FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Redes de co-expressão entre genes codificantes de proteínas mitocondriais e todos os restantes genes nos vários tecidos humanos

João Alexandre Ribeiro de Almeida

Mestrado Integrado em Engenharia Informática e Computação

Orientadora: Dra. Luísa Pereira (I3S / IPATIMUP)

Co-orientador: Prof. Rui Camacho (FEUP)

27 de Junho de 2017

Redes de co-expressão entre genes codificantes de proteínas mitocondriais e todos os restantes genes nos vários tecidos humanos

João Alexandre Ribeiro de Almeida

Mestrado Integrado em Engenharia Informática e Computação

27 de Junho de 2017

Resumo

Avanços na sequenciação de genomas permitem estudar, para diferentes contextos, a actividade de todos os genes humanos identificados (cerca de 22.000 genes codificantes de proteínas). Contudo, o conhecimento actual sobre relações entre genes encontra-se longe de estar totalmente adquirido, nomeadamente quando pelo menos um dos elementos é um gene que codifica uma proteína mitocondrial (cerca de 1500). As proteínas mitocondriais são codificadas quer pelo DNA mitocondrial (*mtDNA*; 13 proteínas) quer pelo *DNA* nuclear (*nDNA*; as restantes), o que implica uma comunicação controlada entre os dois genomas. Uma vez que as mitocôndrias coordenam várias actividades celulares essenciais à vida, nomeadamente a produção de energia e a morte celular, a desregulação desta comunicação está implicada em muitas doenças complexas como doenças neurodegenerativas, cancro e diabetes.

Assim, este trabalho teve como objetivo identificar os grupos de co-expressão elevada (significativa) entre os pares de genes mitocondrial-todos os genes e as redes proteicas associadas em tecidos humanos. Os dados de expressão de genes em tecidos humanos saudáveis foram recolhidos da base de dados *Genotype-Tissue Expression* (https://www.gtexportal.org/home/), contabilizando 49 tecidos (um total de 8527 amostras, média de 174 por tecido). Os dados foram curados para inclusão de apenas genes codificantes de proteínas e fisicamente não sobreponíveis (só um dos genes sobreponíveis foi mantido). Os valores de correlação de *Pearson* foram calculados em todos os pares gene mitocondrial-todos os genes proteicos, tendo-se eliminado os *outliers* que não se incluíam no intervalo $[u_x - 4SD, u_x + 4SD]$ ou $[u_y - 4SD, u_y + 4SD]$, em que SD corresponde ao desvio padrão (*standard deviation*). Todos os pares com valores de correlação acima de 0.9 e 0.8, o que corresponde a elevada quantidade de informação, foram representados em estrutura de grafos e analisados com técnicas de *Data Mining*, nomeadamente *clustering*, de modo a extrair informação útil. Para a análise das redes foi utilizada a ferramenta Cytoscape, que permitiu avaliar vários parâmetros de extensão e conexão das redes de genes correlacionados nos vários tecidos humanos. Estas redes foram enriquecidas com dados funcionais (*pathways*) das bases de dados *Kyoto Encyclopedia of Genes and Genomes* (http://www.genome.jp/kegg/) e *Gene Ontology* (http://www.geneontology.org/), que permitem inferir acerca da possível função exercida pelos genes correlacionados. De modo a comparar os dados funcionais entre tecidos, procedeu-se à técnica de *clustering* hierárquico, pela construção de matrizes binárias, matrizes de semelhança pelo método *Jaccard* e aplicação dos métodos de aglomeração *UPGMA* (*Unweighted Pair Group Method with Arithmetic Mean*) e *NJ* (*Neighbor Joining*). Foi desenvolvida uma plataforma *web* para visualizar e analisar, de forma interactiva, as árvores resultantes deste métodos.

Em termos biológicos, constatamos que existem pares de genes mitocondrial-todos os genes proteicos altamente correlacionados e que estes estão incluídos em *pathways* de elevada importância funcional como a produção de energia e síntese de metabolitos. As redes são maiores e mais densas para os tecidos do cérebro, enquanto tecidos como o rim, sangue e fibroblastos apresentam também um elevado número de genes correlacionados mas não tão interconetados. Em geral, as elevadas correlações entre genes mitocondriais codificados pelo *mtDNA* limitam-se a genes

codificados por este genoma, enquanto que genes mitocondriais codificados pelo *nDNA* se correlacionam significativamente com outros genes (mitocondriais ou não) codificados pelo *nDNA*. O que prova que a correlação entre genes codificados pelo mesmo genoma é mais eficiente.

Toda a pipeline desenvolvida neste trabalho bem como a plataforma *web* será disponibilizada na plataforma *GitHub* em *open source* acompanhada da documentação de instalação para que possa ser facilmente utilizada ou adaptada a outras análises semelhantes nos muitos dados que vão sendo publicados, no contexto de doenças ou de outras espécies.

Abstract

Recent advances in genome sequencing allow the study, at different contexts, of all identified human gene activities ($\approx$ 22.000 protein encoding genes). However, current knowledge on gene interactions lags behind, especially when one of the elements is a mitochondrial protein encoding gene ($\approx$ 1500). Mitochondrial proteins are encoded either by mitochondrial DNA (mtDNA; 13 proteins) or by nuclear DNA (nDNA; the remaining), which implies a coordinated communication between the two genomes. Since mitochondria coordinate several life-critical cellular activities, namely energy production and cell death, deregulation of this communication is implicated in many complex diseases such as neurodegenerative diseases, cancer and diabetes.

Thus, this work aimed to identify high co-expression groups between mitochondrial genes-all genes, and associated protein networks in human tissues. Gene expression data for tissues were collected from the Genotype-Tissue Expression database (https://www.gtexportal.org/home/) counting 49 tissues (a total of 8527 samples, an average of 174 per tissue). The data was filtered to include only protein-encoding and physically non-overlapping genes (only one of the overlapping genes was maintained). Pearson's correlation values were calculated on all pairs of mitochondrial genes-all protein encoding genes, and outliers in the range $[u_x - 4SD, u_x + 4SD]$ or $[u_y - 4SD, u_y + 4SD]$ (SD stands for standard deviation) were excluded. Gene pairs with a correlation higher than 0.9 and 0.8, corresponding to big datasets, were represented in graph structures and analyzed by Data Mining clustering techniques in order to help extracting important information. Cytoscape software was used for graph analysis, allowing to evaluate complex network parameters and identify connection properties on the biological networks. The networks were enriched with functional data (pathways) from two different biological databases: Kyoto Encyclopedia of Genes and Genomes (https://www.genome.kp/kegg) and Gene Ontology (http://www.geneontology.org). This network enrichment helped to infer biological functions of the correlated genes. Functional data comparison between tissues was conducted through hierarchical clustering techniques, by building binary matrices, similarity matrices using Jaccard index and applying agglomeration methods such as UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and NJ (Neighbor Joining). A web platform was built to interactively visualize and analyze the trees resulting from these methods.

Biologically, we confirmed the existence of highly correlated pairs of mitochondrial-all protein encoding genes, which are included in pathways of functional importance such as energy production and metabolite synthesis. Brain tissues have the largest and most dense networks, while kidney cortex, whole blood and fibroblasts had large but sparser networks. Generally, the strongest correlation between mitochondrial genes encoded by mtDNA belong to genes encoded by this genome, while mitochondrial genes encoded by nDNA are significantly correlated with other genes (mitochondrial or not) encoded by nDNA. This proves that correlation among genes encoded by the same genome is more efficient.

The pipeline and the web tree viewer developed in this work will be available at GitHub under open source distribution along with installation documentation. This will make it possible to use

and adapt the tools to the analyses of datasets being released to the public, in the context of diseases or other species.

Agradecimentos

Gostaria de agradecer primeiramente à minha família que me apoiou desde o início ao término do meu percurso académico e aos meus amigos que sempre me apoiaram nos momentos menos bons. Uma palavra de agradecimento especial à minha orientadora Dr. Luísa Pereira e ao meu co-orientador Prof. Rui Camacho por todo o trabalho incansável que tiveram em orientar-me durante todo o percurso deste trabalho. Por último mas não menos importante, agradecer a todos os colaboradores do i3S, mais especificamente ao grupo de diversidade genética, por terem tornado o meu dia-a-dia de trabalho mais agradável.

João Almeida

"There is a crack in everything,
that is how the light comes in."

Leonard Cohen

Conteúdo

Lista de Figuras

Lista de Tabelas

Abreviaturas e Símbolos

CSV	Comma-separated values
DBSCAN	Density-based spatial clustering of applications with noise
DM	Data Mining
DNA	Deoxyribonucleic Acid
gct	Gene Cluster Test
gmt	Gene Matrix Transposed
GTEx	The Genotype-Tissue Expression project
GO	Gene Ontology
ID3	Iterative Dichotomiser 3
JSON	JavaScript Object Notation
mRNA	Messenger Ribonucleic Acid
mtDNA	Mitochondrial Deoxyribonucleic Acid
nDNA	Nuclear Deoxyribonucleic Acid
NJ	Neighbor Joining
K-NN	K-Nearest Neighbors
KDD	Knowledge Discovery in Databases
KEGG	Kyoto Encyclopedia of Genes and Genomes
PPI	Protein-protein Interactions
RNA	Ribonucleic Acid
TCGA	The Cancer Genome Atlas
tRNA	Transfer Ribonucleic Acid
UPGMA	Unweighted Pair Group Method with Arithmetic Mean

Capítulo 1

Introdução

1.1 Contexto

O facto de actualmente já ser possível a sequenciação total do genoma humano permite um estudo mais aprofundado do mesmo, levantando novas questões científicas. A grande quantidade de informação gerada pelas sequenciações em larga escala trazem novos desafios na área da Bioinformática. O estudo destes dados permite avaliar de uma forma mais refinada a influência do genoma humano no fenótipo de um indivíduo e qual a relação dos genes com as doenças denominadas complexas. O genoma humano está organizado em genoma nuclear (*nDNA*) concentrado no núcleo e genoma mitocondrial (*mtDNA*) localizado nos organelos citoplasmáticos designados mitocôndrias. As mitocôndrias são responsáveis pela produção da energia celular e, como tal, desempenham um papel essencial na vida celular, estando já implicadas em muitas doenças complexas como doenças neurodegenerativas, cancro, diabetes, etc. Todos os genes proteicos mitocondriais (13) codificam proteínas mitocondriais, mas 99% das proteínas mitocondriais são codificadas pelos genes nucleares (entre 1000-2000 genes). Assim, os dois genomas têm que estar finamente coordenados por mecanismos ainda largamente desconhecidos. É por isso importante compreender como os genes do *nDNA* e *mtDNA* se relacionam entre si nos vários tecidos do organismo, pela análise de correlações de expressão génica, em estado saudável e em situação de doença.

A existência de bases de dados públicas como o *GTEx* e o *TCGA* permite abordar esta questão. Estas bases de dados disponibilizam dados de expressão génica de todos os genes humanos codificados em vários tecidos do organismo, em indivíduos saudáveis (que morreram maioritariamente devido a acidentes) no *GTEx* e em pacientes de cancro no *TCGA*.

1.2 Motivação e Objetivos

O acesso público a grandes quantidades de dados genómicos e transcriptómicos abre portas para estudos aprofundados. A motivação deste trabalho surge da falta de conhecimento em como o *nDNA* se coordena com o *mtDNA* na codificação de proteínas mitocondriais. A identificação

de grupos de co-expressão entres pares de genes pode fornecer informações importantes sobre as redes proteicas e como estas diferem de tecido para tecido. O estudo, em específico, de pares em que um dos membros é uma proteína mitocondrial (seja esta codificada pelo *mtDNA* ou *nDNA*) pode contribuir com informação essencial sobre como a interacção entre os dois genomas.

Assim, este trabalho tem como objectivo identificar grupos de co-expressão entre pares de genes que codificam proteínas mitocondriais *versus* todas as proteínas nos vários tecidos do organismo em situação saudável, de modo a perceber de que forma o *nDNA* e o *mtDNA* se encontram coordenados na expressão fenotípica do indivíduo. É também objectivo deste trabalho a realização de uma *pipeline* genérica que permita aplicar o mesmo tipo de análise a outros dados, relativos a outros organismos ou doenças. Os resultados destas análises serão utilizados para identificação e ponto de partida de possíveis casos estudo.

1.3 Estrutura da Dissertação

Para além da introdução, esta dissertação é composta por mais 4 capítulos. No capítulo 2, são apresentados conceitos biológicos básicos e fundamentais para perceber o trabalho realizado e é descrito o estado de arte em *Data Mining*, Teoria de Grafos e *Big Data*. No capítulo 3, é referida a implementação do trabalho, detalhando todas as etapas da análise dos dados e discutindo os métodos e técnicas utilizadas. Os detalhes da plataforma *web BioTree Viewer* são também discutidos neste capítulo. No capítulo 4 são apresentados os resultados da análise dos dados e são demonstrados o conjunto de funcionalidades da plataforma desenvolvida e potenciais usos. Finalmente, no capítulo 5 são apresentadas as conclusões deste trabalho.

Capítulo 2

Biologia Molecular, Genómica e Análise de Dados

Neste capítulo apresentam-se os conceitos básicos relevantes para o nosso estudo e é também descrito o estado de arte. São referidos trabalhos relacionados para mostrar o que existe no mesmo domínio e quais os problemas em aberto. Por fim, são descritas tecnologias de Análise de Dados relevantes para o estudo.

2.1 Biologia Molecular

As células são a unidade estrutural dos seres vivos. Nos organismos eucarióticos, existem compartimentos celulares designados organelos, como o núcleo, mitocôndrias, cloroplastos (só existentes nas células vegetais), retículo endoplasmático, ribossomas, vacúolos, etc. O núcleo é um dos organelos mais importantes pois é neste que estão situados os cromossomas que contêm o ácido desoxirribonucleico (*DNA*). O *DNA* contém as instruções genéticas que coordenam o funcionamento e o desenvolvimento dos seres vivos.

2.1.1 *DNA*

O *DNA* é constituído por unidades de bases azotadas denominadas nucleótidos (Figura 2.1). Cada nucleótido contém uma molécula de açúcar (do tipo desoxirribose), um grupo fosfato e ainda uma base azotada. Existem quatro tipos de bases azotadas: Citosina (C), Guanina (G), Adenina (A) e Timina (T). Os nucleótidos ligam-se entre si através dos grupos fosfato, formando uma longa molécula ou cadeia e é a ligação de duas cadeias que forma a estrutura em dupla-hélice. Esta ligação em dupla-hélice faz-se através das bases azotadas, sendo altamente específica ou complementar: a Adenina só se liga à Timina e a Citosina à Guanina.

A informação contida no *DNA* tem que ser transformada em proteínas de modo à célula conseguir interpretar essa informação. A sequência de nucleótidos que contém a informação para uma

determinada proteína recebe a designação de gene. A espécie humana tem cerca de 22.000 genes [VAM+01], distribuídos ao longo de cerca de 6.469,66 Mega pares de bases, organizadas em 23 pares de cromossomas, localizados no núcleo. O Homem é um organismo diplóide, o que significa que possui dois pares de cromossomas homólogos, um par recebido da mãe e o outro recebido do pai. Destes, 22 pares são designados por autossomas (o cromossoma paterno é igual ao materno) e um par é designado por cromossomas sexuais (os cromossomas são iguais e designados por X no sexo feminino; são diferentes, um X e um Y, no sexo masculino). O conjunto de cromossomas de um organismo é também conhecido como genoma.

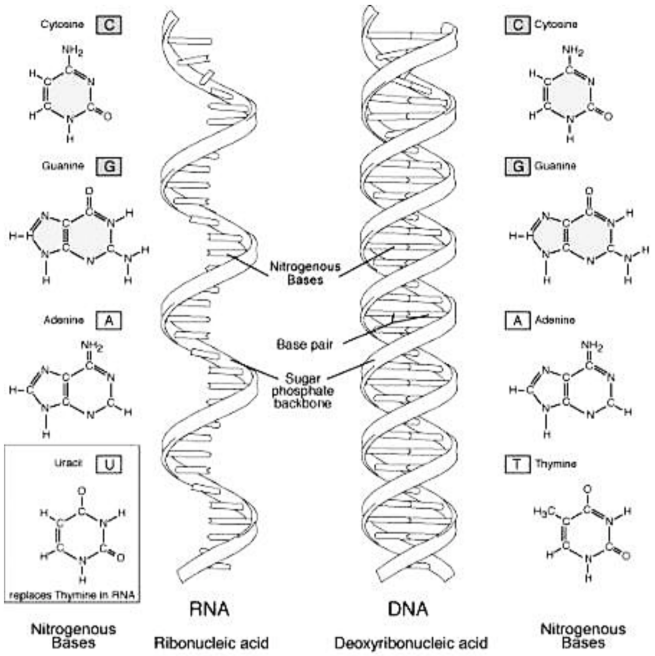


Figura 2.1: Comparação entre *DNA* e *RNA* [Suc].

2.1.2 *DNA* Mitocondrial

Apesar de um genoma de um organismo eucariota estar localizado principalmente no núcleo, alguns organelos do citoplasma contêm o seu próprio *DNA*. Este é o caso das mitocôndrias e, nas plantas, também dos cloroplastos.

As mitocôndrias são organelos intracelulares compostos por duas membranas e estão presentes em quase todas as células eucarióticas, nomeadamente nas células nucleadas dos mamíferos. As mitocôndrias estão muito relacionadas com a homeostasia celular. Estas exercem funções muito importantes na apoptose (morte celular), na conversão de nutrientes em componentes celulares (metabolismo intermediário) mais especificamente o ciclo de Krebs, e no metabolismo de aminoácidos, lípidos entre outras funções. As mitocôndrias são as principais responsáveis pela produção de energia celular, uma vez que são capazes de sintetizar energia *ATP* (*Adenosine triphosphate*) essencial para as células, num processo designado por cadeia respiratória (Figure 2.2).

