UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

ANÁLISE DAS TÉCNICAS DE MICROARRANJOS E RNASEQ ATRAVÉS DA ONTOLOGIA ONTOCANCRO

DISSERTAÇÃO DE MESTRADO

Karlise Soares Nascimento

Santa Maria, RS, Brasil
2013

ANÁLISE DAS TÉCNICAS DE MICROARRANJOS E RNASEQ ATRAVÉS DA ONTOLOGIA ONTOCANCRO

Karlise Soares Nascimento

Dissertação apresentada ao Curso de Mestrado do Programa de Pós-Graduação em Informática da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Mestre em Informática**.

Orientador: Dr. Giovani Rubert Librelotto

Santa Maria, RS, Brasil
2013

Ficha catalográfica elaborada através do Programa de Geração Automática
da Biblioteca Central da UFSM, com os dados fornecidos pelo(a) autor(a).

Soares, Karlise
Análise das técnicas de microarranjos e RNAseq
através da Ontologia Ontocancro / Karlise Soares.-2013.
84 p.; 30cm

Orientador: Giovani Rubert Librelotto
Dissertação (mestrado) - Universidade Federal de Santa
Maria, Centro de Tecnologia, Programa de Pós-Graduação em
Informática, RS, 2013

1. Ontocancro 2. Ontologias 3. Barreira anticâncer 4.
Microarranjos 5. RNAseq I. Rubert Librelotto, Giovani
II. Título.

UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

A Comissão Examinadora, abaixo assinada, aprova a Dissertação de Mestrado.

ANÁLISE DAS TÉCNICAS DE MICROARRANJOS E RNASEQ ATRAVÉS DA ONTOLOGIA ONTOCANCRO

elaborada por
Karlise Soares Nascimento

como requisito parcial para obtenção do grau de Mestre em Informática

Comissão Examinadora

Dr. Giovani Rubert Librelotto
(Presidente/Orientador)

Deise de Brum Saccol
(Doutor/Universidade Federal de Santa Maria)

Luís Álvaro de Lima Silva
(Doutor/Universidade Federal de Santa Maria)

Santa Maria, 11 de setembro de 2013.

À minha família e amigos.

AGRADECIMENTOS

Ao Prof. Dr. Giovani Rubert Librelotto, por ter me aceitado como sua aluna e orientanda, pelos seus ensinamentos e pelo suporte em todos os momentos difíceis, que passei durante o período do mestrado.

Aos professores Dr. Eduardo Kessler Piveta e Dra. Deise de Brum Saccol pelos ensinamentos, pela disponibilidade e pelas orientações, sempre que precisei.

À Prof Dra. Marialva Sinigaglia, por ter aceitado o convite para fazer parte da banca de defesa, e pelas contribuições com a pesquisa.

Aos meus colegas do grupo de pesquisa da Ontocancro, em especial, ao Éder Simão, por todas as vezes que precisei da sua opinião e dos seus esclarecimentos, durante e após a realização do seu doutorado em física, sob orientação do Prof Dr. José Carlos Merino Mombach, o qual também agradeço os ensinamentos e orientações, pois foram de fundamental importância para a elaboração desta dissertação.

Aos meus colegas do mestrado, em especial, ao Ederson Bastiani e ao Miguel Bauermann, por todas as vezes que me incentivaram e auxiliaram na elaboração de artigos e na busca de eventos para publicação.

Ao Josmar Nuernberg, secretário do PPGI, por todas as vezes que precisei da sua ajuda.

Aos demais professores do PPGI pelos ensinamentos compartilhados.

À toda minha família, principalmente, à minha mãe Leila Soares, minhas irmãs Alessandra e Katiane, meus sobrinhos Marion, Nicolas, Rafael e Henrique, por sempre estarem ao meu lado, me incentivando e consolando, sempre com muito amor.

Aos amigos, que mantiveram-se ao meu lado, mesmo na minha constante ausência.

E a Deus, por dar-me esperança de que todo sacrifício tem um objetivo, o do nosso crescimento espiritual, sem Ele todo o esforço desprendido não faria sentido.

RESUMO

Dissertação de Mestrado
Programa de Pós-Graduação em Informática
Universidade Federal de Santa Maria

ANÁLISE DAS TÉCNICAS DE MICROARRANJOS E RNASEQ ATRAVÉS DA ONTOLOGIA ONTOCANCRO

Autora: Karlise Soares Nascimento
Orientador: Dr. Giovani Rubert Librelotto
Data e Local da Defesa: Santa Maria, 11 de setembro de 2013

A investigação sobre as interações moleculares, que causam doenças genéticas têm sido foco de muitos estudos nos últimos anos. Com a inclusão da informática em pesquisas biológicas, houve um grande avanço na descoberta de novas soluções e técnicas de diagnósticos para o tratamentos de doenças. O câncer é uma das doenças que mais tem causado interesse entre cientistas do mundo inteiro, e entre suas variações, estão aquelas causadas por defeitos genéticos que afetam os mecanismos de manutenção do genoma (GMM), que promovem o bom funcionamento do organismo de um indivíduo. Halazonetis e colaboradores, em 2008, publicou um estudo envolvendo as vias metabólicas GMM, que identifica o surgimento de uma barreira que evita a propagação do câncer, mas é rompida dando início a reprodução descontrolada de células defeituosas que causam os tumores cancerígenos. A motivação desta dissertação consiste em investigar o processo de ativação da barreira anticâncer, usando métodos quantitativos para análise da expressão de genes e vias. Para isso, utilizou-se a ontologia Ontocancro como ferramenta de análise, sendo efetuadas diversas alterações em sua arquitetura para a realização deste estudo.

Palavras-chave: Ontologia. Câncer. RNAseq. Microarranjos.

ABSTRACT

Dissertação de Mestrado
Programa de Pós-Graduação em Informática
Universidade Federal de Santa Maria

ANALYSIS OF TECHNICAL AND MICROARRAY AND RNASEQ THROUGH ONTOCANCRO ONTOLOGY

Autora: Karlise Soares Nascimento
Orientador: Dr. Giovani Rubert Librelotto
Data e Local da Defesa: Santa Maria, 11 de setembro de 2013

The research on molecular interactions which cause genetic diseases has been the focus of many studies in recent years. With the inclusion of informatics in biological research, there was a great advance in the discovery of new solutions and diagnostic techniques for the treatment of diseases. Cancer is one of the diseases that have caused more interest among scientists around the world, and between their variations, are those caused by genetic defects that affect the genome maintenance mechanisms (GMM), which promote the proper functioning of the organism of an individual. Halazonetis and colaborators (2008) published a study involving of the metabolic pathways GMM, which identifies the appearance of a barrier that prevents the spread of cancer, but it is broken by initiating uncontrolled reproduction of defective cells that cause the cancerous tumors. The motivation of this work is to investigate the process of activation of anticancer barrier, using quantitative methods to analysis of the expression of genes and pathways. For this reason, it was used the Ontocancro ontology as analysis tool, being made several changes in its architecture for the realization of this study.

