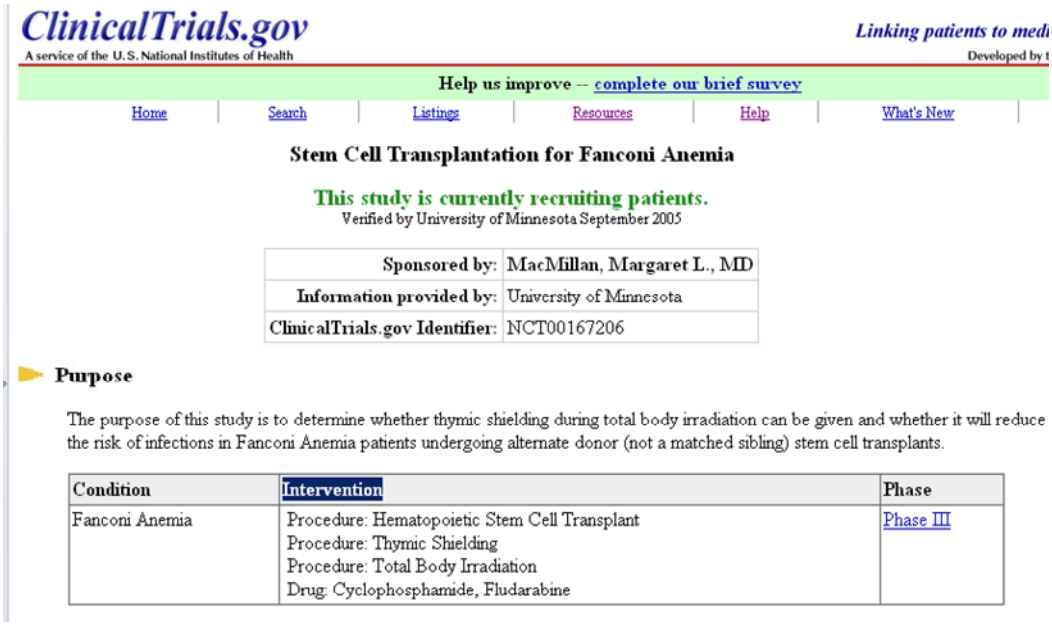


Figura 4.4 – Detalhe de um teste clínico no *site ClinicalTrials*.

Cada teste é caracterizado por uma condição, por uma intervenção, e pela fase actual de desenvolvimento. A condição refere o motivo do teste, ou seja, a doença que se pretende estudar. A intervenção descreve o processo do teste em forma de uma lista de itens. Cada item pode ser, por exemplo, um procedimento clínico, a administração de um fármaco ou a utilização de um determinado equipamento. A fase descreve o estado de estudo do teste. A

maioria dos testes divide-se em três grupos, fase I, fase II e fase III, sendo que cada uma difere no tipo de questões que se pretende obter resposta. Assim sendo, a fase I refere-se a testes de novos fármacos ou tratamentos num pequeno grupo de pessoas (20-80), pela primeira vez, para avaliar a segurança, determinar a gama de dosagem e identificar efeitos secundários. Na fase II o teste é efectuado num grupo maior de pessoas (100-300), pretendendo-se avaliar a efectividade do tratamento e aprofundar os testes de segurança. A fase III envolve um ainda maior número de pessoas (1000-3000) e espera-se confirmar a efectividade do tratamento, monitorizar os efeitos secundários, compará-lo com tratamentos comuns e recolher informação necessária para proporcionar a administração correcta do fármaco ou do tratamento.

Além desta informação, outros detalhes são disponibilizados como os critérios de inclusão e de exclusão nos testes clínicos e os contactos e localização dos organismos que realizam os testes.

A figura seguinte apresenta o fluxo de informação numa consulta à base de dados *ClinicalTrials*.

Figura 4.5 – Fluxo de informação na base de dados *ClinicalTrials*.

4.1.3 OMIM

O projecto *Mendelian Inheritance in Man* é uma base de dados baseada no livro com o mesmo nome, da autoria do Dr. *Victor A. McKusick* e colegas do instituto *Johns Hopkins* que cataloga todas as doenças humanas que tenham uma componente genética. Quando possível faz a ligação dessas doenças aos respectivos genes [80]. A cada doença e gene é atribuído um identificador com seis algarismos em que o primeiro reflecte o tipo de hereditariedade.

Tabela 4.1 – Atribuição da numeração dos códigos OMIM aos diferentes tipos de hereditariedade.

Primeiro Dígito	Código MIM	Hereditariedade
1	100000-199999	*Loci* autossómicos dominantes ou fenótipos
2	200000-299999	*Loci* autossómicos recessivos ou fenótipos
3	300000-399999	*Loci* no cromossoma X ou fenótipos
4	400000-499999	*Loci* no cromossoma Y ou fenótipos
5	500000-599999	*Loci* mitocondriais ou fenótipos
6	600000-699999	*Loci* autossómicos ou fenótipos (adicionados a partir de 15-5-94)

Além da numeração referida na tabela anterior, a preceder a sequência de seis dígitos existe também um prefixo (*, #, +, %, ^) que reflecte a natureza do código OMIM conforme a Tabela 4.2.

Tabela 4.2 – Significado dos prefixos dos códigos OMIM.

Prefixo	Significado
*****	Gene com uma sequência conhecida
#	Entrada descritiva para um fenótipo, não representando um locus único.
+	Entrada que contém a descrição de um gene com sequência conhecida e o fenótipo
%	Entrada que descreve um fenótipo mendeliano confirmado ou um locus fenotípico com base molecular não conhecida.
Sem prefixo	Entrada que descreve um fenótipo cuja base mendeliana, embora suspeita, não está claramente estabelecida ou então a separação deste fenótipo com outro numa outra localização não é clara.
^	Indica que este código já não está associado a nenhuma entrada, a qual foi removida da base de dados ou associada a outro código

Esta base de dados é actualizada diariamente e disponibilizada pelo NCBI (*National Center for Biotechnology Information*) Em Junho de 2006 continha cerca de 16800 registos, fornecendo informação textual e referências assim como uma lista abundante de *links* para a MEDLINE (referências de literatura) bem como para registos no sistema *Entrez* e outros sistemas do NCBI.

A base de dados permite pesquisar variantes alélicas (sequências alternativas para o mesmo gene) para um dado fenótipo sendo estas classificadas através de um número de 10 dígitos em que os primeiros seis indicam o *locus* pai e os restantes quatro identificam o alelo. A maioria das variações alélicas produz mutações que resultam em doenças.

Contém também informação acerca da localização citogenética dos genes associados a cada código OMIM. Esta informação está presente em *OMIM Gene Map* sob a forma de

tabela (também disponibilizada numa interface gráfica) listando todos os genes desde o telómero *p* do cromossoma 1 até ao telómero *q* do cromossoma 22, seguido dos cromossomas X e Y.

Além desta informação, a base de dados OMIM disponibiliza ainda a tabela *OMIM Morbid Map* que consiste numa lista alfabética de todas as doenças descritas no catálogo OMIM incluindo as respectivas localizações citogenéticas.

A Figura 4.6 apresenta alguns pormenores de páginas do *site* OMIM incluindo a página central contendo, num conjunto vasto de temas, uma descrição textual da doença e diversas ligações a outras localizações no *site*. Uma dessas localizações consiste numa página com informação sobre a localização citogenética do gene associado à doença (*NCBI Map Viewer*), onde está disponível uma representação gráfica do cromossoma e de outros genes próximos conhecidos.

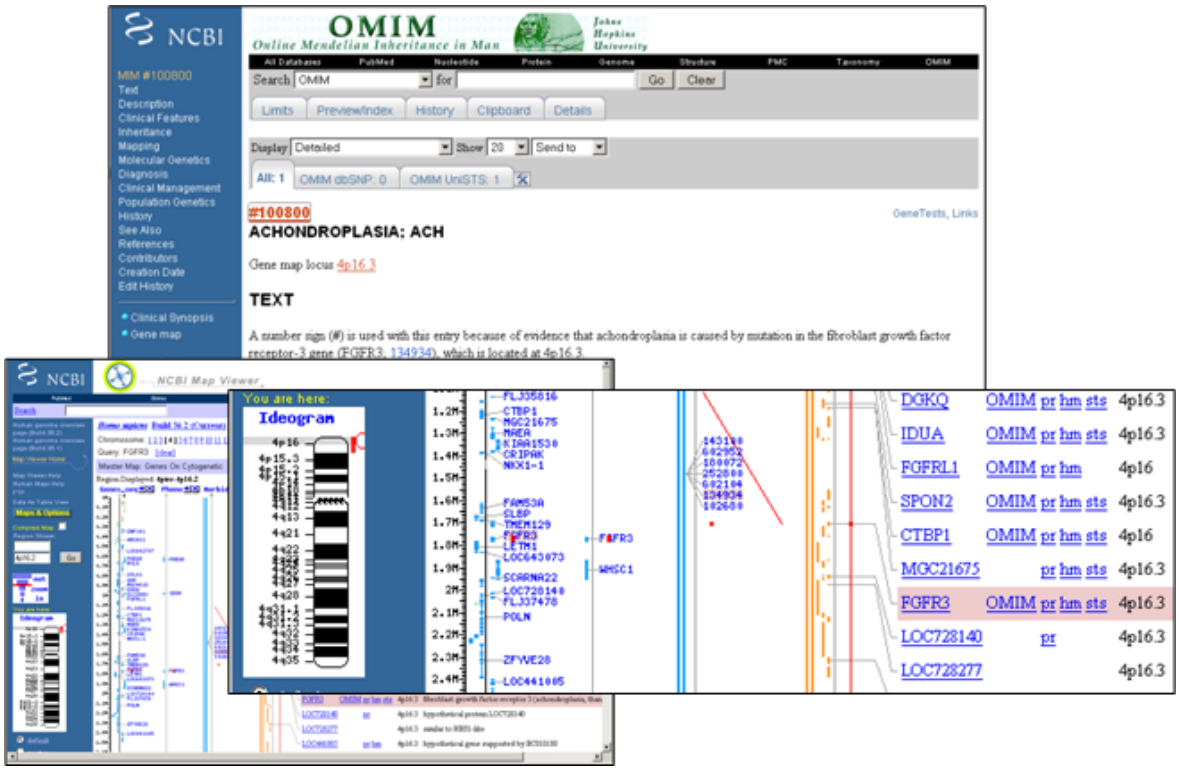


Figura 4.6 – Detalhe de páginas do site OMIM.

A base de dados OMIM pode ser pesquisada tanto a partir da sua página de entrada bem como a partir de qualquer página das bases de dados do NCBI Entrez. A sua informação pode ser obtida essencialmente a partir de consultas baseadas no código OMIM, doença e nome ou símbolo do gene. Além disto, permite ainda a utilização de uma série de

operadores lógicos sobre as consultas de modo a impor limites aos resultados das pesquisas.

Em cada entrada OMIM (página de fundo na Figura 4.6), à esquerda na barra vertical existe um menu de acesso directo a todos os sub capítulos da página central. Além destas ligações são também disponibilizadas ligações para páginas do *site* OMIM bem como páginas de *sites* externos. *Clinical Synopsis* por exemplo liga a uma página que contém uma lista exaustiva das características clínicas da doença. *Gene Map* liga directamente a uma página de síntese do mapa de genes humanos, indicando a localização citogenética dos genes associados à doença. Como já foi dito anteriormente, uma entrada OMIM tanto pode estar associada a uma doença ou fenótipo como a um gene e, quando ambas estão disponíveis na base de dados, é apresentada uma ligação para uma página de testes genéticos. Uma outra ligação disponível é a das variantes alélicas contendo um conjunto de mutações genéticas para associadas ao gene em questão [81].

A Figura 4.7 ilustra o fluxo de informação numa consulta à base de dados OMIM. As categorias de informação aqui ilustradas são uma pequena amostra daquilo que se pode obter através da página OMIM dado que este *site* faz parte de um vasto conjunto incluído no sistema NCBI.

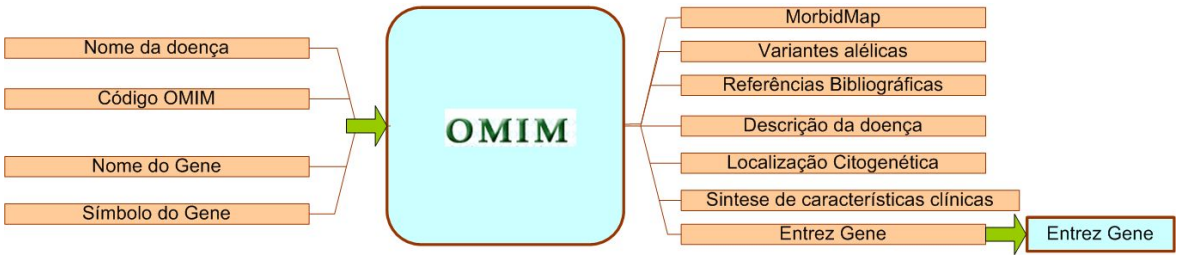


Figura 4.7 – Fluxo de informação na base de dados OMIM.

4.1.4 dbSNP

A base de dados dbSNP (*Single Nucleotide Polymorphism Database*) [82] foi desenvolvida em 1998 pelo NCBI (*National Center for Biotechnology Information*) em colaboração com o *National Human Genome Research Institute* servindo como um repositório central público de armazenamento de uma vasta colecção de polimorfismos genéticos (ver anexo 8.1 - Polimorfismos). Actualmente a dbSNP classifica as variações nas seguintes categorias:

- Sequências de substituições de nucleótidos simples (SNPs), correspondendo a 99,7% dos registos na base de dados;
- Inserção/deleção de polimorfismos (DIPs – *Deletion Insertion Polymorphisms*), 0,21%;
- Regiões de sequências invariantes, 0,02%;
- Microsatélites – 0,001% [83].

Uma vez descobertos, estes polimorfismos podem ser associados a fenótipos hereditários, permitindo a partir de diferenças específicas na população, a identificação célere de genes causadores de doenças.

A pesquisa de informação directamente sobre a base de dados dbSNP pressupõe o conhecimento prévio dos identificadores numéricos para cada SNP. Outra forma mais intuitiva de pesquisar esta base de dados passa pelo seu acesso através do sistema Entrez SNP. Este último permite efectuar pesquisas tendo como ponto de partida um conjunto de identificadores tais como organismo, cromossoma, tipo de snp, etc. Além disso permite que o utilizador construa expressões lógicas sobre todos eles. A Figura 4.8 ilustra os diferentes modos de acesso à base de dados dbSNP.

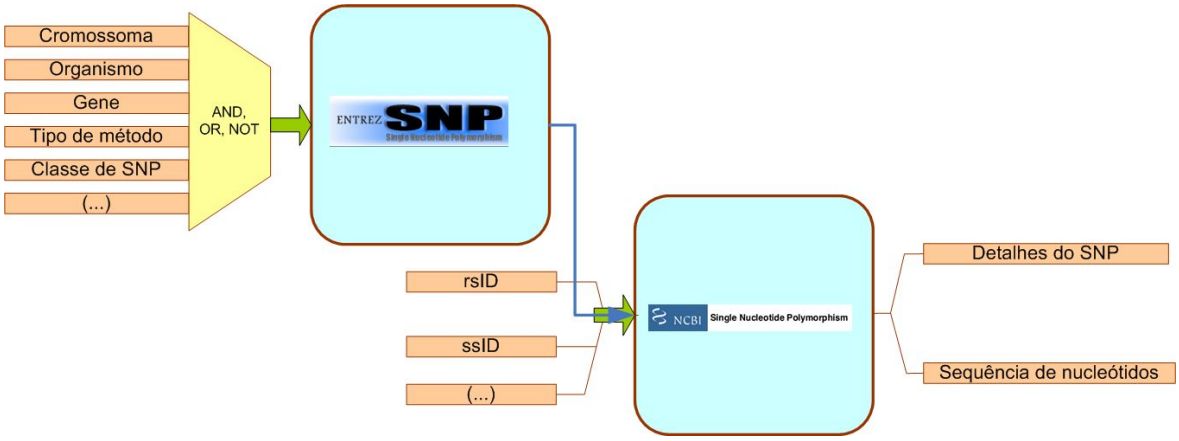


Figura 4.8 – Fluxo de informação da base dados dbSNP.

Este método de acesso através do Entrez SNP foi o escolhido para implementar no modelo de dados uma vez que permite obter as sequências dos SNPs a partir do respectivo gene. A Figura 4.9 mostra um detalhe da página Entrez SNP para o gene FGFR3. Cada item apresentado corresponde a um SNP, contendo o respectivo identificador com ligação a uma entrada na base de dados dbSNP, a sequência com o SNP em causa e um conjunto de ligações adicionais a diversas bases de dados (MapView, GeneView, SeqView, etc.).

O primeiro item apresenta um SNP que origina duas sequências alternativas ou seja, a posição assinalada com a expressão [C/G] pode ser ocupada com a base C ou G.

Figura 4.9 – Detalhe de uma página do Entrez SNP relativa à pesquisa de snps para o gene humano FGFR3.

Seleccionando o primeiro item acede-se à página da base de dados dbSNP apresentada na Figura 4.10.

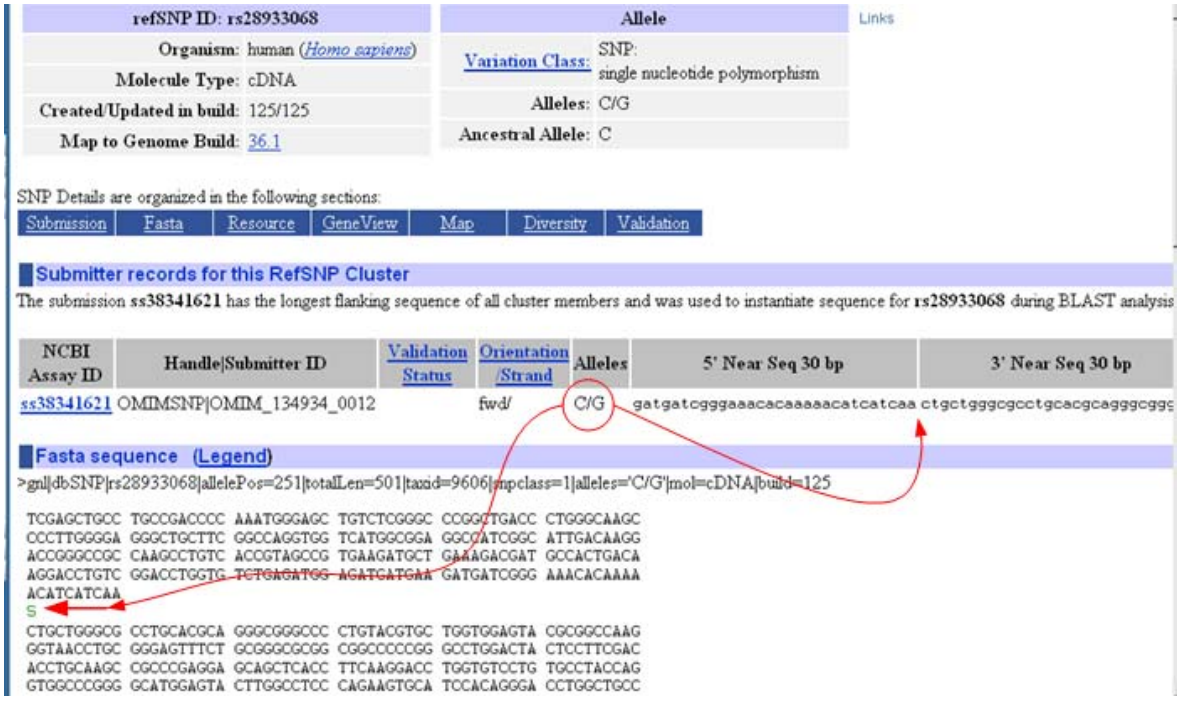


Figura 4.10 – Detalhe de uma página da base de dados dbSNP para um snp do gene FGFR3.

4.1.5 Entrez-Gene

Entrez-Gene [84] é a base de dados do NCBI para armazenar informação específica dos genes. Ao invés de incluir todos os genes actualmente conhecidos, esta base de dados concentra-se totalmente em genes pertencentes a genomas que estão completamente sequenciados e que tenham suporte activo por parte da comunidade científica de forma a contribuir com informação específica.

Para cada gene ou local genético (*locus*), a base de dados associa um identificador único com o qual é possível aceder ao sistema. Para cada gene existe informação que inclui nomenclatura do gene, localização citogenética, produtos genéticos, interacções das proteínas, marcadores associados, fenótipos, *links* para citações na área da saúde, etc. Além do NCBI, esta integração resulta de uma mistura e análise de várias bases de dados como o EMBL, GenBank, DDBJ, GeneRIFs, etc. No fundo, esta base de dados representa o acesso central e integral a toda a informação específica dos genes do NCBI. A relação gene – sequência fornecida pelo Entrez-gene é utilizada por outros recursos do NCBI como o BLAST, Geo, HomoloGene, MapViewer, UniGene [85, 86].

Figura 4.11 – Detalhe da página de entrada do sistema Entrez com ligações a todas a bases de dados que a integram.

4.1.6 Nucleotide (NCBI)

Nucleotide é uma base de dados que armazena sequências de nucleótidos provenientes de várias fontes entre as quais GenBank, RefSeq e PDB. Estando a crescer a um ritmo exponencial, o número se bases presentes nestas fontes ultrapassava os 130 milhões em Abril de 2006.

4.1.7 Swiss-Prot

Swiss-Prot é uma base de dados de sequências de proteínas de várias espécies criada em 1986 por *Amos Bairoch* e desenvolvida pelo Instituto Suíço de Bioinformática e pelo Instituto Europeu de Bioinformática. Além das entradas para as sequências, a base de dados disponibiliza também informação sobre a função das proteínas, a estrutura dos domínios, modificações pós tradução (*post-translational modifications*), informação sobre a taxonomia da proteína (descrição da fonte biológica da proteína), referências bibliográficas, etc. Toda esta informação suporta um elevado nível de integração com outras bases de dados.

Fundada pelo NIH e produzida para juntar informação proveniente da Swiss-Prot, da PIR (*Protein Information Resource*) e da TrEMBL, existe também a base de dados UniProt (*UniProt Knowledgebase*) que consiste no catálogo de proteínas mais abrangente a nível mundial [87] [62].

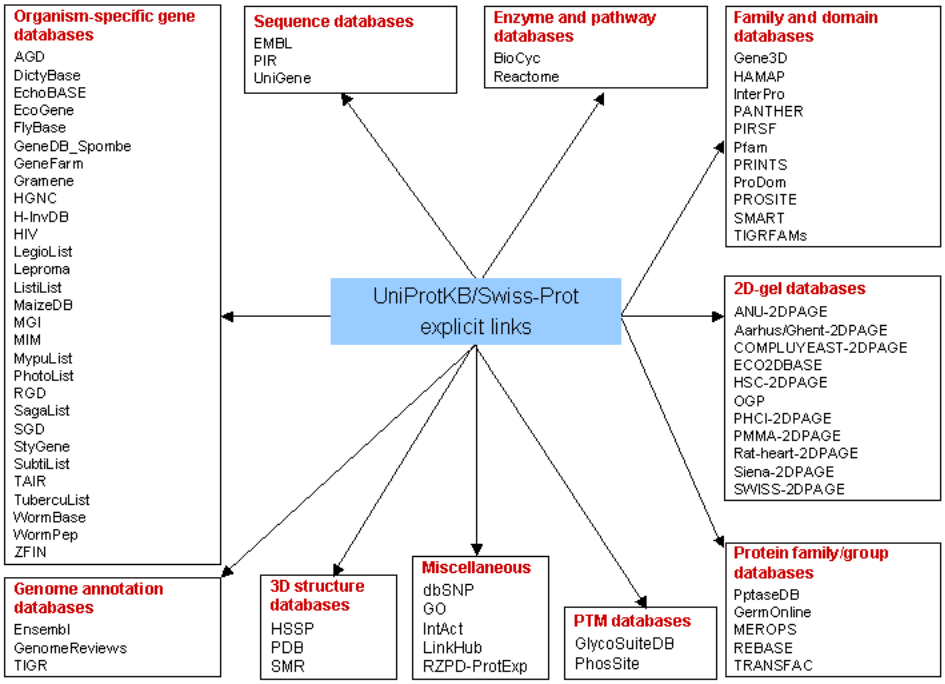


Figura 4.12 – Diagrama ilustrativo do estado actual das referências cruzadas entre a Swiss-Prot e outras bases de dados biomoleculares.

4.1.8 Inter-Pro

InterPro é uma fonte integrada de documentação sobre famílias e domínios de proteínas. Esta integração é conseguida a partir da combinação de múltiplas bases de dados (membros) que possuem metodologias de acesso diferentes e uma grande diversidade de informação biológica sobre proteínas. Actualmente integra informação das bases de dados PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, CATH e PANTHER [88]. A versão 13.0 da base de dados InterPro contém 13147 entradas, representando 3760 domínios, 9080 famílias de proteínas.