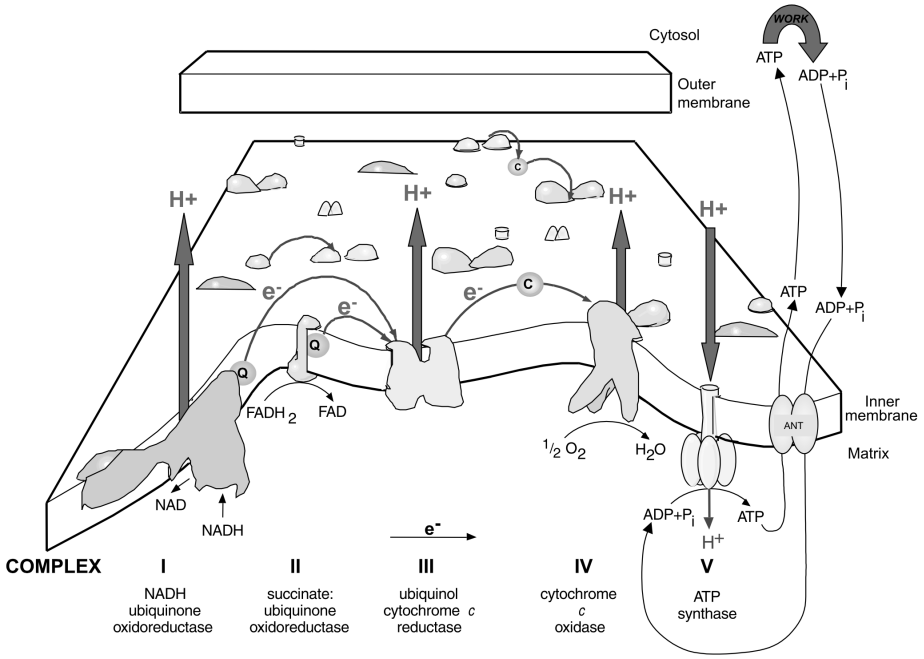


Figura 2.2: Detalhe da cadeia respiratória presente nas mitocôndrias [Chi].

A primeira sequência de *DNA* mitocondrial (*mtDNA*) humano foi publicada em 1981 [TAK+99]. Este acontecimento impulsionou o interesse no estudo do genoma mitocondrial de forma a perceber qual a sua relação com a evolução humana e doenças. Actualmente, já é possível sequenciar genomas mitocôndriais completos em larga escala trazendo novos desafios bioinformáticos devido à grande quantidade de informação gerada. O *mtDNA* humano contém 16569 pares de bases e codifica 13 proteínas da cadeia respiratória, bem como dois *RNAs* ribossómicos (*rRNA*) e 22 *tRNAs*. Mas a quantidade de proteínas mitocondriais é muito superior – pelo menos entre 1000-2000 proteínas estão já catalogadas nas mitocôndrias, muitas delas fazendo também parte da cadeia respiratória, formando complexos com as proteínas codificadas pelo *mtDNA*. Essas cerca de 98% proteínas mitocondriais são codificadas pelo *DNA* nuclear (*nDNA*), traduzidas no citoplasma e depois transportadas para as mitocôndrias [Sho01]. Sabe-se que os genes nucleares têm um papel importante nomeadamente no controlo do genoma mitocondrial, como os que codificam a *mtDNA* polimerase γ (POLG1) [GDL+01], produtos que garantem um equilíbrio de nucleótidos livres dentro da mitocôndria [Nis99], a proteína *Twinkle* (sintetizada pelo gene C10orf2) que parece intervir na manutenção do mtDNA [SLT+01]. O *nDNA* também produz alguns factores de transcrição e tradução mitocondrial [FGR+02] . Pode-se afirmar que uma mutação quer nos genes do *nDNA* ou do *mtDNA* podem gerar disfunção nas mitocôndrias.

2.1.3 Mutações no *mtDNA* e doenças

Estudos epidemiológicos recentes demonstraram que o aparecimento de doenças devido a mutações ao nível do *mtDNA* é mais comum do que se imaginava [MMU+98].

Este tipo de mutações pode ser encontrado em pelo menos 1 em cada 8000 Europeus e uma doença causada por uma mutação no *mtDNA* em 1 em cada 15000 adultos. A maior parte das crianças com uma disfunção mitocondrial têm uma mutação num gene nuclear que se revela importante na cadeia respiratória [URR+00]. Mas as mutações também ocorrem no *mtDNA*, dividindo-se em dois grupos: mutações pontuais (alteração das bases azotadas no gene) e rearranjos (inserções, deleções e inversões de partes da molécula) [SBD97, Wal99, DS01].

No caso das mutações serem no *mtDNA*, há um factor adicional a ter em conta devido a todas as células diplóides humanas conterem milhares de cópias de *mtDNA*, em comparação com as duas cópias no *nDNA*. À nascença, o material genético das diferentes cópias é idêntico (homoplasmia intracelular), mas ao longo da vida vão-se acumulando mutações somáticas que levam a diferentes populações de *mtDNA* na mesma célula ou tecido (heteroplasmia) [LC95]. A proporção de *mtDNA* mutante varia de célula para célula. Estudos demonstraram que a proporção de *mtDNA* mutante tem que exceder um limite crítico (normalmente entre 50% e 85%) para que exista um defeito bioquímico na cadeia respiratória de célula [SBD97]. O valor limite varia de mutação para mutação e consoante o tecido em que é aplicável. Estes fatores foram induzidos através da observação de relações genéticas entre diferentes indivíduos que pertencem a determinada descendência (*pedigrees*) em que estes transmitiram *mtDNA* mutante para os seus descendentes [WBM+98]. Estudos recentes apontam que o *nDNA* pode ser responsável por regular os níveis de heteroplasmia presentes na mitocôndria, desempenhando um papel adicional na determinação do fenótipo de uma doença no *mtDNA* [BS01].

2.1.4 Expressão Génica

A expressão génica faz-se através de um processo denominado por síntese proteica, que é constituído por duas fases, em que a informação contida no *DNA* é convertida na proteína que codifica. Este processo é essencial para que a célula funcione. A primeira fase é a transcrição do *DNA*. Nesta fase, a informação que se encontra no *DNA* é copiada para um *RNA* (ácido ribonucleico) mensageiro (*mRNA*) com tempo de vida curto. A estrutura do *RNA* é diferente da do *DNA*. O *RNA* contém ribose em vez de desoxirribose e as 4 bases azotadas que se ligam à ribose são a Citosina (C), Guanina (G), Adenina (A) e Uracilo (U) e forma uma molécula em cadeia simples. Durante a fase de transcrição o *DNA* desemparelha a cadeia dupla e o *mRNA* é gerado através do emparelhamento das bases do *mRNA* de forma complementar com as bases do *DNA* (com a diferença que a Timina é substituída por Uracilo: C-G e T-U). Após a produção do *mRNA*, este é transportado do núcleo para o citoplasma, onde vai ser traduzido num organelo denominado ribossoma. Durante esta segunda fase, a leitura do *mRNA* faz-se de três em três bases, em que cada tripleto ou codão é traduzido num aminoácido que irá constituir a proteína. O tradutor é um outro tipo de *RNA*, o *RNA* transportador ou *tRNA*, que possui numa extremidade o aminoácido que corresponde a um determinado conjunto de três bases azotadas, o anti-codão que vão ser complementares ao *mRNA* que está a ser traduzido. Este código, que permite a tradução da informação do *mRNA* a proteína é conhecido por código genético.

A síntese proteica das proteínas mitocôndriais codificadas pelo *nDNA* decorre da maneira descrita, mas engloba um passo adicional que consiste no transporte dessas proteínas do citoplasma para as mitocôndrias. Maioritariamente, estas proteínas têm uma sequência inicial que funciona como sinal identificador, sendo posteriormente removida quando a proteína já se encontra no interior do organelo. A síntese proteica das proteínas mitocondriais codificadas pelo *mtDNA* decorre no interior das mitocôndrias e não no citoplasma. O código genético mitocondrial é ligeiramente diferente do código genético nuclear, mas o processo é totalmente idêntico [Chi].

Nos organismos multicelulares, com tecidos diferenciados e especializados, os genes não são expressos na sua totalidade em todos os tecidos. Há genes, denominados *housekeeping*, que são expressos em todos os tecidos, mas a maioria tem padrão de expressão variável conforme a especialização do tecido [HDY+01]. Por exemplo, um neurónio precisa de transmitir informação eléctrica, enquanto uma célula renal tem que excretar substâncias tóxicas.

2.1.4.1 Base de dados de expressão génica *GTEx*

O *Genotype-Tissue Expression* é uma base de dados com informação da expressão génica em diferentes tecidos humanos. Esta informação quantitativa foi obtida através da sequenciação de *mRNA* extraído de tecidos *post-mortem* de indivíduos acidentados. Os indivíduos elegíveis tinham idades compreendidas entre os 21 e os 70 anos, numa proporção semelhante entre os géneros e eram maioritariamente caucasianos [LTS+13]. A condição essencial era que a extracção fosse realizada até 24 horas após morte, dado que a degradação do *mRNA* é muito rápida, variando consoante a zona do tecido e o tempo isquémico da amosta (intervalo de tempo em que a amostra não tem irrigação sanguínea). Através da disponibilização pública destes dados, é possível aos investigadores aprofundar o estudo na expressão génica e identificar através dos diferentes níveis de expressão quais as regiões mais influentes do genoma e de que maneira podem afectar a expressão dos genes.

2.1.5 Redes de genes/proteínas

As redes de genes/proteínas representam o conjunto de processos físicos e metabólicos que determinam as propriedades fisiológicas e bioquímicas da célula. Como tal, estas redes compreendem as reacções químicas do metabolismo, as vias metabólicas, bem como as interacções regulatórias que coordenam estas reacções. Em termos de processamento de informação, a comparação de listagens de centenas de genes pode ser pouco esclarecedora em termos biológicos, mas quando a organização em redes é sobre-imposta, a inferência dos processos funcionais é muito mais fácil e robusta.

2.1.5.1 Bases de dados *Gene Ontology* e *Kyoto Encyclopedia of Genes and Genomes*

Dois exemplos de redes de genes/proteínas informativas são as bases de dados *Gene Ontology* (*GO*) e *Kyoto Encyclopedia of Genes and Genomes* (*KEGG*).

A base de dados *GO* fornece informações sobre a funções e interacções de genes e proteínas e classifica-as em diferentes ontologias: processos biológicos (*GO BP*), funções moleculares (*GO MF*) e componentes celulares (*GO CC*) [ABB+00]. A ontologia de processos biológicos refere-se ao objectivo biológico do gene ou proteína. Um processo é atingido através de uma conjunto de funções moleculares. Na maior parte das vezes os processos envolvem uma transformação física ou química, ou seja, o que entra como *input* de um processo é transformado e resulta em algo diferente. Exemplo de alguns processos biológicos são manutenção e crescimento celular (*cell growth and maintenance*), transdução de sinal (*signal transduction*), translação (*translation*) ou biosíntese de *cAMP* (*cAMP biosynthesis*). A função molecular é definida como a actividade bioquímica de um gene da proteína que este codifica. Esta ontologia descreve apenas o resultado sem especificar onde ou qual a situação em que o evento ocorre. Exemplos de alguns termos funcionais são enzima (*enzyme*), transportador (*transporter*) ou ligando (*ligand*). A componente celular indica o local na célula em que o gene se encontra activo. Estes termos refletem o conhecimento actual acerca da estrutura da célula eucariótica e como exemplo podem-se encontrar os termos ribossoma (*ribosome*), proteossoma (*proteasome*) ou membrana nuclear (*nuclear membrane*). Esta base de dados é composta por 41775 termos, em que 27284 referem-se a processos biológicos (*GO BP*), 10733 a funções moleculares (*GO MF*) e 3758 a componentes celulares (*GO CC*) [C+15]. A estrutura de uma ontologia compreende termos e relações entre eles bem definidas. A sua estrutura representa o conhecimento biológico actual e serve também como guia para organizar novos dados. Os dados podem ser classificados em diferentes níveis, do mais geral ao particular. Os termos do *GO* são representados por vértices numa rede e as arestas entre os seus vértices pais (termos de nível mais geral) e filhos (termos de nível mais particular) são conhecidas e formam grafos acíclicos directos.

O *KEGG* é uma enciclopédia que tem como principal objectivo associar genes e genomas a determinas funções desde o nível molecular a níveis mais gerais. O *KEGG* foi originalmente desenvolvido em 1995 como uma base de dados integrada para a interpretação biológica de sequenciação completa de genomas através do mapeamento de vias metabólicas. A base de dados de vias metabólicas *KEGG PATHWAY* é a base de dados principal do *KEGG* e compreende diferentes categorias: metabolismo (*metabolism*), processamento de informação genética (*genetic information processing*), processamento de informação do meio celular (*environmental information processing*), processos celulares (*cellular processes*), sistemas do organismo (*organismal systems*), doenças (*human diseases*) e desenvolvimento de medicamentos (*drug development*) [KFT+17].

2.2 *Big Data*

Big Data é o termo atribuído aos conjuntos de dados aos quais as aplicações de processamento tradicionais não são adequadas devido ao facto de serem grandes ou demasiado complexos. Muitas vezes estes *datasets* trazem um conjunto de desafios como, por exemplo, a sua análise, visualização, transferência, etc. *Big Data* pode ser definida por um conjunto de características [Hil16]:

- Variabilidade - inconsistência dos dados dificulta os processos de tratamento e gestão;

- Veracidade - a qualidade da informação pode ser variável influenciando negativamente a sua análise;
- Variedade - o tipo e a natureza da informação ajuda os especialistas que analizam a informação a obter resultados mais concisos;
- Velocidade - a velocidade com que a informação é gerada e processada de forma a resolver os desafios em tempo adequado;
- Volume - quantidade de informação que é gerada e guardada. O volume de informação determina se efectivamente a informação é considerada *Big Data*.

Foi com esta necessidade de trabalhar com grandes quantidades de informação que começaram a surgir novas tecnologias. Estas tecnologias utilizam diversos *clusters* para a análise dos dados de forma mais rápida e eficaz.

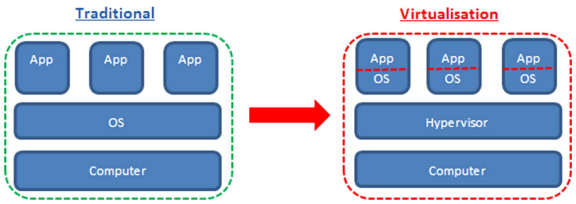


Figura 2.3: Conceptualização da via tradicional de computação em que as aplicações interagem com o hardware através de uma instância do sistema operativo e da evolução para ambientes virtuais onde várias imagens partilham recursos (*CPU, RAM*, armazenamento e rede) que são geridas por *software* de virtualização (*hypervisor* ou *virtual machine monitor*) [ODS13].

2.3 Análise de Dados

2.3.1 Estatística

Para conseguir determinar a relação entre os diferentes genes existentes, é necessário recorrer a ferramentas estatísticas.

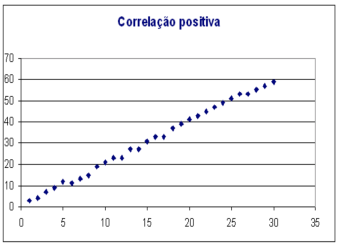
2.3.1.1 Coeficiente de correlação de Pearson

A intensidade da associação linear existente entre as duas variáveis pode ser quantificada através do coeficiente de correlação linear de Pearson, dado pela equação:

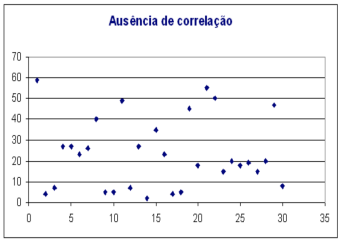
$$r = \frac{C_{X,Y}}{S_X, S_Y}, r \in [-1, 1] \tag{2.1}$$

O valor $C_{X,Y}$ corresponde a covariância ou variância conjunta das váriaveis X e Y, S_X e S_Y correspondem ao desvio padrão das variáveis X e Y, respectivamente. O cálculo da covariância ($C_{X,Y}$) é dado pela equação:

$$C_x, C_Y = \frac{\sum_{i=1}^{n}(x_i - u_x)(y_i - u_y)}{n-1} \tag{2.2}$$



(a) Correlação positiva.



(b) Ausência de correlação.

Figura 2.4: Possíveis resultados entre a correlação de duas variáveis X e Y.

Dependendo do valor final de r é possível determinar se as variáveis se encontram de alguma maneira correlacionadas ou não.

Tabela 2.1: Valor do r e correlação.

Coeficiente de correlação (r)	**Correlação**
$r = 1$	Perfeita positiva
$0,8 \leq r < 1$	Forte positiva
$0,5 \leq r < 0,8$	Moderada positiva
$0,1 \leq r < 0,5$	Fraca positiva
$0 < r < 0,1$	Ínfima positiva
0	Nula
$-0,1 < r < 0$	Ínfima negativa
$-0,5 < r \leq -0,1$	Franca negativa
$-0,8 < r \leq 0,5$	Moderada negativa
$-1 < r \leq -0,8$	Forte negativa
$r = -1$	Perfeita negativa

2.3.1.2 *Outliers*

Algumas observações podem, por vezes, apresentar grandes afastamentos da maioria ou serem inconsistentes. Estas observações são denominadas de *outliers*. O estudo de *outliers*, quaisquer que sejam as suas causas, pode ser realizado em várias etapas.