Keywords: Ontology. Cancer. RNAseq. Microarrays.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BER	*Base Excision Repair*
CS	*Chromosome Stability*
cDNA	DNA complementar
cRNA	RNA complementar
DDR	*Reparo ao Dano do DNA*
DNA	Ácido desoxirribonucleico
DSB	*Double-Strand Break* ou Quebras em fitas duplas no DNA
EBI	*European Bioinformatics Institute*
EJ	*Non-homologous End Joining*
FA	*Fanconi Anemia Repair*
GEO	*Gene Expression Omnibus*
GMM	Mecanismos de manutenção do genoma
HGNC	*HUGO Gene Nomenclature Committee*
HR	*Homologous Recombination* ou Recombinação homóloga
HRR	*HR Repair of Replication-Independent DSB*
MMR	*Mismatch Repair (MMR)* ou Reparo de bases mal pareadas
mRNA	RNA Mensageiro
NCBI	*National Center for Biotechnology Information*
NCI	*National Cancer Institute*
NHEJ	*Non-Homologous end Joining* ou Junção de extremidades homólogas
NER	*Nucleotide Excision Repair* ou Reparo por excisão de nucleotídeos
RNA	Ácido ribonucleico
rRNA	RNA Ribossômico
tRNA	RNA de Transferência

SUMÁRIO

1. INTRODUÇÃO

Uma das doenças humanas mais pesquisadas cientificamente, nos últimos anos, tem sido o câncer, e a informática tem tido um papel fundamental nesse sentido. As células cancerosas possuem uma grande variedade de anormalidades produzidas a partir de erros dos mecanismos reguladores da célula e essa compreensão dos processos biológicos que operam dentro da célula impõe, aos cientistas da computação e biólogos, a procura por métodos inovadores para tratar destes dados (BARABÁSI; OLTVAI, 2004). O tamanho e a complexidade dos dados biológicos coletados durante os últimos anos incluem informações que requerem uma abordagem integradora (UETZ; IDEKER; SCHWIKOWSKI, 2005), uma vez que as informações encontram-se dispersas em bases de dados públicas de diferentes instituições.

A Ontocancro foi desenvolvida para realizar estudos de transcriptômica de câncer, ou seja, investigar o processo de formação de um tumor cancerígeno. Para isso, integra dados biológicos dispersos, que devido a falta de padronização de termos e nomenclaturas, fazem com que o pesquisador busque tais informações em diversas locais. Desta forma, a Ontocancro disponibiliza em um único local, os dados necessários para pesquisas relacionados ao processo carcinogênico.

Inicialmente, a Ontocancro armazenava somente informações sobre genes e redes genéticas, que possuem alguma conexão com a formação do câncer, pois fazem parte dos Mecanismos de Manutenção do Genoma (GMM). No entanto, a quantidade de dados não permitia uma análise mais precisa.

Em um estudo desenvolvido por Halazonetis e colaboradores (2008), foi constatada a existência de uma barreira contra formação do câncer, que quando rompida dá início a proliferação de células defeituosas que causam câncer.

Deste estudo originou o objetivo geral deste trabalho, que consiste na atualização da Ontocancro para incluir métodos quantitativos que permitem avaliar valores de expressão de genes e vias produzidos por técnicas de microarranjos e RNAseq, tais como os métodos de atividade relativa, diversidade relativa e mudança de expressão entre os genes (*fold change*). Essas alterações permitiram uma melhor análise do processo de ativação da barreira anticâncer, proposta por Halazonetis.

Os objetivos específicos incluem:

- Integração das informações obtidas em pesquisas das áreas de conhecimento que formam a ontologia;
- Remodelagem da base de dados, adicionando informações extraídas das técnicas de microarranjos e RNASeq;
- Importação dos novos dados;
- Implementação do algoritmo para cálculo estatístico da atividade relativa e diversidade relativa;
- Produção de estudos de caso e análise dos resultados.

Portanto, o presente trabalho contribui para a atualização da Ontocancro para a versão 2.0, acrescentando informações sobre transcriptomas de alguns tipos de câncer, como da tireoide e do cólon, e sua integração com as redes genéticas relacionadas ao desenvolvimento do câncer. Já a inclusão dos métodos quantitativos nesta versão, permitem traçar o perfil das redes de manutenção do genoma e identificar as redes genéticas que estão mais ativas na fase inicial do câncer.

A ontologia foi desenvolvida de forma interdisciplinar e o grupo de pesquisadores participantes na Ontocancro 2.0 é formado por especialistas na área da física, biologia molecular, bioinformática e informática, cada área de conhecimento contribuiu para a construção da atual base de dados, conforme esquematizado pela Figura 1.

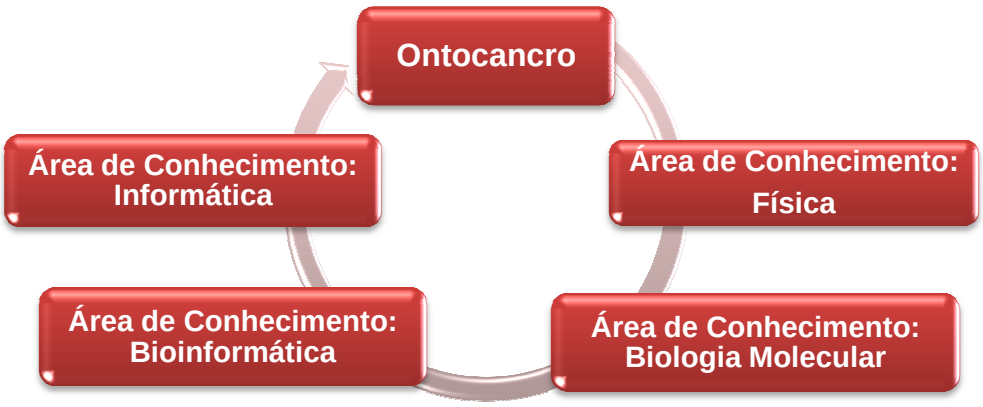


Figura 1 - Esquema das áreas que agregam informações a Ontocancro 2.0

1.1. Organização da Dissertação

A dissertação está organizada conforme o mapa conceitual da Figura 2.