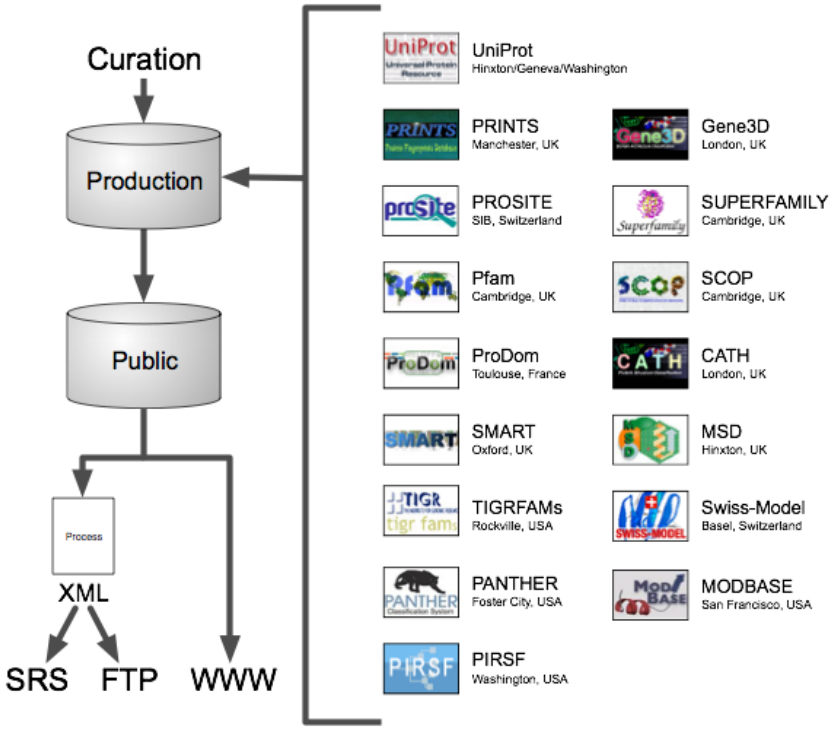


Figura 4.13 – Esquema ilustrativo do fluxo de dados da InterPro desde as bases de dados membro até à distribuição da informação.

4.1.9 KEGG

KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [89] é um conjunto de bases de dados que entre outras funções, visa integrar conhecimento nas redes de interacção molecular inerentes a processos biológicos (vias metabólicas). Um dos grandes desafios da era pós–genómica é obter uma representação computacional completa da célula e do organismo que permita calcular e prever processos celulares de elevada complexidade e o comportamento dos organismos partindo da informação genómica. Tendo isto em conta, o KEGG permite, partindo de um conjunto completo de genes no genoma, prever e observar as redes de interacção das proteínas, responsáveis por vários processos celulares. Para tal, integra informação sobre redes de interacção molecular tais como vias e complexos metabólicos (base de dados PATHWAY), informação sobre genes e proteínas (GENES, SSDB, KO) e informação sobre compostos e reacções bioquímicas (COMPOUND, GLYCAN, REACTION). Estes três tipos de bases de dados representam três grafos que compõem o KEGG: Rede de Proteínas, Universo dos Genes e Universo Químico.

O universo dos genes é um grafo conceptual que representa relações entre genes; o universo químico é outro grafo conceptual e representa as reacções químicas e as relações

funcionais e estruturais ao nível dos metabolitos ou outros compostos bioquímicos; finalmente, a rede de proteínas, descreve os fenómenos biológicos, representando redes de interacções moleculares conhecidas em diversos processos celulares [90]. A Figura 4.14 é um exemplo de um grafo da rede de proteínas, descrevendo um via metabólica do processo do metabolismo da Timina.

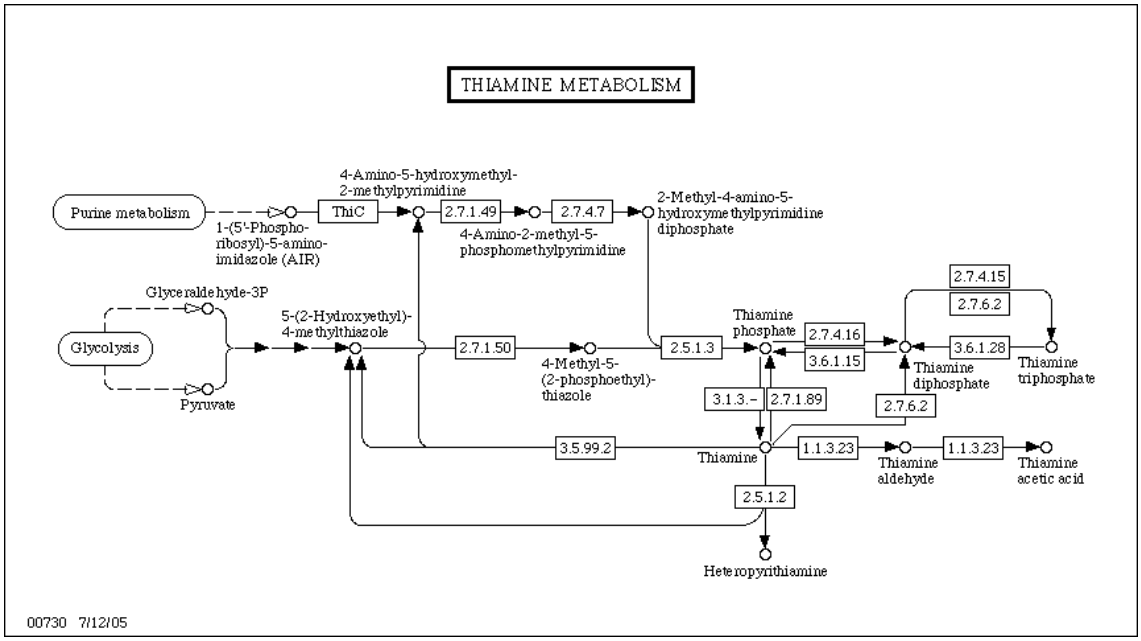


Figura 4.14 – Exemplo de um grafo de via metabólica, retirado da base de dados KEGG PATHWAY. A legenda dos símbolos utilizados nestes diagramas está presente em anexo no capítulo 8.2 (Mapas das vias metabólicas KEGG).

4.1.10 GeneCards

A base de dados GeneCards, fundada em 1997 pelo *Crow Human Genome Center* pertencente ao Instituto de Ciência *Weizmann* em Israel, procura integrar num único ponto de acesso informação orientada a genes humanos, proteínas e doenças, informação essa que se encontra dispersa por várias bases de dados especializadas por assuntos [91] [92]. A sua informação é acedida a partir do gene (pelo símbolo, nome alternativo, identificador GC ou por texto livre) e apresentada sob a forma de um cartão virtual, dividido por várias áreas como por exemplo a localização citogenética, proteínas associadas, função do gene, sequência de nucleótidos, etc.

4.1.11 PubMed

É um serviço que permite a pesquisa de literatura médica, contendo citações de artigos biomédicos desde 1950. Actualmente esta base de dados faz parte do sistema de informação *Entrez*, abrangendo mais de 4800 revistas publicadas nos Estados Unidos e em mais 70 países. Além do acesso à MEDLINE, o serviço PubMed permite também aceder a artigos da OLDMEDLINE a qual fornece citações de artigos anteriores a 1966 [55].

4.1.12 PharmGKB

PharmGKB (*Pharmacogenetics Knowledge Base*) é uma ferramenta de pesquisa desenvolvida pela Universidade de *Stanford*, cujo principal objectivo é fornecer aos investigadores informação que permita perceber o modo como as variações em genes humanos levam à variação da resposta a medicamentos, estabelecendo relações entre genes e medicamentos. Esta base de dados contém informação genómica, fenotípica e clínica associada a estudos farmacogenéticos. A variabilidade na resposta aos medicamentos tem vindo a preocupar médicos, pacientes e companhias farmacêuticas. Este fenómeno deve-se a várias causas como factores ambientais, condições médicas e diferenças genéticas hereditárias, sendo estas últimas identificadas através dos polimorfismos simples (SNPs) nas pessoas [93].

4.1.13 HGNC

A descoberta de um mesmo gene por mais do que um grupo de investigadores é uma situação bastante comum, resultando daqui que o mesmo gene passa a ser identificado por vários nomes. Com o intuito de evitar esta situação caótica surgiu a necessidade de criar uma terminologia sistemática, consistente e hierárquica que não só facilitasse a ligação e criação de bases de dados integrados como também aumentasse a confiança nos dados daí extraídos. Há mais de vinte anos que o HGNC (*Human Gene Nomenclature Committee*) é reconhecido pela organização HUGO (*Human Genome Organization*) como a autoridade para atribuir nomes e símbolos aos genes humanos. Assim sendo, os objectivos do HGNC passam pela promoção da aceitação e utilização universal de uma nomenclatura normalizada e também pela coordenação da atribuição de terminologias pelas várias espécies, particularmente os mamíferos.

A base de dados HGNC contém actualmente cerca de 23000 símbolos aprovados correspondendo maioritariamente a genes que codificam proteínas. Além destes também inclui símbolos para pseudogenes, ARN não codificante, entre outros [94] [95].

4.1.14 GO (Gene Ontology)

Este projecto começou em 1998 com a colaboração de três bases de dados de organismos sendo elas a *FlyBase* (para a *Drosophila*), a *Saccharomyces Genome Database* (SGD) e a *Mouse Genome Informatics* (MDG). Desde então, este projecto tem alargado a lista de bases de dados a outros organismos incluído plantas, animais e genomas microbiais [96].

O *Gene Ontology* tem por função produzir e divulgar vocabulário controlado e estruturado para descrever atributos associados a genes de qualquer organismo. Para tal, o projecto GO desenvolveu três sistemas de ontologias distintos para descrever os produtos genéticos em termos do Processo Biológico, do Componente Celular e da Função Molecular. Os termos referentes ao Processo Biológico referem-se ao propósito do ponto de vista biológico para o qual o produto genético contribui. Neste contexto, um processo envolve uma transformação química ou física. Exemplos de processos biológicos são o crescimento e manutenção da célula e transdução de sinal e o metabolismo da pirimidina. A Função Molecular é definida como a actividade bioquímica de um produto genético. A *adenilato ciclase* é um exemplo de uma função molecular. O Componente Celular define o lugar na célula onde o produto genético está activo. Termos como *ribossoma*, *proteassoma*, membrana nuclear ou aparelho de *Golgi* fazem parte do Componente Celular e descrevem onde os produtos genéticos podem ser encontrados [97].

Entre as ferramentas desenvolvidas pelo *Gene Ontology* encontra-se a base de dados AmiGO a qual permite pesquisar todos os produtos genéticos associados a um termo GO.

4.1.15 EDDNAL

O EDDNAL *(European Directory of DNA Diagnostic Laboratories*) é um serviço sedeado na Bélgica, no Centro de Genética Humana, *Institut de Pathologie et de Génétique* que fornece uma vasta lista de laboratórios de testes genéticos bem como os serviços por eles oferecidos. Esta lista contém endereços e contactos de mais de 300 laboratórios de genética.

4.1.16 PDB

O PDB (*Protein Data Bank*) é um repositório de processamento e distribuição de estruturas biológicas macromoleculares tridimensionais. Criado em 1971 pelos laboratórios *Brookhaven National Laboratories* começou por conter apenas sete estruturas. Contudo este número cresceu drasticamente durante os anos oitenta devido a melhoramentos significativos nas técnicas de cristalografia. Actualmente possui cerca de 38000 estruturas sendo uma referência fundamental para a maioria das revistas do ramo [57] [98]. A Figura 4.15 é um exemplo de uma estrutura tridimensional de uma proteína, retirada desta base de dados.

Figura 4.15 – Exemplo de uma estrutura tridimensional da proteína 3PTE retirada da base de dados PDB.

O sistema PDB permite fazer uma série de consultas sobre uma estrutura ou conjuntos de estruturas moleculares. Os resultados destas consultas são apresentados numa página *web* em forma de sumário e, adicionalmente a estes resultados, são disponibilizados *links* a outras bases de dados contendo informação associada

4.1.17 Prosite (Expasy)

Prosite é uma base de dados de famílias e domínios de proteínas que faz parte dos servidores EXPASY pertencentes ao Instituto Suíço de Bioinformática. O seu principal objectivo é identificar possíveis funções de sequências de proteínas recém descobertas com base em estruturas de proteínas já conhecidas.

Apesar de existir um número enorme de proteínas diferentes, a maior parte delas pode ser agrupada num número limitado de classes ou famílias, tendo em conta as semelhanças nas suas sequências. Esta classificação é importante porque as proteínas ou domínios de proteínas que pertençam a uma mesma família partilham geralmente as mesmas funções e derivam de um antepassado comum.

Prosite fornece um serviço chamado *ScanProsite* que permite aos utilizadores examinarem sequências de proteínas contra todas as entrada da *Prosite* [99].

4.1.18 Enzyme (Expasy)

A base de dados *Enzyme* faz também parte do *Expasy* e é um repositório de informação relativa a nomenclatura de enzimas. Inicialmente foi baseada nas recomendações do comité IUBMB (*Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*). Para cada enzima a base de dados contém informação como o identificador de acesso EC (*Enzyme Commission*), nome recomendado e nomes alternativos para a enzima, actividade catalítica, apontadores para entradas de sequências de proteínas associadas na base de dados *Swiss-Prot*, apontadores para entradas na *Prosite*, etc.

As entradas na base de dados *Enzyme* estão estruturadas de tal forma que permitam uma leitura fácil tanto por humanos como por programas de computador (Figura 4.16). Cada entrada é composta por várias linhas, em que cada uma regista os vários tipos de dados que compõem a entrada [100].

```
ID   1.14.17.3
DE   Peptidylglycine monooxygenase.
AN   Peptidyl alpha-amidating enzyme.
AN   Peptidylglycine 2-hydroxylase.
CA   Peptidylglycine + ascorbate + O(2) = peptidyl(2-
     hydroxyglycine) +
CA   dehydroascorbate + H(2)O.
CF   Copper.
CC   -!- Peptidylglycines with a neutral amino acid
     residue in the penultimate
CC   position are the best substrates for the enzyme.
CC   -!- The enzyme also catalyses the dismutation of
     the product to
CC   glyoxylate and the corresponding desglycine
     peptide amide.
CC   -!- Involved in the final step of biosynthesis of
     alpha-melanotropin
CC   and related biologically active peptides.
PR   PROSITE; PDOC00080;
DR   P08478, AMD1_XENLA; P12890, AMD2_XENLA; P10731,
     AMD_BOVIN ;
DR   P19021, AMD_HUMAN ; P97467, AMD_MOUSE ; P14925,
     AMD_RAT ;
```

Figura 4.16 – Exemplo de uma amostra de uma entrada da base de dados *Enzyme*.

4.2 Modelo de Navegação

Nas subsecções anteriores foram apresentadas e descritas algumas das principais bases de dados seleccionadas para preencher todos os conceitos do protocolo conceptual ilustrado na Figura 4.1. Como resultado final, obteve-se o mapeamento completo destas bases de dados para cada conceito do mesmo diagrama. O resultado está ilustrado na Figura 4.17.

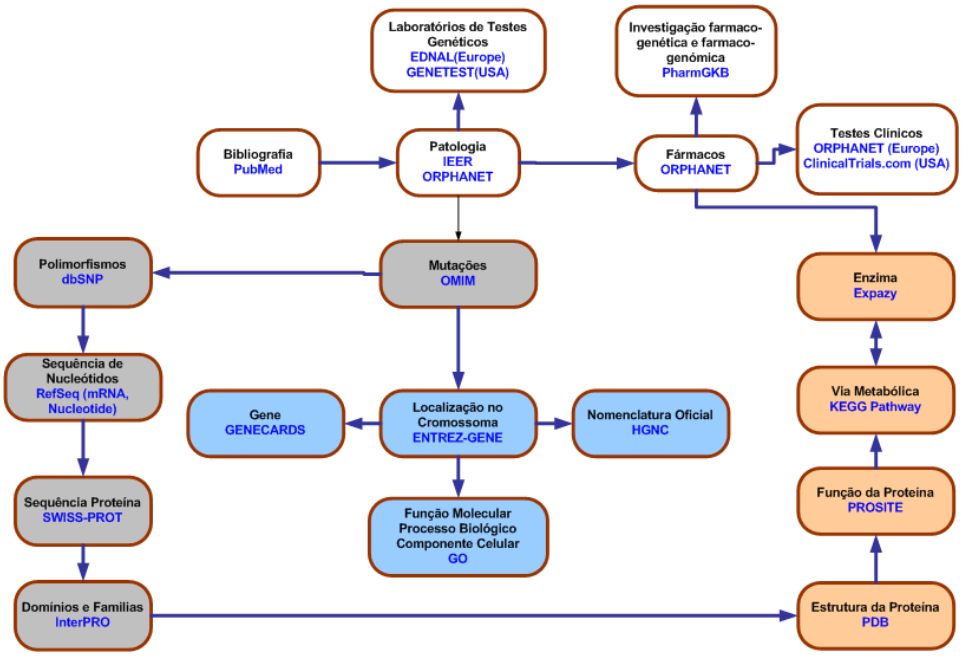


Figura 4.17 – Protocolo conceptual contendo as bases de dados seleccionadas para o efeito.

Estando a tarefa da selecção e análise das bases de dados terminada, é necessário agora descrever um processo para as percorrer na totalidade e assim retirar a informação relevante de modo a preencher todos os conceitos apresentados na Figura 4.17.

Assumindo que um dado utilizador deseja obter informação alargada acerca de uma dada doença genética rara, ele tem de explorar vários *sites* e bases de dados públicas, seguindo um conjunto de caminhos e obedecendo uma determinada sequência, de forma a preencher todos os conceitos que constituem o mapa conceptual. Esta exploração de informação obedece a um protocolo de navegação cujo exemplo é ilustrado na Figura 4.18. Nesta figura está representada uma possível rede de navegação através das fontes de informação que permitem preencher os conceitos do mapa conceptual anteriormente apresentado e descrito. Este diagrama foi obtido com base no mapa conceptual das doenças genéticas

raras e tendo em conta os percursos físicos seguidos através das páginas disponíveis na Internet.

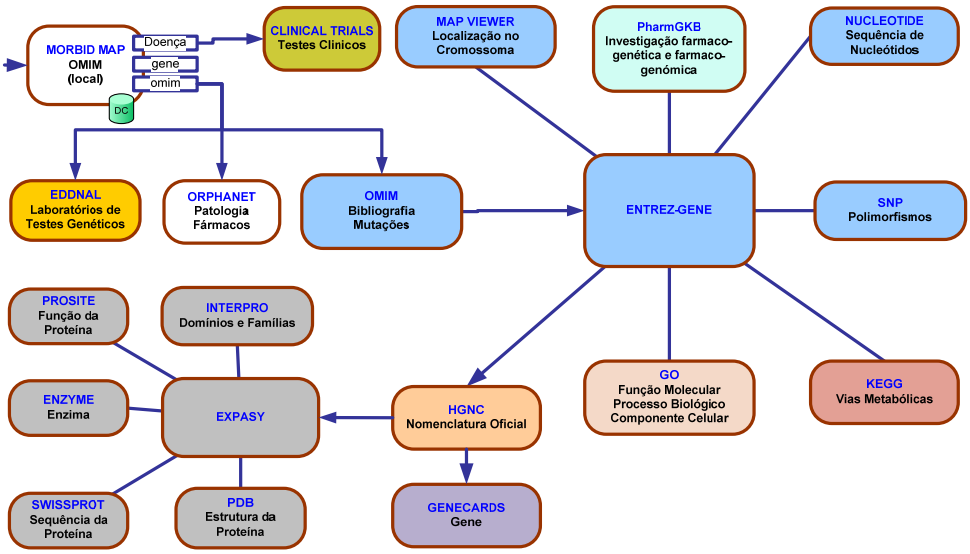


Figura 4.18 – Exemplo de um protocolo de navegação, uma rede de fontes de informação associada às doenças genéticas raras. Cada bloco representa uma fonte de dados da qual se retira informação relativa aos respectivos conceitos. Por exemplo a base de dados *Orphanet* contém informação para os conceitos Patologia e Fármacos. Existem blocos que além de fornecer informação para os respectivos conceitos são também pontos de acesso a outras bases de dados como é o caso por exemplo do bloco *OMIM* que permite alcançar o *site Entrez-Gene.*

O ponto de entrada neste protocolo de navegação é a tabela *Morbid Map* da base de dados OMIM (NCBI). A tabela *Morbid Map* é uma lista que associa as doenças descritas na base de dados OMIM a um código de acesso (código *omim*) e aos respectivos genes responsáveis pela doença. Com base no nome da doença ou no respectivo código *omim* é possível obter informação para todos os conceitos do protocolo de navegação. Como exemplo, através do código *omim* é possível aceder ao *site Orphanet* e obter informação sobre a patologia e medicamentos existentes. Com o mesmo código acede-se também à página *Omim* onde se obtém informação sobre mutações e referências bibliográficas. Por sua vez, esta página permite aceder ao *Entrez-Gene* e assim sucessivamente ao longo de todos os ramos do protocolo.

Uma das principais dificuldades inerentes à operação de obtenção e integração de informação prende-se com a necessidade do utilizador em conhecer todo este protocolo. O sucesso desta operação ou seja, o preenchimento de todos os conceitos do mapa com informação respectiva, depende da navegação e pesquisa correctas entre as diferentes bases de dados disponíveis. Outro factor importante é a especificidade de cada fonte de

informação que obriga a um conhecimento prévio da parte dos utilizadores o que exige algum tempo de adaptação. Este factor torna-se cada vez mais importante dado que o rápido crescimento das bases de dados, tanto em número bem como em tamanho dificulta o trabalho dos investigadores na obtenção eficaz de informação concreta [11]. A Figura 4.19 ilustra uma pequena parte do processo de exploração de conceitos relacionados com uma doença através de várias páginas de bases de dados biomédicas.

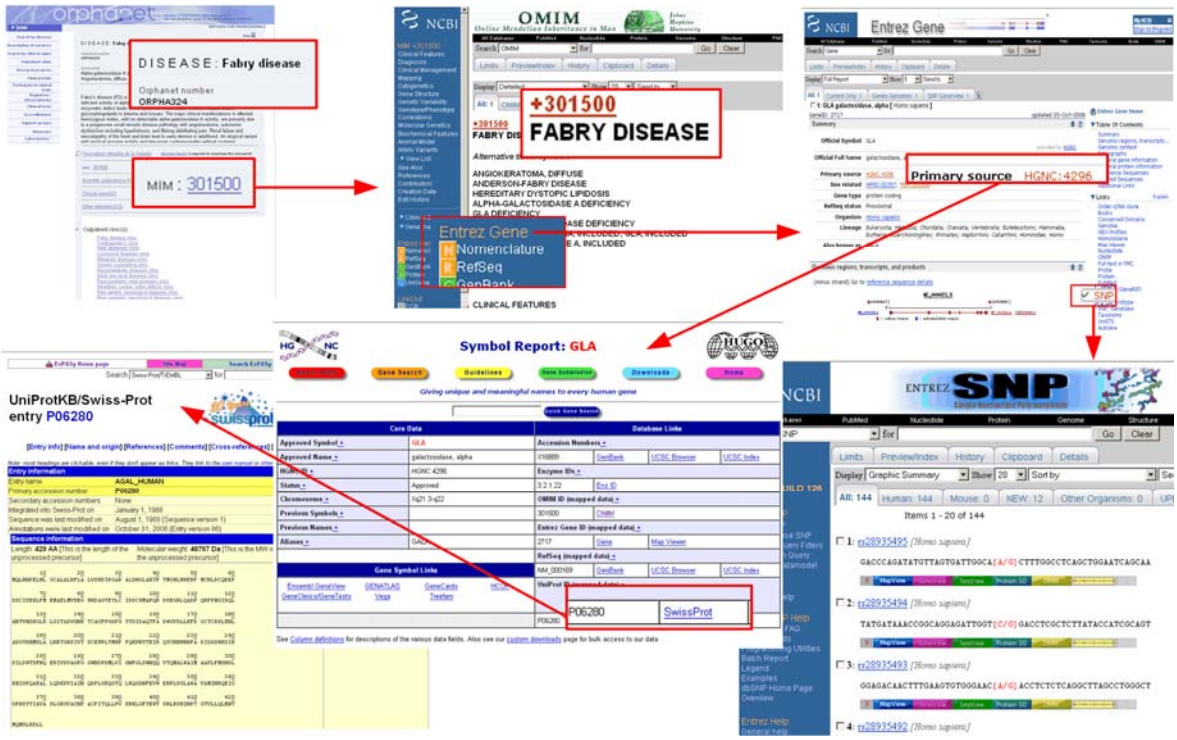


Figura 4.19 – Ilustração dos passos que um médico tem de seguir para obter informação de uma doença.

Grande parte das bases de dados existentes foram desenhadas para suportar e processar pesquisas que necessitam de envolvimento humano directo. O utilizador tem de localizar a fonte de dados correcta tendo para tal de estar bastante à vontade com as convenções de interacções entre bases de dados tendo além disso de formular pesquisas de acordo com sintaxes específicas. Por outro lado a informação que o investigador obtém das bases de dados é apenas um subconjunto da informação disponível se a pesquisa for dirigida apenas para uma base de dados específica [101]. Pelas razões acima referidas, para construir este protocolo de navegação são necessários conhecimentos consideravelmente profundos no que respeita ao domínio de exploração em questão. No caso das doenças genéticas raras cabe esta tarefa a utilizadores experientes em áreas como medicina, farmacologia, biologia molecular e farmacologia.

4.3 Sumário

Neste capítulo foi apresentado um modelo de navegação que permite sistematizar o processo de exploração e recolha de conhecimento associado às doenças genéticas raras. Para tal, foi previamente desenhado um mapa que reúne categorias consideradas importantes no contexto das doenças raras, que possibilite a sua integração, cobrindo conceitos associados à medicina em geral, à genómica e à farmacologia.

Ainda neste capítulo foi também descrita uma lista de bases de dados biomédicas seleccionadas das quais é possível extrair informação para cada conceito do protocolo de navegação.

Estando o mapeamento entre cada conceito e a respectiva base de dados estabelecido, a fase seguinte consiste na análise e desenvolvimento da automatização do processo de exploração de informação através do protocolo de navegação.

5 Um Modelo para Integração Virtual de Dados Heterogéneos

A aplicação *Diseasecard* surge como uma proposta de integração e ligação de informação associada a doenças genéticas raras, tendo em conta o modelo de navegação e o conjunto de bases de dados descritos nos capítulos anteriores.