A fase inicial é responsável pela identificação de potenciais *outliers*. A identificação geralmente é feita por análise gráfica ou no caso de o conjunto de dados ser pequeno pode ser por observação directa dos mesmos. A segunda fase tem como papel principial determinar se os potenciais *outliers* são efectivamente *outliers*. São escolhidos testes adequados para a situação em estudo. Na última fase decide-se o que fazer com os *outliers* previamente identificados. Geralmente, a abordagem utilizada é a eliminação dessas observações, contudo só se justifica no caso de os *outliers* serem provocados por erros cuja correcção é inviável.

Existem um conjunto de técnicas e métodos que permitem identificação de *outliers*. Nos métodos baseados em *clustering* tentam-se formar grupos de dados e os que não se encaixam em nenhum dos grupos são considerados como excepções.

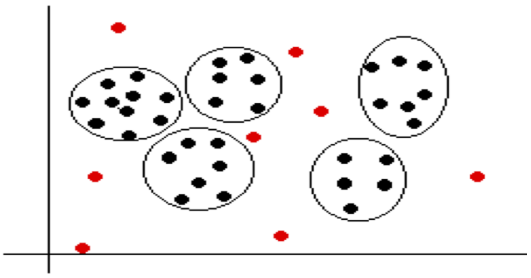


Figura 2.5: Identificação de *outliers* em métodos baseados em *clustering*.

Uma das vantagens destes métodos é que não requer conhecimento prévio de distribuição, no entanto como limitação este optimiza agrupamentos e não diretamente a detecção de excepções.

Nos métodos baseados em estatística assume-se a distribuição ou modelo probabilístico para o *dataset*. Através da realização de testes de discordância é possível identificar os *outliers* com respeito ao modelo probabilístico escolhido. Este tipo de método permite avaliar a significância de uma excepção, contudo o modelo escolhido influencia a identificação dos *outliers*.

Existem também métodos baseados no desvio padrão em que as excepções são definidas como pontos cujo valor desviam da maioria ao longo das dimensões. Mais em específico a busca genética com uma função de selecção, *crossover* e mutação específica para o problema permite encontrar, com um custo muito menor, a maioria das excepções.

Por fim, os métodos baseados na distância vêm resolver algumas limitações dos métodos estatísticos. Um *outlier* é determinado baseado na distância $D^k(p)$, isto é, a distância de p ao seu k-ésimo vizinho. Este método evita a suposição sobre a distribuição dos dados, tem um custo computacional menor, no entanto não é escalável para mais do que 5 dimensões.

2.3.2 Teoria dos Grafos

Um grafo é um par ordenado (V,A) em que V é um conjunto qualquer não vazio e A é um subconjunto de V, o conjunto de todos os pares não-ordenados de V. O conjunto dos elementos de V é denominado de vértice ou nó e o conjunto dos elementos de A é designado de arestas. Se o grafo possuir arestas com direcção, este é designado de grafo orientado.

(a) Grafo orientado, as arestas têm direcção. (b) Grafo não dirigido.

Figura 2.6: Diferenças entre um grafo orientado e um grafo não orientado.

As arestas, independentemente de terem uma direcção ou não, podem ter um peso associado. Este valor pode ter significados diferentes consoante aquilo que o grafo em questão representa. Se um grafo estiver a ser utilizado, por exemplo, como representação de uma rede biológica em que os vértices representam os genes e as arestas as interacções entre eles, o peso de uma aresta pode corresponder ao valor de correlação entres os 2 genes.

Figura 2.7: Grafos em que as arestas possuem um peso associado.

Na teoria dos grafos existem parâmetros, uns mais complexos que outros, que permitem o estudo mais aprofundado das redes nomeadamente o grau de um vértice (*degree*), medidas de centralidade como a intermediação (*betweenness centrality*), proximidade (*closeness centrality*) e radialidade (*radiality*), coeficiente de *clustering*, caminho mais curto entre outros. A análise destes parâmetros permitem tirar conclusões importantes sobre as redes.

O grau de um vértice v indica o número de vértices adjacentes a v. Seja $G = (V,E)$ um grafo não dirigido, o somatório dos graus dos vértices de G pode ser dado por:

$$\sum_{v \in [V]} deg(v) = 2|E| \tag{2.3}$$

O coeficiente topológico (*topological coefficient*) de um nó é uma medida relativa que permite avaliar o número de vizinhos partilhados entre si e os vértices vizinhos. O coeficiente topológico T_n de um vértice n com k_n vizinhos, é calculado por:

$$T_n = \frac{avg(J(n,m))}{K_n} \tag{2.4}$$

em que $J(n,m)$ é definido para todos os vértices m que partilham pelo menos um vértice com n. $J(n,m)$ é o número de vizinhos partilhados entre os nós n e m e se estes forem adjacentes é adicionada uma unidade. Os vértices que possuem menos de 2 nós adjacentes têm um valor de coeficiente topológico nulo.

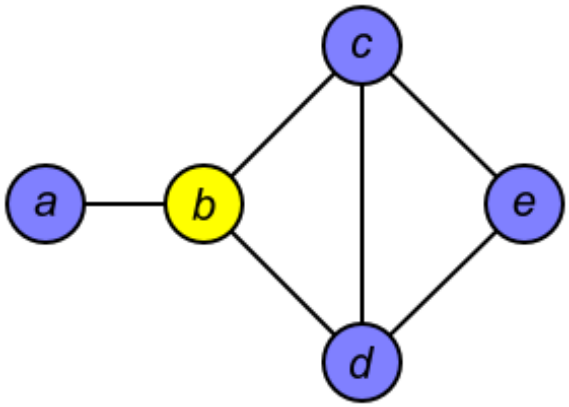


Figura 2.8: Exemplo de rede não dirigida com 5 vértices e 6 arestas.

Utilizando como exemplo a rede da figura 2.8, o coeficiente topológico do vértice b, é dado por:

$$J(b,c) = J(b,d) = J(b,e) = 2$$
$$T_b = 2/3$$

A intermediação (*betweenness centrality*) de um vértice b é um parâmetro que indica a centralidade de b na rede. Calcula o número de caminhos mais curtos de todos os vértices para quaisquer outros que passem por b. Este parâmetro só é calculado para redes que não possuem arestas paralelas. O valor de *betweenness centrality* $C(b)$ de um vértice b é dado por:

$$C(b) = \sum_{s \neq b \neq t} \frac{\sigma_{st}(b)}{\sigma_{st}} \tag{2.5}$$

em que s e t são vértices pertencentes à mesma de b, σ_{st} corresponde ao número total de caminhos mais curtos do vértice s para t e $\sigma_{st}(b)$ é o número de caminhos mais curtos que passam pelo vértice b. Este valor final é normalizado através da seguinte expressão:

$$C_{Normalizado}(b) = \frac{C(b)}{(N-1)(N-2)}, C_{Normalizado}(b) \in [0,1] \tag{2.6}$$

onde N é o número de vértices existentes no componente grande (sub-rede ou não) em que o vértice b se encontra. O valor de $C_{Normalizado}(b)$ na rede da figura 2.8 pode ser calculado por:

$$C_{Normalizado}(b) = ((\sigma_{ac}(b)/\sigma_{ac}) + (\sigma_{ad}(b)/\sigma_{ad}) + (\sigma_{ae}(b)/\sigma_{ae}) + (\sigma_{cd}(b)/\sigma_{cd}) + (\sigma_{ce}(b)/\sigma_{ce}) + (\sigma_{de}(b)/\sigma_{de}))/((5-1)-(5-2)/2) = ((1/1)+(1/1)+(2/2)+(1/2)+0+0)/6 = 3.5/6 \approx 0.583$$

A radialidade é outra medida que, tal como a anterior, avalia a centralidade de um vértice numa rede tendo em conta o diâmetro da componente em que um vértice se localiza. Este parâmetro indica a tendência média de proximidade ou isolamento de um vértice. A radialidade de um vértice n, é dado por:

$$R(n) = \sum_{n \neq b} L(n,b) - N + 1$$

$$R_{Normalizado}(n) = \frac{R(n)}{(N-1)}, R_{Normalizado}(n) \in [0,1]$$

em que N corresponde ao diâmetro do componente onde o vértice n se encontra, b corresponde a todos os vértices diferentes de n e $L(n,b)$ corresponde ao valor do caminho mais curto entre o vértices n e b.

2.3.2.1 *Cytoscape*

O *Cytoscape* é um *software open-source* que permite visualizar e analisar dados através de redes (grafos) [SMO+03]. A facilidade de integração de qualquer tipo de dados e posterior visualização na forma de redes é uma das melhores funcionalidades deste *software*. O *Cytoscape* permite a análise de variados parâmetros complexos das redes e posterior exportação da informação em formato *csv* e *JSON*. Além das funcionalidades principais, este pode ser extendido através do desenvolvimento de aplicações (*plugins*). Este *software* conta com uma *App Store* bastante enriquecida que permite adicionar um conjunto de funcionalidades muito variadas ao *core* do programa, permitindo numa análise mais profunda e detalhada das redes.

2.3.3 Data Mining

Data Mining é uma das fases de um processo mais complexo de análise de dados chamado *Knowledge Discovery in Databases* (*KDD*). Este processo consiste em explorar grandes quantidades de *datasets* com o objectivo de identificar padrões consistentes, como regras de associação ou sequências temporais, de forma a detectar possíveis relacionamentos entre variáveis e consequente geração de novos subconjuntos. As aplicações de *Data Mining* podem ser classificadas por vários conjuntos de problemas que possuem características semelhantes nos vários domínios de aplicação. Estas são suportadas por um conjuntos de algoritmos que são utilizados para extrair as relações relevantes dos *datasets*. Os algoritmos diferem consoante o tipo de problema que visam

solucionar. Neste trabalho em específico, o interesse está centrado nas técnicas de classificação e *clustering* uma vez que são estas que se adequam ao problema que é pretendido solucionar.

2.3.3.1 Classificação

A técnica de classificação consiste em prever determinado resultado baseado nos dados que são dados como *input*. Para que seja possível a previsão de resultados, o algoritmo processa um *training set* contendo um conjunto de atributos em que os resultados são conhecidos de antemão. O algoritmo tenta identificar relações entre diferentes atributos que permitam prever o resultado final. Na fase seguinte é fornecido um *dataset* desconhecido denominado de *prediction set* que contém o mesmo conjunto de atributos excepto o resultado, ainda desconhecido. Finalmente, o algoritmo analisa os dados e produz uma previsão. Esta técnica recorre a regras *IF-THEN* para produzir resultados. O antecedente (*IF*) consiste num conjunto de condições e o seu consequente *THEN* prevê um determinado valor no atributo que satisfaça as condições presentes no antecedente.

Excepto para determinado problemas específicos, a técnica de classificação por definição recorre sempre a algoritmos de aproximação [VV07].

2.3.3.1.1 Construção de Árvores de Decisão - ID3 e C4.5

O ID3 induz árvores de decisão a partir de um *dataset*. A árvore resultante é usada para classificar futuras amostras. ID3 separa os dados em vários subconjuntos de forma a que estes contenham exemplos de uma única classe. Os nós folha da árvore de decisão resultante contêm o nome da classe enquanto os outros são nós de decisão. O ID3 é um algoritmo não incremental, ou seja, as classes são derivadas de um conjunto *training sets* que não é variável. As classes criadas por este algoritmo são indutivas e funcionam para todas as futuras classificações. Os *training sets* fornecidos ao ID3 têm que respeitar certos requisitos:

- Descrição atributo-valor - os mesmo atributos devem descrever cada exemplo e terem um número fixo de valores;
- Classes pré-definidas - as classes dos exemplos fornecidos já têm que estar definidas.
- Classes discretas - Não podem existir classes ambíguas. Estas devem ser bem definidas e diferentes entre elas de maneira a diminuir possíveis erros nas induções dos resultados.
- Número suficiente de exemplos - uma vez que o ID3 gera as árvores de decisão a partir de indução é necessário existir um número de exemplos suficientes que permitam destinguir todos os possíveis padrões.

Para determinar quais os atributos mais importantes, o ID3 utiliza a entropia como medida de impureza. A entropia de um nó N é dada pela equação:

$$Entropia(N) = -\sum_{C=1}^{k} p(C|N)\log_2 p(C|N) \tag{2.7}$$

em que $p(C|N)$ é a fracção de elementos que correspondem à classe C, no nó N e k é o número de classes. Para determinar o quanto um atributo é bom, é necessário recorrer ao conceito de ganho que traduz a diferença entre a impureza do nó pai e a soma da impureza das partições resultantes, multiplicadas pelas suas probabilidades. O ganho associado a uma divisão D, é dado pela equação:

$$Ganho(D) = Entropia(N_{pai}) - \sum_{i=1}^{n} \frac{|N_i|}{|N_{pai}|} Entropia(N_i) \tag{2.8}$$

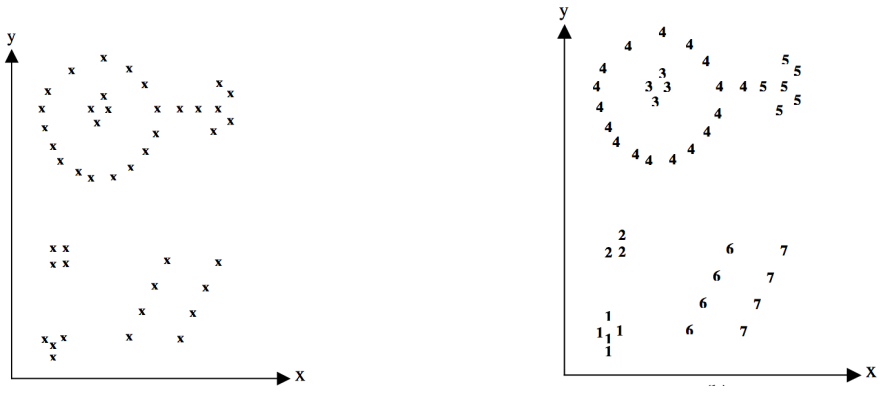
em que n é o número de de nós filhos após divisão, $|N_{pai}|$ é o tamanho do *dataset* associado ao nó pai e $|N_i|$ é o número de elementos associados ao nó filho N_j.

Uma das limitações do ID3 é que é sensível a atributos com um número grande de valores. A entropia para esses atributos é muito pequena e consequentemente não ajuda no processo de selecção dos atributos mais importantes. Para evitar esta limitação, foi criada uma extensão do ID3, o algoritmo C4.5 .

O algoritmo C4.5 resolve o problema da entropia usando outra propriedade estatística denominada ganho de informação. O ganho de informação mede o quão correcto é que um conjunto de amostras é dividido noutro conjunto tendendo a uma determinada classe.

2.3.3.2 Clustering

Clustering (agrupamento) é uma técnica de modelação de dados que se baseia na construção de *clusters*. *Clusters* são *datasets* que gozam da seguinte propriedade: os elementos pertencentes a um mesmo conjunto apresentam maior semelhança entre si que os elementos pertencentes a qualquer outro conjunto, com relação a um certo critério de similaridade [LdR05]. Na Figura 2.9 é apresentado um exemplo de um *clustering*. Na sub-figura 2.9a encontra-se um *dataset* de entrada, onde cada elemento é representado pelo símbolo 'x'. Na sub-figura 2.9b pode-se ver o resultado após a realização do *clustering* sobre o *dataset* inicial. Como se pode verificar, a cada elemento 'x' foi atribuído um identificador do conjunto final pertencente.



(a) *Dataset* de entrada (b) Resultado do *clustering* com 7 *clusters*.

Figura 2.9: Técnica de *clustering*. Os dados que pertencem a um mesmo *cluster* apresentam o mesmo rótulo [JMF99].

As técnicas de *clustering* podem ser divididas em dois principais conjuntos de algoritmos existentes:

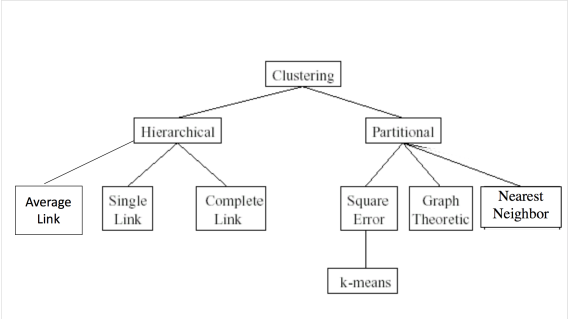


Figura 2.10: Taxonomia para as diferentes técnicas de *clustering* [JMF99].

2.3.3.2.1 Algoritmos de *clustering* hierárquicos

Este tipo de algoritmos, tal como o nome indica, constroem uma hierarquia de *clusters*, ou seja, uma árvore de *clusters*. Nesta árvore, cada *cluster* pode conter outros *clusters* filhos. Se um *cluster* não tiver nenhum filho é denominado uma folha da árvore. Os algoritmos *single-link* [SS+73] e *complete-link* [Kin67] são os algoritmos mais populares, mas existe ainda o *average link*.

O algoritmo *single-link* inicia-se com todos os padrões como *clusters* individuais e à medida que este é executado vai recursivamente agrupando-os. A principal característica deste algortimo é o cálculo da distância entre 2 *clusters* como sendo a menor das distâncias existentes entre todos os pares de padrões pertencentes aos 2 *clusters*. Este algoritmo resume-se nas seguintes etapas:

- Definição de cada padrão como sendo um *cluster*;
- Construção de uma lista com as distâncias entre padrões para todos os pares de padrões;
- Ordenação da lista por ordem crescente de distâncias;
- Para cada valor c_k da lista de distâncias, constrói uma grafo nos quais os pares de padrões cujos valores são mais próximos de c_k são conectados por uma aresta;
- Se todos os padrões pertencem a um grafo conexo o processo termina. Caso contrário repetem-se todos os passos do início.

(a) Cada padrão definido como um *cluster*

(b) Resultado do *clustering* com 7 *clusters*.

Figura 2.11: Árvore obtida após aplicação do algoritmo de *clustering* hierárquico *single-link*.