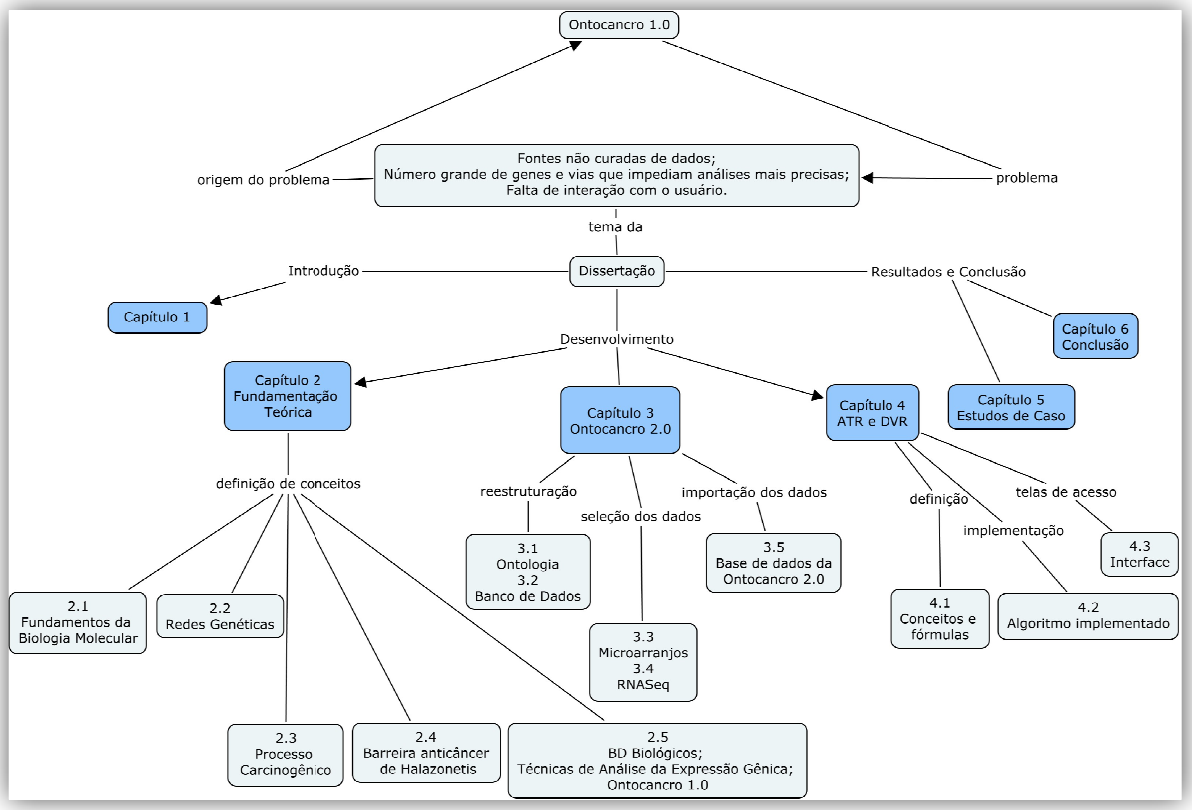


Figura 2 - Mapa conceitual da dissertação

Ela contempla os seguintes assuntos: no Capítulo 2 encontra-se a revisão bibliográfica dos principais conceitos abordados, tais como fundamentos da Biologia Molecular, funcionamento das redes metabólicas envolvidas com o câncer, processo de desenvolvimento inicial do câncer (pré-câncer) e uma descrição do artigo de Halazonetis. Também apresenta os bancos de dados biológicos que foram usados na construção da base de dados da Ontocancro, assim como a sua versão 1.0. O Capítulo 3 descreve a nova estrutura da ontologia, denominada Ontocancro 2.0, abordando sua arquitetura e como os dados foram importados. O Capítulo 4 detalha a implementação dos cálculos estatísticos no banco de dados e no site. *O* Capítulo 5 apresenta os resultados obtidos através de estudos de caso. E por fim, a conclusão, no Capítulo 6.

2. FUNDAMENTAÇÃO TEÓRICA

Uma das doenças humanas mais observadas cientificamente tem sido o câncer. Suas variações evoluem de acordo com fatores ambientais e genéticos. Na maioria dos casos em humanos, os tumores demoram em torno de 20 anos para saírem do estado carcinogênico (potencial para o desenvolvimento do câncer) para o estado de exposição podendo assim, serem detectados clinicamente. Durante este tempo as células cancerosas adquirem uma capacidade de divisão, invasão e metástase (LOEB; LOEB; ANDERSON, 2003).

As células cancerosas possuem uma grande variedade de anormalidades, tais como: diferença na quantidade de genes, inatividade de genes ou falta de cromossomos. Essa instabilidade genética pode ser encontrada em vários tipos de câncer. A perda da estabilidade do genoma associada com a deterioração genética é um dos muitos aspectos importantes dos carcinomas (JEFFORD; IRMINGER-FINGER, 2006).

Para a devida compreensão do tema desta dissertação este capítulo abordará os conceitos fundamentais da biologia molecular (seção 2.1), que incluem uma revisão sobre o dogma central, sobre o DNA (ácido desoxirribonucleico) e sobre o RNA (ácido ribonucleico), uma vez que os dados analisados pela Ontocancro referem-se, também, a tais componentes.

A seção 2.2 apresenta o funcionamento das redes genéticas e sobre as redes responsáveis pela manutenção do genoma. Estas redes são mecanismos de identificação e reparo de erros durante o processo de replicação celular, quando há uma falha nesta verificação, ocorre um crescimento desordenado de células que podem evoluir para um tumor cancerígeno.

A formação do câncer (processo carcinogênico) está descrita na seção 2.3, que aborda a transição do estágio inicial (pré-câncer) para o estágio em que as células estão em metástase, já com o câncer formado.

A seção 2.4 trata da barreira anticâncer, proposta por Halazonetis e colaboradores (2008), responsável por impedir a transformação tumoral.

A seção 2.5 apresenta os principais bancos de dados biológicos que forneceram os dados usados na Ontocancro e as técnicas para análise da expressão de genes. Também será apresentado o desenvolvimento da Ontocancro

1.0, em trabalhos anteriores, e sua base de dados relacional. Por fim, explica-se os motivos que levaram a atualização da Ontocancro 2.0 neste trabalho.

2.1. Fundamentos da Biologia Molecular

O intenso estudo sobre os seres vivos, nos últimos anos, resultou na descoberta de importantes componentes do seu organismo, com funções diversas existentes, porém utilizando-se de um mesmo conjunto para a maioria de suas funções básicas (ALBERTS *et al.*, 2010). Como exemplo têm-se as células, que representam a menor unidade de vida, contendo as características morfológicas e fisiológicas dos organismos vivos. De forma a entender o funcionamento de um organismo vivo, primeiramente é necessário conhecer não apenas a sua estrutura celular, mas também as interações entre os constituintes moleculares que formam as células (ZAHA; FERREIRA; PASSAGLIA, 2012).

As células são encontradas em todos os organismos vivos das mais diversas formas e espécies. Uma única célula pode formar um organismo inteiro, como no caso dos protozoários, ou juntar-se a muitas outras células do mesmo tipo para formar tecidos ou órgãos de um organismo multicelular (DE ROBERTIS, 2006).

Um dos principais estudos da Biologia Molecular compreende a estrutura das moléculas de DNA (ácido desoxirribonucleico) e RNA (ácido ribonucleico), material encontrado no núcleo e no citoplasma das células, como mostrado na Figura 3.