Vários conceitos estão associados a este tema desde a descrição da patologia, fármacos existentes, testes clínicos até à área genética visto que este tipo de doenças se deve a mutações em determinados genes do paciente. A partir daqui pode-se determinar polimorfismos, sequências de nucleótidos, genes associados, proteínas codificadas enzimas etc. Um dado utilizador, por exemplo um profissional da área da saúde, que pretenda obter informação acerca de uma dada doença rara tem de pesquisar todos estes conceitos e em grande parte dos casos cada conceito tem associada uma base de dados ou uma página *web*. É neste ponto que surge o problema da integração da informação do ponto de vista do utilizador. Muitos profissionais de saúde não estão familiarizados com as fontes de informação disponíveis na Internet. Estas fontes são muito específicas, utilizando, além de nomenclaturas próprias, modelos de organização e pesquisa distintos. Outra razão centra-se na grande quantidade e diversidade de fontes de informação disponíveis sendo complicado para o utilizador manter-se ao corrente das últimas novidades na área.

Para resolver este problema desenvolveu-se uma plataforma de pesquisa e acesso a informação dispersa em fontes de dados na Internet de modo a tornar a navegação entre os inúmeros *sites* transparente para o utilizador. A sua funcionalidade é obtida através de uma camada de informação (lógica) que homogeneíza o acesso a diferentes fontes de dados.

5.1 Arquitectura do Software

Desde o início da sua implementação, a aplicação *Diseasecard* foi fruto de uma evolução que se reflectiu tanto a nível da sua funcionalidade como a nível da arquitectura. A Tabela 5.1 apresenta os principais passos da evolução da aplicação *Diseasecard* desde que foi criada estando as principais funcionalidades divididas por versões.

Tabela 5.1 – Principais funcionalidades do *Diseasecard* ao longo das suas versões.

Versões Diseasecard	Principais funcionalidades
Diseasecard 1.0	Ferramenta *web* colaborativa
Diseasecard 2.0	Ferramenta *web* colaborativa + Criação automática de cartões
Diseasecard 3.0	Criação automática de cartões

O grande objectivo por detrás da concepção do *DiseaseCard* prendeu-se inicialmente com o desenvolvimento de uma ferramenta, disponível através da Internet, que permitisse a colaboração *on-line* de vários grupos de pessoas experientes e especialistas em áreas específicas como genética, genómica, medicina e farmacologia, de modo a partilharem, armazenarem e disseminarem o seu conhecimento em torno das doenças raras (*Diseasecard 1.0*).

Deste modo um dado utilizador autenticado no sistema, organiza e delega a um grupo de investigadores especialistas nas diversas áreas o preenchimento com informação relevante (por meio de *links* e referências a páginas *web*) dos conceitos associados a doenças genéticas raras para os quais ele é especialista ou está mais familiarizado. Assim, estando o cartão preenchido, é publicado pelo coordenador da equipa ficando disponível para consulta de todos os utilizadores. A partir daqui toda a informação relativa a uma dada doença rara fica disponível de uma forma integrada e intuitiva para utilizadores menos experientes na pesquisa de informação na Internet. A Figura 5.1 ilustra os principais casos de utilização da aplicação na versão 1.0 bem como os perfis de utilização do sistema.

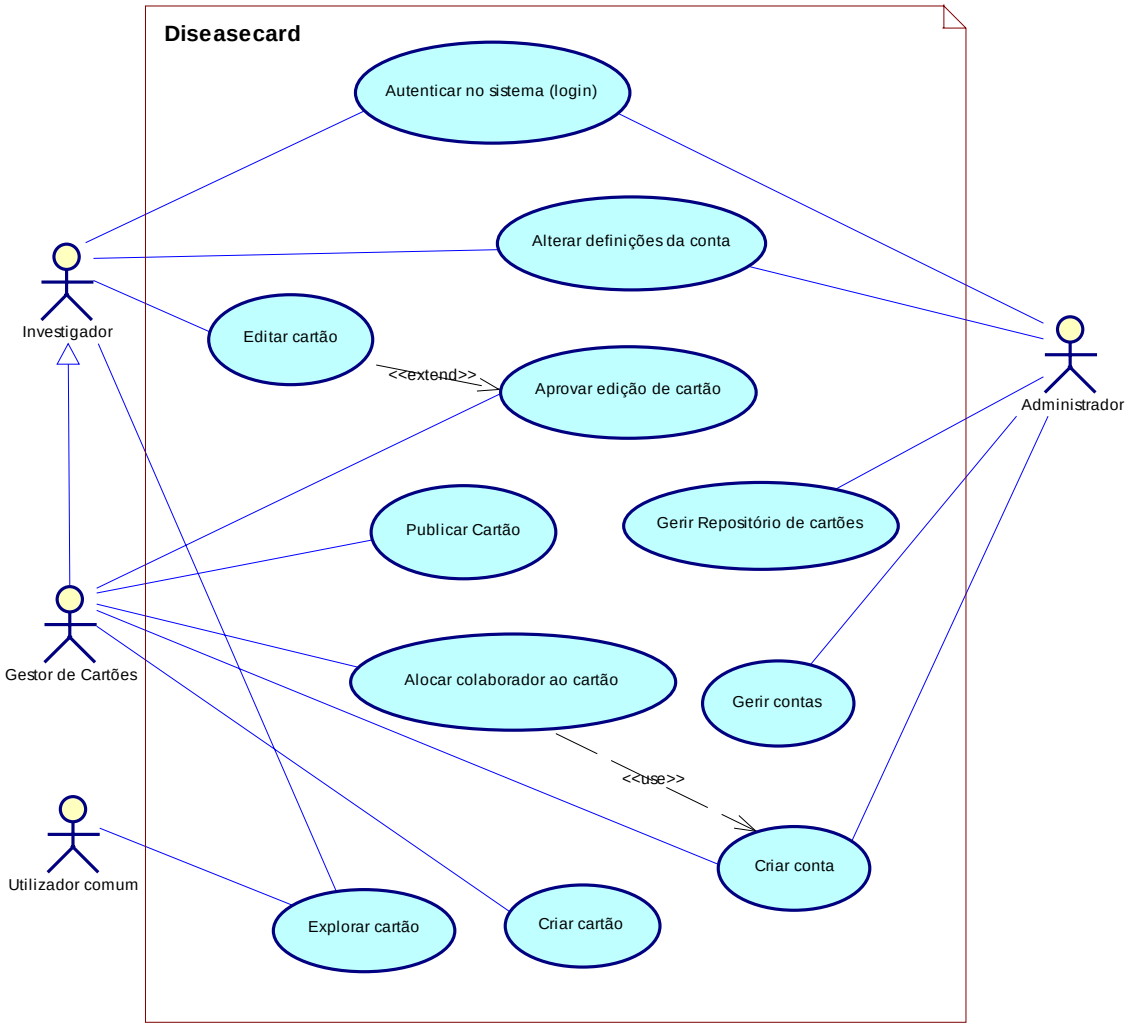


Figura 5.1 – Diagrama de casos de utilização da aplicação Diseasecard na versão 1.0.

Esta implementação assenta num repositório de "cartões" de doença cuja criação e desenvolvimento depende da investigação de conteúdos disponíveis em páginas *web* por um grupo de pessoas experientes na área. Para tal, esta solução comporta um sistema de autenticação de contas privadas e um repositório de dados estruturados – os cartões de doença. A interagir com este sistema existem quatro tipos de utilizadores, sendo eles os administradores da aplicação, os gestores de cartões, os investigadores e, por fim, os utilizadores comuns. Os administradores, ocupando o topo da hierarquia, encarregando-se de gerir as contas dos utilizadores bem como a colecção de cartões. Os gestores de cartões podem instanciar novos cartões e nomear outros utilizadores para colaborar no seu preenchimento com informação relativa a cada categoria do cartão. Outra competência importante dos utilizadores com este perfil é a avaliação e publicação do cartão. O utilizador nomeado pelo gestor de cartões assume o perfil de investigador, cabendo-lhe o

papel de reunir informação em torno do cartão para o qual foi solicitado. O investigador pode editar o cartão de doença, adicionar e remover *links* e informação. A publicação será da competência do gestor do cartão. A partir deste momento, a informação do cartão estará disponível para todos os utilizadores, sendo estes representados pelo perfil de utilizador comum. A estes é apenas permitida a consulta e exploração dos cartões publicados.

A Figura 5.2 ilustra o processo de produção de um cartão de doença, desde a sua instanciação por um gestor de cartões até à sua publicação. Quando der por finalizada a pesquisa, o investigador submete as suas alterações ao cartão de doença para serem avaliadas pelo gestor de cartões. Este, se achar pertinentes as alterações ou inserções do investigador, aprova o cartão, tornando-o público a partir deste momento. Caso não aceite as alterações, o cartão volta ao estado de desenvolvimento.

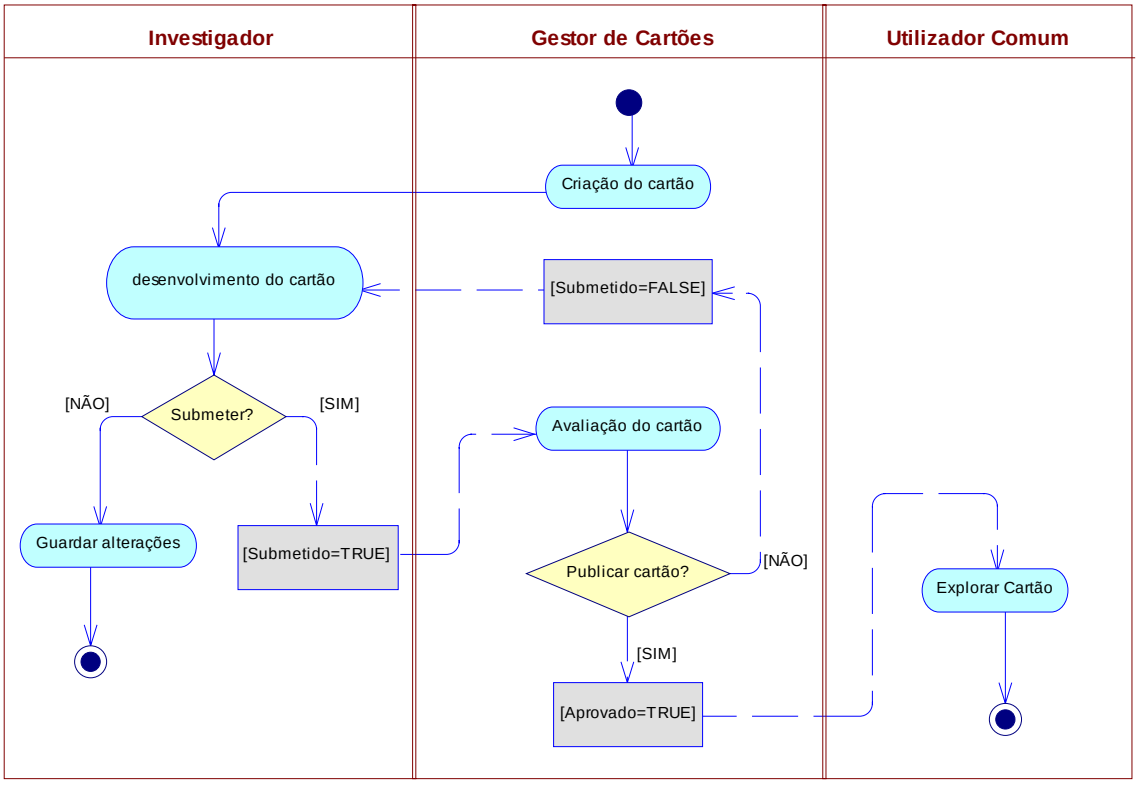


Figura 5.2 – Diagrama de actividade do processo de produção de um cartão de doença, distinguindo as competências dos diferentes utilizadores.

O passo seguinte do *Diseasecard* surgiu da constatação de que a informação da grande maioria das doenças raras pode ser obtida a partir de um conjunto bem definido de bases de dados públicas. Daqui se concluiu que a tarefa de explorar e navegar pelos recursos de dados na Internet, anteriormente delegada para os conjuntos de especialistas, pode ser

automatizada. Assim surgiu a versão (*Diseasecard 2.0*) a qual é ainda uma ferramenta colaborativa mas que permite a alternativa de criar cartões automaticamente. Foi criado para o efeito um módulo de *software* chamado *cardmaker*, incluído no *DiseaseCard* que, mediante um protocolo previamente estabelecido, percorre um conjunto de bases de dados, procurando e armazenando os *links* para as páginas com informação relevante para cada conceito (patologia, fármaco, genes, etc.). Apesar do *cardmaker* reunir informação num cartão de doença autonomamente, é ainda necessária a aprovação dos utilizadores especializados para que este fique visível ao público. Esta funcionalidade estava apenas disponível para os utilizadores autenticados no sistema ou seja, aqueles que podem reunir informação. Depois de melhoramentos significativos no pacote *cardmaker* constatou-se que a ferramenta apresentava bons resultados tanto em termos de conteúdo como de performance. Com efeito decidiu-se delegar totalmente esta tarefa ao sistema, desde a pesquisa e armazenamento dos cartões até à sua publicação (*Diseasecard 3.0*). Com esta abordagem a utilização do *Diseasecard* tornou-se mais intuitiva reduzindo significativamente os casos de utilização iniciais do sistema. Além disso a criação de cartões tornou-se totalmente transparente para o utilizador.

A decisão de se optar pela abordagem da criação automática de cartões não se deveu somente às consequentes vantagens das quais as mais importantes são a grande rapidez de criação de novos cartões e a não necessidade de pessoal especializado para reunir informação. Durante a fase colaborativa do *Diseasecard*, a criação de cartões de doença pelos colaboradores revelou-se uma tarefa morosa e complicada. Por um lado, o número de doenças raras catalogadas é muito grande, da ordem dos milhares, o que introduziria um atraso significativo na recolha e disponibilização da informação das doenças no *Diseasecard*. Por outro, a informação presente nas bases de dados externas é dinâmica, ou seja, a cada dia que passa, surge nova informação cuja actualização no sistema pelos colaboradores seria impraticável para uma lista de doenças tão extensa.

5.2 Componentes da Aplicação

Relativamente às versões mais antigas do sistema, a actual apresenta-se mais simples em termos de casos de utilização pelas razões apresentadas na secção anterior. A Figura 5.3 mostra as principais funcionalidades da aplicação.

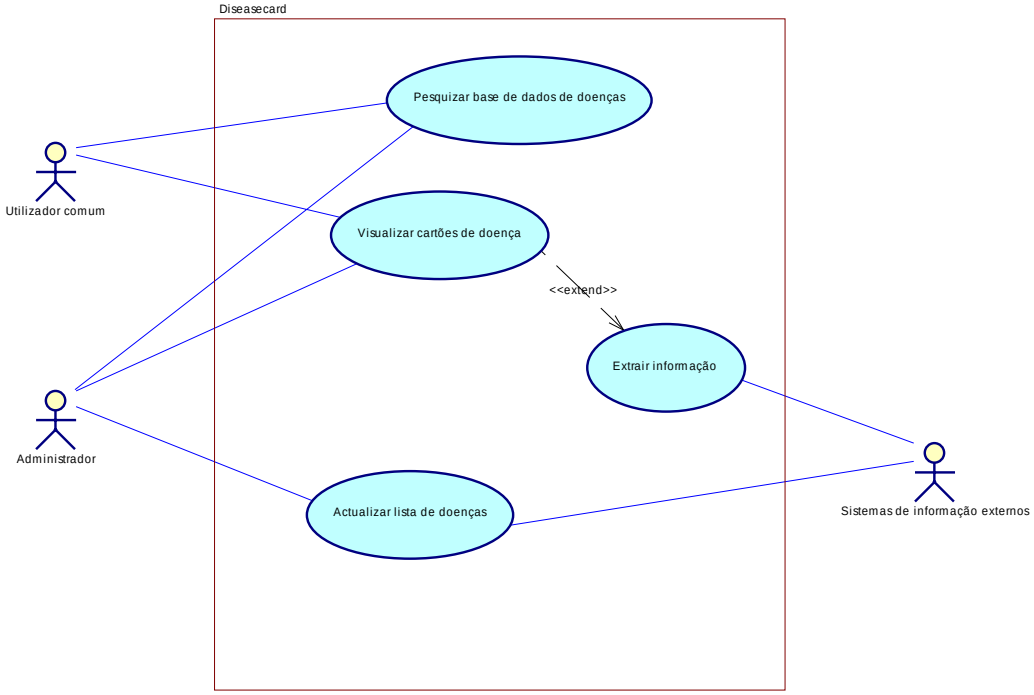


Figura 5.3 – Diagrama de casos de utilização da aplicação *Diseasecard*.

5.2.1 Descrição dos utilizadores e casos de utilização

Na actual implementação do *Diseasecard* existem três papéis diferentes em termos de perfis de utilização e de relacionamento com o sistema:

- **Utilizador Comum:** Este papel destina-se a todos os utilizadores que pretendem consultar e pesquisar informação de doenças raras presentes na base de dados do sistema. Neste perfil estão incluídos utilizadores profissionais da área da saúde (médicos de clínica geral, geneticistas, farmacologistas, etc.) como também os utilizadores em geral. A estes últimos poderá interessar-lhes a camada superior do protocolo conceptual dos cartões de doença (Figura 4.1) cujo conteúdo da informação está associado à descrição fenotípica da doença, os seus sintomas, laboratórios e centros clínicos, fármacos existentes etc. As camadas inferiores do protocolo são mais específicas dos utilizadores com experiência nas áreas da genética, genómica, etc.
- **Administrador/Gestor de dados**: Na presente implementação cabe a este utilizador gerir a lista de doenças presente na base de dados local. A principal tarefa consiste em ordenar um processo de actualização desta lista com base na informação de bases de dados externas.

- **Sistemas de informação externos:** Tratam-se de todas as fontes de dados públicas sobre as quais o sistema executa consultas e obtém informação para o preenchimento dos cartões de doença. Estes componentes participam nas acções do sistema na medida em que é deles que se extrai informação relevante no contexto do *Diseasecard*.

Para estes três tipos de utilizadores o sistema dispõe dos seguintes casos de utilização:

- **Pesquisar base de dados de doenças:** O utilizador preenche um formulário com uma expressão associada à doença que pretende pesquisar. Além da expressão, o utilizador pode também iniciar a pesquisa com base no símbolo (código) do gene ou no código OMIM da doença. Mediante a sua consulta o sistema pesquisa todas as doenças numa tabela armazenada na base de dados local retornando uma sub lista com doenças candidatas que o utilizador pode seleccionar. Além da pesquisa por palavra ou código, o utilizador pode também escolher a doença a partir da listagem completa organizada por ordem alfabética.
- **Visualizar cartão de doença:** Escolhendo previamente uma doença da lista, o utilizador consulta e visualiza toda a informação presente na base de dados do sistema relativa a essa doença. O sistema apresenta uma série de detalhes da doença estruturados num mapa visual de conceitos e numa árvore que contém *links* para as páginas das fontes de dados externas que contêm informação relevante.
- **Extrair informação:** O sistema explora uma colecção de bases de dados externas reunindo para cada uma delas informação que de alguma forma está associada. Esta informação é dividida em categorias ou conceitos. Esta acção fica completa quando a informação encontrada é armazenada na base de dados local sob a forma de um cartão de doença. O termo “extrair” não é o mais correcto para descrever este caso de utilização uma vez que este processo não extrai informação textual das fontes de dados mas antes identificadores e *links* directos para as páginas de Internet que contêm a informação. Este processo é transparente para o utilizador na medida em que a decisão de o executar depende unicamente da avaliação que o sistema faz ao seu pedido. Isto significa que quando o utilizador selecciona uma doença de uma lista, o sistema verifica se o cartão de doença que lhe está associado está

preenchido. O processo de extracção só é despoletado caso o cartão esteja vazio ou caso a sua informação esteja desactualizada.

- **Actualizar listas de doenças:** Mediante uma ordem de um agente com perfil de administrador, o sistema acede a uma lista de doenças genéticas presente no *site* ncbi/OMIM chamada *MorbidMap* descarregando-a via *ftp* e estruturando-a de acordo com o modelo de dados da base de dados local.

Destes casos de utilização, será dado um maior relevo à extracção de informação pois nele reside toda a funcionalidade de interpretação e navegação automáticas baseadas no protocolo anteriormente descrito. Assim, no capítulo seguinte é apresentado o processo de aquisição de informação e alguns detalhes da ferramenta desenvolvida para o executar.

5.3 Caracterização funcional do processo de aquisição

Nesta secção é apresentada a ferramenta de pesquisa e extracção de informação incluída no *Diseasecard*. Esta ferramenta foi concebida para ser modular ou seja, para ser programável e adaptável a diferentes contextos de exploração de acordo com um plano predefinido por cada utilizador ou aplicação, para diferentes interesses e, além disso, poder ser integrada como uma API num outro sistema. O objectivo deste sistema de um ponto de vista genérico centra-se na integração e ligação de informação disponível em páginas de Internet, associada a um dado contexto temático definido de acordo com os requisitos de um utilizador ou grupo de utilizadores que partilham interesse nesse mesmo contexto.

Sendo possível descrever o processo de exploração de conteúdos num protocolo de navegação e uma vez que para um dado contexto ou domínio de exploração, neste caso as doenças raras, as bases de dados a explorar são sempre as mesmas independentemente da consulta, foi desenvolvido um sistema que automatiza todo o processo de extracção e integração de informação, baseado na utilização de *web crawlers* controlados a partir de um protocolo configurável pelos utilizadores. Segue-se de seguida uma descrição do sistema desenvolvido para implementar esta funcionalidade bem como o seu enquadramento na aplicação *Diseasecard*.

Diseasecard é uma aplicação *web* de pesquisa, integração e apresentação de informação relativa a doenças genéticas raras, cuja arquitectura básica está ilustrada na Figura 5.4.

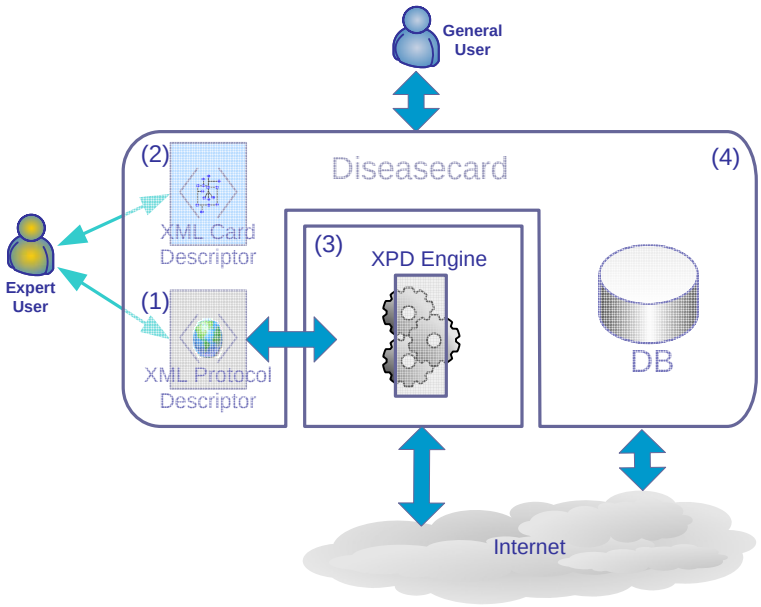


Figura 5.4 – Principais blocos funcionais do sistema de extracção e integração de informação proveniente de fontes de dados heterogéneas

Esta arquitectura está dividida em vários blocos funcionais. O *XML Protocol Descriptor* (XPD), bloco (1), representa um protocolo ou um conjunto de protocolos de navegação desenhados por utilizadores com conhecimentos nos diversos domínios de extracção. Com o intuito de tornar o sistema mais flexível e abrangente, estes protocolos são descritos com base numa sintaxe XML. *XML Card Descriptor* (XCD), bloco (2), é também um ficheiro XML que faz associar cada item extraído do protocolo de navegação *XPD* a um nó no mapa de representação gráfica do cartão de doença na página do *diseasecard*. *XPD Engine* (XPDE) (3) corresponde ao motor de busca de informação cujo processo de extracção está descrito nas instruções do *XPD*. Este módulo é responsável por aceder a fontes de dados públicas de uma forma dinâmica e organizada. Dado que este bloco tem como entradas ficheiros configuráveis pelo utilizador (protocolos *XPD*), toda a sua funcionalidade está contida num módulo permitindo assim a sua utilização e integração nos mais diversos cenários e aplicações. O bloco *Diseasecard* (4) corresponde à camada de aplicação que serve de interface entre o utilizador comum e o bloco funcional XPDE. A informação é extraída com base em consultas simples introduzidas pelo utilizador comum, que são cruzadas com as instruções e configurações especificadas no protocolo *XPD*. Depois de o XPDE obter toda a informação, processa todos os itens de acordo com o mapeamento especificado no ficheiro *XCD* criando finalmente um cartão de doença.

5.3.1 Protocolo de Extracção de Informação

Um factor importante no desenvolvimento desta abordagem de extracção de informação é que a navegação ao longo dos conceitos do protocolo é feita com base nos *hyperlinks* (URLs) contidos em cada página. O objectivo do sistema não é armazenar informação mas antes integrá-la num ponto único de acesso retornando para a camada de aplicação apenas um conjunto estruturado de *links* directos para as páginas *web* que contêm informação explorada pelo protocolo.

A Figura 5.5 ilustra o processo de extracção de informação através de uma rede de fontes de dados públicas. Cada bloco representa uma base de dados contendo vários itens que representam os conceitos para os quais se pode recolher informação. Alguns destes itens são pontos de acesso que permitem alcançar outras fontes de dados.