Exemplo de um algoritmo *single-link* muito utilizado na criação de árvores filogenéticas é o *Neighbor Joining* [SN87]. Este método de aglomeração requer como *input* uma matriz de distâncias. As árvores geradas por este algoritmo não têm raíz. O algoritmo inicia-se com uma árvore não resolvida, com uma topologia correspondente à rede estrela e itera sobre os passos seguintes até que todas variáveis da matriz de distância estejam presentes e sejam conhecidos todos os comprimentos das ramificações.

- É calculada uma matriz Q, tal que:

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

  em que $d(i,j)$ corresponde à distância entre a varíavvel i e j e n o número de variáveis presentes na matriz distância fornecida.

- Encontrar o menor valor de Q(i,j) tal que $i \neq j$. As variavéis i e j formam um novo nó que está conectado ao nó central da árvore.

- É calculada uma nova distância do par gerado (i,j) ao nó adicionado. Considerando u o novo nó criado na árvore e f e g o par de folhas correspondentes, a nova distância pode ser calculada através da seguinte expressão:

$$\delta(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)]$$
$$\delta(g,u) = d(f,g) - \delta(f,u)$$

- Calcular a distância de todas as variáveis, que não pertencam a par anterior, ao novo nó u através da expressão:

$$\delta(u,k) = \frac{1}{2}[d(f,k) + d(g,k) - d(f,g)]$$

em que u corresponde ao novo nó adicionado, k é o nó que pretendemos calcular a distância e f e g são os membros do novo par que se juntou ao nó.

O algoritmo *complete-link*, à semelhança do anterior, também se inicia com os *clusters* individuais, os quais vão sendo agrupados recursivamente à medida que o algoritmo é executado. A diferença em relação ao anterior reside no processo de concatenação, a distância calculada entre 2 *clusters* agora é a maior das distâncias existentes em dois *clusters*. As diferentes etapas deste algoritmo são iguais às do anterior excepto na ordenação da lista de distâncias que agora contém a maior distância entre 2 *clusters*.

O método *average linking clustering* ou também *Unweighted Pair-Group method with arithmetic mean* (UPGMA) é o método de aglomeração mais simples para a construção de árvores. Tal como os métodos de *clustering* hierárquico abordados anteriormente, este também requer uma matriz de distâncias. Ao contrário do algoritmo *Neighbor Joining*, este constrói uma árvore com raíz (dendograma). Este algoritmo resume-se nas seguintes etapas:

- É seleccionado da matriz o par de variavéis com menor distância entre eles e são agrupados num *cluster* que corresponde a um nó no dendograma
- É calculada a distância do novo nó u que contém o par de variáveis i e j através da seguinte expressão:

$$\delta(i,j) = \delta(j,i) = \frac{d(i,j)}{2}$$

- A matriz de distância é actualizada e reduzida numa linha e numa coluna resultante do *clustering* das variáveis i e j. As distâncias da matriz são atualizadas calculando a média das distâncias do *cluster* anterior (i,j) e cada uma das outras variáveis da matriz.
- Se i e j forem as 2 últimas variáveis na matriz o algoritmo termina, senão volta à primeira iteração criando outro nó.

2.3.3.2.2 Algoritmos de *clustering* particionais

Ao contrário dos algoritmos hierárquicos em que é construída uma árvore de *clustering*, os algoritmos particionais constroem uma partição simples dos padrões. Este tipo de algoritmos apresenta vantagem nas aplicações que envolvem uma grande quantidade de conjuntos. Este tipo de algoritmos produzem frequentemente os *clusters* através da optimização de uma função.

Os algoritmos de *clustering Graph Theoretic* baseiam-se na teoria de grafos para o seu funcionamento. O melhor algoritmo deste tipo é baseado na construção de uma *Minimum Spanning Tree* (*MST*) que liga um conjunto de padrões em que as arestas representam a maior distância euclidiana entre os mesmo [Zah71].

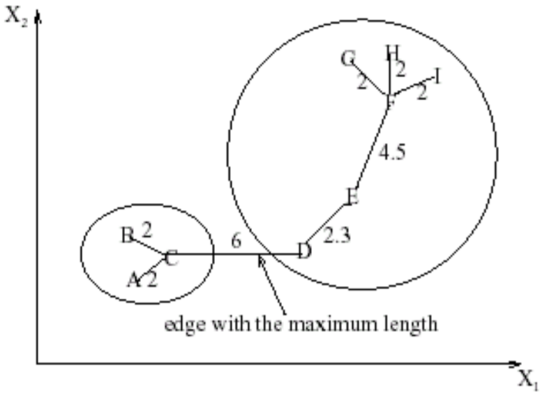


Figura 2.12: Ilustração da identificação de arestas com a maior distância euclidiana no algoritmo de *clustering Graph Theoretic*.

Após a criação da *MST* são removidas as arestas com maior valor para serem gerados os *clusters*.

Os algoritmo de *clustering k-means* utliza uma função para dividir um conjunto de padrões. A função mais frequentemente optimizada neste algoritmo é a função *square-error*, dada pela equação:

$$e^2(X,L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} ||x_i^{(j)} - c_j||^2 \tag{2.9}$$

O valor $x_i^{(j)}$ é o i-ésimo padrão pertencente ao j-ésimo *cluster* e c_j é o centróide do j-ésimo *cluster*. X representa o conjunto de padrões e K o número de *clusters* pretendidos. Resulta um *clustering* L.

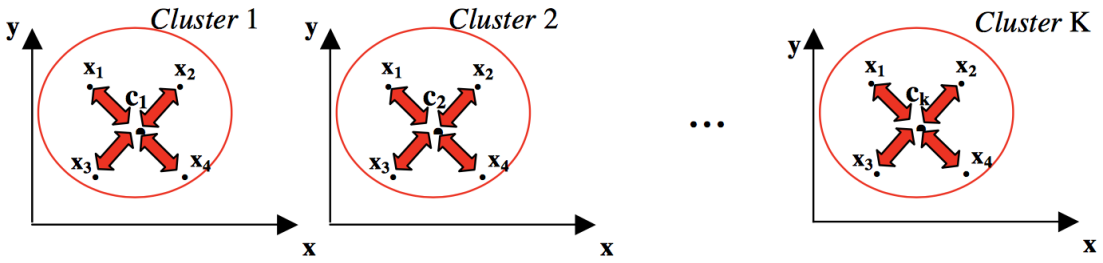


Figura 2.13: Distância entre os pontos dos vários *clusters* e o respectivo centróide, no qual se baseia o algoritmo *k-means*.

É a partir da fórmula da equação 2.9 que o algoritmo *k-means* se baseia. As etapas deste algoritmo são:

- Escolhe k centros de *clusters* para coincidir com k padrões escolhidos de forma aleatória;
- Atribui cada padrão ao centro do *cluster* mais próximo;

- Recomputa os centros dos *clusters* usando os padrões actualmente existentes nesses *clusters*;
- Volta à 2ª etapa até o critério de convergência ser alcançado.

Os critérios de paragem mais comuns neste algoritmo são a ocorrência de um decréscimo mínimo da função e ou a não re-atribuição de um padrão a um novo *cluster*.

Este algoritmo é muito popular porque tem uma implementação fácil e a sua complexidade é $O(n)$ em que n corresponde ao número de padrões, contudo é sensível à selecção das divisões iniciais.

2.3.3.3 Matriz de distâncias

Para a criação de árvores através de métodos de aglomeração de *clustering* hierárquico é necessário fornecer *a priori* uma matriz de distância. Esta matriz contém as distâncias, aos pares, de um conjunto de pontos ou variáveis. Estas matrizes podem representar, por exemplo, a distância entre os vértices de um grafo ou então as não similaridades entre tecidos.

Existem vários índices para medir a similaridade ou distância entre objectos. Estes índices são úteis para expressar a diferença entre pares de amostras de uma população.

O índice de *Jaccard* é utilizado para calcular distâncias em matrizes de presência-ausência, ou seja, matrizes binárias.

Sejam $x = (x_1, x_2, ..., x_n)$ e $y = (y_1, y_2, ..., y_n)$ tal $x_i, x_j \geq 0$, o índice de similaridade de *Jaccard* é calculado por:

$$J(x,y) = \frac{\sum_i min(x_i, y_i)}{\sum_i max(x_i, y_i)} \tag{2.10}$$

e a distância de *Jaccard*:

$$d_J(x,y) = 1 - K(x,y), d_J \in [0,1] \tag{2.11}$$

2.3.3.4 R - Linguagem de Programação

O *R* é um sistema para computação de estatísticas e gráficos [Tea00]. Destaca-se por fornecer uma linguagem de programação própria, um ecossistema rico em bibliotecas para geração de gráficos e análises estatísticas, uma grande comunidade e ainda a possibilidade de interface com outras linguagens de programação. Esta linguagem já contém implementada, por exemplo, algoritmos de *clustering* hierárquico *UPGMA* e *NJ* utilizados na análise, pacotes para a geração de árvores, dendogramas e gráficos, bem como métodos disponíveis para calcular matrizes de distâncias utilizando diferentes índices, mais especificamente o índice de *Jaccard* abordado anteriormente. Esta ferramenta poderosa é um auxílio para os demais cálculos estatísticos, evitando a necessidade de voltar a implementar algoritmos já conhecidos e que se encontram optimizados.

Capítulo 3

Implementação

A implementação deste trabalho compreende diferentes etapas. Neste capítulo são mencionadas e abordados os detalhes de implementação das mesmas. Como etapas deste trabalho destacam-se a extracção de dados de sequenciação de *RNA* (*RNA-Seq*), a sua filtragem, o cálculo de correlações entre os pares de genes, o posterior enriquecimento e visualização das redes de tecidos, a sua análise topológica e por fim a análise biológica dessas mesmas redes de forma interactiva através de uma plataforma *web* (Figura (3.1).

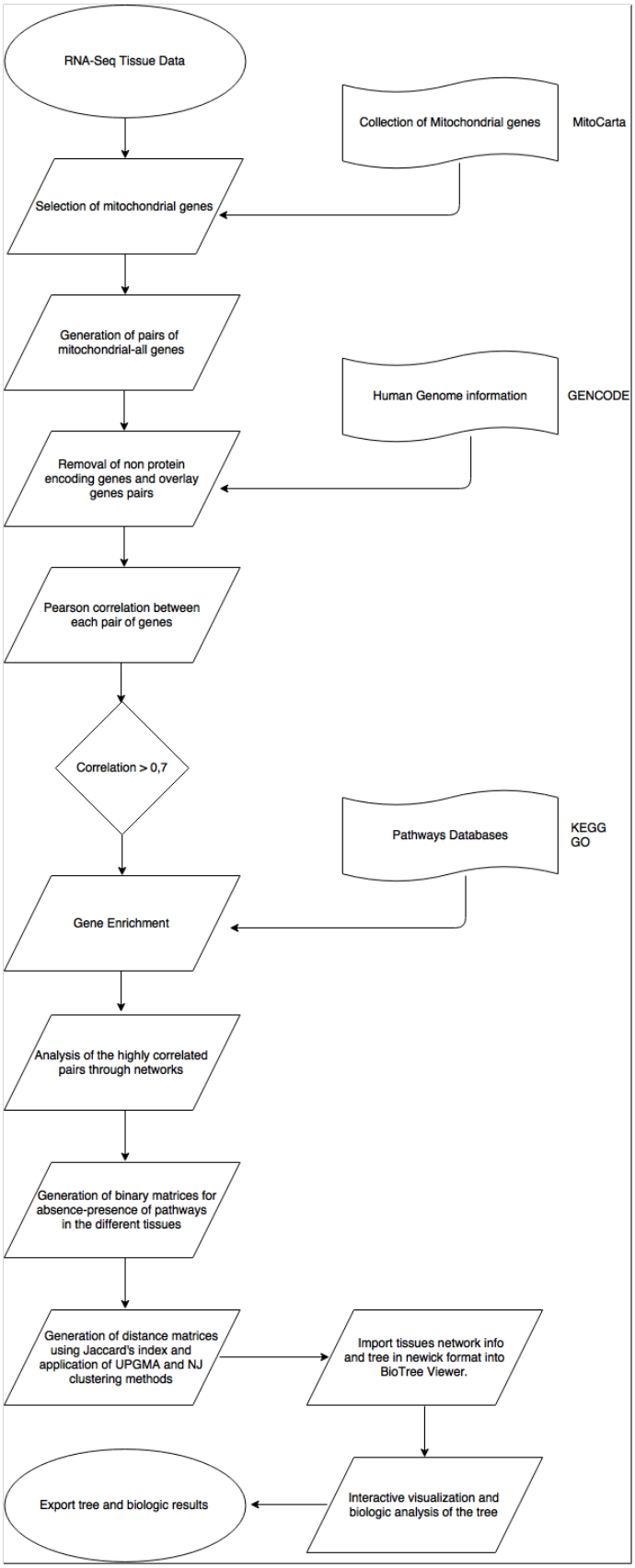


Figura 3.1: Fases principais da implementação deste trabalho.

3.1 Extracção de dados

Numa primeira etapa, foi implementado um *parser* responsável por ler a informação gerada pelos *RNA-Seq* e colocá-la numa estrutura adequada ao problema. Separou-se a informação de todos os tecidos que se encontrava contida num ficheiro único com a extensão *.gct* em vários ficheiros, cada um representando um tecido diferente. Os tecidos com um número de amostras inferior a 10 são, *a priori*, excluídos da extração e não fazem parte do conjunto de resultados gerado pelo *parser*.

```
1 ##NUMERO_GENES NUMERO_SAMPLES
2 56318   286
3 ##GENE SAMPLE1 SAMPLE2 ... SAMPLE286
4 DDX11L1 0.0397965013980865 0.0611212588846684 ...  0
5 RP5-857K21.1 0.0614116229116917 0.0554628632962704  ... 0.100065223872662
6 (...)
7 MTND1P23 10.0158805847168 22.8889236450195  ...  12.4281549453735
```

Listing 3.1: Exemplo ficheiro output após extração dos dados.

Este *parser* encontra-se implementado em *c++* gozando da rapidez e performance que esta linguagem permite em operações *I/O* (*Input/Output*) de ficheiros.

3.2 Filtragem de dados e correlacão entre pares de genes

Como o objectivo era o estudo de pares de genes em que ambos são genes mitocondriais ou em que pelo menos um dos genes é mitocondrial e o outro codifica uma proteína, foi fornecido *a priori* a base de dados *Mitocarta* que contém uma colecção de 1158 genes que codificam proteínas mitocondriais e um ficheiro (no formato *.gtf*) com informações detalhadas do genoma humano baseado em evidências, nomeadamente a posição em que os genes se encontram, qual o cromossoma a que pertence, se codificam proteínas etc.

Através da informação da *Mitocarta* foi possível filtrar as interações de interesse em que pelo menos um dos genes codifica uma proteína mitocondrial.

O facto de dois genes poderem estar sobrepostos numa determinada posição do cromossoma pode levar a uma correlação de expressão falsamente positiva entre os mesmos. Para evitar a sobreposição de genes, foram determinadas quais as combinações de genes sobreponíveis com base no ficheiro em formato *.gtf* que contém informações do genoma humano nomeadamente a posição em que os genes se encontram, qual o cromossoma a que pertence, se codificam proteínas etc. Foi gerado um ficheiro para cada cromossoma de forma a permitir a paralelização.

Para aglomerar todas as listas resultantes de cada cromossoma num único ficheiro, foi desenvolvido um *bash script*, resultando apenas um ficheiro com toda a informação.

```
1 ##GENE1 GENE2
```

```
2 CYC1  PCBD2
3 CYC1  SLC35A4
4 CYC1  WWC1
5 CYC1  FARS2
6 (...)
7 PDHA1 PREP
```

Listing 3.2: Excerto de *output* gerado após identificação de pares de genes sobrepostos.

Uma vez preparada e filtrada a informação, foram calculadas as correlações de *Pearson* entre os pares de genes através da leitura dos ficheiros resultantes da fase anterior. Além da filtragem ocorrida na fase anterior, foi utilizada uma técnica de remoção de *outliers* baseada no cálculo do desvio padrão. Se em determinada correlação os valores não respeitarem o intervalo $[u_x - 4SD, u_x + 4SD]$ ou $[u_y - 4SD, u_y + 4SD]$, em que SD corresponde ao desvio padrão (*standard deviation*), esse valores são excluídos da amostra. Desta etapa resultam 3 ficheiros por tecido correspondentes às correlações fortemente negativas ($C_{X,Y} < -0,7$), correlações fortemente positivas ($C_{X,Y} > 0,7$) com genes sobrepostos e correlações fortemente positivas sem sobreposição de genes.

```
1 ##GENE_MITOCONDRIAL GENE_CODIFICA_PROTEINA  CORRELACAO
2 CYC1 CDK16 0.705952
3 CYC1 HSD17B10 0.803689
4 CYC1 NAA10 0.711389
5 SDHB AURKAIP1 0.722465
6 SDHB PARK7 0.705339
7 SDHB MRPS15 0.703715
8 (...)
9 SDHB UQCRH 0.816855
```

Listing 3.3: Excerto de *output* gerado após o cálculo de correlações e filtragem dos dados de interesse.

3.3 Enriquecimento dos dados

Para que a análise dos dados não seja apenas de natureza topológica foi feito um enriquecimento dos dados com informação biológica, nomeadamente a identificação das *pathways* a que os genes pertencem. A informação provém das bases de dados *Kyoto Encyclopedia of Genes and Genomes (KEGG)* e *Gene Ontology (GO)*. Como já foi mencionado anteriormente, a informação pode incidir sobre processos biológicos, componentes celulares ou funções moleculares.

```
1 KEGG_GLYCOLYSIS_GLUCONEOGENESIS http://www.broadinstitute.org/gsea/msigdb/cards/
2 KEGG_GLYCOLYSIS_GLUCONEOGENESIS ACSS2 GCK PGK2  PGK1  PDHB  PDHA1 ...   PGAM2
3 (...)
```

```
4 KEGG_VIRAL_MYOCARDITIS  http://www.broadinstitute.org/gsea/msigdb/cards/
5 KEGG_VIRAL_MYOCARDITIS  CASP9 LOC100418883  CASP8 HLA-DOA HLA-DOB CD80  CD86  CD28
      EIF4G3  ITGAL ICAM1 CXADR MYH13 ... HLA-G
```

Listing 3.4: Exemplo de ficheiro *input* com informação biológica da base de dados *KEGG*.