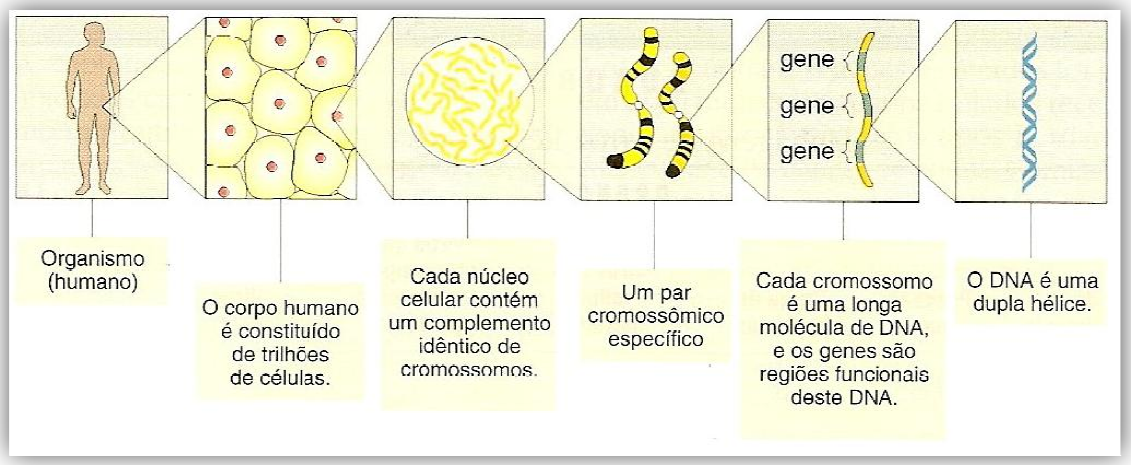


Figura 3 – Localização do material genético no organismo humano (GRIFFITHS *et al.*, 2001)

Isso porque são elas que guardam toda a informação genética responsável pela transmissão da hereditariedade das espécies que ocorre através de processos e fases das moléculas de DNA e RNA. O DNA produz uma cópia de si mesmo em um processo denominado replicação, onde a informação genética é conservada e passada para os descendentes (GIBAS; JAMBECK, 2001). O DNA também serve como molde para sua transcrição através da síntese do RNA. E finalmente, ocorre a tradução do RNA, ou seja, a síntese da proteína. Estes processos caracterizam o Dogma Central da Biologia Molecular ilustrado na Figura 4.

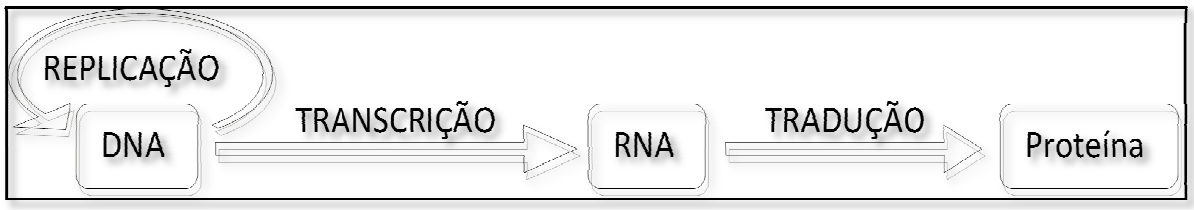


Figura 4 – Representação do Dogma Central da Biologia Molecular

Segundo De Robertis (2006), o DNA é um polímero (molécula muito grande) formado pela associação de quatro unidades químicas diferentes, chamadas nucleotídeos. O DNA é formado por duas cadeias helicoidais de ácidos nucleicos com giro à direita, que formam uma dupla hélice em torno de um mesmo eixo central. Este modelo foi proposto por James Watson e Francis Crick em 1953, com base nos dados obtidos por Maurice Wilkins e Rosalind Franklin e que contemplava as propriedades químicas e biológicas já conhecidas, assim como a capacidade de duplicação desta molécula (WATSON; CRICK, 1953).

Os ácidos nucleicos são constituídos por moléculas de açúcar (pentoses), bases nitrogenadas (purinas e pirimidinas) e ácido fosfórico. As pentoses podem ser de dois tipos: Ribose, no RNA e Desoxirribose, no DNA. A diferença entre elas está na quantidade de átomos de oxigênio, pois na ribose há um átomo a mais (DE ROBERTIS, 2006). As bases pirimidinas são de dois tipos: Timina (T) e Citosina (C), e as purinas são de três tipos: Adenina (A), Guanina (G) e a Uracila (U).

A sequência linear das bases e de seus ácidos nucleicos configura toda a informação genética dos organismos vivos. A estrutura primária das proteínas é codificada por um alfabeto de quatro letras, o que torna a interpretação deste código

genético uma das descobertas mais importantes da biologia molecular (DE ROBERTIS, 2006).

O DNA também tem a função de servir de modelo para originar a molécula de RNA. Existem três tipos de RNA como mostra o Quadro 1.

Tipo	Sigla	Função
RNA Mensageiro	mRNA	Originam da transcrição do DNA, e carregam a informação do genoma para o ribossomo.
RNA de Transferência	tRNA	São moléculas não traduzidas que carregam aminoácidos para o ribossomo. Os aminoácidos servem de base para a síntese das proteínas.
RNA Ribossômico	rRNA	São moléculas não traduzidas dos ribossomos, compostos por proteínas e RNA. Fixam as moléculas de mRNA e surgem em processos da tradução.

Quadro 1 – Tipos de RNA

A etapa final do repasse da informação genética para as células é o momento da tradução do mRNA em proteína. O mRNA é decodificado em uma cadeia de aminoácidos chamado polipeptídio, que formam as proteínas funcionais da célula.

Os ribossomos fazem a leitura do código genético se movendo ao longo da molécula de mRNA. Esta leitura é realizada a cada três nucleotídeos, grupo chamado de *códon* (GRIFFITHS *et al.*, 2001). Existem 64 códons diferentes, que formam o Código Genético utilizado por todos os organismos e representado na Figura 5.