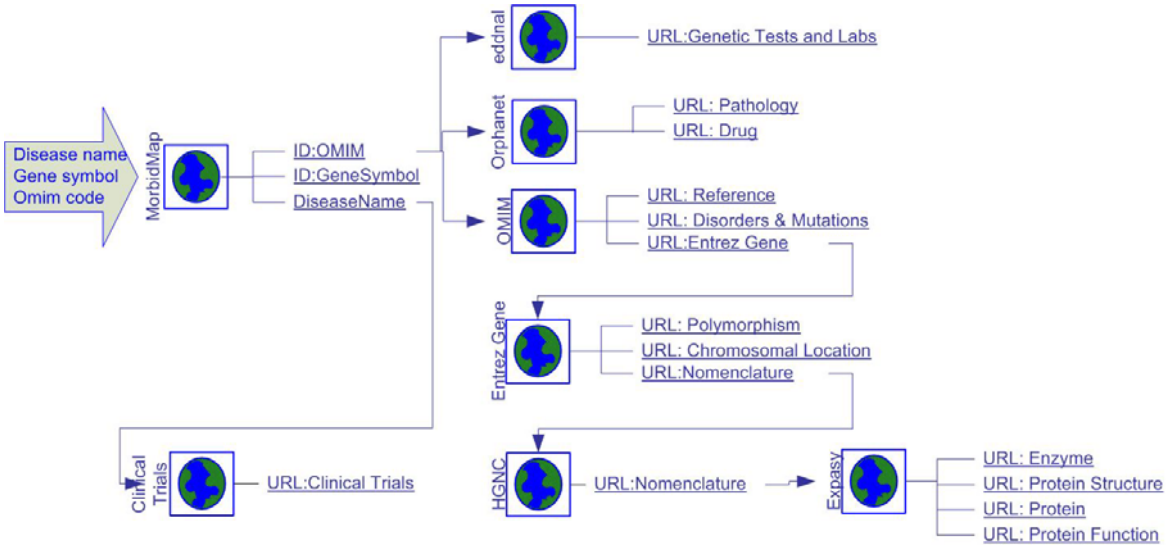


Figura 5.5 – Processo de extracção de informação descrito num protocolo para doenças genéticas raras. Cada caixa representa uma fonte *web* contendo itens com informação disponível. O motor XPDE inicia a pesquisa na fonte *MorbidMap* baseado numa palavra-chave que tanto pode ser uma palavra associada ao nome da doença, um gene ou um código *omim*. A partir daqui o motor pesquisa todas as fontes de dados presentes no protocolo varrendo todas as ligações entre elas.

A Figura 5.6 mostra uma representação completa do protocolo *XPD* utilizado actualmente no *Diseasecard* para produzir cartões de doenças. Os elementos do desenho com letras maiúsculas representam páginas *web* das várias fontes de dados exploradas pela aplicação. Os elementos com letras minúsculas representam os conceitos (categorias da informação) retirados que por sua vez, poderão dar acesso a outras fontes de dados como acontece por exemplo em *pathology*. Os conceitos cujo nome é iniciado por '_' não fazem parte da

estrutura final do cartão sendo de certa forma "invisíveis", servindo apenas como elementos auxiliares para aceder a fontes com dados relevantes.

Como se pode observar, este mapa está dividido em dois ramos não existindo qualquer interligação entre eles. Os conceitos *_gomim* e *_domim* são os pontos a partir dos quais se inicia a pesquisa de informação. Como já foi dito anteriormente, o utilizador inicia a pesquisa de uma doença com base numa palavra associada ao nome, num gene ou num código *omim*. Com base numa destas entradas e com base na tabela *morbidMap* (armazenada localmente no sistema *Diseasecard*) determina-se se existe uma doença associada e, caso exista, qual o seu código *omim*. Sabendo o código *omim*, o sistema inicia a pesquisa percorrendo todos os nós do protocolo.

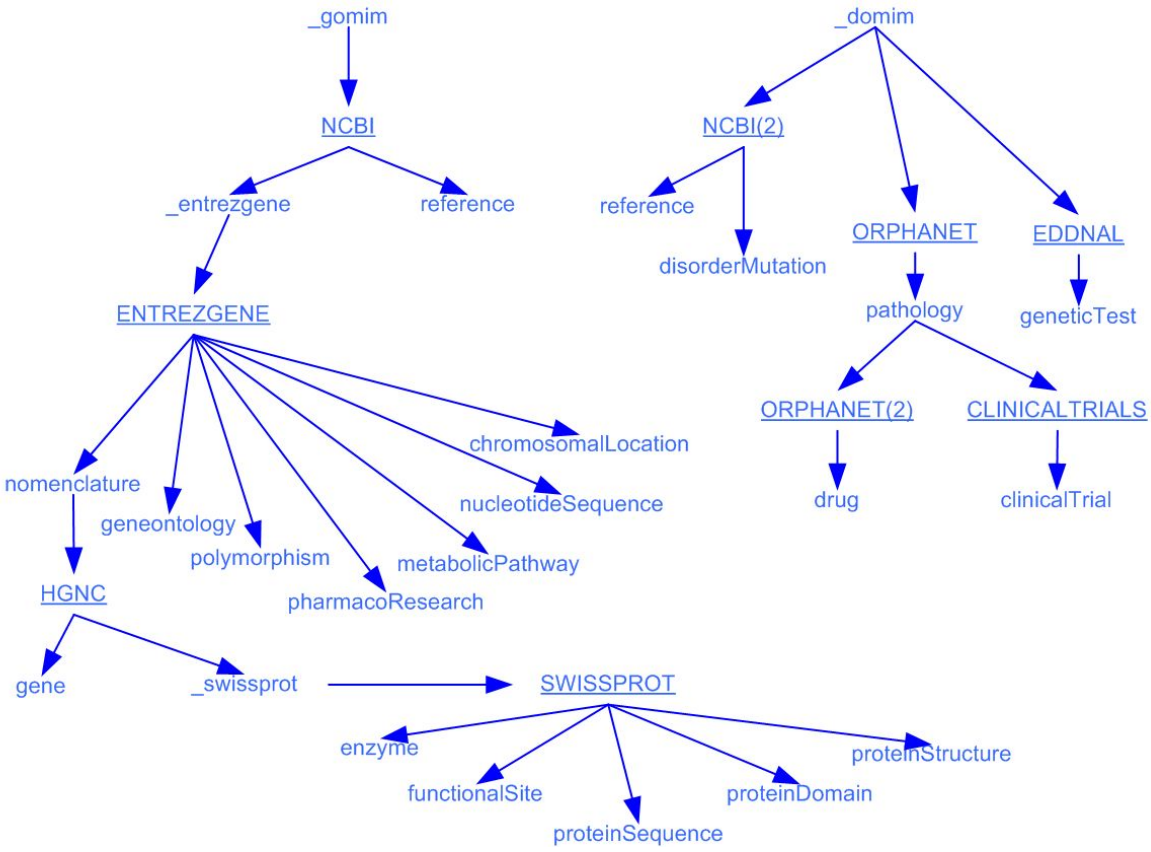


Figura 5.6 – Representação do mapa *XML Protocol Descriptor*. Os elementos com letras maiúsculas e sublinhados representam as fontes de dados na Internet e os elementos com letras minúsculas representam os conceitos ou categorias de informação que se obtêm de cada fonte de dados. Conceitos iniciados por '_' como por exemplo "*_entrezgene*" servem apenas de pontos de acesso a outras fontes de dados não aparecendo na representação final do cartão de doença.

Para se perceber a razão dos códigos *_gomim* e *_domim* convém salientar que, na perspectiva da aplicação, o código OMIM contém uma ambiguidade na medida em que

tanto pode representar uma doença genética (fenótipo) como apenas um gene [80]. Por essa razão se representa a nível do protocolo o código *omim* da doença por *_domim* e o de um gene por *_gomim*. Como exemplo temos o caso da doença *Achondroplasia* que segundo a tabela *morbidMap* contém dois códigos *omim* associados sendo eles o 100800 que representa o código da doença (prefixo #) e o 134934 (prefixo *) que representa o gene FGFR3. Assim sendo o primeiro corresponde a *_domim* no mapa da Figura 5.6, o qual produz informação respeitante à doença (referências bibliográficas, testes genéticos, testes clínicos, fármacos etc.) e o segundo, correspondendo a *_gomim* iniciará a pesquisa de informação do foro genético (nome do gene, sequência de nucleótidos, proteínas, enzimas, etc.).

A tabela *MorbidMap* é disponibilizada pelo *site* OMIM sob a forma de um ficheiro cujo aspecto é apresentado no extracto seguinte.

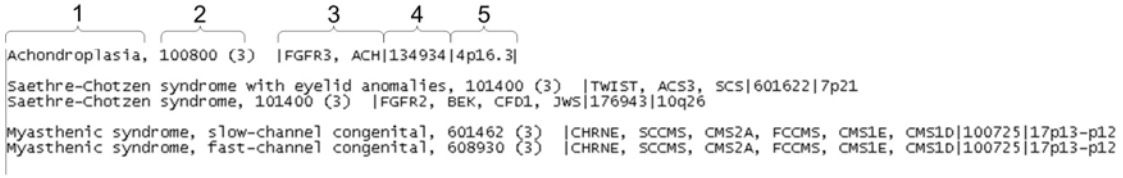


Figura 5.7 – Extracto da tabela *MorbidMap* retirada da base de dados OMIM.

Cada linha da tabela *MorbidMap* corresponde a uma doença ou fenótipo existente na base de dados OMIM. Com base no exemplo da Figura 5.7, a primeira linha define a doença/fenótipo *achondroplasia* (posição 1), cujo código *omim* da doença (*domim* no protocolo *XPD*) é 100800 (posição 2). A posição (3) contém o símbolo do gene associado à doença (FGFR3) e um sinónimo (ACH). A posição (4) contém o código *omim* do gene (*gomim* no protocolo XPD) e a posição (5) contém a localização citogenética do gene (a expressão 4p13.3 significa que o gene FGFR3 encontra-se no cromossoma 4; *p*: braço superior do cromossoma; 16.3: localização especifica no cromossoma). A Figura 5.8 ilustra a posição deste gene no cromossoma.

Como se pode ver nas doenças seguintes, a mesma doença pode estar associada a genes diferentes (um código *domim* corresponde a *n* códigos *gomim*) e o mesmo gene pode estar associado a doenças diferentes (um *gomim* para n *domim*). Tendo em conta os vários tipos de multiplicidade entre estas entidades, a aplicação *Diseasecard* extrai este ficheiro e

converte-o numa estrutura de dados que associa uma lista de doenças raras aos respectivos códigos. A tabela *MorbidMap* pode ser obtida no endereço descrito em [102].

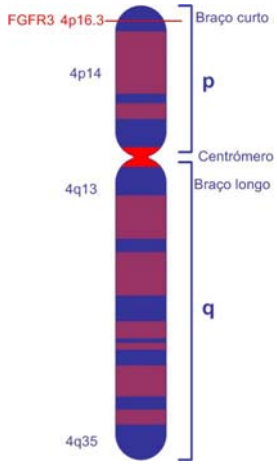


Figura 5.8 – Posição do gene FGFR3 no cromossoma 4.

O protocolo *XPD* é descrito utilizando a sintaxe XML. Neste documento o utilizador incumbido de especificar o protocolo de navegação descreve todo o *workflow* executado durante o processo de pesquisa. Basicamente esta tarefa reflecte todo o processo manual de um investigador ao percorrer uma série de páginas na Internet em torno das bases de dados propostas. A Figura 5.9 apresenta o esquema de dados para os protocolos *XPD*. Cada protocolo é escrito em XML e tem como raiz o elemento *protocol*, contendo uma colecção de elementos *wrapper*. Cada *wrapper* é caracterizado pelos seguintes elementos.

1. O nome da fonte de dados a explorar: *resource-name*;
2. O objecto de pesquisa: Pode ser um *hyperlink* para uma fonte de dados (*resource-url*) como um conceito obtido por outro *wrapper* (*follow*) previamente adquirido no processo de pesquisa. Quando o objecto de pesquisa é do tipo *resource-url* este é caracterizado por dois atributos além o *hyperlink*.
    a. *Key-origin*: É uma chave que identifica um dado registo de uma fonte de dados como por exemplo, código *omim*, número de acesso a uma proteína numa base de dados, símbolo de um gene etc. Este identificador é inserido no URL especificado em *resource-url* (ver Figura 5.11).
    b. *Request-method*: Especifica o método de *httprequest* a utilizar na pesquisa do URL. Os valores possíveis são *get* e *post*. A maioria dos *sites* aceita o

método *get*. Contudo existem alguns como o *EDDNAL* que funcionam somente em *post*.

3. *search-for*: Este elemento especifica um conjunto de termos de filtragem que servem para pesquisar e extrair informação relevante em cada página bem como o conceito ao qual a informação obtida é associada. Estes termos de pesquisa e de *matching* são baseados em expressões regulares [103].

    a. *regex*: Contém uma expressão regular para um padrão de *hyperlink* que se pretende encontrar na página.

    b. *put-into*: especifica o nome do conceito no qual será armazenada a informação encontrada na página.

    c. *concept-home*: Em alguns casos, o *hyperlink* obtido na pesquisa não está completo, ou seja, o URL é relativo não contendo o endereço do domínio. Noutros casos, o *wrapper* procura apenas um identificador o qual será posteriormente adicionado a um URL. Nestes dois casos, *concept-home* serve para especificar o URL completo. O atributo *url-encoded* serve para definir a codificação do URL. Na versão actual do *xprotocol* não é necessário utilizá-lo bastando defini-lo como *false*.

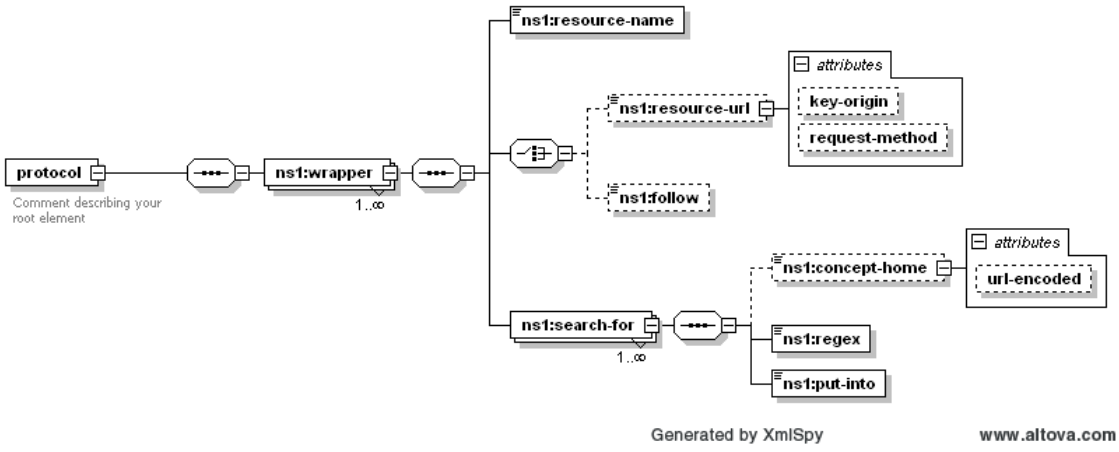


Figura 5.9 – Esquema de dados para os protocolos XPD.

Quando um protocolo é executado, cada elemento *wrapper* é interpretado por um *parser* que instancia em *run-time* um agente (*wrapper*) que extrai da informação na página. Os exemplos seguintes correspondem a excertos de um ficheiro XPD baseado no protocolo ilustrado na Figura 5.6, tendo em conta o esquema de dados da Figura 5.9.

```
<wrapper>
    <resource-name>NCBI</resource-name>
    <resource-url key-origin="_gomim" request-method="get">
        <![CDATA[http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=KEY&cmd=toc]]></resource-url>
    <search-for>
        <regex><![CDATA[$regex1$]]></regex>
        <put-into>_entrezgene</put-into>
    </search-for>
    <search-for>
        <regex><![CDATA[$regex2$]]></regex>
        <put-into>reference</put-into>
    </search-for>
</wrapper>
```

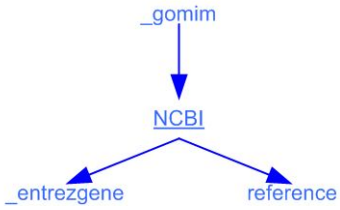


Figura 5.10 – Sub mapa do protocolo XPD correspondente ao wrapper NCBI.

Neste exemplo estão definidas as instruções de pesquisa para uma página OMIM (NCBI). Esta página corresponde a um dos pontos de entrada do protocolo. A pesquisa é feita no URL definido em *resource-url* e com base identificador *_gomim* que é inserido no URL na posição KEY (ver figura seguinte).

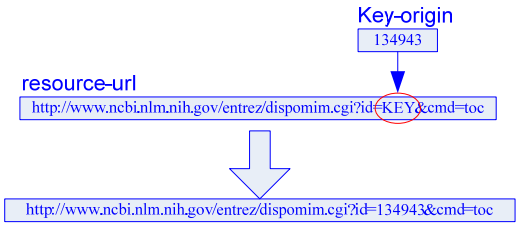


Figura 5.11 – O processo de substituição da chave KEY pelo valor de *key-origin*.

Convém relembrar que este identificador corresponde ao código *omim* relativo a um gene e que foi obtido a partir da tabela *morbidMap*.

Os elementos *search-for* contêm informação acerca dos parâmetros de pesquisa associados a um dado conceito. Neste caso, o *wrapper* irá procurar expressões ou URLs associadas aos conceitos *_entrezgene* e *reference* com base nas expressões de pesquisa contidas nos elementos *regex* (*Regular Expression*). As palavras *regex1* e *regex2* representam expressões regulares que, por serem muito extensas não são apresentadas nestes exemplos.

O próximo exemplo corresponde a seguinte extracto do mapa.

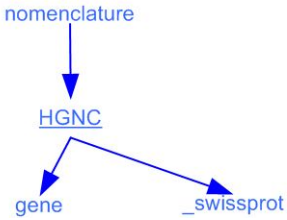


Figura 5.12 – Sub mapa do protocolo XPD correspondente ao *wrapper* HGNC

```
<wrapper>
        <resource-name>HGNC</resource-name>
        <follow>nomenclature</follow>
        <search-for>
                <regex><![CDATA[(http\:\/\/bioinfo\.weizmann\.ac\.il\/cards-bin\/carddisp\?(.*?))\"]]></regex>
                <put-into>gene</put-into>
        </search-for>
        <search-for>
                <regex><![CDATA[(http\:\/\/www\.expasy\.org\/cgi-bin\/niceprot\.pl\?(.*?))\"]]></regex>
                <put-into>_swissprot</put-into>
        </search-for>
</wrapper>
```

Este último caso difere do anterior por causa do elemento *follow*. Para aceder à informação associada aos conceitos *gene* e *_swissprot*, o *wrapper* utiliza directamente o URL extraído no conceito *nomenclature* especificado no excerto anterior. Como exemplo demonstrativo, suponhamos que o conceito *nomenclature* explorado anteriormente contém o URL *http://www.nomencalture.gov/query.fcgi?list_uids=**2261***. Utilizando o elemento *follow* o *wrapper* utilizaria este URL inteiro para aceder à página.

A pesquisa por padrões é processada através de *regular expressions*. O elemento *regex* define o padrão a pesquisar. Segundo as regras de criação de protocolos *XPD*, cada padrão tem de estar obrigatoriamente dividido em dois grupos. A Figura 5.13 ilustra um exemplo de uma expressão no formato correcto. O *wrapper* vai pesquisar um URL que tenha este padrão. Caso o encontre, o *toolkit* de *Regular Expressions* pode retornar a expressão completa ou dividida em grupos. Os parênteses na expressão indicam como será dividida.

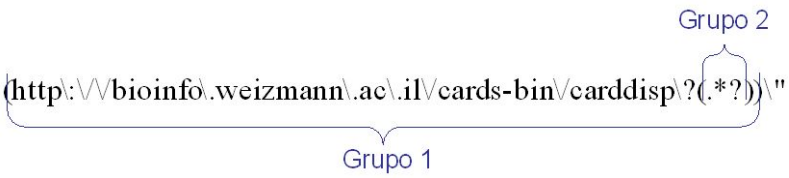


Figura 5.13 – Formato do padrão de pesquisa no elemento *regex* no protocolo XPD.

Assim sendo, o Grupo 1 retorna a expressão inteira e o Grupo 2 retorna apenas a sub expressão que estiver entre o primeiro "?" e as aspas. No contexto do *diseasecard* o Grupo

1 é o URL para a página *web* com informação para o conceito "*_swissprot"* e o Grupo 2 será utilizado como *caption* e como identificador (ID) do conceito.

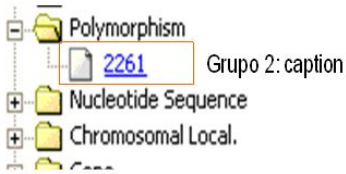


Figura 5.14 – O Grupo 2 da expressão regular será utilizado como legenda no cartão de doença.

5.3.2 Protocolo de descrição de um cartão de doença

A informação extraída com base do protocolo XPD retornada pelo *xprotocol* é devolvida à camada superior da aplicação sob a forma de uma colecção de objectos (conceitos). Cada conceito é definido por um nome (enumerado nos elementos *put-into* do protocolo XPD) e por uma colecção de *itens*. Cada item corresponde a um elemento de informação extraída por um *wrapper* e que consiste no URL da página encontrada, numa descrição e num identificador. O URL e o identificador provêm respectivamente do Grupo1 e Grupo2 do padrão de pesquisa especificado no elemento *regex* do protocolo XPD (ver Figura 5.14).

Tendo esta informação disponível o passo seguinte é processá-la e armazená-la sob a forma de um cartão de doença. Os conceitos dos cartões de doença estão interligados sob a forma de um grafo. A associação de cada conceito à respectiva posição no mapa bem como a interligação entre conceitos são definidas no ficheiro *XML Card Descriptor* (XCD) ilustrado na Figura 5.4. Esta informação é importante apenas para representar os cartões de doença sob a forma de uma interface gráfica que facilite de alguma forma a pesquisa do utilizador. Assim sendo, cada conceito de um cartão está definido na base de dados como sendo um nó de um grafo possuindo informação relativa à sua posição espacial (coordenadas x,y), nós (conceitos) a que está ligado etc.

A Figura 5.15 mostra os dois tipos de representação de um cartão de doença apresentados no *Diseasecard*. À esquerda está representada a estrutura em árvore em que o símbolo da "pasta" corresponde aos conceitos e a "folha" corresponde à informação associada (URL para uma página *web*). À direita está apresentado o grafo de conceitos com as suas interligações. Apesar de representarem os mesmos conceitos, estas duas vistas apresentam perspectivas complementares para a mesma informação. Enquanto que a primeira vista lista os conceitos do cartão com os respectivos *links*, o grafo apresenta as relações lógicas

entre conceitos, ou seja, o fluxo conceptual que permite “navegar do sintoma ao gene” e “da proteína ao fármaco”, o mesmo apresentado no capítulo anterior (modelo de navegação Figura 5.17).

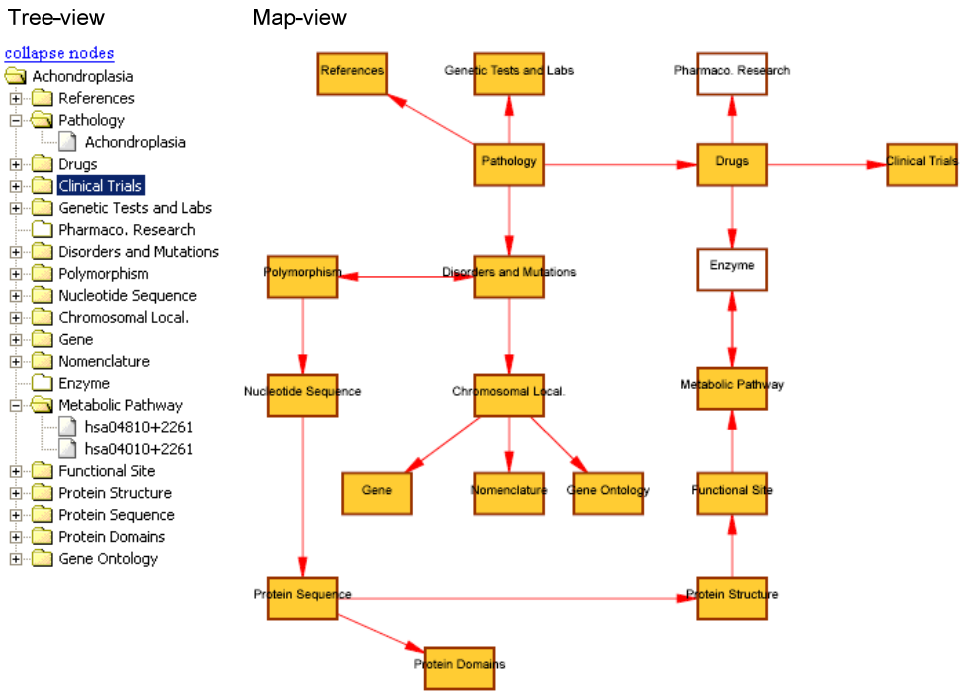


Figura 5.15 – Aspecto de um cartão de doença na aplicação *Diseasecard*. Além a informação relevante, cada conceito contém ainda informação sobre a sua posição no mapa.

Cada objecto nas duas vistas do cartão é representado como um nó ou seja, tanto um conceito como a ligação entre dois conceitos são definidas como um nó. Contudo cada um possui algumas características distintas.

A figura seguinte mostra o esquema de dados do protocolo *XCD* actualmente utilizado no *Diseasecard*.