O programa desenvolvido verifica se os genes estão contidos em alguma *pathway* e acrescenta essa informação ao ficheiro gerado na fase anterior (3.3). Como resultado deste enriquecimento, são gerados 4 novos diretórios cada um correspondente à utilização de uma base de dados diferente. A separação em 3 diretórios com a base de dados *Gene Ontology (GO)* foi propositada, uma vez que faz sentido para o tipo de análise desejada ter as *sub-ontologias* separadas, pois a informação é redundante. De salientar que é possível enriquecer os dados com informação de outras bases de dados desde que estas respeitem o formato *.gmt*.

```
1 ##GENE_MITO GENE_PROT Correlation PathwayList PathwayList
2 CYC1  HSD17B10  0.803689    KEGG_ALZHEIMERS_DISEASE,
    KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION
3 IARS2 NDUFS1  0.724112  KEGG_AMINOACYL_TRNA_BIOSYNTHESIS,
    KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_BIOSYNTHESIS  KEGG_ALZHEIMERS_DISEASE,
    KEGG_HUNTINGTONS_DISEASE,KEGG_OXIDATIVE_PHOSPHORYLATION,KEGG_PARKINSONS_DISEASE
4 (...)
5 SDHB  PARK7 0.705339    KEGG_PARKINSONS_DISEASE
```

Listing 3.5: Resultado do enriquecimento biológico dos tecidos.

3.4 Geração e análise de redes de genes correlacionados

Após a estruturação dos dados de forma adequada ao problema e o cálculo de correlações entre os genes constituintes dos tecidos, estes foram integrados em forma de redes de interacção. Foi utilizado o *Cytoscape*[SMO+03] para a integração dos dados, visualização e posterior análise topológica destas redes. Seja G um grafo simples que representa as interacções presentes num tecido:

$$G = (V, E) \tag{3.1}$$

V representa o conjunto de todos os genes existentes na rede do tecido e E o conjunto de interacções entre genes tendo como peso associado a correlação entre os mesmos (Figura 3.2).

Através do *Cytoscape* foram feitas análises topológicas das redes, destacando-se o conhecimento do grau de todos os vértices dos grafo (nós), a centralidade (*betweenness centrality*) dos mesmos e ainda a informação do número total de genes e interacções presentes nas redes de tecidos. O grau de um vértice representa o número de vértices que ele está conectado diretamente enquanto o valor *betweenness centrality* representa o número de caminhos mais curtos de todos os nós para quaisquer outros nós que passem pelo vértice. Estas medidas são indicadores importantes em redes biológicas uma vez que indicam que o gene (vértice) pode ter uma função importante na

rede. Após a análise topológica das redes efectuadas pelo *Cytoscape*, os *dados* foram exportados para serem utilizados como *input* em posteriores análises de *pathways*. Num primeiro passo são produzidas listagens de todas as interacções presentes nas redes (Exemplo 3.6).

```
1 ##TECIDO GENE_MIT PATHWAY GENES_CORRELACIONADOS
2 Adipose_Visceral_Omentum  ATP5B KEGG_ALZHEIMERS_DISEASE ATP5A1,SDHB
3 Adrenal_Gland OGDHL KEGG_TRYPTOPHAN_METABOLISM  DDC
4 Brain_Amygdala  UQCRC1  KEGG_HUNTINGTONS_DISEASE  CYC1,NDUFV1
5 (...)
6 Whole_Blood ATP5D KEGG_OXIDATIVE_PHOSPHORYLATION  NDUFS7,NDUFS8
```

Listing 3.6: Listagem de interacções presentes nas redes de todos os tecidos para a base de dados *KEGG*.

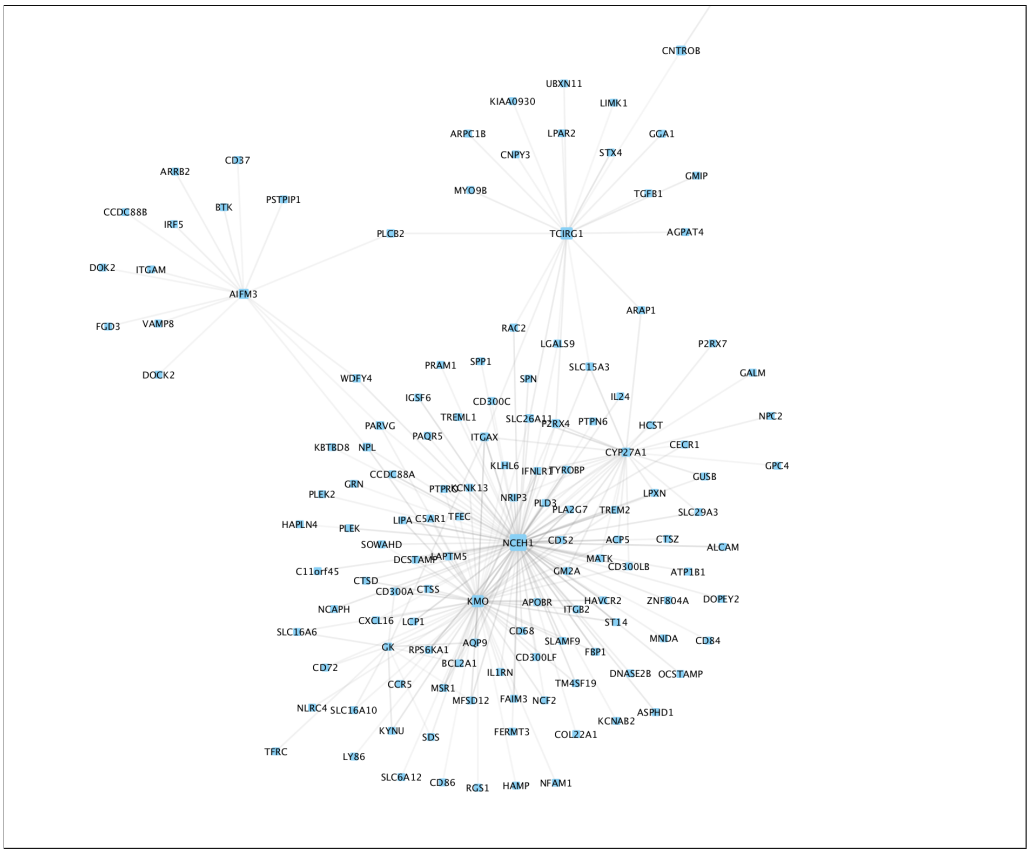


Figura 3.2: Detalhe da rede do tecido *Adipose - Subcutaneous* com 2310 nós e 12282 arestas. Quanto menor a transparência das arestas mais forte a correlação entre os genes. Quanto maior o tamanho do vértice, maior o seu valor de *Betweenness Centrality*.

Em seguida, são geradas matrizes binárias com informação de presença e ausência de *pathways* dos tecidos. Seja A uma matriz binária para cada uma das quatro bases de dados biológicas, tal

que:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \ddots & a_{3n} \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, (a_{ij}) \in \{0,1\}^{m*n} \tag{3.2}$$

m corresponde ao número de tecidos existentes para análise e n ao número de *pathways* existentes na base de dados. O valor de a_{ij} é 0 quando a rede do tecido i não contém a pathway j em nenhuma das suas interações ou 1 quando a rede do tecido i contém pelo menos uma interacção com a *pathway* j.

Calculou-se a matriz de distância de A através do método *vegdist* do pacote *vegan* da linguagem *R*, juntamente com o índice de *Jaccard* (Listagem 3.7). Já que a matriz de entrada é uma matriz binária e contém informação de presença e ausência, considerou-se a utilização do índice de *Jaccard* para o cálculo das distâncias o mais adequado.

```
1 distance_matrix <- vegdist(A, method = "jaccard")
2
3 ##UPGMA
4 hc <- hclust(distance_matrix, method="average")
5
6 ##Neighbor Joining
7 neighbor_joining <- nj(distance_matrix)
```

Listing 3.7: Métodos e parâmetros utilizados na realização do *clustering*.

Foi feita uma análise qualitativa dos tecidos recorrendo à técnica de *data mining* de *clustering* hierárquico a partir da matriz de distâncias da fase anterior, através de 2 métodos de aglomeração. Recorrendo ao método *Unweighted Pair Group Method with Arithmetic Mean* (*UPGMA*) utilizando a função *hclust* e ao método *Neighbor Joining* (*NJ*) utilizando a função *nj* do pacote *ape* (Listagem 3.7). Posteriormente, geraram-se árvores para os dois métodos como forma de visualização dos resultados representando os tecidos mais similares entre si consoante as *pathways* que têm em comum (Listagem 3.8 e Figura 3.3). Estas árvores são exportáveis em formato imagem e *newick* (formato que usa correspondencia entre árvores e parêntesis) e representam exactamente a mesma informação. Na árvore (Figura 3.8) é possível observar que os tecidos estão coloridos de acordo ao grupo/sistema a que pertencem, por exemplo, os tecidos pertencentes ao sistema nervoso encontram-se coloridos a violeta.

```
(Whole_Blood:0.3122783844,(((((Skin_Not_Sun_Exposed_Suprapubic:0,Thyroid_:0)
    :0.0588429397,Muscle_Skeletal:0.08401420316):0.02715287273,Esophagus_Mucosa
    :0.09784712727):0.06007071719,Brain_Cerebellar_Hemisphere:0.2250483304)
    :0.01358281476,Adipose_Visceral_Omentum:0.1849020337):0.03387484965,(((((((
    Cells_EBV_Transformed_Lymphocytes:4.554761127e-17,Esophagus_Muscularis
    :-4.554761127e-17):0.05721279794,Brain_Cerebellum:0.1427872021):0.07666146583,
    Brain_Cortex:0.2042909151):0.04273579268,Colon_Sigmoid:0.2231372232)
    :0.01312059902,Pancreas_:0.3155229796):0.03896596302,(((((
    Cells_Transformed_Fibroblasts:0.2975492236,Minor_Salivary_Gland:0.2738793478)
    :0.007444286867,(((Colon_Transverse:0.1927909699,Small_Intestine_Terminal_Ileum
    :0.1960979189):0.03078848839,Kidney_Cortex:0.2739918734):0.06710338717,Stomach_
    :0.3290634305):0.03237672963):0.02974809493,(((((Brain_Amygdala:0.1709528692,
    Brain_Hypothalamus:0.1711523939):0.006996241515,((
    Brain_Anterior_Cingulate_Cortex_BA24:0.1577640757,Brain_Putamen_Basal_Ganglia
    :0.1570507391):0.0105046254,(Brain_Caudate_Basal_Ganglia:0.08793372619,
    Brain_Nucleus_Accumbens_Basal_Ganglia:0.08349484524):0.06110236365)
    :0.041760572):0.03760045987,(Brain_Frontal_Cortex_BA9:0.1559370874,
    Brain_Hippocampus:0.1419352531):0.0827731844):0.06591116231,
    Brain_Substantia_Nigra:0.257543565):0.02835499508,Brain_Spinal_Cord_cervical_c1
    :0.238000371):0.08452862554):0.02758828851,(((((Adrenal_Gland:0.2991347122,
    Ovary_:0.3008652878):0.07135326715,(((((((Adipose_Subcutaneous:7.558965274e-17,
    Artery_Aorta:-7.558965274e-17):7.723290606e-17,Artery_Tibial:-7.723290606e-17)
    :7.894919286e-17,Skin_Sun_Exposed_Lower_Leg:-7.894919286e-17):1.614869854e-16,
    Breast_Mammary_Tissue:-1.614869854e-16):1.652424967e-16,Nerve_Tibial
    :-1.652424967e-16):1.691768418e-16,Pituitary_:-1.691768418e-16):0.2504856698,
    Prostate_:0.2495143302):0.1203133995):0.03938844092,Uterus_:0.3522782257)
    :0.03146627181,Testis_:0.4565793631):0.0300065198,Artery_Coronary:0.4451953616)
    :0.07312386146):0.02239159627,(((Lung_:0.0837956809,Vagina_:0.1162043191)
    :0.06685143421,Spleen_:0.3831485658):0.04138303333,Liver_:0.1661344492)
    :0.1058172833):0.02526210345):0.02537542596,((Heart_Atrial_Appendage
    :0.143981632,Heart_Left_Ventricle:0.2677830739):0.02452577644,
    Esophagus_Gastroesophageal_Junction:0.2161159348):0.04338147834):0.00399547284)
    ;
```

Listing 3.8: Árvore gerada pelo método *Neighbor Joining* no formato *newick*.

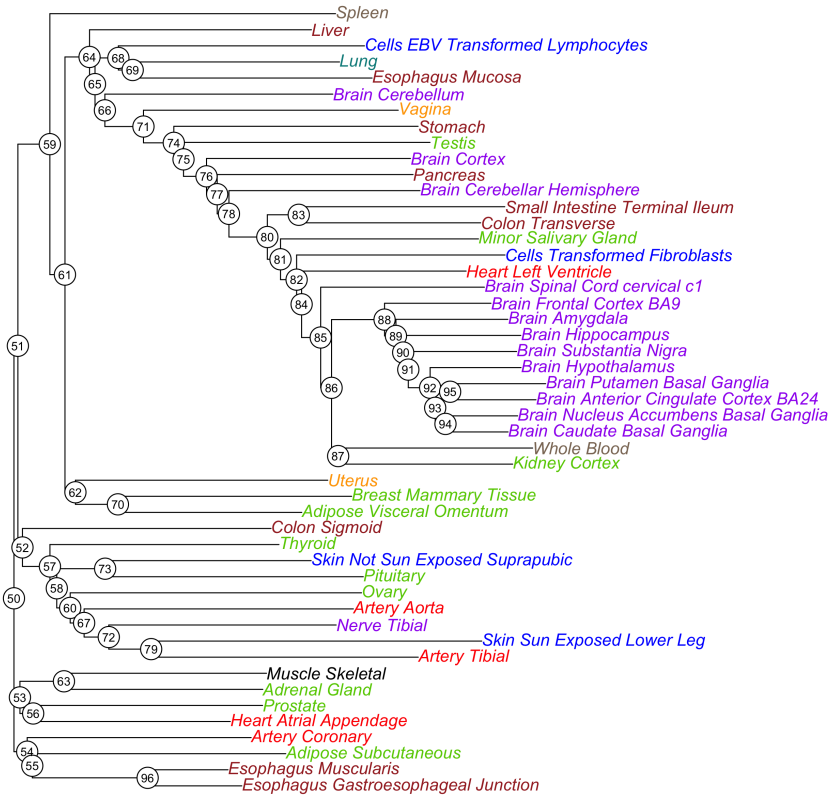


Figura 3.3: A mesma árvore representada em 3.8
mas agora em formato imagem.

Uma vez que se verificou uma dependência de diversos *softwares* para a visualização, tanto das árvores como dos resultados das análises, foi criada uma plataforma *Web* capaz de manipular e observar de forma interactiva os dados biológicos relativos aos tecidos. A plataforma denominada BioTree Viewer é baseada no *software* phylotree.js (`https://github.com/veg/phylotree.js`) para a visualização das árvores, mas as funcionalidades biológicas bem como a possível exportação e importação de dados foram criadas no contexto deste trabalho, tornando-a diferenciadora entre as soluções já existentes.

Foi desenvolvido um módulo de importação de dados biológicos capaz de interpretar os dados biológicos das redes de interacção dos diversos tecidos através de uma sintaxe adequada (Listagem 3.6). Esta pode ser gerada através da *pipeline* descrita anteriormente a partir da exportação de redes do *Cytoscape*. Os dados foram mapeados e organizados pelos nomes das redes de interacções dos vários tecidos e das *pathways* contidas na base de dados em análise, em que estas possuem um conjunto de genes representando as interacções da rede. Assim, de uma forma fácil, é possível

verificar se uma determinada rede de interacções de um tecido tem a presença de determinada *pathway* nas suas interacções e ainda identificar quais os genes que fazem parte dessas interacções. A possibilidade de exportação das árvores foi feita através da conversão de objectos *svg* para *png* e a exportação de dados biológicos é inerente à tecnologia utilizada para mostrar os dados em tabelas, *Datatables*.

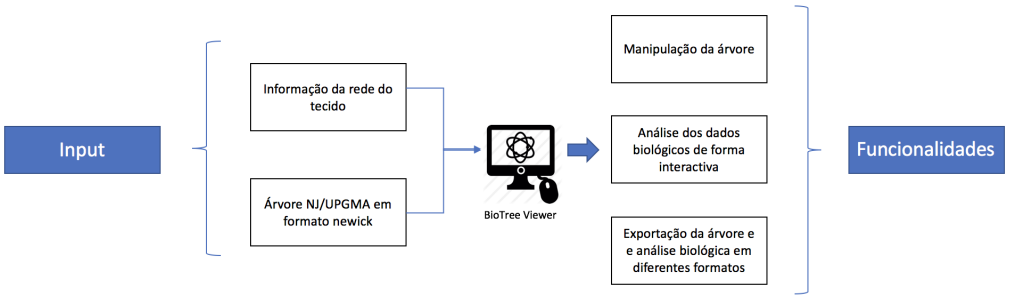


Figura 3.4: Diagrama de processos da plataforma *web* depositada num servidor.

Toda a informação é processada do lado do cliente e nunca é para nenhum servidor. O único requisito necessário para correr esta plataforma é um *web browser* compatível com *JavaScript*.