PRIMEIRO NUCLEOTÍDEO DO CÓDON	SEGUNDO NUCLEOTÍDEO DO CÓDON: U				SEGUNDO NUCLEOTÍDEO DO CÓDON: C				SEGUNDO NUCLEOTÍDEO DO CÓDON: A				SEGUNDO NUCLEOTÍDEO DO CÓDON: G				TERCEIRO NUCLEOTÍDEO DO CÓDON
	CÓDON	ABREVIAÇÕES	AMINOÁCIDO		CÓDON	ABREVIAÇÕES	AMINOÁCIDO		CÓDON	ABREVIAÇÕES	AMINOÁCIDO		CÓDON	ABREVIAÇÕES	AMINOÁCIDO		
U	UUU	Fen	F	*Fenilalanina*	UCU	Ser	S	*Serina*	UAU	Tir	Y	*Tirosina*	UGU	Cis	C	*Cisteína*	U
	UUC	Fen	F	*Fenilalanina*	UCC	Ser	S	*Serina*	UAC	Tir	Y	*Tirosina*	UGC	Cis	C	*Cisteína*	C
	UUA	Leu	L	*Leucina*	UCA	Ser	S	*Serina*	UAA	Códon de Parada			UGA	Códon de Parada			A
	UUG	Leu	L	*Leucina*	UCG	Ser	S	*Serina*	UAG	Códon de Parada			UGG	Trp	W	*Triptofano*	G
C	CUU	Leu	L	*Leucina*	CCU	Pro	P	*Prolina*	CAU	His	H	*Histidina*	CGU	Arg	R	*Arginina*	U
	CUC	Leu	L	*Leucina*	CCC	Pro	P	*Prolina*	CAC	His	H	*Histidina*	CGC	Arg	R	*Arginina*	C
	CUA	Leu	L	*Leucina*	CCA	Pro	P	*Prolina*	CAA	Gln	Q	*Glutamina*	CGA	Arg	R	*Arginina*	A
	CUG	Leu	L	*Leucina*	CCG	Pro	P	*Prolina*	CAG	Gln	Q	*Glutamina*	CGG	Arg	R	*Arginina*	G
A	AUU	Ile	I	*Isoleucina*	ACU	Tre	T	*Treonina*	AAU	Asn	N	*Asparagina*	AGU	Ser	S	*Serina*	U
	AUC	Ile	I	*Isoleucina*	ACC	Tre	T	*Treonina*	AAC	Asn	N	*Asparagina*	AGC	Ser	S	*Serina*	C
	AUA	Ile	I	*Isoleucina*	ACA	Tre	T	*Treonina*	AAA	Lis	K	*Lisina*	AGA	Arg	R	*Arginina*	A
	AUG	Met	M	*Metionina*	ACG	Tre	T	*Treonina*	AAG	Lis	K	*Lisina*	AGG	Arg	R	*Arginina*	G
G	GUU	Val	V	*Valina*	GCU	Ala	A	*Alanina*	GAU	Asp	D	*Ácido Aspártico*	GGU	Gli	G	*Glicina*	U
	GUC	Val	V	*Valina*	GCC	Ala	A	*Alanina*	GAC	Asp	D	*Ácido Aspártico*	GGC	Gli	G	*Glicina*	C
	GUA	Val	V	*Valina*	GCA	Ala	A	*Alanina*	GAA	Glu	E	*Ácido Glutâmico*	GGA	Gli	G	*Glicina*	A
	GUG	Val	V	*Valina*	GCG	Ala	A	*Alanina*	GAG	Glu	E	*Ácido Glutâmico*	GGG	Gli	G	*Glicina*	G

Figura 5 – Código Genético

O sistema de códigos utilizados pelas células tem base na sequência dos nucleotídeos no DNA, que influencia o sequenciamento no RNA, e sendo este um mRNA, determinará o sequenciamento dos aminoácidos na proteína (DE ROBERTIS, 2006). Existem 20 aminoácidos conhecidos, contendo características únicas, que ocorrem nas proteínas.

A próxima seção trata das interações moleculares que ocorrem através das redes metabólicas, um dos temas de investigação deste trabalho.

2.2. Funcionamento das Redes Genéticas

Os sistemas biológicos possuem um número muito grande de componentes: proteínas, genes, compostos químicos que estão organizados em redes complexas de interação. Assim sendo, qualquer espécie é o resultado de uma infinidade de acidentes históricos que forjaram suas particularidades. Hoje, existem disponíveis na Internet dados referentes a muitos desses processos biológicos.

O metabolismo de todos os organismos é caracterizado por uma rede complexa de reagentes conectados por reações químicas. As reações são organizadas em módulos chamados mapas metabólicos que realizam funções específicas. O conjunto completo destes mapas caracteriza a rede metabólica de um dado organismo.

Para a perpetuação de uma espécie, além de um mecanismo preciso para efetuar a replicação do DNA, também é necessário um controle da estabilidade genética capaz de corrigir falhas acidentais que ocorrem no DNA (ALBERTS *et al.*, 2010). Este controle é realizado pelos mecanismos de manutenção do genoma (GMM), formados por um conjunto de genes divididos em redes metabólicas.

Para a investigação do processo de ativação da barreira anticâncer, proposta por Halazonetis *et. al.* (2008) foram selecionadas três importantes vias GMM, que são: Reparo do DNA, Ciclo Celular e Apoptose (SIMÃO, 2012), tratadas respectivamente, nas seções a seguir. Estas vias dão sustentabilidade aos Mecanismos de Manutenção do Genoma (GMM), e auxiliam na barreira contra propagação do câncer, que será detalhado na seção 2.4 (HALAZONETIS; GORGOULIS; BARTEK, 2008).

2.2.1. Funcionamento das Redes de Reparo de DNA

As redes de proteção não funcionam corretamente em células cancerosas, ocasionando uma frequência extremamente elevada de mutações. Sabe-se que os genes de uma das redes de reparo, chamada de Reparo por Excisão de Nucleotídeos (NER), não possui mutações catalogadas relacionadas a câncer somático. Assim, acreditava-se que ela não estaria envolvida no aparecimento de mutações em células cancerosas (FUTREAL *et al.*, 2004). Investigou-se, então, a expressão gênica dos genes desta rede de reparo usando os dados públicos do Projeto Genoma do Câncer Humano disponíveis na Internet (CASTRO, MAURO A. A. *et al.*, 2007).

Neste banco são disponibilizados dados sobre o funcionamento de células cancerosas e normais. Utilizando a entropia da distribuição de ativação dos genes de todas as redes de reparo, redes energéticas e da rede envolvida em apoptose celular (morte celular programada) verificou-se que a rede NER, embora estruturalmente conservada (sem mutações), é a que apresenta a maior alteração funcional em relação às outras redes. Além disso, com a construção da rede de interação entre estes genes foi proposto que o mau funcionamento da rede NER era ocasionado pela disfunção da rede de apoptose que se comunica com esta via através do gene TP53. Este gene é o gene mutante mais comum em câncer, e exerce influência sobre a transcrição, o ciclo celular, a apoptose e a angiogênese.

Na Figura 6, onde o grafo de interações entre os genes da rede de reparo e da rede de apoptose pode ser visto, os genes da apoptose estão representados em azul. Os genes da rede NER estão representados em vermelho e os genes de reparo estão representados em outras cores. É possível observar que a comunicação entre a rede NER e a apoptose se dá principalmente através do gene TP53 (FUTREAL *et al.*, 2004).