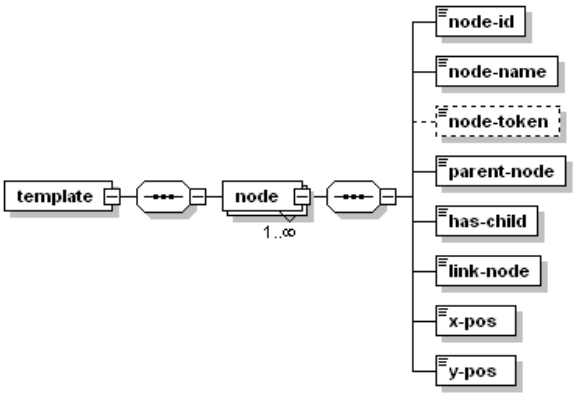


Figura 5.16 – Esquema de dados de um protocolo *XCD*.

Como se pode observar, cada nó é definido com uma série de elementos. Esta estrutura é suficiente para caracterizar todos os tipos de nós das duas vistas.

1. *node-id*: Número de ordem que identifica o nó;
2. *node-name*: Nome do nó. Este nome corresponde à designação do conceito tanto na vista em árvore como na vista em mapa de nó;
3. *node-token*: Este é o elemento que indica qual o conceito extraído pelo *xprotocol* que este nó representa. É este elemento que estabelece a relação entre o protocolo de extracção *XPD* e a sua representação gráfica no cartão de doença.
4. *parent-node*: Indica o nó pai com o *node-id* correspondente. Quando o nó representa um conceito então *parent-node* é igual a 0. Quando é diferente de 0 representa uma ligação entre dois conceitos e o valor de *parent-node* indica o conceito origem;
5. *has-child*: Indica se o nó tem nós filhos. Assume os valores *true* ou *false*;
6. *link-node*: Indica o *node-id* do nó ao qual este se liga. Este elemento é utilizado para a representação em mapa;
7. x-pos e *y-pos*: Coordenadas x, y do nó na representação em mapa.

A Figura 5.17 apresenta um extracto de um cartão de doença para o qual se vão definir os nós.

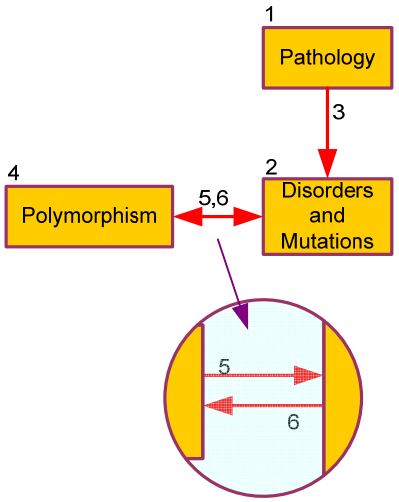


Figura 5.17 – Extracto do mapa de conceitos da Figura 5.15.

```
<node>
        <node-id>1</node-id>
        <node-name>Pathology</node-name>      <!—o nome que aparece na caixa-->
        <node-token>pathology</node-token>    <!— associação ao wrapper pathology-->
        <parent-node>0</parent-node>  <!—o valor 0 indica que este nodo é um conceito-->
        <has-child>true</has-child>   <!—Este nodo tem filhos-->
        <link-node>0</link-node>
        <x-pos>35</x-pos>
        <y-pos>18</y-pos>
</node>
```

Este nó (caixa nº1 da Figura 5.17) refere-se ao conceito *Pathology* e irá armazenar informação gerada no *xprotocol* pelo *wrapper pathology* (ver mapa da Figura 5.6).

```
<node>
        <node-id>2</node-id>
        <node-name>Disorders and Mutations</node-name>
        <node-token>disorderMutation</node-token>
        <parent-node>0</parent-node>  <!—o valor 0 indica que este nodo é um conceito-->
        <has-child>true</has-child>   <!—Este nodo tem filhos-->
        <link-node>0</link-node>
        <x-pos>35</x-pos>
        <y-pos>34</y-pos>
</node>
```

Este nó (caixa nº2 da Figura 5.17) refere-se ao conceito *Disorders and Mutations* e irá armazenar informação gerada no *xprotocol* pelo *wrapper disorderMutation* (ver mapa da Figura 5.6).

```
<node>
        <node-id>3</node-id>
        <node-name>Disorders and Mutations</node-name>
        <parent-node>1</parent-node>
        <has-child>false</has-child>
        <link-node>2</link-node>
        <x-pos>-1</x-pos>
        <y-pos>-1</y-pos>
</node>
```

Este nó (seta nº3 da Figura 5.17) refere-se a uma ligação do conceito *Pathology* (*parent-node:1*) para o conceito *Disorders and Mutations* (*link-node:2*).

Destes exemplos é fácil concluir que existem algumas regras que permitem ao sistema actual identificar o tipo de nó mediante os valores dos seus elementos. Estas regras são representadas na Tabela 5.2 e na Tabela 5.3.

Tabela 5.2 – Regras para a definição de um conceito.

Conceito	
node-id	[número inteiro]
node-name	[palavra ou expressão]
parent-node	“0”
has-child	“true”
link-node	“0”
x-pos	[número inteiro]
y-pos	[número inteiro]

Tabela 5.3 – Regras para a definição de uma ligação entre dois conceitos.

Ligação	
node-id	[número inteiro]
node-name	[palavra ou expressão] Por defeito pode assumir o nome do nodo destino

parent-node	[número inteiro correspondente ao nodo base (conceito)]
has-child	"false"
link-node	[número inteiro correspondente ao nodo alvo]
x-pos	"-1"
y-pos	"-1"

Nota: Os valores entre [] são variáveis de acordo com o tipo de nó e os valores entre "" são fixos dependendo do tipo de nó.

5.3.3 O processo de extracção de informação

Uma vez mapeado todo o processo de pesquisa num protocolo *XPD* é necessário um motor que interprete a linguagem do protocolo *XPD*, execute as suas instruções e que controle todo o processo de busca de informação. *XPD Engine* (bloco *XPD Engine* na Figura 5.4) é o bloco funcional responsável por estas operações. A Figura 5.18 apresenta em detalhe os elementos mais importantes deste bloco.

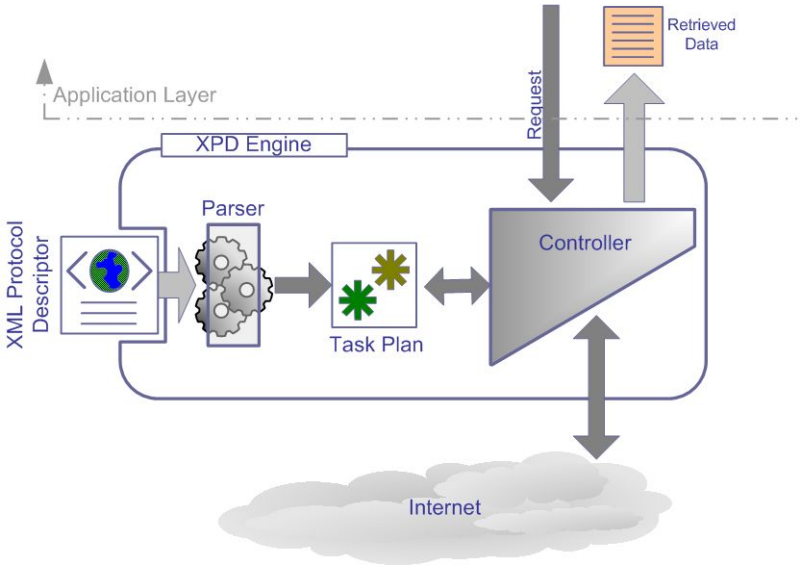


Figura 5.18 – Arquitectura do bloco funcional *XPD Engine*. Este módulo extrai informação tendo por base o protocolo *XPD* e um pedido (*token*) vindo da camada *Application (Diseasecard)* que por sua vez o recebe do utilizador (ver Figura 5.4). O bloco *parser* converte as instruções do protocolo *XPD* num conjunto de tarefas/wrappers (*Task Plan*). Através de *multithreading*, o *Controller* gere os *wrappers* que exploram páginas *web* e extraem informação relevante. Os dados resultantes são retornados à camada superior.

O *XPD Engine* extrai informação nos recursos *web* com base num pedido vindo da camada *Application (Diseasecard)* (que por sua vez o recebe do utilizador) e com base no protocolo descrito no ficheiro *XPD*. Um *parser* converte em tempo real a linguagem do *XPD* num conjunto de tarefas (objectos) designado por *TaskPlan*. Estas tarefas serão posteriormente solicitadas pelo bloco *Controller* para instanciar dinamicamente agentes de

pesquisa em páginas *web* (*wrappers*). Com base no *TaskPlan* o *Controller* gere os *wrappers* recolhendo a informação por eles extraída e retornando-a à camada superior.

O *Parser* como já foi referido, converte as instruções descritas no ficheiro *XPD* numa colecção de objectos (Figura 5.19). Para cada elemento *wrapper* especificado no *XPD* é gerado um objecto da classe *Task*. Este objecto pode instanciar *wrappers* mediante solicitação do *Controller*. Durante a operação de *parsing* todos os objectos *Task* são incluídos numa estrutura chamada *TaskPlan*.

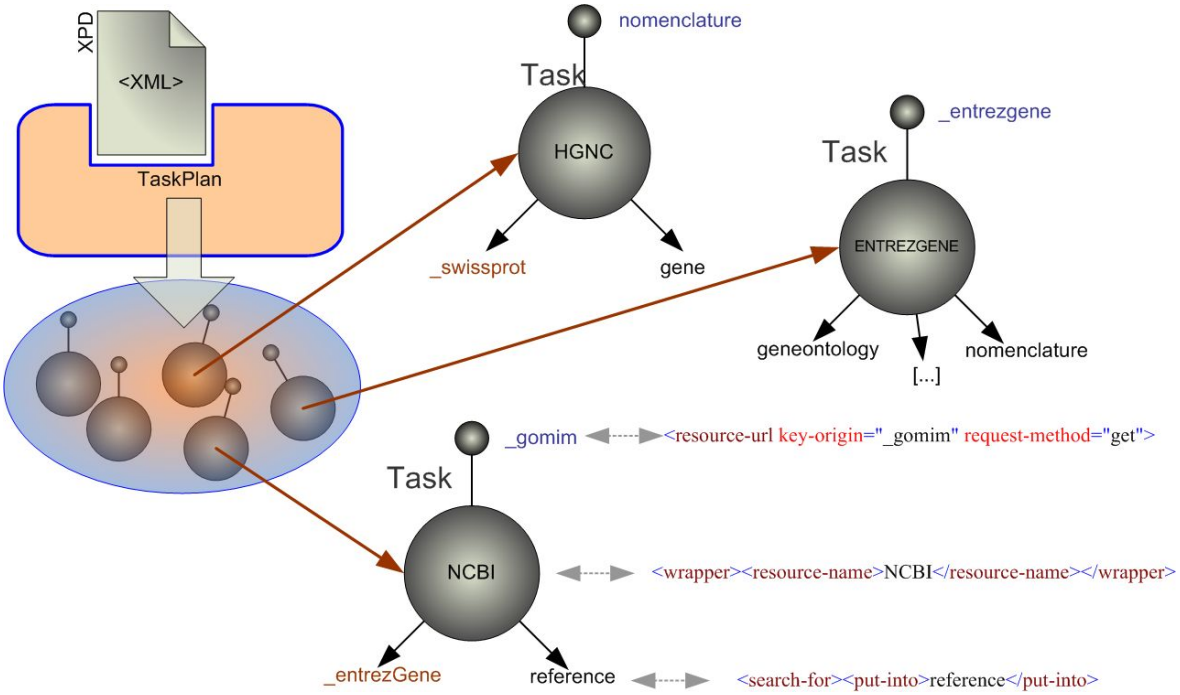


Figura 5.19 – Representação gráfica da conversão do protocolo definido num ficheiro *XPD* numa colecção de tarefas. O *TaskPlan* armazena a colecção de tarefas (*Task*) que são geradas pelo *parser* a partir do protocolo *XPD*. Cada tarefa instancia *web wrappers* que, quando invocados, são responsáveis pela extracção de informação nas páginas *web* definidas no protocolo.

A Figura 5.19 ilustra algumas tarefas geradas pelo *XPD*. Cada tarefa instancia um ou mais *web wrappers*. Os *wrappers* criados pela tarefa *NCBI* são iniciados pelo *token* *_gomim*. Finalizada a sua tarefa de extracção, o *wrapper NCBI* retorna informação que, além de produzir conceitos para o cartão de doença, servirá também de entrada para os próximos *wrappers* a actuarem, ou seja, aqueles que responderem aos *tokens* *_entrezGene* e *reference*. Assim sendo, o conceito *reference* servirá de *token* para o *wrapper Reference* e o conceito *_entrezGene* servirá de *token* para o *wrapper EntrezGene* (ver Figura 5.6).

O *Controller* é o módulo que gere e invoca todas as tarefas do plano de extracção. Inicia o processo de extracção com base num ou mais *tokens* enviados pela camada *Application*

que, no caso do protocolo *XPD* actual, correspondem aos dois códigos *omim* extraídos da tabela *morbidMap* (*_gomim* e *_domim* conforme a ilustração da Figura 5.6). O *Controller* recebe de entrada a colecção *TaskPlan* de todas as tarefas do protocolo *XPD*. No início do processo o *Controller* invoca o *TaskPlan* a procurar todas as tarefas que respondam aos *tokens* iniciais (*_gomim* ou *_domim*). A tarefa (*Task*) que satisfaz o pedido ao *token* *_gomim* é neste caso NCBI. Com base nesta tarefa o *TaskPlan* gera um *wrapper* e um *thread* retornando-os ao *Controller*. Este *thread* será utilizado pelo *Controller* para executar o *wrapper* assincronamente de forma a optimizar o processo. As figuras seguintes ilustram o processo de extracção tendo em conta as configurações do protocolo *XPD* apresentado na Figura 5.6. Para efeitos de simplificação dos diagramas, o processo abaixo ilustrado (Figura 5.20) refere-se apenas aos dois primeiros nós do protocolo (NCBI e ENTREZ-GENE).

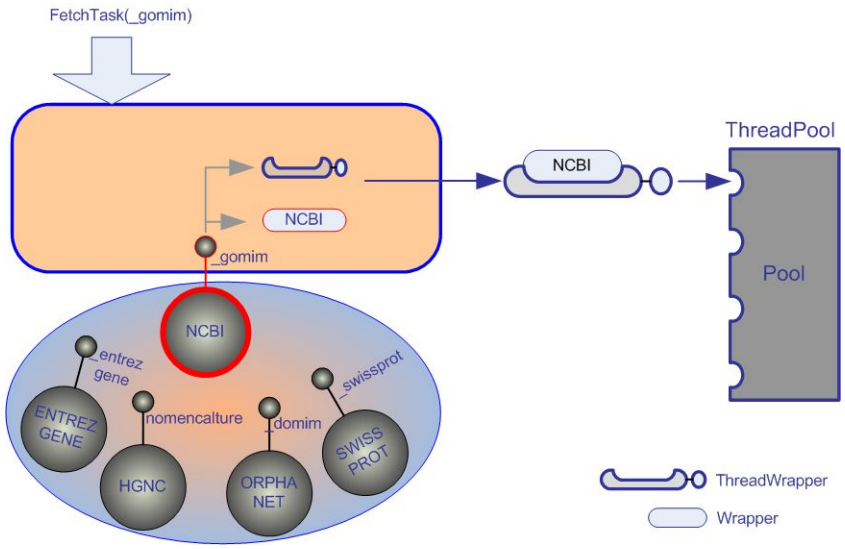


Figura 5.20 – O *TaskPlan* procura por tarefas que respondam ao *token* *_gomim*. A partir das tarefas retornadas gera um *wrapper* e um *thread* que serão lançados num *threadpool*.

Quando o *thread* é lançado, o *wrapper* é executado explorando a página *web* e extraindo informação para os conceitos que contém. A informação dos conceitos é retornada ao *controller* e, além de produzir os respectivos conceitos para o cartão de doença, será utilizada para despoletar os próximos *wrappers*.

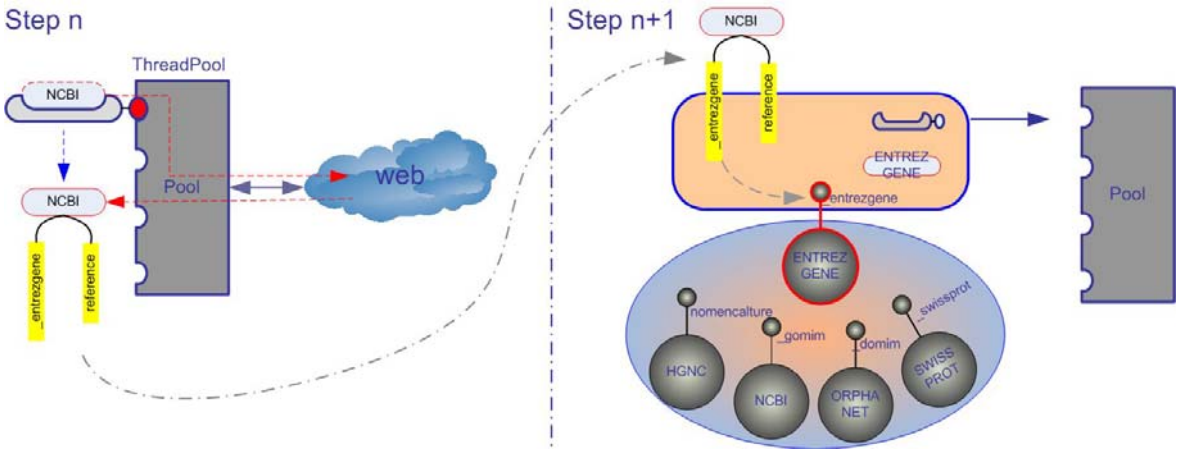


Figura 5.21 – O *wrapper NCBI* é lançado pelo *threadpool*. Este *wrapper* extrai informação associada aos conceitos *_entrezgene* e *reference*. Estes dois itens serão utilizados para continuar o processo. Agora o *TaskPlan* procurará tarefas que respondam às chaves *_entrezgene* e *reference*. Neste caso, só uma tarefa é retornada (*ENTREZGENE*). O processo repete-se em cascata até ao fim do protocolo.

O bloco *Pool* (*ThreadPool*) é uma classe implementada pelo *Controller* que gere o lançamento dos *threads* que contêm os *wrappers*. Desta forma é possível executar vários *wrappers* em simultâneo resultando daí um aumento da performance em termos de duração do processo de extracção.

5.4 Tecnologias de Desenvolvimento

Do ponto de vista físico, a aplicação *Diseasecard* está organizada de acordo com a Figura 5.22. Trata-se de uma aplicação *web* desenvolvida em JSP (*Java Server Pages*), Java e Struts [104] sobre uma base de dados *SQL Server 2000*.

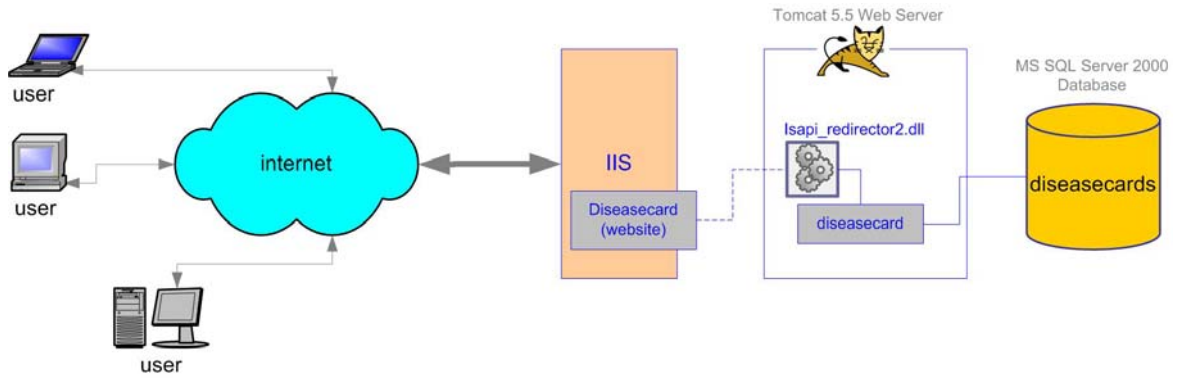


Figura 5.22 – Arquitectura física da aplicação Diseasecard.

O Diseasecard é uma aplicação *web* disponível na Internet através do endereço http://www.diseasecard.org. É composta por uma colecção de páginas JSP e vários pacotes de classes Java.

Apesar da aplicação Diseasecard estar contida num servidor *Tomcat*, é disponibilizada para o exterior a partir de um servidor *Microsoft* IIS (*Internet Information Services*) por razões de portabilidade e integração de todas as aplicações do grupo de bioinformática. A interligação entre o servidor IIS e o *Tomcat* (servidor que contém a aplicação Diseasecard) é realizada por uma *DLL* de redireccionamento (isapi_redirector2.dll) que está localizada no Tomcat.

A utilização do Struts [104] deve-se às suas vantagens na estruturação e organização das aplicações. Trata-se de uma *framework* aberta pertencente ao projecto *Jakarta* [105] que auxilia a construção de aplicações web. É uma solução construída sobre uma camada de controlo flexível baseada em tecnologias Java Servlets, JavaBeans, ResourceBundles e XML. A simples utilização de páginas JSP leva a que se misturem os vários níveis da arquitectura da aplicação, ou seja, misturam-se o código de acesso à base de dados, o código do desenho das páginas e o código de controlo da lógica da aplicação, o que, para aplicações complexas, torna a sua manutenção bastante complicada.

O Struts surge para dar solução a este problema na medida em que sugere o desenvolvimento das aplicações baseando-se numa arquitectura MVC (*Model-View-Controller*). Deste modo, a estrutura da aplicação é dividida em três tipos de competências: o modelo, as vistas e o controlo. O modelo representa a camada relativa à base de dados e ao código de acesso. As vistas representam a camada de apresentação em termos das páginas e o respectivo design. Por fim, o controlo representa o código que gere o fluxo de navegação da aplicação.

A utilização desta arquitectura veio simplificar significativamente o desenvolvimento da aplicação *Diseasecard*, tanto a nível da redução do número de páginas como a nível da sua legibilidade. A Figura 5.23 representa de um modo gráfico os componentes principais do *Diseasecard* na perspectiva da arquitectura *Struts*. Basicamente este diagrama representa todas as páginas web e as classes Java que as processam.

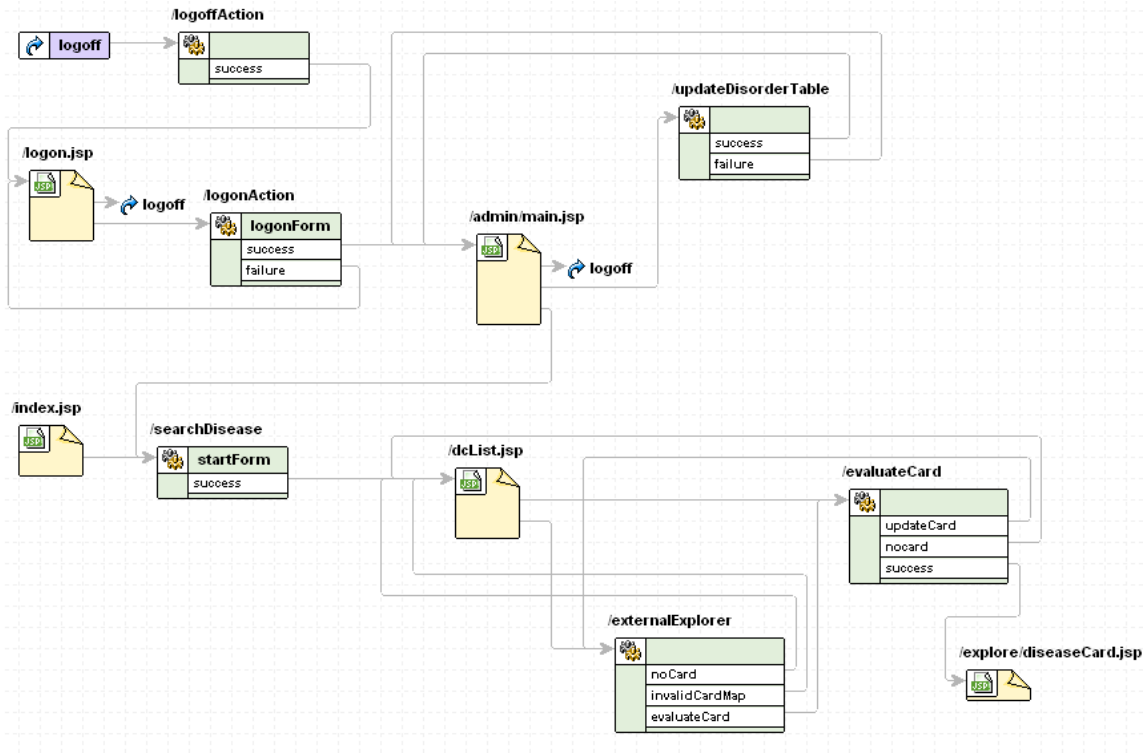


Figura 5.23 – Representação visual da aplicação Diseasecard em termos de arquitectura *Struts*.

De um modo simplificado e para perceber a dinâmica geral do funcionamento desta aplicação no âmbito do *Struts* é suficiente referir os blocos principais do diagrama e o respectivo significado. A Figura 5.24 distingue as páginas JSP comuns das classes do tipo *Action*.