De forma a validar os resultados desta plataforma foi ainda criado um módulo que permite, passando o número identificador de nó existente na árvore, observar quais as *pathways* e genes que os tecidos descendentes desse nó partilham em comum. Este módulo é constituído por dois programas. O primeiro, desenvolvido em *R*, recebe como parâmetros de entrada uma árvore e o número de um nó e devolve um ficheiro com os nomes dos tecidos a que os descendentes do nó seleccionado correspondem. O segundo programa foi implementado em *Prolog* e recebe como *input* um ficheiro com a informação das redes e o nome dos tecidos a serem analisados. Este é responsável pela detecção de dados funcionais que os tecidos partilham gerando ficheiros que contêm as *pathways* , os genes mitocondriais e os interactomas comuns entre os tecidos passados como parâmetro de entrada.

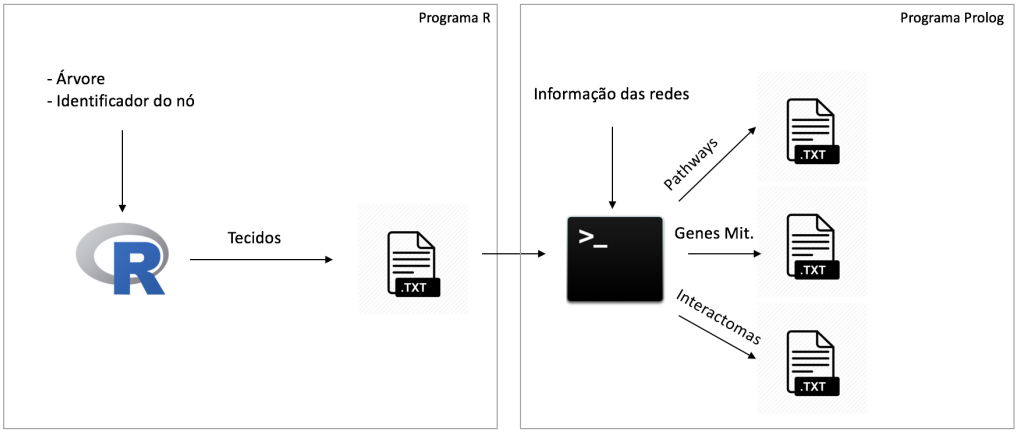


Figura 3.5: Esquematização do módulo, representado por um programa desenvolvido em *R* e outro em *Prolog*.

Capítulo 4

Resultados

4.1 Resultados biológicos

Foram fornecidos dados de *RNA-Seq* de 53 tecidos diferentes, no entanto, apenas foram considerados para o estudo os que tinham um número de amostras superiores a 10. Tal como é possível verificar no Anexo A os tecidos *Cervix - Ectocervix*, *Certix - Endocervix*, *Fallopian Tube* não respeitaram o número mínimo de amostras e, por isso, foram excluídos da análise. Decidimos manter o nome dos tecidos em inglês, uma vez que a nomenclatura usada na base de dados *GTEx* e que a codificação foi feita inglês para poder ser o mais genérica possível.

Uma vez que, mesmo após a filtragem de correlações superiores a $0,7$ a quantidade de dados era demasiado grande considerou-se útil filtrar ainda mais os pares de genes altamente correlacionados. Deste modo, foram analisadas as correlações superiores a 0,9 e as superiores a 0,8.

O tecido da bexiga (*bladder*) foi considerado como um *outlier* devido ao seu grande número de genes e interações, 11165 e 281978 respectivamente para a rede com correlação superior a 0,9 e 16855 e 1125749 respectivamente na rede com correlações superiores a 0,8. O facto de este tecido conter apenas 11 amostras (consultar Tabela A) e uma rede biológica com um tamanho muito superior em comparação à dos outros tecidos levou a que fosse excluído das análises (Tabela B).

A figura 4.1 representa alguns parâmetros das redes de genes com correlação $> 0,9$. Como pode ser observado, os tecidos que se encontram na lateral esquerda dos gráficos são tecidos correspondentes ao cérebro, à excepção do cerebelo (*brain-cerebellum*), juntamente com os rins (*kidney-cortex*), sangue (*whole blood*) e fibroblastos (*cells-transformed fibroblasts*). Isto significa que os tecidos cerebrais apresentam mais genes mitocondriais (entre 247 e 493) altamente correlacionados com outros genes (mitocondriais ou não). Na maioria, os genes presentes nos tecidos cerebrais apresentam um elevado número de correlações (*degree*), especialmente nos tecidos cérebro-cortex cingulado anterior (*brain-anterior cingulate cortex*) e putamen (*brain-putamen (basal ganglia)*). Já o tecido dos rins (*kidney-cortex*), apesar de ter um elevado número de genes mitocondriais altamente correlacionados com outros genes (570 genes, quase metade dos genes mitocondriais existentes) comparado com todos os tecidos cerebrais, estes contém um número

inferior de correlações, formando *clusters* independentes (aumentando o valor de *betweenness centrality*). Isto pode ser verificado nas figuras 4.2 e 4.3.

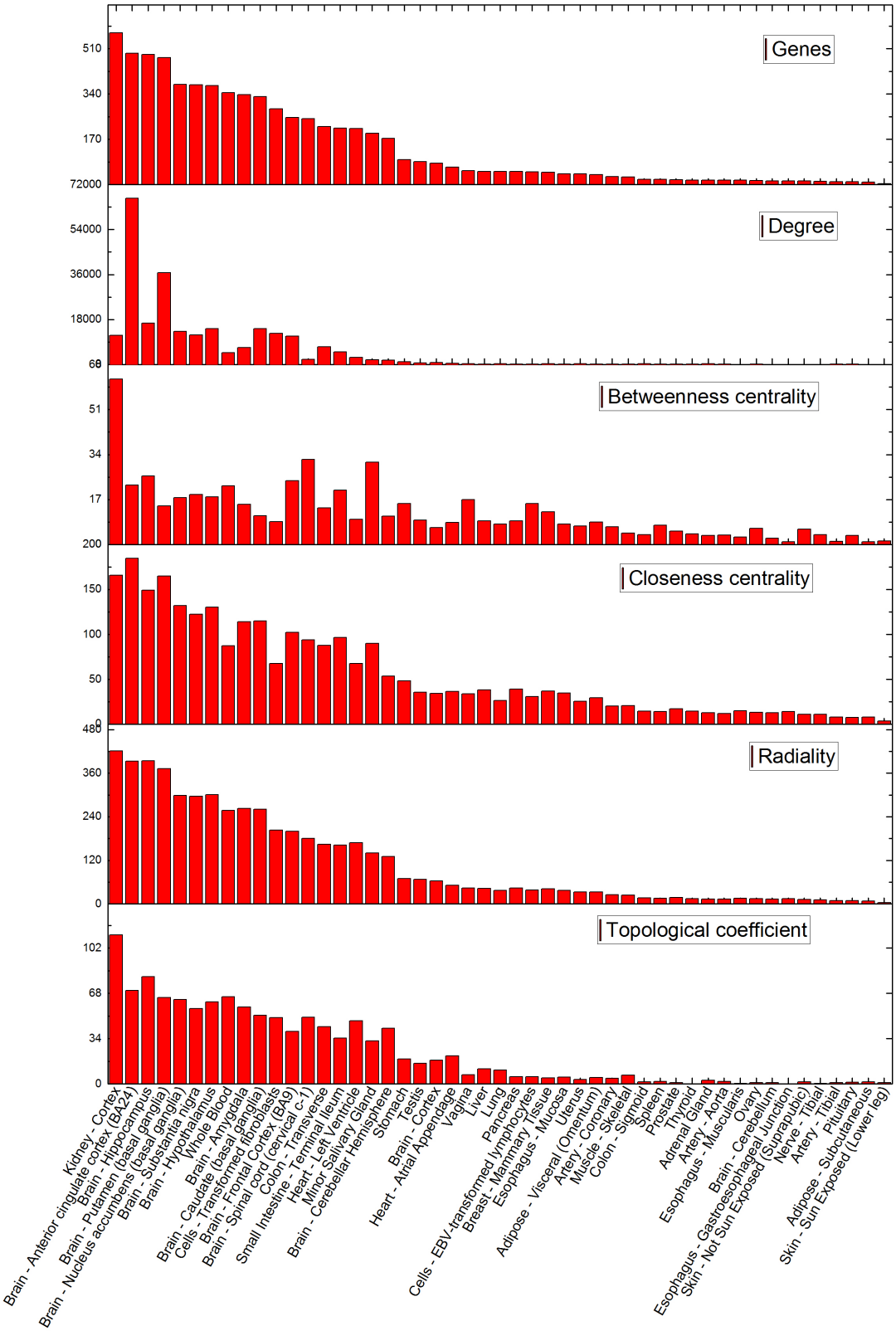


Figura 4.1: Parâmetros das redes de genes com correlação $> 0,9$ em 49 tecidos.

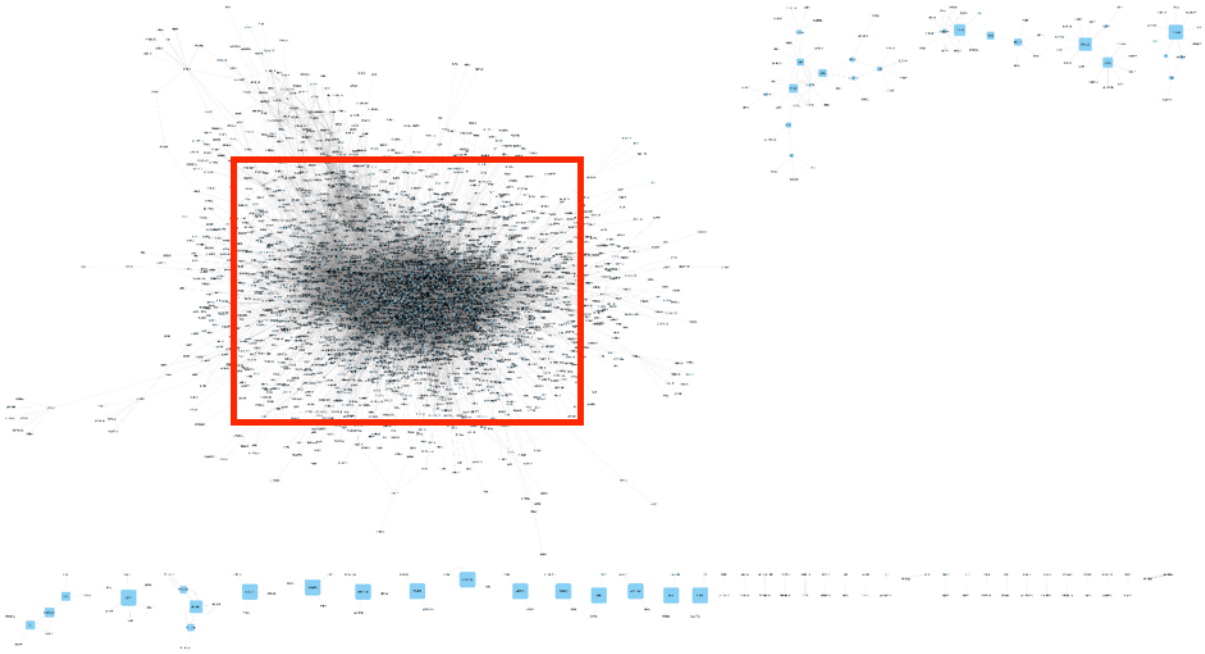


Figura 4.2: Redes de genes com correlação $> 0,9$ no tecido *Brain - Anterior cingulate cortex*. Assinalado o elevado número correlações com genes mitocondriais, resultando uma rede muito densa.

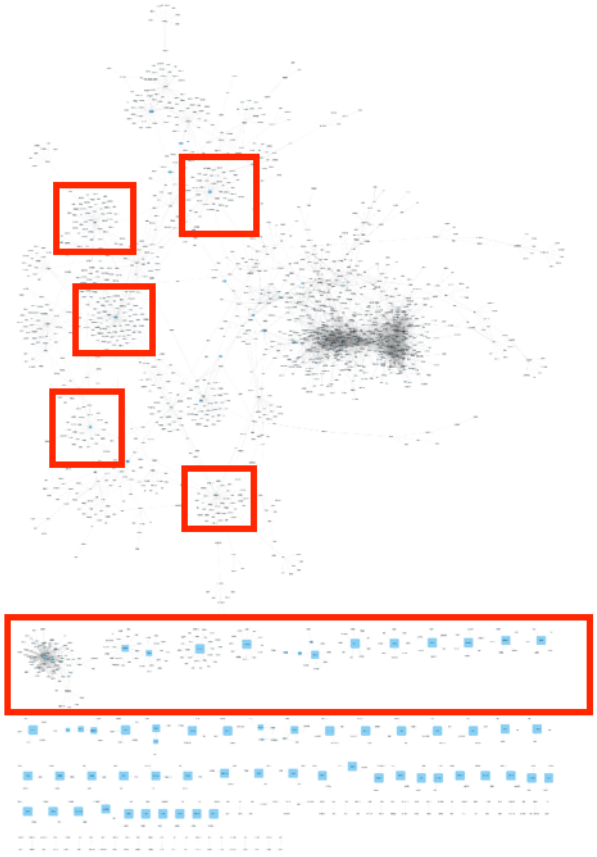


Figura 4.3: Redes de genes com correlação $> 0,9$ no tecido *Kidney (cortex)*. Assinalados grupos de sub-redes existentes, aumentando o valor de centralidade dos vértices.

Verifica-se em quase todos os tecidos que os genes codificados pelo *mtDNA* estão, na maior parte das vezes, altamente correlacionados com genes codificados pelos mesmo genoma. Nos tecidos cerebrais, estes encontram-se particularmente em pares de genes ou em pequenas redes (Figura 4.4). As únicas excepções entre pares de genes codificados pelo *mtDNA*-codificados pelo nDNA e correlação $> 0,9$ ocorre nos fibroblastos (*cells transformed fibroblasts*) para o gene *MT-ND2* (Figura 4.4) com os genes *CFP* (*complement factor properdin*; sistema imunitário inacto), *CHTOP* (*chromatin target of PRMT1*; activação de genes responsivos a estrogénios), *CYP27B1* (*cytochrome P450 family 27 subfamily B member 1*; reacção envolvida no metabolismo de medicamentos e síntese de colesterol, esteróides e outros lípidos), *GPBP1L1* (*GC-rich prometer biding protein 1 lik 1*; possível factor de transcrição), *PIGC* (*phosphatidylinositol glycan anchor biosynthesis class C*; envolvido na síntese da proteína com âncora lipídica *GPI* que liga proteínas à superfície celular de várias células sanguíneas), *RABIF* (*RAB interacting factor*; regulação do transporte intracelular de vesículas) e *YY1* (*YY1 transcription factor*; envolvido na supressão e activação de vários promotores). Só o gene *CYP27B1* é um gene mitocondrial codificado pelo *nDNA*. Como estes fibroblastos foram transformados em laboratório e não são células naturais, podemos concluir que não foram observados pares de genes *mtDNA-nDNA* com valores muito altos de correlação nos tecidos.

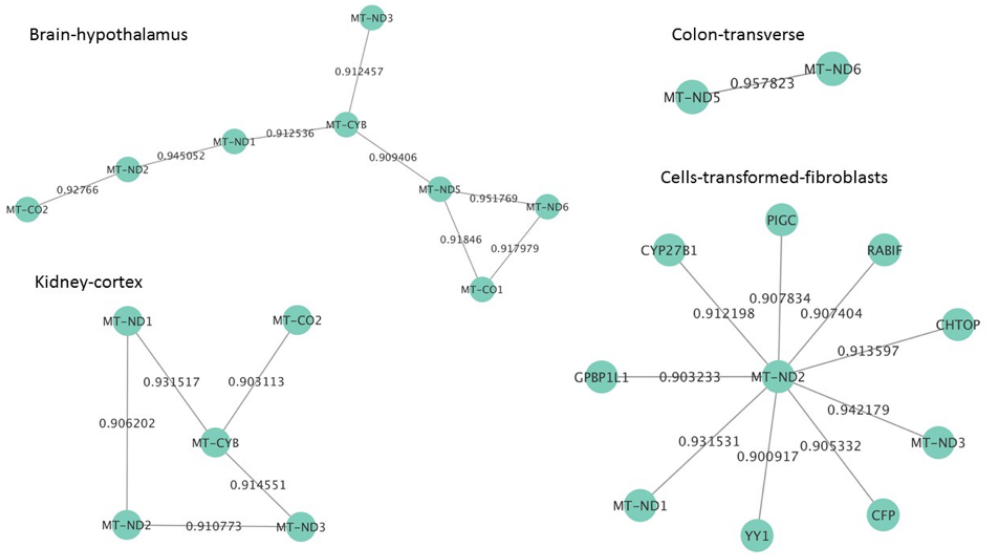


Figura 4.4: Redes de correlação $> 0,9$ que envolvem pares de genes *mtDNA-nDNA* nos tecidos *brain-hypothalamus*, *colon-transverse*, *kidney-cortex* e *cells-transformed-fibroblasts*.

Relativamente às redes de genes de correlação $> 0,8$ (4.5), é possível observar nos tecidos cerebrais, rins (*kidney-cortex*), sangue (*whole blood*) e glândulas salivares (*minor salivary glands*) que entre 70% a 86% dos genes mitocondriais estão altamente correlacionados com outros genes, enquanto na maior parte dos outros tecidos os valores variam entre 18% e 52% .