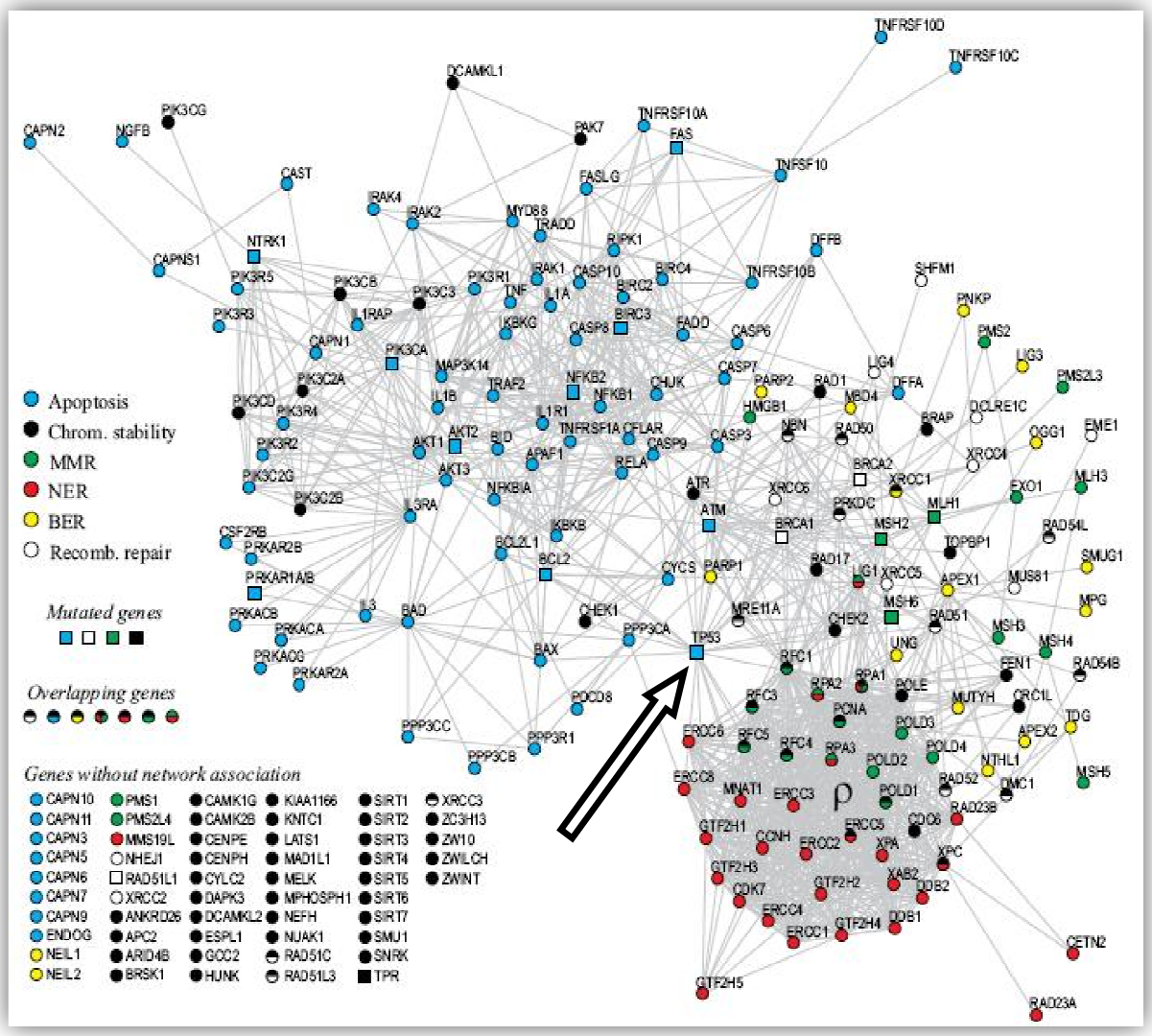


Figura 6 – Redes genéticas envolvidas na apoptose e reparo de DNA.

Em uma análise topológica adicional da conectividade e clusterização dos genes da rede NER, Mombach (2008) propôs que ela possui características que a fazem essencial para a sobrevivência das células cancerosas e por esta razão não se encontram mutações nestes genes (MOMBACH *et al.*, 2008).

Na Ontocancro 2.0, a rede de Reparo em Danos do DNA foi dividida em 9 (nove) subvias. Estas subvias atuam em diferentes tipos de danos, no entanto ocorrem durante o ciclo celular.

O Quadro 2 apresenta as subvias de reparo que estão armazenadas na Ontocancro 2.0.

Nome da subvia de reparo	Base de Dados	Total de Genes
(BER) Base Excision Repair	Ontocancro	44
DSB Repair - Double-Strand Break Repair	Reactome	22
(FA) Fanconi Anemia Pathway	Reactome	24
(HR) Homologous Recombination	Ontocancro	34
(HRR) HR Repair of Replication-Independent DSB	Reactome	16
(MMR) Mismatch Repair	Ontocancro	28
(NHEJ) Non-Homologous end Joining	Ontocancro	14
(NER) Nucleotide Excision Repair	Ontocancro	51
Processing of DNA DSB ends Recruitment of Repair and Sig. Proteins	Reactome	12

Quadro 2 – Subvias da rede de reparo em danos do DNA

Quando há quebra da fita simples de DNA, duas subvias podem ser acionadas para corrigir o dano: a BER ou a NER. Elas retiram o dano e a sequência de DNA original é restaurada pela ação da enzima DNA-polimerase, que utiliza a fita intacta como molde para preencher a região afetada com a base correta. A quebra resultante na dupla hélice é então ligada pela enzima DNA-ligase (KAO *et al.*, 2005).

E quando ocorre a quebra da fita dupla do DNA, o reparo torna-se mais complexo, uma vez que não há fita intacta para servir como molde. Neste caso, outras duas subvias podem ser acionadas: a NHEJ e a HR. A NHEJ repara o dano através da união aleatória de quaisquer extremidades do DNA (ALBERTS *et al.*, 2010), enquanto que a subvia HR obtém informações de cromossomos homólogos para reparar o DNA duplamente quebrado.

A subvia MMR é responsável pela correção de erros ocorridos durante a replicação do DNA e na recombinação de genes que resulta em um mau pareamento dos nucleotídeos. Já a subvia FA possui o gene ATR ativo que atua como sensor de danos no DNA e auxilia na ativação dos pontos de verificação do DNA (STRACKER; PETRINI, 2011).

As quebras de fita dupla (DSB) representam uma grande ameaça para a estabilidade do genoma porque a cadeia complementar não pode ser usada como um modelo fiel para a correção. As quebras podem ocorrer durante a replicação do DNA, interrompendo completamente o processo de replicação, o que pode levar a morte celular. O reparo consiste na utilização de um cromossomo homólogo para o processamento do DNA a ser reparado. Outras subvias também são responsáveis pelo reparo em DSB, tais como a subvia *Homologous Recombination Repair of*

Replication-Independent DSB e Processing of DNA DSB ends Recruitment of Repair and Signaling Proteins.

2.2.2. Redes do Ciclo Celular

Para que ocorra a multiplicação celular, uma série de processos bem definidos e regulados devem ser executados. A célula duplica seu material genético e o divide de forma igual às células-filhas garantindo assim, a renovação celular e o crescimento normal de um organismo. O Ciclo Celular define os estágios em que esses processos ocorrem (COOPER; HAUSMAN, 2007).