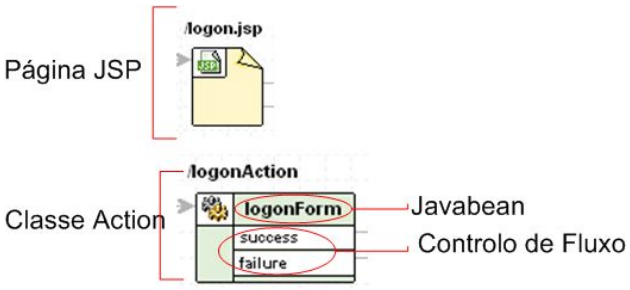


Figura 5.24 – Legenda da simbologia utilizada na Figura 5.23

Basicamente, para aceder à aplicação Diseasecard, existem duas páginas de entrada: *logon.jsp* e *índex.jsp*. A primeira dá acesso a diversas operações de administração tendo para isso o utilizador de se autenticar no sistema. Assim, depois de preencher um formulário com os dados *login/password*, a operação de autenticação é delegada para a classe *logonAction*. O objecto *logonForm* é um *javabean* que confere persistência aos dados de autenticação do utilizador. Mediante o resultado da autenticação, o fluxo de

controlo é dirigido ou para uma página de menu (*admin/main.jsp*) ou, na condição de falha, para a página *logon.jsp* onde é apresentada uma mensagem de erro de autenticação.

Na outra página de acesso, *índex.jsp* o acesso não é submetido a autenticação. Neste caso é apresentado o formulário de início da pesquisa de doenças raras no qual o utilizador introduz uma palavra-chave, o símbolo de um gene ou um código omim de cuja doença ele pretenda obter informação. A classe *searchDisease* processa o pedido do utilizador procurando na base de dados do sistema nomes de doença que, de alguma forma se associem ao pedido inicial. Estando esta pesquisa terminada, a classe *searchDisease* retorna como resultado a lista de doenças para a página *dcList.jsp*. Cada item da lista corresponde a uma doença. Contudo isto não significa que já exista um cartão (de doença) criado e preenchido para essa doença. Assim, o fluxo da aplicação será dirigido para *evaluateCard* caso o cartão já exista para a doença escolhida em *dcList.jsp* ou, para *externalExplorer* caso o cartão ainda não tenha sido criado. A Figura 5.25 ilustra a sequência das principais actividades envolvidas neste processo.

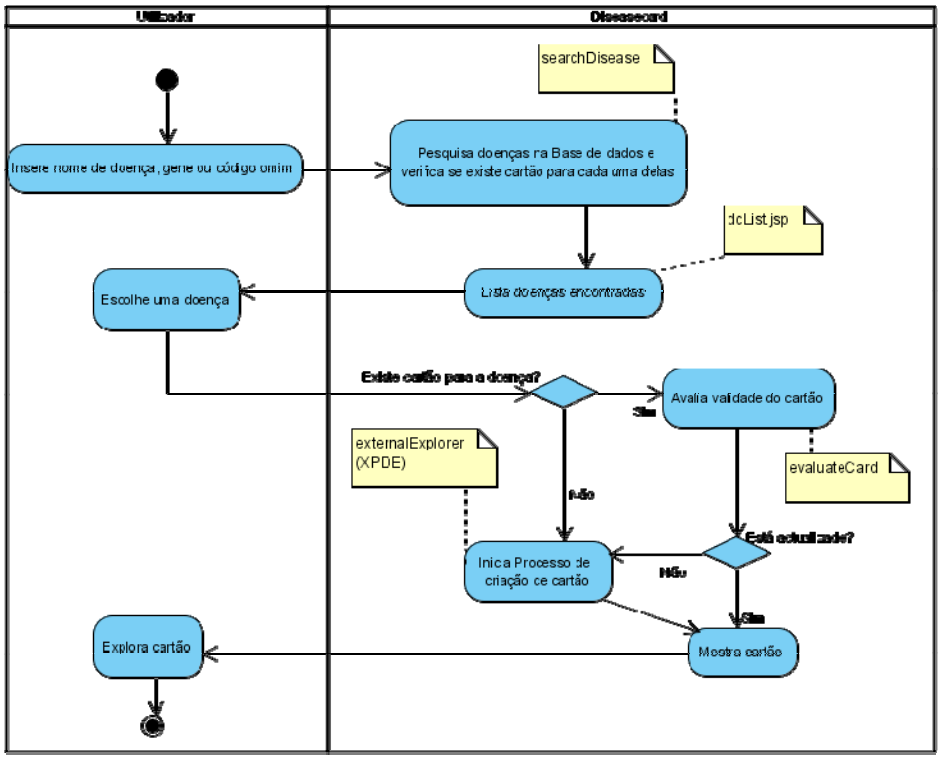


Figura 5.25 – Diagrama de actividade da tarefa de pesquisa de um cartão de doença e respectivo processamento do pedido.

Como se pode observar na Figura 5.25, a actividade que processa a criação de um novo cartão está plenamente associada à funcionalidade do bloco XPDE, cujo funcionamento é

abordado no capítulo 5.3. Como também se pode observar, este processo é despoletado em duas situações concretas: quando o cartão procurado pelo utilizador ainda não foi criado e consequentemente não existe na base de dados ou quando o cartão já existe mas a sua informação encontra-se desactualizada. Neste contexto, um cartão desactualizado significa que, desde a sua criação ou anterior actualização, já se passou um dado período de tempo predefinido.

5.4.1 Modelo de dados da aplicação

A base de dados do *Diseasecard* divide-se actualmente em duas secções: Aplicação de gestão de cartões e Base de dados de doenças.

A primeira é relativa aos dados directamente associados aos cartões de doença que o sistema armazena ou seja, informação contida nos cartões, a sua estrutura e informação dos utilizadores autenticados. A segunda é relativa à lista de doenças importada periodicamente da base de dados *OMIM* (*MorbidMap*). Apesar destes dois grupos residirem na mesma base de dados, podiam pertencer a sistemas separados. A sua separação faz sentido na medida em que a camada de gestão dos cartões pode, além das doenças raras, ser utilizada noutros contextos. Assim sendo, num projecto futuro, caso se pretenda expandir o paradigma do Diseasecard para outros cenários, as tabelas de gestão de cartões ficam localizadas numa base de dados diferente das bases de dados específicas de cada cenário.

A Figura 5.26 representa o modelo conceptual da base de dados do *DiseaseCard*, relativo à parte da gestão de cartões. Como se pode confirmar na figura, a estrutura deste modelo foi herdada das primeiras versões da aplicação *Diseasecard*, uma vez que contém informação orientada ao utilizador (entidades *User* e *Level*).

Cada utilizador pode ter associada uma colecção de cartões (entidade *Card*). Cada cartão é constituído por uma estrutura de nós ou conceitos (entidade *Node*) que, por sua vez, contêm informação útil, sob a forma de ligações a páginas de bases de dados biomédicas (entidade *NodeLeaf*).

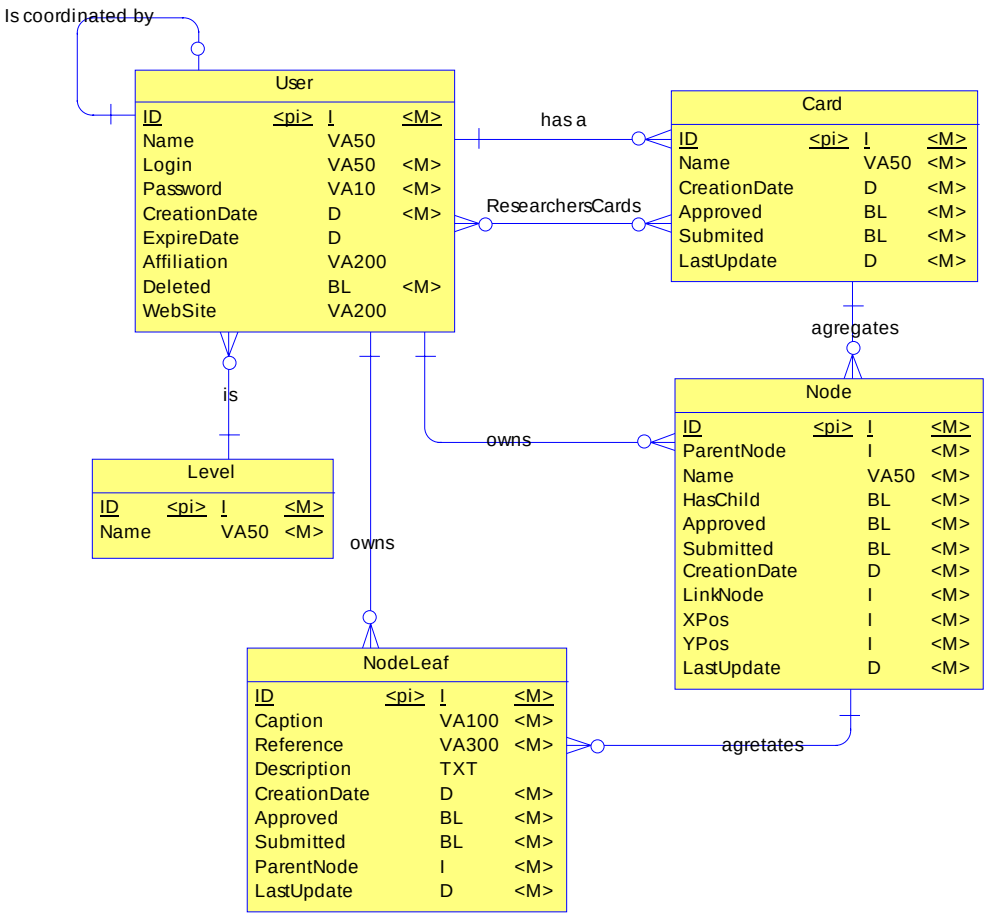


Figura 5.26 – Modelo conceptual da base de dados relativo à gestão de cartões.

A descrição pormenorizada de todas as entidades e relacionamentos deste modelo de dados é apresentada nas tabelas seguintes.

Tabela 5.4 – Descrição da entidade *Users*

Users	Entidade que armazena os dados das contas dos utilizadores do sistema.
ID	Identificador único do utilizador
Login; Password	Login e password do utilizador
Name	Nome completo do utilizador
CreationDate	Data de criação da conta do utilizador
ExpireDate	Data em que a conta expira
Deleted	*Flag* que marca se a conta foi apagada do sistema. Em vez de se apagar efectivamente o registo, optou-se por o guardar, marcando-o como removido.
Affiliation	Organização à qual o utilizador pertence
Website	Site da organização.

Tabela 5.5 – Descrição da entidade *Level*.

Level	Cada utilizador tem um nível de permissão de acesso ao sistema.
ID	Identificador numérico do nível de permissão do utilizador ao sistema.
Name	Nome do nível de permissão (*user*, *researcher*, *cardt manager*, *administrator*)

Tabela 5.6 – Descrição da entidade *Card*.

Card	Armazena definições gerais de cada cartão.
ID	Identificador numérico do cartão.
Name	Nome do cartão.
CreationDate	Data da criação do cartão.
Approved	*Flag* que identifica se o cartão está aprovado pelo respectivo gestor de cartões. Este atributo era utilizado nas versões anteriores do Diseasecard, na altura em que este era uma ferramenta colaborativa em que um gestor de cartões nomeava uma equipa de desenvolvimento para uma dada doença. Depois de preenchido o cartão este teria de ser aprovado pelo gestor de cartões.
Submitted	Depois do cartão estar aprovado, o passo seguinte é a sua publicação para que este seja visível por todos os utilizadores. Tal como o anterior, este atributo não está a ser utilizado.
LastUpdate	Data da última actualização do cartão

Tabela 5.7 – Descrição da entidade *Node*.

Node	Cada cartão é composto por uma colecção de nós referentes aos conceitos que o constituem. Na altura da criação de um cartão, a estrutura de nós é copiada do ficheiro XCD (XML Card Descriptor) descrito no capítulo 5.3.2
ID	Identificador numérico do nó. Equivalente ao elemento <node-id> do XCD
ParentNode	Nó pai. Equivalente ao elemento <parent-node> do XCD
Name	Nome do nó. Equivalente ao elemento <node-name> do XCD
HasChild	Equivalente ao elemento <has-child> do XCD
Approved	*Flag* que identifica se o nó está aprovado pelo gestor do cartão.
Submitted	*Flag* que identifica se o nó está submetido.
CreationDate	Data da criação do nó.
LinkNode	Equivalente ao elemento <link-node> do XCD
XPos	Equivalente ao elemento <x-pos> do XCD
YPos	Equivalente ao elemento <y-pos> do XCD
LastUpdate	Data da última actualização do nó.

Tabela 5.8 – Descrição da entidade *NodeLeaf.*

NodeLeaf	Cada nó (conceito) tem associados items (NodeLeaf) que contêm a informação relevante do ponto de vista do utilizador.
ID	Identificador numérico do nó de informação.
Caption	Nome do item de informação.
Reference	URL para a página que contém informação.
Description	Contém uma descrição adicional (opcional).
CreationDate	Data da criação do item de informação.
Approved	Flag que identifica se o item está aprovado pelo gestor do cartão.
Submitted	Flag que identifica se o item está submetido.
ParentNode	Indica o nó ao qual este item está associado.
LastUpdate	Data da última actualização do item.

As tabelas seguintes descrevem os principais relacionamentos entre as entidades descritas anteriormente.

Tabela 5.9 – Descrição dos principais relacionamentos do modelo de dados para a gestão de cartões.

Relacionamentos	
"Is coordinated by"	Existem diferentes perfis de utilizadores e cada um deles tem uma posição específica numa hierarquia. Este auto relacionamento indica para cada utilizador o seu superior hierárquico.
"Has a"	Indica qual é o utilizador gestor do cartão.
"ResearchersCards"	Este relacionamento origina uma nova tabela no diagrama físico (ResearchersCards). Nas versões colaborativas do diseasecard cada cartão tinha associado um grupo de utilizadores, nomeados pelo *manager* (gestor do cartão). Esse grupo é identificado na tabela resultante deste relacionamento.
"Owns"	Estes relacionamentos identificam os criadores de cada nodo e cada item nas entidades Node e NodeLeaf respectivamente.

A Figura 5.28 mostra as entidades da segunda secção do modelo de dados, referente às listas de doenças extraídas da tabela *MorbidMap* do OMIM. Como já foi referido anteriormente, o *MorbidMap* é disponibilizado *online* na forma de um ficheiro de texto com o formato apresentado na Figura 5.27.

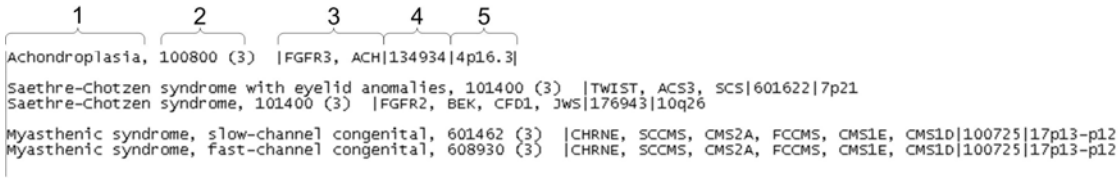


Figura 5.27 – Extracto de um ficheiro da tabela MorbidMap.

Este modelo está orientado às doenças ou, mais concretamente aos fenótipos (entidade *Disorder*), em que cada um é definido univocamente através de um código omim (domim, campo 2 da Figura 5.27). Cada doença pode ser conhecida por várias designações (campo 1) as quais serão armazenadas na entidade *DisorderName*. Por outro lado, tratando-se de uma doença genética, está associada a um ou mais genes, os quais são identificados com um código omim (gomim, campo 4 da Figura 5.27). A entidade *Omim* armazena estes códigos omim para cada gene bem como a sua localização citogenética (campo 5). O relacionamento "*has*" entre as entidades *Disorder* e *Omim* é do tipo muitos para muitos uma vez que a doença pode estar associada a mais do que um gene e, por outro lado, um dado gene pode ser responsável por mais do que uma doença. O símbolo do gene bem como os nomes alternativos (campo 3) são armazenados na entidade *Gene*. Por último, a entidade, *omimDiseasecard* é responsável por mapear cada doença da tabela *Disorder* a um cartão de doença da base de dados da secção de gestão de cartões, cujo modelo está ilustrado na Figura 5.26.

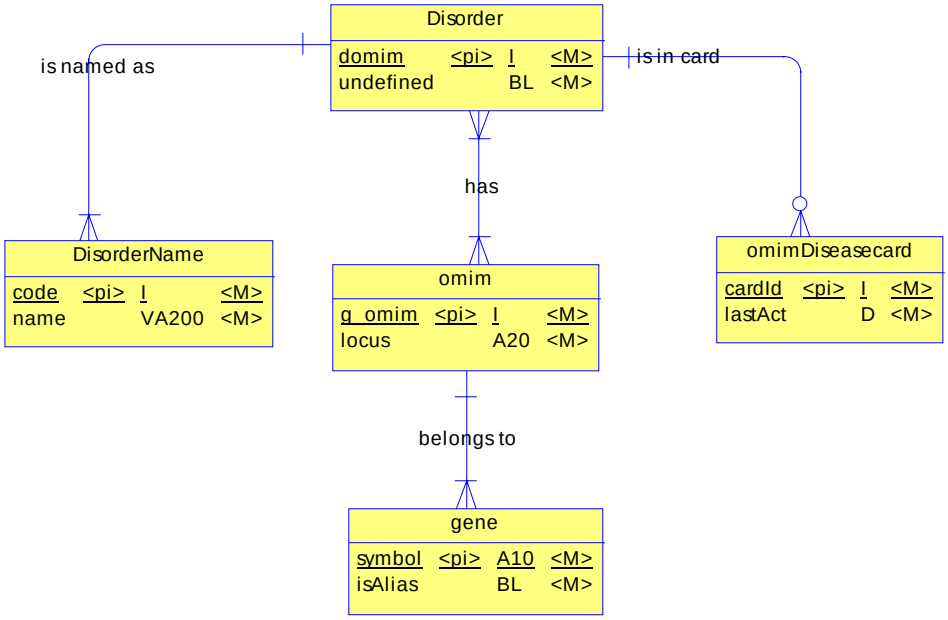


Figura 5.28 – Modelo conceptual da base de dados relativo à informação de doenças extraída da tabela *MorbidMap*.

As tabelas seguintes descrevem as entidades e respectivos atributos do modelo de dados da Figura 5.28.

Tabela 5.10 – Descrição da entidade *Disorder*.

Disorder	Tabela de doenças (fenótipos)
domim	Código OMIM para o fenótipo correspondente à doença
undefined	Alguns items não contêm qualquer informação na posição (1). Estes são marcados com o valor *true*.

Tabela 5.11 – Descrição da entidade *DisorderName*.

DisorderName	Tabela de nomes para as doenças. Esta tabela existe porque uma dada doença pode ter diversos nomes ou seja o mesmo domim pode corresponder a várias designações.
code	Código correspondente à doença.
name	Nome da doença.

Tabela 5.12 – Descrição da entidade *Omim*.

Omim	Esta tabela associa a cada código domim os respectivos códigos gomim.
gomim	Código omim do gene associado à doença.
locus	Localização citogenética. Cada gomim está associado a um gene cuja localização se encontra descrita em *locus*..

Tabela 5.13 – Descrição da entidade *Gene*.

Gene	Esta tabela associa a cada domim os respectivos gomim.
symbol	Símbolo para o gene associado à doença..
isalias	Flag que estando a *true* indica que este símbolo é um *alias* (nome alternativo)

Tabela 5.14 – Descrição da entidade *OmimDiseasecard*.

OmimDiseasecard	Tabela que associa as doenças do MorbidMap à tabela de cartões de doença
cardid	Identificador do cartão de doença da tabela Card (Figura 5.26)
lastAct	Data da última actualização.

Devido à sua simplicidade, não é necessário descrever os relacionamentos presentes no modelo, uma vez que todos eles são intuitivos e o seu significado subentende-se na descrição apresentada anteriormente.

5.5 Resultados

A Figura 5.29, anteriormente apresentada no capítulo 5.3.1, é uma representação parcial das regras básicas interpretadas pelo bloco XPDE de modo a construir cartões de doença ou seja, de modo a explorar e reunir os vários conceitos (genes, sequências de proteínas, sintomas da doenças etc.) associados às doenças genéticas raras. O processo é iniciado com base numa das três chaves (nome da doença, código *omim* ou símbolo do gene) que é cruzada com os dados da tabela *MorbidMap*, obtendo assim as chaves fundamentais para dar início ao processo de extracção.

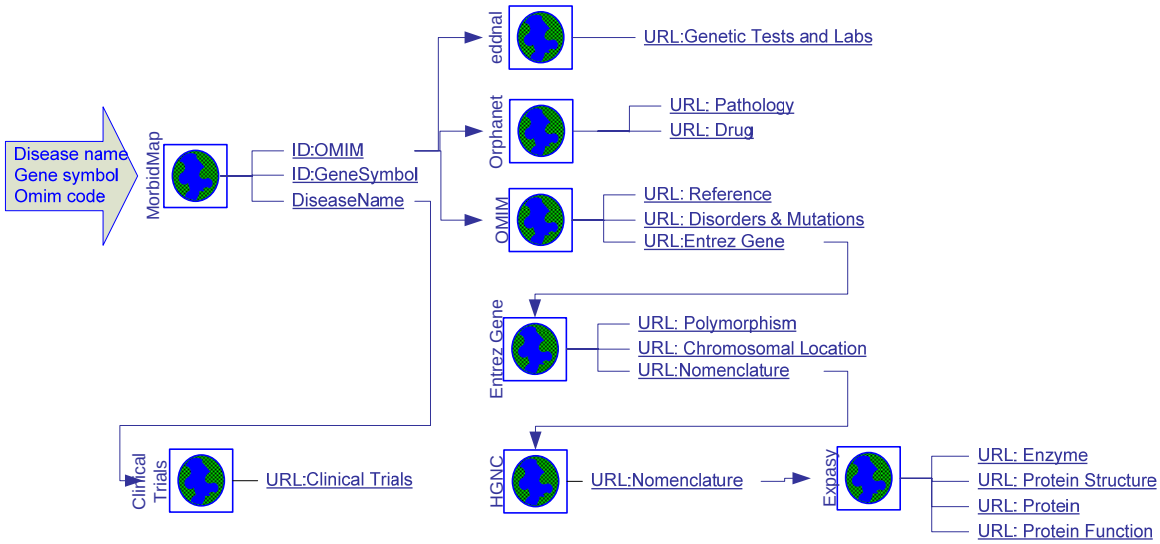


Figura 5.29 – Representação de uma vista parcial do protocolo XPD para doenças raras.

Como resultado deste sistema e tendo em conta o protocolo XPD actual, o *DiseaseCard* pode fornecer respostas a várias questões relevantes tanto para o diagnóstico de doenças genéticas como para o seu tratamento e acompanhamento. Algumas destas questões são:

- Quais as principais características da doença?
- Existem medicamentos para a tratar?
- Existem terapias genéticas ou testes clínicos envolvendo a doença?
- Que laboratórios realizam testes genéticos para a doença?
- Quais os genes responsáveis pela doença?
- Em que cromossomas estão estes genes localizados?
- Que mutações podem ser encontradas nestes genes?
- Quais os nomes oficiais para os genes?
- Que proteínas são codificadas a partir destes genes?
- Quais as funções destas proteínas?
- Qual é a estrutura tridimensional destas proteínas?
- Que enzimas estão associadas a estas proteínas?

Utilizando este conjunto de questões, são apresentados a seguir, em forma de sumário os resultados do *DiseaseCard* para três doenças diferentes: Síndrome de X-Frágil, Acondroplasia e Doença de *Fabry*. Tendo em conta que um cartão de doença actual contém 19 conceitos, pode-se concluir que o sistema pode responder no mínimo a 19 questões, já que um dado conceito pode ter associada uma página *web* que responde a mais do que uma questão.

As respostas para as seguintes questões estão apresentadas na língua original da página de onde foram extraídas. Assim, para manter a coerência, as questões estão também formuladas em inglês. Convém ainda referir que estes dados são susceptíveis de mudar visto que ocorrem periodicamente actualizações nas fontes de dados de origem e consequentemente na tabela de cartões.

Nome da Doença: Síndrome de X-Frágil:

Data de extracção: 18-5-2006

1. What are the main features of the Fragile X Syndrome?
   - *Fragile X syndrome is the most frequent cause of inherited mental retardation. It is caused by a dynamic mutation i.e. the progressive expansion of polymeric (CGG)n trinuleotide repeats located in the non coding region at the 5' end of the FMR1 gene at Xq 27.3.[…];*

2. Are there any drugs for this disease? [Empty];
3. Are there any gene therapies or clinical trials for this disease? [Empty];
4. What laboratories perform genetic tests for this disease?
   - *University of Leipzig - Medical Faculty - Institute of Human Genetics;*
   - *Department of Clinical Genetics - Lund University Hospital;*
   - *[…];*
5. What genes cause this disease?
   - *This disease is caused by a mutation in the FMR1 gene. The official name is fragile X mental retardation 1;*
6. On which chromosome is FMR1 located?
   - *The FMR1 is located at locus Xq27.3;*
7. What mutations have been found in gene FMR1?
   - *5 entries have been found in the literature for this gene with omim code = 309550 which are:*
     - *0001 FRAGILE X MENTAL RETARDATION SYNDROME [FMR1, ILE304ASN]*
     - *0002 FRAGILE X MENTAL RETARDATION SYNDROME [FMR1, 1-BP DEL, ACT125CT, FS159TER]*
     - *0003 FRAGILE X MENTAL RETARDATION SYNDROME [FMR1, IVS1, G-T, -1 AND G-A, +1]*
     - *0004 FRAGILE X MENTAL RETARDATION SYNDROME [FMR1, (CGG)n EXPANSION]*
     - *0005 FMR1 POLYMORPHISM [FMR1, IVS10, C-T, +14]*
8. What names are used to refer to this gene?
   - *FMR1 is the official symbol for the gene;*
   - *FRAXA and MGC87458 are alias;*
9. What are the proteins coded by this gene?
   - *1 protein has been found in SwissProt with accession number Q06787, entry name FMR1_Human and name Fragile X mental retardation 1 protein;*
10. What are the functions of the gene product?
    - *RNA-binding protein. Associated to polysomes and might be involved in the transport of mRNA from the nucleus to the cytoplasm.*
11. What is the 3D structure for this protein?
    - *The structure for this protein is available in PDB with the code 2FMR.*
12. What are the enzymes associated to these proteins? [Empty]

Nome da doença: Acondroplasia:

Data de extracção: 18-5-2006

1. What are the main features of the Achondroplasia?
   - *Achondroplasia is the most frequent form of chondrodysplasia with a prevalence of one child in every 15,000. This type of dwarfism is characterized by short limbs, hyperlordosis, short hands, and macrocephaly with high forehead and saddle nose […];*
2. Are there any drugs for this disease?
   - *Norditropin;*
3. Are there any gene therapies or clinical trials for this disease?