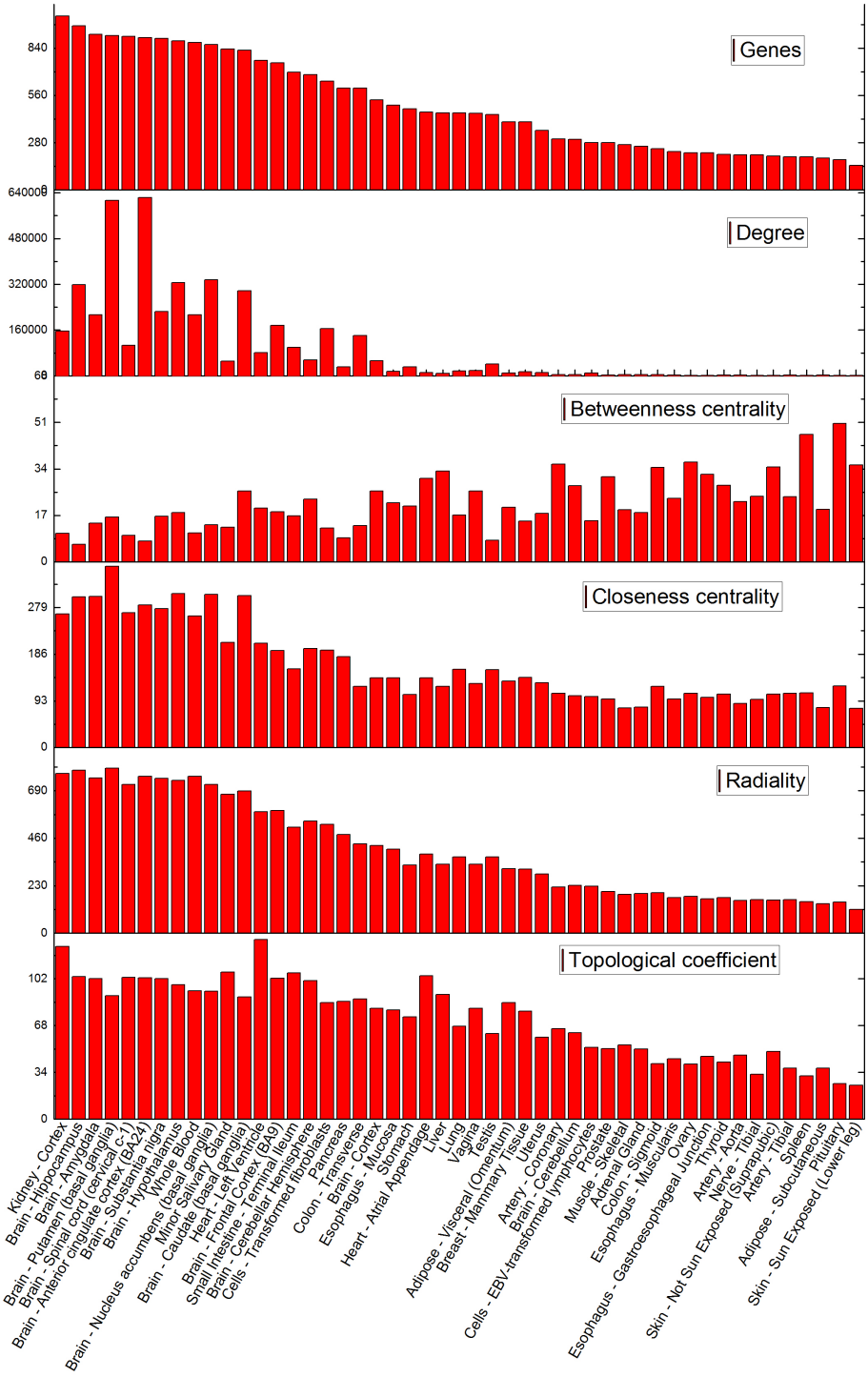


Figura 4.5: Parâmetros das redes de genes com correlação $> 0,8$ em 49 tecidos.

As redes tornam-se, tal como esperado, mais complexas quando consideradas as correlações superiores a $0,8$ (Figura 4.6). Os tecidos do cérebro-hipotálamo (*brain-hypothalamus*) e rins (*kidney-cortex*) continuam a envolver apenas interacções entre genes mitocondriais codificado pelo *mtDNA*, mas o cólon transverso (*colon-transverse*) já inclui os genes *ANKDD1B* (*Ankyrin Repeat and Death Containing 1B*; função desconhecida), *BAIAP2L2* (*BAI1 Associated Protein 2 Like 2*; induz a formação estrutural de membranas planas e curvas), *EPHA10* (*EPH Receptor A10*; mediadores importantes na comunicação entre células que a regulam a sua ligação, forma, e mobilidade nas células neuronais e epiteliais), *RPL10L* (*Ribosomal Protein L10 Like*; desconhecido se funciona como proteína ribossómica funcional ou apresenta outra função) e *TTC6* (*Tetratricopeptide Repeat Domain 6*; função desconhecida) a interagir com genes codificados pelo *mtDNA*. Existe uma grande quantidade de genes do *nDNA* correlacionados com os genes *MT-ND1*, *MT-ND2*, *MT-ND3* com um valor superior a $0,8$ nos fibroblastos (*cells-transformed-fibroblasts*).

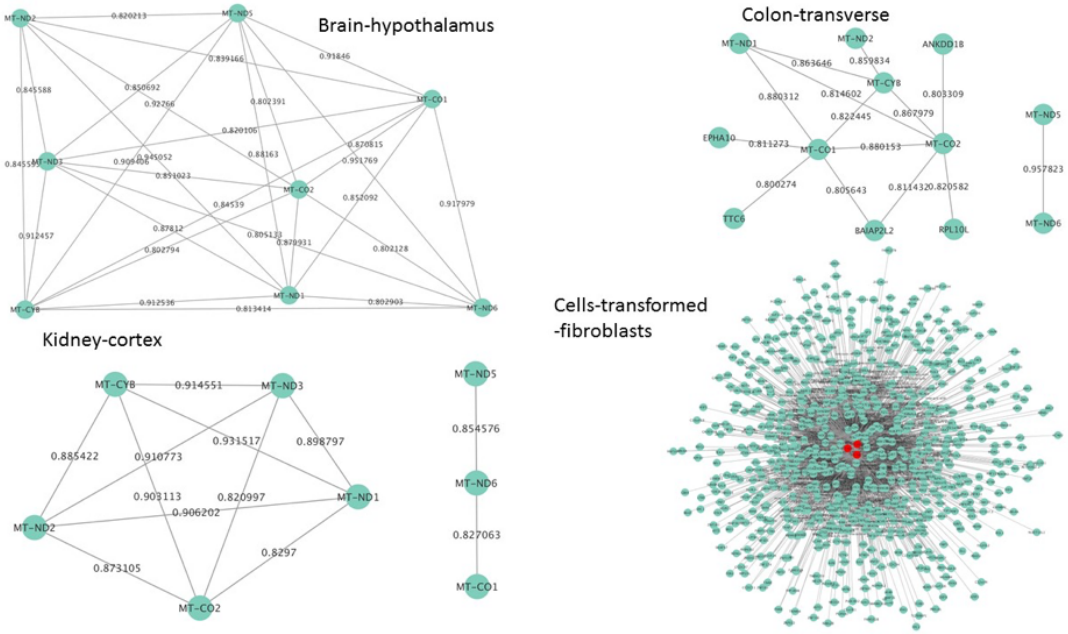
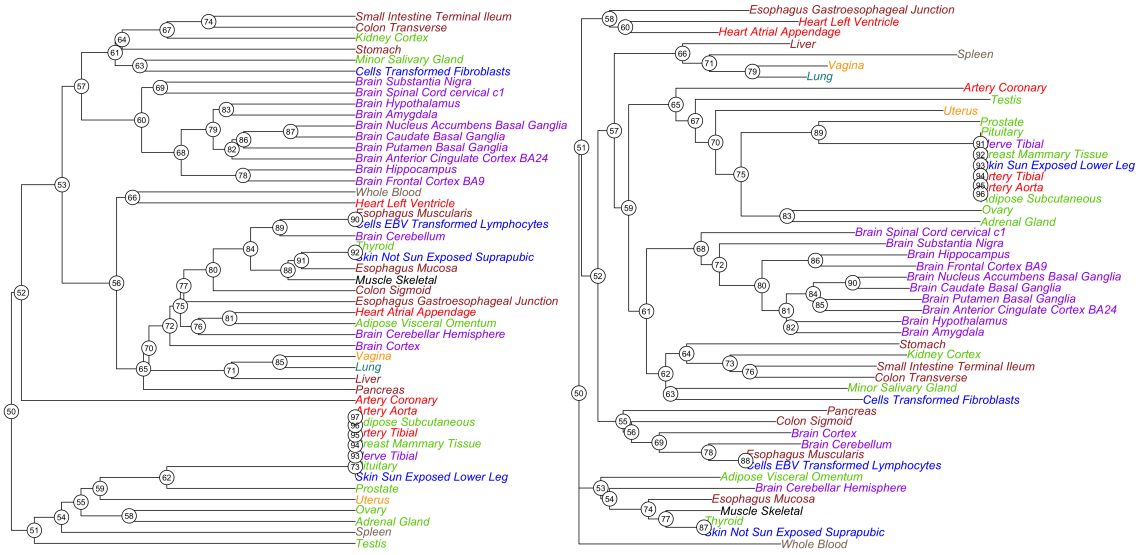


Figura 4.6: Redes de correlação $> 0,8$ que envolvem pares de genes *mtDNA-nDNA* nos tecidos *brain-hypothalamus*, *colon-transverse*, *kidney-cortex* e *cells-transformed-fibroblasts*. Os vértices a vermelho na rede do tecido *cells-transformed-fibroblasts* representam os genes *MT-ND1*, *MT-ND2* e *MT-ND3*.

O enriquecimento com dados funcionais (biológicos) das redes anteriores revelou alguns resultados interessantes. Serão apenas demonstrados os resultados de enriquecimento com a base de dados *KEGG*, para redes com correlação superior a 0,9 (Figura 4.7), uma vez que esta base dados é menos detalhada (alto nível) que a *GO* e serve o propósito de demonstração nesta dissertação.



(a) Resultado obtido da aplicação do método *UPGMA* (b) Resultado obtido da aplicação do método *NJ*.

Figura 4.7: Árvores resultantes do enriquecimento com dados do *KEGG* em redes de correlação $> 0,9$. As cores agrupam os tecidos de acordo com a similaridade de localização/histológica, em que: vermelho - sistema cardiovascular; castanho - sistema digestivo; verde - sistema exócrino e endócrino; castanho claro - sistemas hémico e imune; azul - sistema tegumentar; preto - sistema musculoesquelético; violeta - sistema nervoso; ciano - sistema respiratório; laranja - sistema urogenital.

A *pathway* ribossoma (*ribosome*) está presente em quase todos os tecidos, excepto nos cérebro-cortex (*brain-cortex*), cérebo-cerebelo (*brain-cerebellum*) e estômago (*stomach*). Na árvore com raiz, que corresponde ao método *UPGMA*, a divisão principal encontra-se no primeiro ramo da esquerda que é composto por tecidos que contêm poucas *pathways* ou nenhumas nos genes correlacionados, enquanto o ramo 53 é definido pela fosforilação oxidativa (*oxidative-phosphorylation*) tal como as *pathways* associadas às doenças *Alzheimer, Parkinson* e *Hungtingtons*, mostrando que esses tecidos dependem de um maior número de genes altamente correlacionados para a produção de energia. Este grupo tem um maior número de *pathways* (média de $18,4$; 4 partilhadas entre todos os nós descendentes) que o grupo anterior (média de $2,2$; só a *pathway* ribossoma (*ribosome*) é comum a todos os descendentes), mostrando uma correlação superior entre pares genes mitocondriais-genes que codificam proteínas nas vias metabólicas.

O grupo com mais vias metabólicas com elevada correlação de genes é o grupo exclusivo aos tecidos cerebrais (à excepção do cerebelo (*brain-cerebellum*), hemisfério cerebelar (*brain-cerebellar hemisphere*), cortex (*brain-cortex*) e nervo tibial (*nerve-tibial*)) identificado pelo número 60 (média de $30,9$; 9 *pathways* comuns entre todos os descendentes), partilhando a via

glicólise-gluconeogénese (*glycolysis-gluconeogenesis*), reflectindo a importância das vias de produção de energia. Este ramo dominado pelos tecidos cerebrais (60 na árvore *UPGMA* e 68 na árvore correspondente ao método *NJ*) partilham ainda a *pathway* de sistema sinalização de fosfatidilinositol (*phosphatidylinositol-signaling-system*) mais especificamente através do gene *PI4K4*. O fosfatidilinositol é o mensageiro secundário mais importante no cérebro [WZF+17].

O sistema digestivo encontra-se espalhado em pelo menos três grupos nas árvores, que parece reflectir a sua extensão pelo torso, com excepção da surpreendente proximidade entre o cólon sigmóide (*colon-sigmoid*) e o esófago-mucosa (*esophagus-mucosa*). O sub-grupo 67 no método *UPGMA* que inclui os rins (*kidney-cortex*), cólon-transverso (*colon-transverse*) e íleo (*small-intestine-terminal-illeum*) partilham algumas *pathways*, incluindo (e quase unicamente) a *pathway* de sinalização celular PPAR (*PPAR-signaling-pathway*) que desempenha um papel na libertação de lipídios circulantes ou celulares. Outra *pathway* interessante partilhada neste sub-grupo é o metabolismo do butanoato (*butanoate metabolism*) pois o butirato (*butyrate*) produzido pelas bactérias que habitam no cólon é a principal fonte de energia dos colonócitos; os ratos que são livres de germes, ao não produzir butirato, apresentam uma diminuição de produção de energia no cólon [DGZ+11]. Outro estudo demonstrou que o butirato tem maior concentração na zona do cólon proximal, não sendo tão importante no cólon distal, o que pode explicar o padrão de *clustering* desta porção do cólon.

As principais cavidades do coração, ventrículos e aurículas, encontram-se agrupadas no método *NJ*, identificadas pela *pathway* de sinalização de cálcio (*calcium signaling*), contracção do músculo cardíaco (*cardiac muscle contraction*), ciclo de *Krebs* (*citrate cylcle - TCA cycle*), metabolismo de ácidos gordos (*fatty acid metabolism*), oxidação fosforilativa (*oxidative phosphorylation*) e degradação da valina, leucina e isoleucina (*valine, leucine and isoleucine degradation*). Portanto, estes tecidos também apresentam um perfil de elevada produção energética, como seria expectável, reflectindo-se na elevada correlação de proteínas mitocondriais com outras proteínas envolvidas nestas vias.

4.2 Plataforma *Web* - BioTree Viewer

Foi desenvolvida uma plataforma *web* que permite a visualização de árvores e análise biológica através da inserção de informação de redes biológicas. Esta plataforma está disponível online em https://joaoalmeida.me/dissertacao/viewer/ e requer um *web browser* compatível *Javascript*. A plataforma está acompanhada de exemplos que foram gerados através dos dados fornecidos inicialmente nesta dissertação com auxílio das ferramentas desenvolvidas para a sua análise. A plataforma denominada *BioTree Viewer* é composta por três grandes componentes:

- Importação de árvores no formato *newick*, visualização e manipulação.
- Importação de informação biológica e visualização da mesma de forma interactiva na árvore.
- Exportação da árvore e de dados de análise biológica em formatos adequados.

A plataforma permite visualisar, manipular e analisar, sem qualquer software instalado, uma árvore importada no formato *newick*. Das suas funcionalidades, destacam-se a possibilidade de escolher o formato visual da árvore (vertical ou circular), manipulação da árvore (colapsar ramos, selecionar nós descendentes de um nó, visualizar caminho até à raíz, pesquisar ramo e observar caminho na árvore), identificação de grupos/*clusters* fornecidos *a priori* e exportação da árvore actual para formato *.png*.

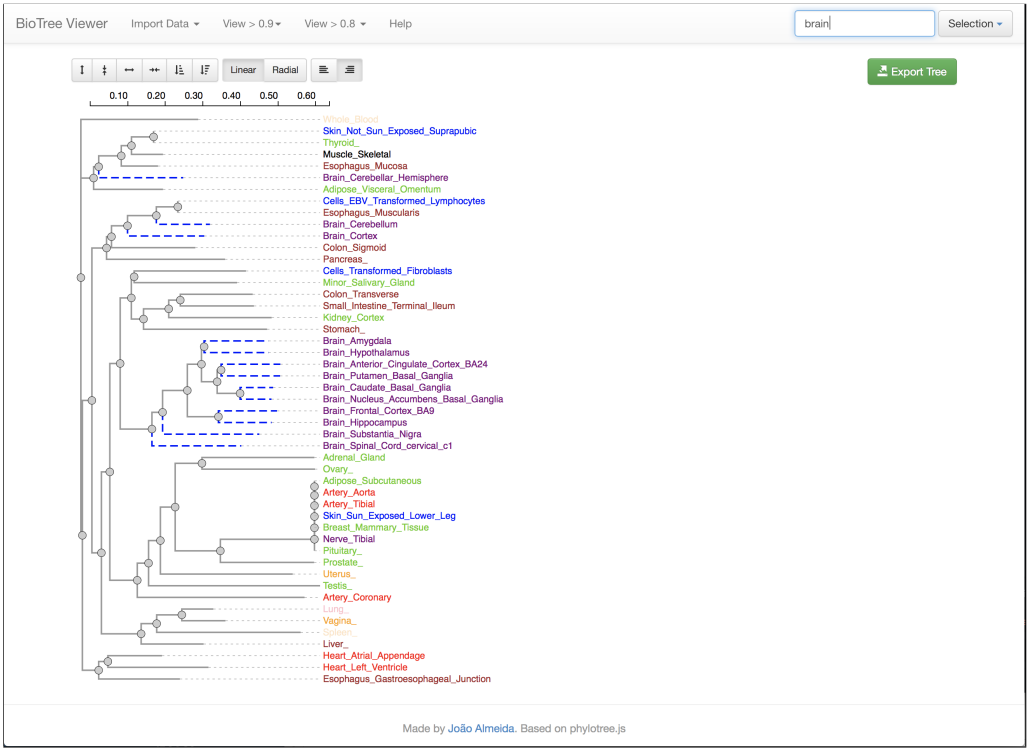


Figura 4.8: Árvore em formato vertical. As cores dos tecidos representam diferentes sistemas do corpo humano. É possível a pesquisa por determinada palavra, os tecidos que a conterem serão assinalados na árvore através de uma linha tracejada.