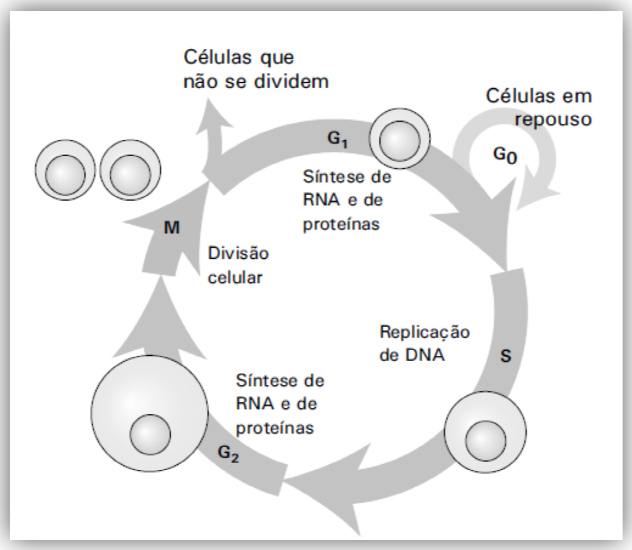


Figura 7 - Fases do Ciclo Celular (LODISH, 2005)

Durante o ciclo celular, surgem mecanismos que controlam as interações moleculares tentando identificar alterações no DNA, conhecido como *checkpoint* ou pontos de verificação. Este mecanismo impede, por exemplo, o início de eventos sem que um evento anterior esteja concluído com sucesso. Se identificada qualquer alteração no genoma celular, este mecanismo interrompe a progressão do ciclo até que seja feito o reparo ou, se o dano for excessivo, até que a célula entre em apoptose (morte celular programada). Alterações do funcionamento de genes controladores do ciclo celular, em decorrência de mutações, são relacionados ao surgimento de um câncer (COOPER; HAUSMAN, 2007).

Segundo Halazonetis e colaboradores (2008), os tecidos propensos a formação de câncer, em geral, apresentam uma diminuição da expressão de genes em vias relacionadas ao índice de proliferação celular (ciclo celular), seguido por um aumento da atividade da via de apoptose e reparo de danos no DNA. Com o aumento da atividade da via de apoptose e reparo do DNA, é definida a barreira de progressão ao câncer. Uma vez que esta barreira é quebrada, há o predomínio do fenótipo mutante, que resulta no acúmulo de mutações em células geneticamente alteradas.

Na Ontocancro 2.0 a via do Ciclo Celular foi dividida em 15 (quinze) subvias, conforme apresenta o Quadro 3.

Nome da subvia do ciclo celular	Base de Dados	Total de Genes
Cell Cycle Checkpoint	Reactome	117
Cell Cycle, Mitotic	Reactome	315
Cell Cycle: G1/S Check Point	BioCarta	21
Cyclins and Cell Cycle Regulation	BioCarta	15
G1/S DNA Damage Checkpoints	Reactome	60
G2/M Checkpoints	Reactome	43
Mitotic G2-G2/M Phases	Reactome	86
Mitotic M-M/G1 Phases	Reactome	178
Mitotic Spindle Checkpoint	Reactome	19
Nemo/NFKB Pathway	(SABATEL *et al.*, 2011)	20
Rb Tumor suppressor/Check. P. Sign. in Response to Damage	BioCarta	13
RB-E2F Pathway	(CALZONE *et al.*, 2008)	169
Regulation of DNA Replication	Reactome	75
Regulation of Mitotic Cell Cycle	Reactome	82
S Phase	Reactome	112

Quadro 3 – Subvias do ciclo celular

As subvias *Nemo/NFKB Pathway* e *RB-E2F Pathway* foram adicionadas durante a realização desta dissertação, com base em artigos publicados, respectivamente, por Sabatel e colaboradores (2011) e Calzone e colaboradores (2008). Todas estas subvias possuem funções específicas dentro do Ciclo Celular.

A subvia *Nemo/NFKB Pathway* apresenta-se alterada tanto em tecidos de adenoma quanto em tecidos já afetados com câncer. Ela tem função de controlar a apoptose e a senescência e está associada com a rede de reparo ao dano além de coordenar a parada do ciclo celular pela ativação dos genes TP53 e NFKB (SABATEL et al., 2011).

Já a subvia *RB-E2F Pathway* regula o ciclo celular na transição da fase G1 para S, também regula a replicação do DNA e está relacionada a outras funções da via da Apoptose (CALZONE *et al.*, 2008).

2.2.3. Redes da Apoptose

A apoptose, ou morte celular programada, tem um papel fundamental, tanto na manutenção de tecidos adultos, com o balanceamento da proliferação celular, como no desenvolvimento embrionário (COOPER; HAUSMAN, 2007). Pode ser acionada como resposta a estímulos internos ou externos à célula, para eliminar células supérfluas ou defeituosas.

Entre as características das células cancerosas está a proliferação sem controle de novas células e de uma crescente incapacidade de morrer por apoptose. Portanto, apesar da enorme variabilidade do câncer, evidências demonstram que a resistência à apoptose é uma das características mais marcantes da maioria dos tumores malignos. A apoptose ocorre nas situações em que o reparo em dano no DNA é irreversível (PEREIRA, 2011).

Na Ontocancro 2.0 a via da apoptose está dividida em 14 (quatorze) subvias, conforme o Quadro 4.

Nome da subvia da Apoptose	Base de Dados	Total de Genes
Apoptosis - Homo sapiens (human)	Ontocancro	83
Apoptotic Execution Phase	Reactome	51
Apoptotic signaling in response to DNA damage	BioCarta	12
Caspase Cascade in Apoptosis	Ontocancro	52
Death Receptor Signalling	Reactome	13
Extrinsic Pathway for Apoptosis	Reactome	13
Granzyme a Mediated Apoptosis Pathway	BioCarta	13
Induction of apoptosis through dr3 and dr4/5 death receptors	BioCarta	20
Intrinsic Pathway for Apoptosis	Reactome	29
Regulation of Apoptosis	Reactome	60
Senescence	Gene Ontology	16
TNF Receptor Signaling Pathway	NCI-Nature	40
tnfr1 Signaling Pathway	BioCarta	13
tnfr2 Signaling Pathway.	BioCarta	10
Cap-dependent Translation Initiation	NCI-Nature	112
Replicative Senescence	Articles	56

Quadro 4 - Subvias da apoptose

2.3. O processo carcinogênico

O estado carcinogênico ou pré-câncer é caracterizado pelo crescimento anormal de células que poderão evoluir ao câncer (carcinomas), podendo ser reversível. Essa fase possui três estágios pré-cancerosos (adenomas), que são definidos como: adenoma primário, intermediário e tardio (LENHARD JR; OSTEEN; GANSLER, 2000).

A Figura 8 mostra as fases da progressão tumoral.