   There are two complete studies sponsored by National Human Genome Research Institute (NHGRI):
   - *Study of Skeletal Disorders and Short Stature*
   - *Issues Surrounding Prenatal Genetic Testing for Achondroplasia*
4. What laboratories perform genetic tests for this disease?
   - *Università degli Studi di Verona - Laboratorio di Genetica Molecolare;*
   - *Department Center of Medical Genetics - University of Antwerp;*

- *Centro de Análisis Genéticos (Zaragoza);*
- *[…];*

5. What genes cause this disease?
   - *This disease is caused by a mutation in the FGFR3 gene. The official name is fibroblast growth factor receptor 3;*
6. On which chromosome is FGFR3 located?
   - *The FGFR3 is located at locus 4p16.3;*
7. What mutations have been found in gene FGFR3?
   - *26 entries have been found in the literature for this gene with omim code =134934;*
     - *0001 ACHONDROPLASIA [FGFR3, GLY380ARG, 1138G-A]*
     - *0002 ACHONDROPLASIA [FGFR3, GLY380ARG, 1138G-C]*
     - *0003 ACHONDROPLASIA [FGFR3, GLY375CYS]*
     - *0004 THANATOPHORIC DYSPLASIA, TYPE II [FGFR3, LYS650GLU]*
     - *0005 THANATOPHORIC DYSPLASIA, TYPE I [FGFR3, ARG248CYS]*
     - *0006 THANATOPHORIC DYSPLASIA, TYPE I [FGFR3, SER371CYS]*
     - *0007 THANATOPHORIC DYSPLASIA, TYPE I [FGFR3, TER807GLY]*
     - *[…]*
8. What names are used to refer to this gene?
   - *FGFR3 is the official symbol for the gene;*
   - *CEK2, JTK4 are alias and ACH is a previous symbol;*
9. What are the proteins coded by this gene?
   - *1 protein has been found in SwissProt with accession number P22607, entry name FGFR3_Human and name Fibroblast growth factor receptor 3 [Precursor];*
10. What are the functions of the gene product?
    - *Receptor for acidic and basic fibroblast growth factors. Preferentially binds FGF1.*
11. What is the 3D structure for this protein?
    - *The structure for this protein is available in PDB with the code 1RY7.*
12. What are the enzymes associated to this protein?
    - *1 enzyme has been found in Expasy with EC=2.7.1.112 called Protein-tyrosine kinase*

Nome da doença: Doença de Fabry:

Data de extracção: 25-5-2006

1. What are the main features of the Fabry Disease?
   - *Fabry's disease (FD) is an X-linked inborn error of glycosphingolipid metabolism due to a deficient activity of alpha-galactosidase A, a lysosomal homodimeric enzyme. The enzymatic defect leads to the systemic accumulation of underivatized neutral glycosphingolipids in plasma and tissues […];*
2. Are there any drugs for this disease?
   - *Fabrazyme: AGALSIDASE BETA;*
   - *Replagal: AGALSIDASE ALFA;*
3. Are there any gene therapies or clinical trials for this disease? [Empty]
4. What laboratories perform genetic tests for this disease?
   - *University Medical Center Utrecht*
   - *Unidade de Enzimologia (Porto)*
   - *Institut de Pathologie et de Génétique asbl (Loverval)*
   - *[…];*
5. What genes cause this disease?
   - *This disease is caused by a mutation in the GLA gene. The official name is galactosidase, alpha;*
6. On which chromosome is GLA located?

- *The FGFR3 is located at locus Xq22;*

7. What mutations have been found in gene GLA?
   - *52 entries have been found in the literature for this gene with omim code =301500;*
8. What names are used to refer to this gene?
   - *GLA is the official symbol for the gene;*
   - *GALA is an alias;*
9. What are the proteins coded by this gene?
   - *1 protein has been found in SwissProt with accession number P06280, entry name AGAL_Human and name Alpha-galactosidase A [Precursor];*
10. What are the functions of the gene product?
    - *Catalytic activity: Hydrolysis of terminal, non-reducing alpha-D- galactose residues in alpha-D-galactosides, including galactose oligosaccharides, galactomannans and galactohydrolase.*
11. What is the 3D structure for this protein?
    - *The structure for this protein is available in PDB with the code 1R47 and 1R46.*
12. What are the enzymes associated to this protein?
    - *1 enzyme has been found in Expasy with EC=3.2.1.22 called Alpha-galactosidase*

Para além destes três exemplos, o sistema *DiseaseCard* permite actualmente a ligação a mais de 2000 doenças. Contudo, muitas delas não contêm as respostas a estas questões pois ainda não existe informação *online* nas bases de dados exploradas através do protocolo.

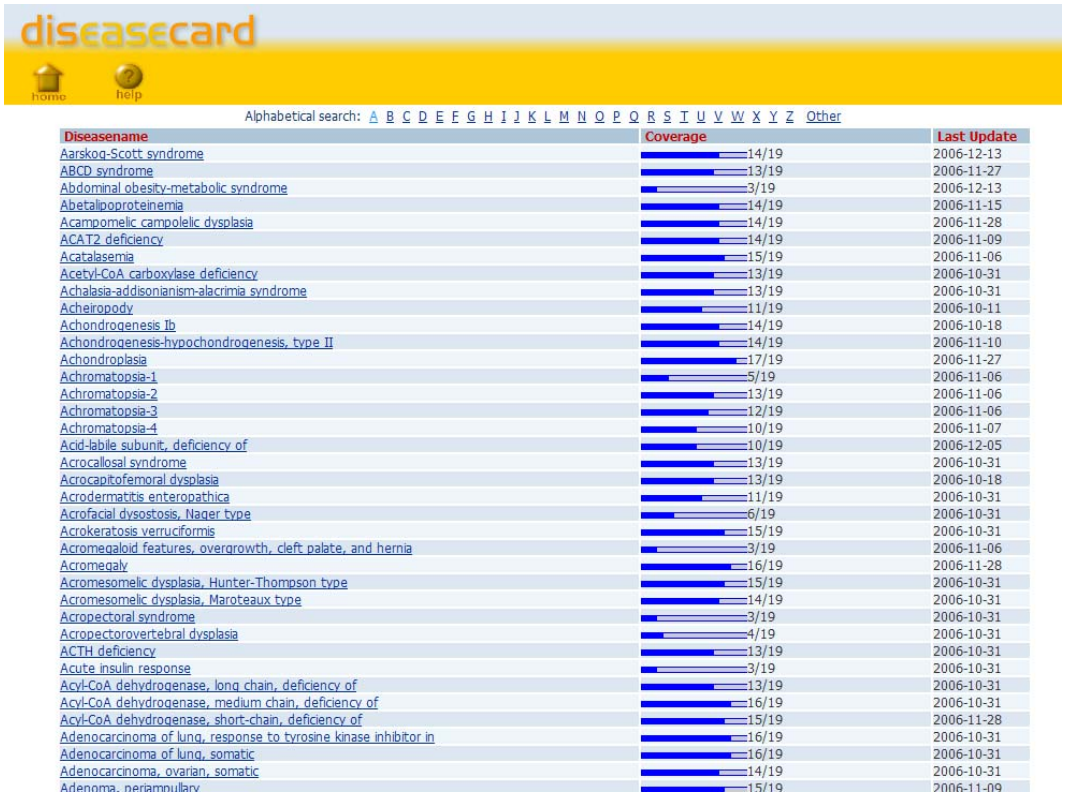
As figuras seguintes mostram o aspecto actual de algumas páginas típicas do portal *Diseasecard*. A Figura 5.30 corresponde à página de entrada no *site* onde o utilizador pode iniciar a sua consulta. A pesquisa é iniciada num formulário onde é inserida uma palavra ou expressão chave que pode ser o nome da doença, o código (símbolo) de um gene ou um código OMIM correspondente à doença ou ao gene associado ("*Search for a disease*"). Além destes três modos de consulta, existe ainda a possibilidade de procurar a partir da lista total das doenças inscritas na base de dados do sistema ("*List of available diseases*").

Figura 5.30 – Página de entrada da aplicação *Diseasecard*.

Como resultado, o *Diseasecard* retorna uma lista de nomes de doenças relacionadas com a consulta onde cada item corresponde a um cartão de doença. A Figura 5.31 mostra uma destas listagens, onde se podem ver as primeiras doenças começadas por "A".

Além do nome da doença, esta listagem fornece também a data da última actualização do cartão e a coluna *Coverage*. Esta coluna é de grande utilidade porque fornece de um modo muito intuitivo o factor de preenchimento do cartão ou seja, representa graficamente o número de conceitos com informação num total de conceitos por cartão.

Como também se pode observar, existem muitos cartões com um factor de preenchimento reduzido, o que indica que para as respectivas doenças a informação procurada é ainda escassa.



Diseasename	Coverage	Last Update
Aarskog-Scott syndrome	14/19	2006-12-13
ABCD syndrome	13/19	2006-11-27
Abdominal obesity-metabolic syndrome	3/19	2006-12-13
Abetalipoproteinemia	14/19	2006-11-15
Acampomelic campolelic dysplasia	14/19	2006-11-28
ACAT2 deficiency	14/19	2006-11-09
Acatalasemia	15/19	2006-11-06
Acetyl-CoA carboxylase deficiency	13/19	2006-10-31
Achalasia-addisonianism-alacrimia syndrome	13/19	2006-10-31
Acheiropody	11/19	2006-10-11
Achondrogenesis Ib	14/19	2006-10-18
Achondrogenesis-hypochondrogenesis, type II	14/19	2006-11-10
Achondroplasia	17/19	2006-11-27
Achromatopsia-1	5/19	2006-11-06
Achromatopsia-2	13/19	2006-11-06
Achromatopsia-3	12/19	2006-11-06
Achromatopsia-4	10/19	2006-11-07
Acid-labile subunit, deficiency of	10/19	2006-12-05
Acrocallosal syndrome	13/19	2006-10-31
Acrocapitofemoral dysplasia	13/19	2006-10-18
Acrodermatitis enteropathica	11/19	2006-10-31
Acrofacial dysostosis, Nager type	6/19	2006-10-31
Acrokeratosis verruciformis	15/19	2006-10-31
Acromegaloid features, overgrowth, cleft palate, and hernia	3/19	2006-11-06
Acromegaly	16/19	2006-11-28
Acromesomelic dysplasia, Hunter-Thompson type	15/19	2006-10-31
Acromesomelic dysplasia, Maroteaux type	14/19	2006-10-31
Acropectoral syndrome	3/19	2006-10-31
Acropectorovertebral dysplasia	4/19	2006-10-31
ACTH deficiency	13/19	2006-10-31
Acute insulin response	3/19	2006-10-31
Acyl-CoA dehydrogenase, long chain, deficiency of	13/19	2006-10-31
Acyl-CoA dehydrogenase, medium chain, deficiency of	16/19	2006-10-31
Acyl-CoA dehydrogenase, short-chain, deficiency of	15/19	2006-11-28
Adenocarcinoma of lung, response to tyrosine kinase inhibitor in	16/19	2006-10-31
Adenocarcinoma of lung, somatic	16/19	2006-10-31
Adenocarcinoma, ovarian, somatic	14/19	2006-10-31
Adenoma, periampullary	15/19	2006-11-09

Figura 5.31 – Detalhe de uma listagem dos cartões de doença por ordem alfabética

Escolhendo um destes itens (nome da doença), o utilizador pode aceder aos detalhes do cartão segundo a estrutura apresentada na Figura 5.32

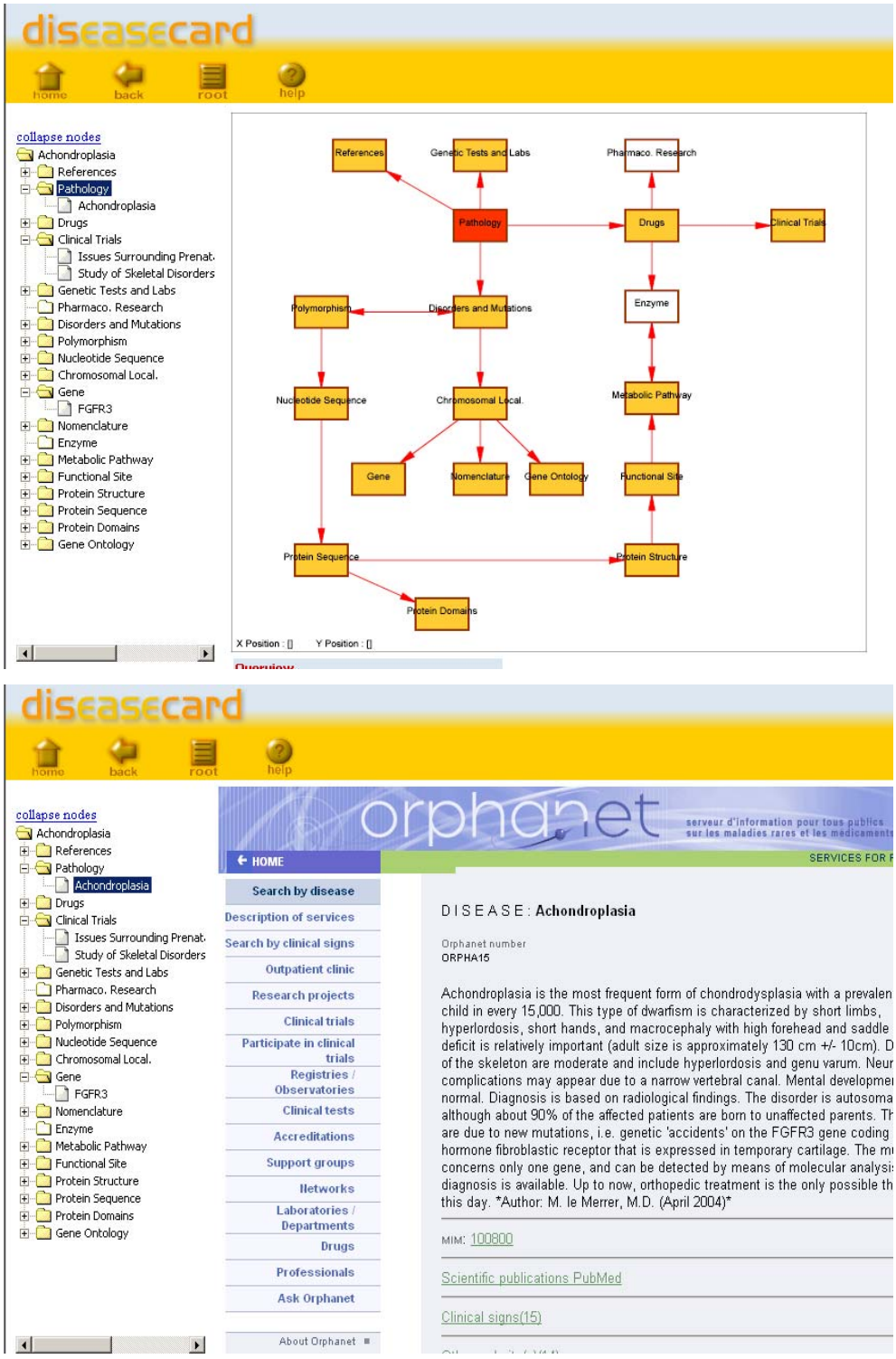


Figura 5.32 – Detalhes de um cartão de doença disponível no *Diseasecard*.

A página inicial do cartão (Figura 5.32, em cima) disponibiliza duas vistas diferentes para a mesma estrutura. A vista em árvore contém todos os conceitos do cartão e, dentro de cada um, as respectivas páginas web. A vista em mapa de nós representa os mesmos conceitos e as respectivas relações entre eles.

O utilizador pode ver os conteúdos de cada conceito no cartão “clicando” nas “folhas” da árvore. Por exemplo, dentro do conceito “Pathology” existe a “folha” “Achondroplasia” que corresponde à página de Internet que contém informação descritiva sobre a doença (*Orphanet* neste caso). Os detalhes desta página são apresentados conforme a Figura 5.32, na imagem inferior.

5.6 Sumário

Tendo em conta o modelo de navegação proposto no capítulo 4, foi aqui abordada uma forma de automatizar o processo de obtenção de informação para todos os seus conceitos no contexto das doenças genéticas raras. Em vez de se extrair e armazenar informação das bases de dados originais, optou-se por desenvolver um sistema que apenas gere as referências URL (*links http*) que ligam à informação relevante dentro das bases de dados.

Neste capítulo é descrita a arquitectura do sistema de extracção de informação, tanto a nível dos protocolos XPD e XCD, criados para mapear no sistema, de um modo configurável, o modelo de navegação de conceitos, como também a nível dos detalhes principais do processo de pesquisa automática e preenchimento dos cartões de doença.

Finalmente, são apresentados alguns exemplos de cartões de doença como resultado do processo de obtenção automática de informação.

6 Conclusões e Trabalho Futuro

DiseaseCard é um portal web público (http://www.diseasecard.org) cujo principal objectivo é proporcionar o acesso de uma forma simples e integrada a informação sobre doenças genéticas raras unindo domínios da medicina, da genética e da farmacologia e cujas fontes de conhecimento se encontram dispersas em inúmeras bases de dados na Internet. Este sistema utiliza as funcionalidades do bloco *XPD Engine* para obter informação relativa ao contexto das doenças raras de um modo transparente para o utilizador.

O módulo *XPD Engine*, responsável pela criação automática de cartões de doença, actua sempre que uma dada doença não esteja catalogada no *DiseaseCard* ou quando a última extracção de informação se encontra desactualizada. O resultado aparece sob a forma de um "cartão" por doença, representado numa árvore de conceitos que permite ao utilizador navegar com facilidade sobre a informação recolhida para essa doença.

Tendo em conta as abordagens de integração de informação apresentadas no capítulo 3.3, esta aqui apresentada é bastante simples, enquadrando-se, numa primeira análise na Integração por *Links* (capítulo 3.3.3). Por outro lado, a funcionalidade do módulo *XPD Engine* confere-lhe algumas propriedades que são características da Tradução de Consultas (capítulo 3.3.2), nomeadamente na utilização de um mediador que submete consultas às bases de dados via HTTP e a informação permanece armazenada somente nas bases de dados intervenientes. Contudo, em detrimento de extrair dados, processá-los e apresentá-los filtrados e cruzados tal como na Tradução de Consultas, na abordagem actual, o *Diseasecard*, limita-se a apresentar as páginas, tal como são disponibilizadas pelas fontes de dados originais. Esta solução comprometeu-se a dar uma resposta rápida e prática ao problema da exploração de um conjunto vasto de bases de dados a utilizadores pouco familiarizados com elas, não necessitando para isso de formular consultas complexas, utilizando nomenclaturas e sintaxes muito específicas. Este objectivo é alcançado na medida em que toda a interface do sistema com o utilizador prima pela simplicidade e pelo reduzido número de controlos. Por outro lado, cada cartão de doença concentra um conjunto de ligações directas às páginas com informação relevante.

Em termos de desempenho, tendo em conta o número de utilizadores actual, o módulo *XPD Engine*, quando executado, produz um cartão num intervalo de tempo compreendido entre 10 e 20 segundos, o que é bastante razoável tendo em conta a quantidade de conceitos a preencher.

Foram feitas algumas campanhas junto da comunidade médica através do envio de e-mails a sugerir a visita ao *Diseasecard*. Apesar da receptividade ter sido satisfatória, constatou-se que a ferramenta neste momento não representa uma grande valia para a maioria. A combinação de conhecimento das áreas da medicina, genómica e farmacologia está agora a dar importantes passos. Contudo a manifestação efectiva de tais avanços no âmbito da prática clínica não é ainda muito significativa.

O sistema aqui apresentado permite descrever um mapa de conceitos e de navegação a partir do qual é possível extrair ou integrar dados de múltiplas fontes disponíveis na Internet com base numa única consulta feita pelo utilizador. O protocolo de navegação pode ser facilmente editado e reconfigurado facilitando a adaptação face ao aparecimento de novos recursos. Por outro lado a sua generalidade permite a sua utilização em qualquer tipo de cenário onde a navegação seja tão importante quanto a própria extracção de informação.

6.1 Trabalho Futuro

Apesar dos resultados obtidos e da disponibilização efectiva da aplicação, existem ainda várias limitações reconhecidas e que num futuro próximo se planeiam colmatar. Estas omissões prenderam-se principalmente com a necessidade que houve em estabelecer prioridades no desenvolvimento, justificando-se também pela natureza demonstrativa do sistema. Alguns aspectos a melhorar são apresentados a seguir.

- Acrescentar redundância ao protocolo de navegação do *Diseasecard*. Actualmente o sistema suporta uma integração do tipo horizontal onde a agregação é feita sobre dados semanticamente complementares. Pretende-se estender a portabilidade do sistema a dados semanticamente similares – integração vertical [3]. Em termos de protocolo de navegação (XPD) esta funcionalidade consiste em adicionar caminhos alternativos no mapa, fornecendo assim múltiplas bases de dados para o mesmo conceito. Além de se introduzir redundância que permite colmatar lacunas de

dados, esta funcionalidade pretende também conduzir o utilizador a várias respostas para o mesmo conceito. O esforço principal desta implementação prende-se com a investigação de novas bases de dados alternativas às já mapeadas no protocolo.

- Alargar o universo de representação dos cartões de doença a outros conceitos de interesse no âmbito das doenças genéticas raras. Este aspecto é de fácil implementação já que basta apenas configurar os ficheiros de configuração XPD e XCD, introduzindo novos nós, um para cada nova base de dados a explorar. O esforço principal desta tarefa exige essencialmente as sugestões de utilizadores e a pesquisa de bases de dados que cumpram os requisitos das respectivas sugestões.
- Estender o modelo XPDE a outros cenários de extracção de informação, cujo processo de procura possa ser representado num protocolo similar ao XPD.
- Implementar uma funcionalidade de monitorização de *logs* para o processo de recuperação de informação executado pelo XPDE. Apesar da implementação actual gerar mensagens, estas não estão ainda disponíveis aos utilizadores e encontram-se num formato pouco legível. O objectivo desta tarefa será manter os utilizadores de perfil mais elevado ao corrente do estado das operações de extracção, gerando alertas quando, por exemplo, uma base de dados para um determinado conceito está em baixo ou um URL de acesso é alterado. Desta forma facilita-se a detecção e localização atempada de possíveis falhas do protocolo de um modo sistemático.
- Permitir a inserção de URLs úteis por utilizadores com conhecimentos em conceitos específicos. Esta funcionalidade vai ao encontro da componente colaborativa das primeiras implementações do *Diseasecard*. Contudo, nesta abordagem parte-se de um cartão pré-preenchido, podendo um dado utilizador autenticado inserir informação relevante que o protocolo não contemple.

7 Referências

1. Altman, R.B., *Bioinformatics in support of molecular medicine.* Proc AMIA Symp 1998, 1998: p. 53-61.
2. Searls, D.B., *Data integration: challenges for drug discovery.* Nature Reviews Drug Discovery, 2005. **4**: p. 45-58.
3. Sujansky, W., *Heterogeneous database integration in biomedicine.* Computers and Biomedical Research, 2001. **34**(4): p. 285-298.
4. EMBL. *EMBL STATISTICS*. 2006 [visitado 2006 8 Fevereiro]; http://www3.ebi.ac.uk/Services/DBStats/.
5. PDB. *PDB Statistics*. Protein Data Bank 2006 21-11-2006 [visitado 2006 23 Novembro]; http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100.
6. Oliveira, I., et al. *On the requirements of biomedical information tools for health applications: the INFOGENMED case study*. in *7th Portuguese Conference on Biomedical Engineering (BioEng'2003)*. 2003. Lisbon, Portugal.
7. Mitsuhara, H., et al., *A web retrieval support system with a comment sharing environment: toward an adaptive web-based IR system.* Proceedings. International Conference on Computers in Education, 2002. **2**: p. 1218- 1222.
8. Martin-Sanchez, F., et al., *Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care* J. of Biomedical Informatics 2004. **37** (1): p. 30-42
9. Galperin, M.Y., *The Molecular Biology Database Collection: 2006 update.* Nucleic Acids Research, 2006. **34**(D3-D5).
10. Pereira, A.S., et al., *The Infogenmed Project.* ICBME 2002: The Bio-Era: New Challenges, New Frontiers, 2002.
11. Nagarajan, R., M. Ahmed, and A. Phatak. *Database Challenges in the Integration of Biomedical Data Sets*. in *Proceedings of the Thirtieth International Conference on Very Large Data Bases*. 2004. Toronto, Canada: Morgan Kaufmann.
12. Babu, M.M., *Integrating Bioinformatics, Medical Sciences and Drug Discovery.* National Conference on Medical Informatics at Vijayawada, 2000.
13. Miller, S.L., *Production of Amino Acids Under Possible Primitive Earth Conditions.* Science, 1953. **117**: p. 528.
14. Wilkins, J.S. *Spontaneous Generation and the Origin of Life*. The Talk.Origins Archive 2004 [visitado 2006 23 Setembro]; The Talk.Origins Archive:[Available from: http://www.talkorigins.org/faqs/abioprob/spontaneous-generation.html.
15. Woese, C.R., O. Kandler, and M.L. Wheelis, *Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya.* Proceedings of the National Academy of Sciences, 1990. **87**(12): p. 4576–4579.