O ponto diferenciador desta plataforma em relação às existentes é a possibilidade de introduzir informação biológica e permitir, de uma forma interactiva, a sua análise, visualização e exportação. Clicando num nó da árvore, podem-se verificar quais as *pathways* comuns aos tecidos (que se encontram dentro do ramo selecionado), bem como os genes (Figura 4.9). É possível exportar os dados desta análise nos formatos *CSV* e *Excel* e ainda filtar a informação através da pesquisa de determinado gene ou *pathway*.

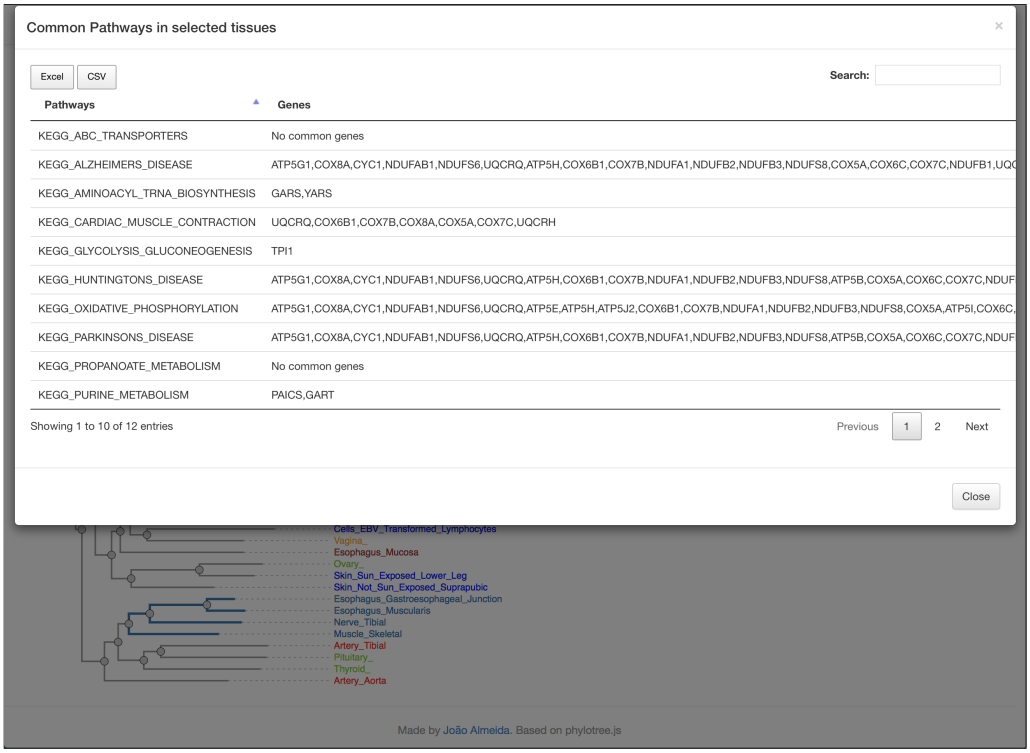


Figura 4.9: Análise de *pathways* e genes em comum entre os tecidos marcados com ramos a azul. É possível a pesquisa e exportação dos dados

Os resultados desta análise biológica foram validados através do módulo desenvolvido em *Prolog* abordado anteriormente na implementação. Através deste, foi possível verificar se os resultados relativos às *pathways* comuns aos diferentes ramos das árvore se encontravam correctos. Para além de permitir a validação dos dados da plataforma, o facto de ter sido desenvolvido o módulo, permite ao utilizador final adoptar uma solução *on-premise* em detrimento da solução *cloud* fornecida pela plataforma *web* disponível *online* (a não ser que corra a plataforma num servidor local).

Uma vez que o público alvo desta plataforma são os investigadores, todo o processamento da informação é feito do lado do cliente (*client-sided*) e nunca é enviada qualquer informação para um servidor, o que é uma característica importante, uma vez que, na maior parte dos projectos de investigação a protecção da informação desempenha um papel fulcral.

Capítulo 5

Conclusões

A aplicação *BioTree Viewer* foi utilizada e testada por profissionais na área de genómica. A análise biológica dos resultados das árvores geradas pelos métodos de *clustering*, *UPGMA* e *NJ* respectivamente, foi feita com o auxílio desta mesma plataforma, provando a sua utilidade.

Espera-se que a plataforma e as ferramentas desenvolvidas neste trabalho venham a ser adoptadas para visualização e estudo de dados de outras análises semelhantes, no contexto de outras espécies ou doenças.

Uma vez que o tempo de realização deste trabalho foi bastante limitado para a complexidade e grande quantidade de dados a processar, a análise não pôde ser tão extensa quanto o desejável. O próximo passo do trabalho seria obtenção de dados de expressão dos mesmo tecidos, em situação tumoral, através da base de dados *The Cancer Genome Atlas* (*TCGA*) e a aplicação do mesmo tipo de análise para poder investigar diferenças devidas ao processo tumoral.

Referências

[ABB+00] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[BS01] Brendan J Battersby e Eric A Shoubridge. Selection of a mtdna sequence variant in hepatocytes of heteroplasmic mice is not due to differences in respiratory chain function or efficiency of replication. *Human Molecular Genetics*, 10(22):2469–2479, 2001.

[C+15] Gene Ontology Consortium et al. Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056, 2015.

[Chi] Patrick F. Chinnery. Mitochondrial DNA in homo sapiens. In *Nucleic Acids and Molecular Biology*, pages 3–15. Springer Nature.

[DGZ+11] Dallas R Donohoe, Nikhil Garge, Xinxin Zhang, Wei Sun, Thomas M O'Connell, Maureen K Bunger e Scott J Bultman. The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell metabolism*, 13(5):517–526, 2011.

[DS01] Salvatore DiMauro e Eric A. Schon. Mitochondrial DNA mutations in human disease. *American Journal of Medical Genetics*, 106(1):18–26, 2001.

[FGR+02] Maria Falkenberg, Martina Gaspari, Anja Rantanen, Aleksandra Trifunovic, Nils-Göran Larsson e Claes M. Gustafsson. Mitochondrial transcription factors b1 and b2 activate transcription of human mtDNA. *Nature Genetics*, 31(3):289–294, jun 2002.

[GDL+01] Gert Van Goethem, Bart Dermaut, Ann Löfgren, Jean-Jacques Martin e Christine Van Broeckhoven. *Nature Genetics*, 28(3):211–212, jul 2001.

[HDY+01] Li-Li Hsiao, Fernando Dangond, Takumi Yoshida, Robert Hong, Roderick V Jensen, Jatin Misra, William Dillon, Kailin F Lee, Kathryn E Clark, Peter Haverty et al. A compendium of gene expression in normal human tissues. *Physiological genomics*, 7(2):97–104, 2001.

[Hil16] Martin Hilbert. Big data for development: A review of promises and challenges. *Development Policy Review*, 34(1):135–174, 2016.

[JMF99] Anil K Jain, M Narasimha Murty e Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[KFT+17] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato e Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.

[Kin67] Benjamin King. Step-wise clustering procedures. *Journal of the American Statistical Association*, 62(317):86–101, 1967.

[LC95] Nils-Göran Larsson e David A Clayton. Molecular genetic aspects of human mitochondrial disorders. *Annual review of genetics*, 29(1):151–178, 1995.

[LdR05] Ricardo Luis Lachi e Heloisa Vieira da Rocha. Aspectos basicos de clustering: conceitos e tecnicas. Technical report, 2005.

[LTS+13] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[MMU+98] Kari Majamaa, Jukka S. Moilanen, Seija Uimonen, Anne M. Remes, Pasi I. Salmela, Mikko Kärppä, Kirsi A.M. Majamaa-Voltti, Harri Rusanen, Martti Sorri, Keijo J. Peuhkurinen e Ilmo E. Hassinen. Epidemiology of a3243g, the mutation for mitochondrial encephalomyopathy, lactic acidosis, and strokelike episodes: Prevalence of the mutation in an adult population. *The American Journal of Human Genetics*, 63(2):447–454, aug 1998.

[Nis99] I. Nishino. Thymidine phosphorylase gene mutations in MNGIE, a human mitochondrial disorder. *Science*, 283(5402):689–692, jan 1999.

[ODS13] Aisling O'Driscoll, Jurate Daugelaite e Roy D Sleator. 'big data', hadoop and cloud computing in genomics. *Journal of biomedical informatics*, 46(5):774–781, 2013.

[SBD97] Eric A. Schon, Eduardo Bonilla e Salvatore DiMauro. *Journal of Bioenergetics and Biomembranes*, 29(2):131–149, 1997.

[Sho01] E. A. Shoubridge. Nuclear genetic defects of oxidative phosphorylation. *Human Molecular Genetics*, 10(20):2277–2284, oct 2001.

[SLT+01] Johannes N. Spelbrink, Fang-Yuan Li, Valeria Tiranti, Kaisu Nikali, Qiu-Ping Yuan, Muhammed Tariq, Sjoerd Wanrooij, Nuria Garrido, Giacomo Comi, Lucia Morandi, Lucio Santoro, Antonio Toscano, Gian-Maria Fabrizi, Hannu Somer, Rebecca Croxen, David Beeson, Joanna Poulton, Anu Suomalainen, Howard T Jacobs, Massimo Zeviani e Catharina Larsson. *Nature Genetics*, 28(3):223–231, jul 2001.

[SMO+03] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski e Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[SN87] Naruya Saitou e Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[SS+73] Peter HA Sneath, Robert R Sokal et al. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

[Suc] Martin Suchara. COS 424: Interacting with Data. Lecture #20.

[TAK+99] Douglass M. Turnbull, Richard M. Andrews, Iwona Kubacka, Patrick F. Chinnery, Robert N. Lightowlers e Neil Howell. *Nature Genetics*, 23(2):147–147, oct 1999.

[Tea00] R Core Team. R language definition. *Vienna, Austria: R foundation for statistical computing*, 2000.

[URR+00] Johanna Uusimaa, Anne M. Remes, Heikki Rantala, Leena Vainionpää, Riitta Herva, Katri Vuopala, Matti Nuutinen, Kari Majamaa e Ilmo E. Hassinen. Childhood encephalopathies and myopathies: A prospective study in a defined population to assess the frequency of mitochondrial disorders. *Pediatrics*, 105(3):598–603, 2000.

[VAM+01] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[VV07] Fabricio Voznika e Leonardo Viana. Data mining classification, 2007.

[Wal99] D. C. Wallace. Mitochondrial diseases in man and mouse. *Science*, 283(5407):1482–1488, mar 1999.

[WBM+98] Douglas C Wallace, Michael D Brown, Simon Melov, Brett Graham e Marie Lott. Mitochondrial biology, degenerative diseases and aging. *Biofactors*, 7(3):187–190, 1998.

[WZF+17] Long Wang, Kai Zhou, Zhi Fu, Di Yu, Hesuyuan Huang, Xiaodong Zang e Xuming Mo. Brain development and akt signaling: The crossroads of signaling pathway and neurodevelopmental diseases. *Journal of Molecular Neuroscience*, 61(3):379–384, 2017.

[Zah71] Charles T Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on computers*, 100(1):68–86, 1971.

REFERÊNCIAS

Appendices

Anexo A

Tabela A.1: Número de amostras e grupos dos tecidos.

Tecido	Grupo	Nº Amostras
Adipose - Subcutaneous	Sistema exó/endócrino	350
Adipose - Visceral (Omentum)	Sistema exó/endócrino	227
Adrenal Gland	Sistema exó/endócrino	145
Artery - Aorta	Sistema cardiovascular	224
Artery - Coronary	Sistema cardiovascular	133
Artery - Tibial	Sistema cardiovascular	332
Bladder	Sistema urogenital	11
Brain - Amygdala	Sistema nervoso	72
Brain - Anterior cingulate cortex (BA24)	Sistema nervoso	84
Brain - Caudate (basal ganglia)	Sistema nervoso	117
Brain - Cerebellar Hemisphere	Sistema nervoso	105
Brain - Cerebellum	Sistema nervoso	125
Brain - Cortex	Sistema nervoso	114
Brain - Frontal Cortex (BA9)	Sistema nervoso	108
Brain - Hippocampus	Sistema nervoso	94
Brain - Hypothalamus	Sistema nervoso	96
Brain - Nucleus accumbens (basal ganglia)	Sistema nervoso	113
Brain - Putamen (basal ganglia)	Sistema nervoso	97
Brain - Spinal cord (cervical c-1)	Sistema nervoso	71
Brain - Substantia nigra	Sistema nervoso	63
Breast - Mammary Tissue	Sistema exó/endócrino	214
Cells - EBV-transformed lymphocytes	Sistema tegumentar	118
Cells - Transformed fibroblasts	Sistema tegumentar	284
Cervix - Ectocervix	Sistema urogenital	6
Cervix - Endocervix	Sistema urogenital	5
Colon - Sigmoid	Sistema digestivo	149
Colon - Transverse	Sistema digestivo	196

Continuação da tabela A.1		
Tecido	**Grupo**	**Nº Amostras**
Esophagus - Gastroesophageal Junction	Sistema digestivo	153
Esophagus - Mucosa	Sistema digestivo	286
Esophagus - Muscularis	Sistema digestivo	247
Fallopian Tube	Sistema urogenital	6
Heart - Atrial Appendage	Sistema cardiovascular	194
Heart - Left Ventricle	Sistema cardiovascular	218
Kidney - Cortex	Sistema exó/endócrino	32
Liver	Sistema digestivo	119
Lung	Sistema Respiratório	320
Minor Salivary Gland	Sistema exó/endócrino	57
Muscle - Skeletal	Sistema musculoesquelético	430
Nerve - Tibial	Sistema nervoso	304
Ovary	Sistema exó/endócrino	97
Pancreas	Sistema digestivo	171
Pituitary	Sistema exó/endócrino	103
Prostate	Sistema exó/endócrino	106
Skin - Not Sun Exposed (Suprapubic)	Sistema tegumentar	250
Skin - Sun Exposer (Lower leg)	Sistema tegumentar	357
Small Intestine - Terminal Ileum	Sistema digestivo	88
Spleen	Sistemas hémico e imune	104
Stomach	Sistema digestivo	193
Testis	Sistema exó/endócrino	172
Thyroid	Sistema exó/endócrino	323
Uterus	Sistema urogenital	83
Vagina	Sistema urogenital	96
Whole Blood	Sistemas hémico e imune	393

Anexo B

Tabela B.1: Número de genes e interacções nas redes biológicas

Tecidos	**Correlação > 0,8**		**Correlação > 0,9**	
	Genes	**Interações**	**Genes**	**Interações**
Adipose - Subcutaneous	480	1291	22	29
Adipose - Visceral (Omentum)	1965	7323	90	119
Adrenal Gland	1298	2616	240	242
Artery - Aorta	809	1809	43	59
Artery - Coronary	1122	2696	107	135
Artery - Tibial	495	1852	34	54
Bladder	16855	1125749	11165	281978
Brain - Amygdala	7423	183590	1580	5696
Brain - Anterior cingulate cortex (BA24)	8550	557848	3438	59073
Brain - Caudate (basal ganglia)	7069	260451	2069	12636
Brain - Cerebellar Hemisphere	4572	46427	574	1232
Brain - Cerebellum	1506	3570	22	18
Brain - Cortex	4245	47698	395	712
Brain - Frontal Cortex (BA9)	6515	159076	1681	10176
Brain - Hippocampus	8684	269190	2288	13019
Brain - Hypothalamus	7795	289529	2158	12785
Brain - Nucleus accumbens (basal ganglia)	8006	299106	2354	11443
Brain - Putamen (basal ganglia)	8513	543360	3328	32214
Brain - Spinal cord (cervical c-1)	7817	91918	1057	1645
Brain - Substantia nigra	7680	191063	1845	9278
Breast - Mammary Tissue	2441	11408	154	223
Cells - EBV-transformed lymphocytes	1190	7260	108	117
Cells - Transformed fibroblasts	4766	142979	1701	10698
Colon - Sigmoid	1241	2956	148	252
Colon - Transverse	5085	115348	970	5158
Esophagus - Gastroesophageal Junction	886	1243	21	16

Continuação da tabela B.1				
Tecidos	**Correlação > 0,8**		**Correlação > 0,9**	
	Genes	**Interações**	**Genes**	**Interações**
Esophagus - Mucosa	2717	12800	86	74
Esophagus - Muscularis	999	2046	25	20
Heart - Atrial Appendage	1204	6270	83	216
Heart - Left Ventricle	3674	57743	466	1690
Kidney - Cortex	10082	120012	2003	7685
Liver	1922	6429	116	120
Lung	1673	11894	106	153
Minor Salivary Gland	6928	42196	921	1553
Muscle - Skeletal	860	2428	59	100
Nerve - Tibial	637	1372	25	18
Ovary	832	1502	63	88
Pancreas	4243	27236	126	118
Pituitary	866	1551	66	81
Prostate	1213	2211	66	62
Skin - Not Sun Exposed (Suprapubic)	804	1275	24	23
Skin - Sun Exposed (Lower leg)	567	735	7	7
Small Intestine - Terminal Ileum	7690	90682	1793	4529
Spleen	955	1525	100	115
Stomach	4087	27680	590	919
Testis	4197	37653	437	603
Thyroid	505	1239	26	26
Uterus	1933	9476	168	259
Vagina	3680	16336	210	216
Whole Blood	7485	185183	1708	4079