16. Lodish, H., et al., *Molecular Cell Biology (5th edition).* 2003.

17. Wolfe, S.L., *An Introduction to Cell and Molecular Biology*. 1995: Thomson Learning.

18. NCBI. *Science Primer - A Basic Introduction to the Science Underlying NCBI Resources*. National Center for Biotechnology Information 2004 30-3-2004 [visitado 2006 23 Novembro]; http://www.ncbi.nlm.nih.gov/About/primer/genetics_cell.html.

19. Jones, M. *Overview of bacterial conjugation in 4 steps*. Wikipedia 2006 [visitado 2006 23-11-2006]; http://en.wikipedia.org/wiki/Image:BacterConjugation.png.

20. Encarta. *Animal Cell*. Microsoft Encarta Encyclopedia 2006 [visitado 2006 23 Novembro]; http://encarta.msn.com/media_461540224/Animal_Cell.html.

21. Lauher, J.W. *Brook Chemistry@The Brook - The Amino Acids*. 2006 [visitado 2006 23 Nov]; http://www.sunysb.edu/chemistry/molecules/aa.html.

22. regalis, O. *Protein*. Wikipedia 2006 18-8-2006 [visitado 2006 23 Nov]; http://en.wikipedia.org/wiki/Image:Proteinviews-1tim.png.

23. Watson, J. and F. Crick, *A structure for Deoxyribose Nucleic Acid.* Nature, 1953. **171**: p. 737 – 738.

24. Ströck, M. *DNA*. Wikipedia 2006 8-2-2006 [visitado 2006 23 Nov]; http://en.wikipedia.org/wiki/Image:DNA_Overview.png.

25. Consortium, I.H.G.S., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**: p. 931 - 945.

26. Brazma, A., et al. *A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays*. 2001 [visitado jan 2006]; http://www.ebi.ac.uk/microarray/biology_intro.html.

27. OGM. *ADN*. OGM Info [visitado 2006 23 Nov]; http://www.ogm-info.com/adn.html.

28. Sayfa, A. *Super Quintet Chemistry I: Introduction to Chemistry*. [visitado 2006 23 Nov]; http://library.tedankara.k12.tr/chemistry/vol1/biochem/trans98.htm.

29. Lander, E., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

30. Pennisi, E., *Gene Counters Struggle to Get the Right Answer.* Science, 2003. **301**(1040–1041).

31. Mitchell, J.A., A.T. McCray, and O. Bodenreider, *From Phenotype to Genotype: Issues in Navigating the Available Information Resources.* Methods of Information in Medicine, 2003. **42**: p. 557-563.

32. Bayat, A., *Science, medicine, and the future: Bioinformatics.* British Medical Journal (BMJ), 2002. **324**(7344): p. 1018-1022.

33. Guttmacher, A.E. and F.S. Collins, *Genomic medicine-a primer.* The New England Journal of Medicine, 2002. **347**(19): p. 1512-1520.

34. Lewontin, R. *The Genotype/Phenotype Distinction*. 2004 [visitado; http://plato.stanford.edu/archives/spr2004/entries/genotype-phenotype/.

35. Orphanet. *Rare diseases*. 2005 August 2 [visitado 2005 12 Dezembro]; http://www.orpha.net/.

36. STRATCARE. *Orphan Drugs, Final Report - European Parliament - Committee on the environment, Public Health and Consumer Protection*. Dick HOLDSWORTH, Head of STOA Unit 1999 [visitado 2006; http://www.europarl.europa.eu/stoa/publi/167780/default_en.htm.

37. Faurisson, F. *EURORDIS - European Organisation for Rare Diseases*. 2004 [visitado 2006; http://www.eurordis.org/sommaire.html.

38. Sheth, A.P. and J.A. Larson, *Federated database systems for managing distributed, heterogeneous, and autonomous databases.* ACM Computing Surveys (CSUR), 1990. **22**(3): p. 183-236.

39. Halevy, A.Y., *Answering queries using views: A survey.* VLDB Journal: Very Large Data Bases, 2001. **10**(4): p. 270-294.

40. Calì, A., et al. *On the Expressive Power of Data Integration Systems*. in *21st Int. Conf. on Conceptual Modeling (ER 2002)*. 2002: Springer.

41. Aparício, A.S., O.L.M. Farias, and N.d. Santos, *Applying Ontologies in the Integration of Heterogeneous Relational Databases*, in *Australasian Ontology Workshop (AOW 2005), Sydney. Conferences in Research and Practice in Information Technology (CRPIT)*, T. Meyer and M.A. Orgun, Editors. 2005, ACS: Sydney, Australia. p. 11-16.

42. Duschka, O.M. and M.R. Genesereth, *Query Planning in Infomaster*, in *SAC '97: Proceedings of the 1997 ACM symposium on Applied computing*. 1997, ACM Press: San Jose, California, USA. p. 109-111.

43. Hull, R., *Managing Semantic Heterogeneity in Databases: A Theoretical Perspective*, in *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. 1997, ACM Press: Tucson, Arizona, USA. p. 51-61.

44. Seth, A., *Changing Focus on Interoperability in Information Systems: From system, Syntax,Structure to Semantics*, in *Interoperating Geographic Information Systems*, M.F. GoodChild, et al., Editors. 1999, Kluwer Publishers. p. 5-29.

45. Wache, H., et al., *Ontology-based integration of information -a survey of existing approaches*, in *IJCAI-01 Workshop: Ontologies and Information Sharing*, H. Stuckenschmidt, Editor. 2001: Seattle. p. 108-117.

46. Hernandez, T. and S. Kambhampati, *Integration of Biological Sources: Current Systems and Challenges.* SIGMOD Record, 2004. **33**(3): p. 51-60.

47. Karp, P.D., *A Strategy for Database Interoperation.* Journal of Computational Biology, 1996. **2**(4): p. 573-583.

48. Hammer, J. and M. Schneider, *Genomics Algebra: A New, Integrating Data Model, Language, and Tool for Processing and Querying Genomic Information*, in *CIDR*

2003, First Biennial Conference on Innovative Data Systems Research. 2003: Asilomar, CA, USA.

49. Davidson, S.B., et al., *K2Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources.* IBM Systems Journal, 2001. **40**(2): p. 512-531.
50. Ouzzani, M. and A. Bouguettaya, *Query Processing and Optimization on the Web.* Distributed and Parallel Databases, 2004. **15**(3): p. 187 - 218.
51. Davidson, S.B., et al., *BioKleisli: A Digital Library for Biomedical Researchers.* International Journal on Digital Libraries, 1997. **1**(1): p. 36-53.
52. Davidson, S., G.C. Overton, and P. Buneman, *Challenges in Integrating Biological Data Sources.* Journal of Computational Biology, 1995. **2**(4): p. 557-572.
53. Buttler, D., et al., *Querying multiple bioinformatics information sources: can semantic web research help?* ACM SIGMOD Record, 2002. **31**(4): p. 59-64.
54. Lacroix, Z., et al., *Exploring Life Sciences Data Sources*, in *Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03)*. 2003: Acapulco, Mexico. p. 203-208.
55. PubMed. *PubMed Overview*. National Center for Biotechnology Information (NCBI) 2006 June, 30 [visitado 2006 17 Agosto]; http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html.
56. Medline. *MEDLINE - Fact Sheet*. U.S. National Library of Medicine 2006 [visitado 2006 9 Março]; http://www.nlm.nih.gov/pubs/factsheets/medline.html.
57. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.
58. Berman, H.M., P.E. Bourne, and J. Westbrook, *The Protein Data Bank: A Case Study in Management of Community Data.* Current Proteomics, 2004. **1**(1): p. 49-57.
59. O'Donovan, C., et al., *High-quality protein knowledge resource: SWISS-PROT and TrEMBL.* Briefings in Bioinformatics 2002, 2002. **3**(3): p. 275-284.
60. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003*, in *Nucleic Acids Research*. 2003, Oxford University Press. p. 365-370.
61. Wu, C.H., et al., *The Protein Information Resource.* Nucleic Acids Research, 2003. **31**(1): p. 345-347.
62. Bairoch, A., et al., *The Universal Protein Resource (UniProt).* Nucleic Acids Research, 2005. **33**(D154-D159).
63. Etzold, T., A. Ulyanov, and P. Argos, *SRS: information retrieval system for molecular biology data banks.* Methods for Enzymology, 1996. **266**: p. 114-128.
64. Zdobnov(2), E.M., et al., *The EBI SRS server—new features.* Bioinformatics, 2002. **18**(8): p. 1149-1150.
65. Stoesser, G., et al., *The EMBL nucleotide sequence database.* Nucleic Acids Research, 2001. **29**(1): p. 17-21.

66. Bairoch, A. and R. Apweiler, *The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.* Nucleic Acids Research, 2000. **28**(1): p. 45-48.

67. Ostell, J. *The NCBI Handbook.* 2003 [visitado 2006 14 Março]; http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.

68. InfoLab. *Infolab, Standford University*. 2006 [visitado 15-3-2006]; http://infolab.stanford.edu/.

69. Chawathe, S., et al., *The TSIMMIS Project: Integration of heterogeneous information sources,* in *16th Meeting of the Information Processing Society of Japan*. 1994: Tokyo, Japan. p. 7-18.

70. Arasu, A., et al., *Searching the Web.* ACM Transactions on Internet Technology, 2001.

71. Srivastava, U., et al., *Query Optimization over Web Services.* Technical Report, Stanford University, 2005.

72. Tomasic, A., L. Raschid, and P. Valduriez, *Scaling Access to Heterogeneous Data Sources with DISCO.* Knowledge and Data Engineering, 1998. **10**(5): p. 808-823.

73. genophen. *Metabolic Knowledge Base*. 2001 [visitado 17-3-2006]; http://integration.genophen.de/.

74. Freier, A., et al., *BioDataServer: A SQL-based service for the online integration of life science data.* In Silico Biology, 2002. **2**(0005).

75. Ambite, J.L., et al., *Ariadne: a system for constructing mediators for Internet sources,* in *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data,* L.M. Haas and A. Tiwary, Editors. 1998, ACM Press: Seattle, Washington, USA. p. 561-563.

76. Knoblock, C.A., et al., *The ARIADNE Approach to Web-based Information Integration.* International Journal of Cooperative Information Systems, 2001. **10**(1): p. 145-169.

77. Arens, Y., et al., *Retrieving and Integrating Data from Multiple Information Sources.* International Journal of Cooperative Information Systems, 1993. **2**(2): p. 127-158.

78. Galperin, M.Y., *The Molecular Biology Database Collection: 2005 update.* Nucleic Acids Research, 2005. **33**(1).

79. ClinicalTrials. *ClinicalTrials*. U.S. National Library of Medicine 2006 20-7-2006 [visitado 2006 24 Nov]; http://www.clinicaltrials.gov/.

80. OMIM. *OMIM Help*. 2005 [visitado 24-3-2006]; http://www.ncbi.nlm.nih.gov/Omim/omimhelp.html.

81. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Research, 2005. **33**(D514-D517).

82. dbSNP. *Single Nucleotide Polymorphism (NCBI).* National Center for Biotechnology Information 2006 25-5-2006 [visitado 2006 25 Nov]; http://www.ncbi.nlm.nih.gov/projects/SNP/.

83. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Research, 2001. **29**(No 1): p. 308-311.

84. EntrezGene. *Entrez Gene - Searchable database of genes*. National Center for Biotechnology Information (NCBI) 2006 [visitado 2006 25 Nov]; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene.

85. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Research, 2005. **33**(D54–D58).

86. Maglott, D., K. Pruitt, and T. Tatusova, *Entrez Gene: A Directory of Genes.* NCBI Handbook, 2005.

87. Wu, C.H., et al., *The Universal Protein Resource (UniProt): an expanding universe of protein information.* Nucleic Acids Research, 2006. **34**(D187-D191).

88. Mulder, N.J., et al., *InterPro, progress and status in 2005.* Nucleic Acids Research, 2005. **33**(D201-D205).

89. Kegg. *KEGG - Kyoto Encyclopedia of Genes and Genomes*. Kanehisa Laboratories 2006 Oct 2006 [visitado 2006 25 Nov]; http://www.genome.jp/kegg/.

90. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Research, 2004. **33**(D277-D280).

91. Rebhan, M., et al., *GeneCards: integrating information about genes, proteins and diseases.* Trends in Genetics, Elsevier Science, 1997. **13**(4): p. 163-163.

92. Safran, M., et al., *GeneCards 2002: towards a complete, object-oriented, human gene compendium.* Bioinformatics, 2002. **18**(11): p. 1542-1543.

93. Hewett, M., et al., *PharmGKB: the Pharmacogenetics Knowledge Base.* Nucleic Acids Research, 2002. **163-165**(1): p. 163-165.

94. HGNC. *About the HGNC.* HUGO Gene Nomenclature Committee 2006 Aug, 1 [visitado 2006 17-Aug]; http://www.gene.ucl.ac.uk/nomenclature/aboutHGNC.html.

95. Wain, H.M., et al., *Guidelines for Human Gene Nomenclature.* Genomics, 2002. **79**(4): p. 464-470.

96. GeneOntology. *An Introduction to the Gene Ontology*. Gene Ontology Consortium 2006 May, 24 [visitado 2006 18 Agosto]; http://www.geneontology.org/GO.doc.shtml.

97. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nature Genetics, 2000. **25**(1): p. 25-29.

98. Westbrook, J., et al., *The Protein Data Bank and structural genomics.* Nucleic Acids Research, 2003. **31**: p. 489 - 491.

99. Hulo, N., et al., *The PROSITE database.* Nucleic Acids Research, 2006. **34**(D227 - D230).

100. Bairoch, A., *The ENZYME database in 2000.* Nucleic Acids Research, 2000. **28**: p. 304 - 305.

101. Palakal, M., et al., *An intelligent biological information management system.* Bioinformatics, 2002. **18**(10): p. 1283-1288.

102. MorbidMap. *Omim Morbidmap table*. 2006 [visitado 2006 15 Março]; ftp://ftp.ncbi.nih.gov/repository/OMIM/morbidmap.

103. regex. *Regular Expressions in Java*. 2005 [visitado 2005 17 Junho]; http://www.regular-expressions.info/java.html.

104. Struts. *The Apache Software Foundation - Struts*. The Apache Software Foundation 2006 15-10-2006 [visitado 2006 30 Outubro]; http://struts.apache.org/.

105. Jakarta. *The Apache Jakarta Project*. The Apache Software Foundation 2006 [visitado 2006 Dez 2006]; http://jakarta.apache.org/.

106. Campbell, M. and L. Heyer, *Discovering Genomics, Proteomics, and Bioinformatics*. 2003.

107. Sowa, J., *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Vol. 13. 2000: Brooks/Cole.

8 Anexos

8.1 Polimorfismos

Polimorfismos ou SNPs (*Single Nucleotide Polymorphism*) são pequenas bases num local genético específico (*Locus*) que variam entre indivíduos da mesma espécie e que são responsáveis por características fenotípicas individuais ou de sub-populações tais como por exemplo a propensão para doenças complexas como o cancro.

Supondo que 90% dos humanos têm a seguinte sequência de nucleótidos num dado local num cromossoma

```
GCATGCATGCATGCAT
||||||||||||||||
CGTACGTACGTACGTA
```

e que apenas 10% têm a sequência seguinte para a mesma localização

```
GCATGCAaGCATGCAT
|||||||||||||||
CGTACGTtCGTACGTA
```

então este local (*Locus*) é um polimorfismo do tipo SNP. A frequência de ocorrência deve ser superior a 1%.

De um modo geral, um SNP pode ser detectado sobrepondo e alinhando sequências de DNA e identificando posições no alinhamento onde as mesmas bases não ocorrem em todos os alinhamentos.

Estão identificados cerca de 1,4 milhões de SNPs no genoma humano com uma periodicidade média de 1 por cada 2 mil bases (2Kb).

Além da identificação de genes causadores de doenças, existem ainda motivações importantes para a pesquisa de SNPs entre as quais se destacam o estudo da evolução com base em mutações em sub-populações específicas, utilização de SNPs em *DNA Fingerprinting* para verificações criminais ou de parentesco, fabrico de marcadores para mapear traços poligenéticos e criação de medicamentos à medida do genótipo do individuo [106].

8.2 Mapas das vias metabólicas KEGG

A base de dados KEGG PATHWAY contém uma colecção de diagramas (*KEGG Pathway Maps*) que representam redes de interacções moleculares em vários processos celulares. Cada diagrama ou via é desenhado segundo a notação apresentada na figura seguinte. KEGG Pathway Maps

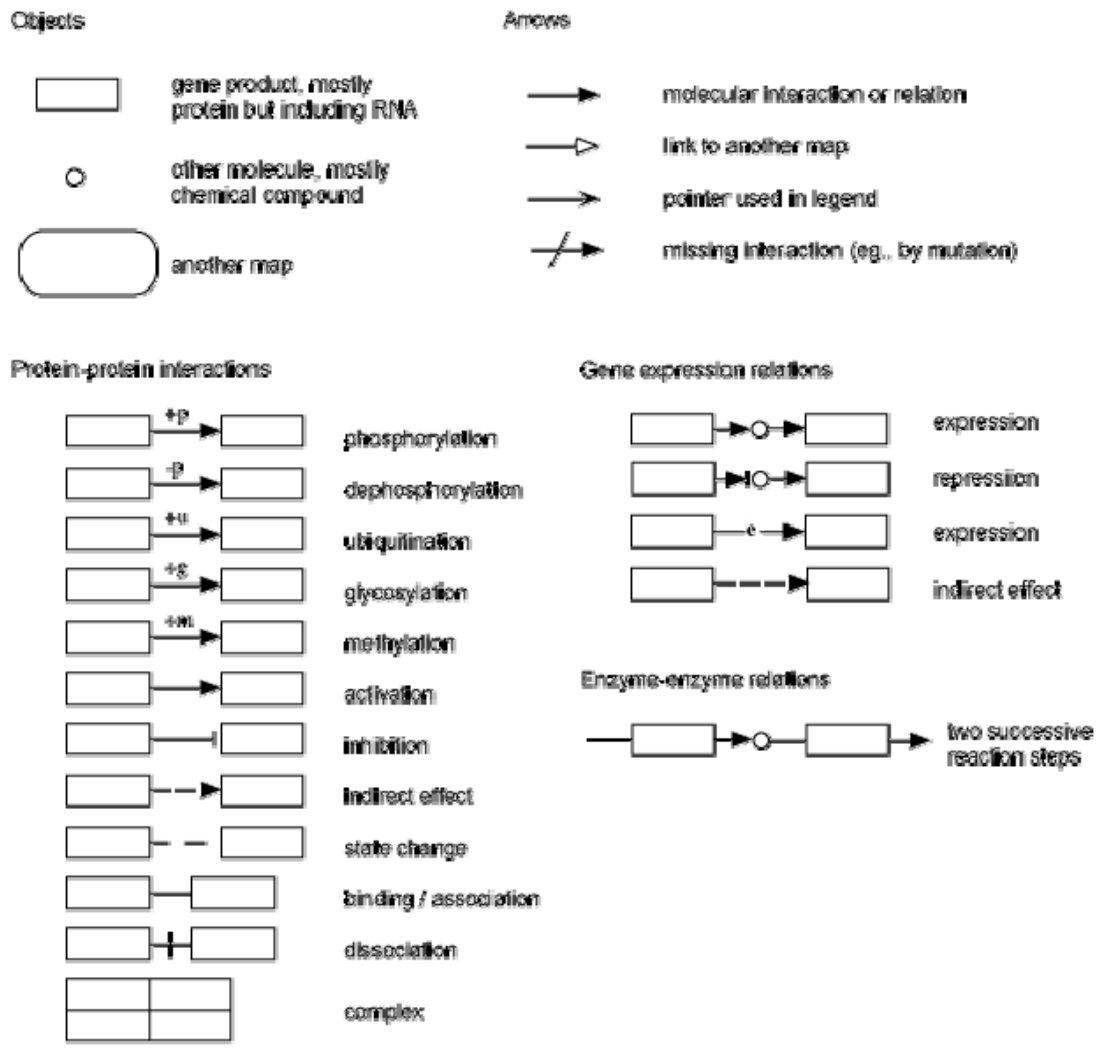


Figura 8.1 – Notação para a representação simbólica das vias metabólicas na base de dados KEGG PATHWAY.

8.3 Glossário

Alelos: Um alelo é cada uma das várias formas alternativas do mesmo gene, ocupando um dado *locus* (posição) num cromossoma. Por exemplo, o gene que determina a cor da flor em várias espécies de plantas – um único gene controla a cor das pétalas, podendo haver diferentes versões desse mesmo gene. Uma dessas versões pode resultar em pétalas vermelhas, enquanto outra versão originará pétalas brancas.

Citocinese: Estrangulamento do citoplasma para fins de divisão celular. Dá-se por meio da interacção actina-miosina.

Codão: Sequência de três bases de ADN que codificam um determinado aminoácido ou seja, cada conjunto de três bases consecutivas é responsável pela codificação de um aminoácido. Existem ainda alguns codões que ou indicam o início (AUG) ou o fim (UAA, UAG ou UGA) da transcrição da respectiva cadeia de ARNm.

Doença órfã – A medicina já identificou mais de 8.000 patologias raras, conhecidas como doenças órfãs devido à pequena quantidade de pacientes e ao interesse diminuto em desenvolver fármacos para elas.

Doença recessiva – Doenças genéticas nas quais duas mutações devem estar presentes (nos cromossomas maternos e paternos) para que o paciente desenvolva a doença.

Domínio: São partições da proteína, as quais são unidades funcionais elementares;

Família: Conjunto de genes num dado genoma, que descende de um gene comum.

Farmacogenómica: Ciência que examina variações nos genes ("*inherited genes*") que influenciam a resposta às drogas, explorando modos de prever o tipo de resposta a uma dada droga por parte de um paciente.

Fenótipo de um organismo é, quer a sua constituição e aparência física, quer a manifestação específica de uma característica, como o tamanho ou a cor dos olhos, que varia entre indivíduos. O fenótipo é determinado até certo ponto pelo genótipo, ou seja pela identidade dos alelos que um indivíduo possui num ou em mais locais dos cromossomas. Grande parte dos fenótipos é determinada por genes múltiplos e influenciados por factores ambientais. Assim, nem sempre a identidade de um ou de alguns alelos conhecidos permite prever o fenótipo.

Fissão binária: Em biologia celular é o nome dado ao processo de reprodução assexuada dos organismos procariótas e que consiste na divisão de uma célula em duas, cada uma com o mesmo genoma da célula mãe.

***Gene Locus*:** A posição do gene no cromossoma.

***Gene Ontology*:** Vocabulário de termos relativos à função molecular, processo biológico ou componentes celulares, desenvolvido pelo *Gene Ontology Consortium*.

Haplótipo: conjunto de alelos para um dado cromossoma. Cada pessoa tem dois haplótipos numa dada região.

Heterozigoto: refere-se a pares de genes que apresentam um gene diferente do outro, sendo sempre um recessivo (possui menor capacidade de manifestar suas características, manifestando-as apenas em homozigose) e outro dominante (possui maior capacidade de manifestar suas características).

Heterozigosidade: A presença de diferentes alelos num ou mais *loci* em cromossomas homólogos.

Homólogos (Cromossomas): Dois cromossomas são homólogos se carregam alelos para as mesmas características.

Homozigoto: Um organismo que tem dois alelos idênticos para um dado gene;

Linkage: especifica quão próximos estão dois *loci* num dado cromossoma. Se forem muito próximos diz-se que os *loci* estão ligados.

Linkage desiquilibrium (LD): descreve alelos em vez de *loci*. Se dois alelos tendem a ser herdados mais do que era previsto diz-se que os alelos estão em LD.

Metabolito: Qualquer produto ou intermediário do metabolismo.

Microsatélites (*Microsatellite loci*): Regiões no genoma de grande variabilidade entre indivíduos da mesma espécie e que podem ser detectadas utilizando PCR (*Polymerase Chain Reaction*) e CE (*capillary electrophoresis*).

MIM (número): um número único de 6 dígitos que representa uma entrada para o catálogo de genes humanos e de doenças genéticas. O primeiro dígito do número OMIM descreve o tipo de hereditariedade do respectivo gene.

Mitose: Divisão nuclear na célula que produz dois núcleos filhos idênticos ao original. Mitose é o processo da divisão nuclear, duplicando os cromossomas. A mitose é seguida pela divisão da membrana celular e do citoplasma, denominada citocinese. A mitose em conjunto com a citocinese gera duas células idênticas.

Ortólogo(s): Genes em espécies diferentes que evoluíram de um ancestral comum. Estes podem divergir mas usualmente possuem semelhanças ao nível da sua estrutura, sequência genética e função.

Parálogo(s): Genes relacionados pelo processo de duplicação dentro do mesmo genoma. Enquanto que os ortólogos retêm a mesma função no decurso da evolução, os parálogos evoluem para funções diferentes

PCR (Reacção de Polimerização em Cadeia) (*Polymerase Chain Reaction*): é um método de amplificação (de criação de múltiplas cópias) de ADN (ácido desoxirribonucleico) sem o uso de um organismo vivo, por exemplo, *E. coli* ou leveduras.

Plasmídeo: Molécula circular dupla de ADN que está separada do ADN cromossómico e capaz de se replicar autonomamente. Os plasmídeos contêm geralmente um ou dois genes que conferem uma vantagem selectiva à bactéria que os abriga, por exemplo, a capacidade de construir uma resistência aos antibióticos.

Semântica: determina como é que as constantes e variáveis no domínio de uma aplicação estão associadas às coisas que representam no domínio real [107].

SNP's (*Single Nucleotide Polymorphism*): Também chamados de *snips* são pequenas variações encontradas nos genes do genoma humano.

Transdução: Processo de reprodução através do qual o ADN bacteriano é transferido de uma bactéria para a outra por intermédio de um vírus (bacteriófagos).