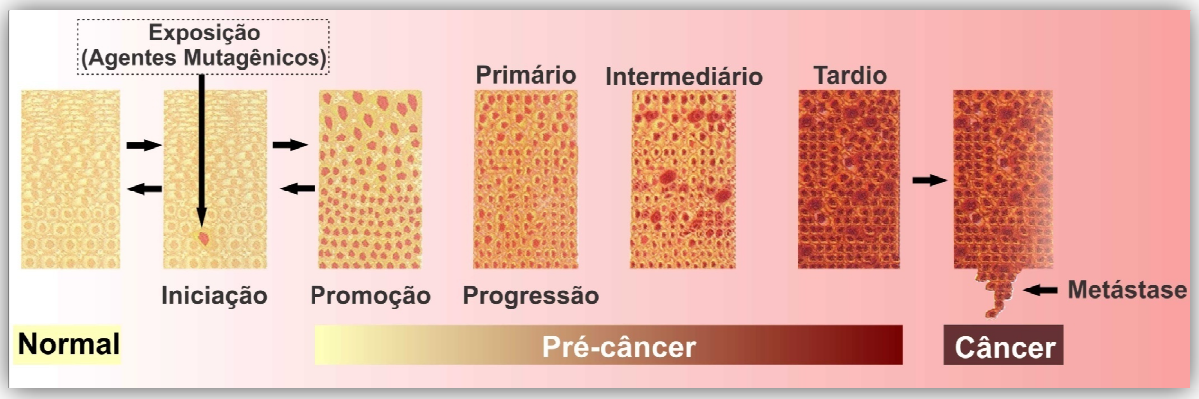


Figura 8 – Progressão tumoral. Adaptada de (WEINBERG, 2007).

Em um organismo, existem os mecanismos de manutenção do genoma (GMM) responsáveis por corrigir falhas e garantir a integridade do DNA e a sobrevivência da célula.

Os GMM envolvem um conjunto de vias e proteínas que atuam na Resposta aos Danos (DDR). Na presença de qualquer alteração que envolva instabilidade genômica são ativados os pontos de verificação do Ciclo Celular, que irão identificar os defeitos e ativar proteínas supressoras de tumor ligadas aos mecanismos de Resposta aos Danos tais como: a morte celular programada (Apoptose), o reparo e a recombinação do DNA em resposta aos danos (CASTRO, MAURO A. A. *et al.*, 2007).

2.4. A barreira anticâncer proposta por Halazonetis e colaboradores (2008)

Com base em um estudo realizado por Halazonetis e colaboradores (2008), os tecidos pré-cancerosos, em geral, apresentam uma diminuição da expressão em vias de manutenção do genoma relacionada ao índice de proliferação celular, seguida por um aumento da atividade da via de apoptose (morte celular programada) e reparo de danos no DNA (DDR), surgindo, assim, uma barreira de progressão anticâncer que resiste a transformação tumoral. Uma vez que esta barreira é rompida, origina-se o tecido maligno.

Segundo Halazonetis, a ativação de oncogenes[1] causa as alterações dos tecidos pré-cancerosos, desencadeando uma proliferação de mutações. Em resposta, o tecido inicia a DDR ativando as vias especializadas que criam a barreira anticâncer.

Na Figura 9 é mostrada a barreira anticâncer em tecidos pré-cancerosos (PreCA) do pulmão, melanoma e colorretal. O índice de proliferação celular (P.I.) está representado pela área verde.

Em pulmão e melanoma, os tecidos normais possuem um índice pouco significante; e em colorretal o índice PI diminui durante a ação da barreira. Porém, em ambos os casos, o índice PI aumenta após a quebra da barreira.

A barreira é formada pela ação das redes de reparo em dano do DNA, representada pela linha laranja (índice DDR.I.), e pela ação das redes de apoptose e senescência (índice A./S.I.) representada pela linha lilás. Tais índices identificam um aumento da morte celular programada e da DDR.

---

[1] Oncogenes denominam os genes relacionados com o surgimento de tumores, sejam benignos ou malignos, assim como genes que quando deixam de funcionar normalmente, transformam uma célula normal em uma célula cancerosa.

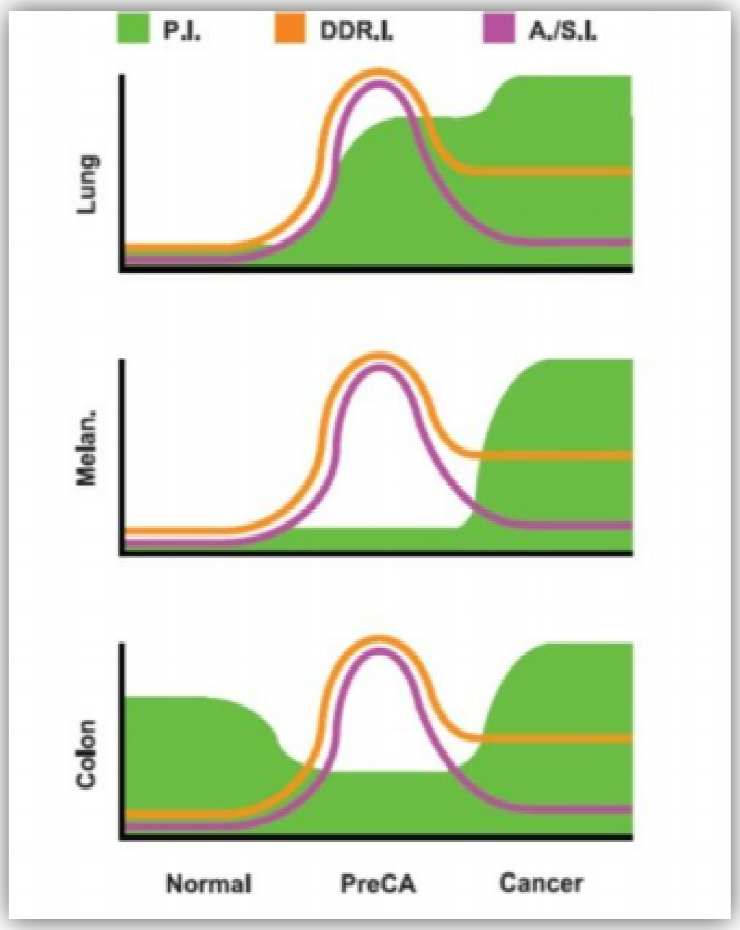


Figura 9 - Barreira de evolução do câncer em lesões pré-cancerosas (HALAZONETIS et al., 2008).

Na Figura 10 observa-se o modelo de danos ao DNA provocado por oncogenes na progressão e no desenvolvimento do câncer. O modelo pode ajudar a explicar muitas características do câncer através da ativação de oncogenes que levam os tecidos sadios a uma proliferação aberrante.

A instabilidade genômica é um resultado direto dos oncogenes que são ativados por estresse na replicação ou por DSB no inicio do desenvolvimento do câncer, a instabilidade genômica é seguida pela instabilidade cromossômica (CIN) em câncer provocando uma adição de genes mutantes o que facilita o desenvolvimento do câncer (HALAZONETIS; GORGOULIS; BARTEK, 2008).

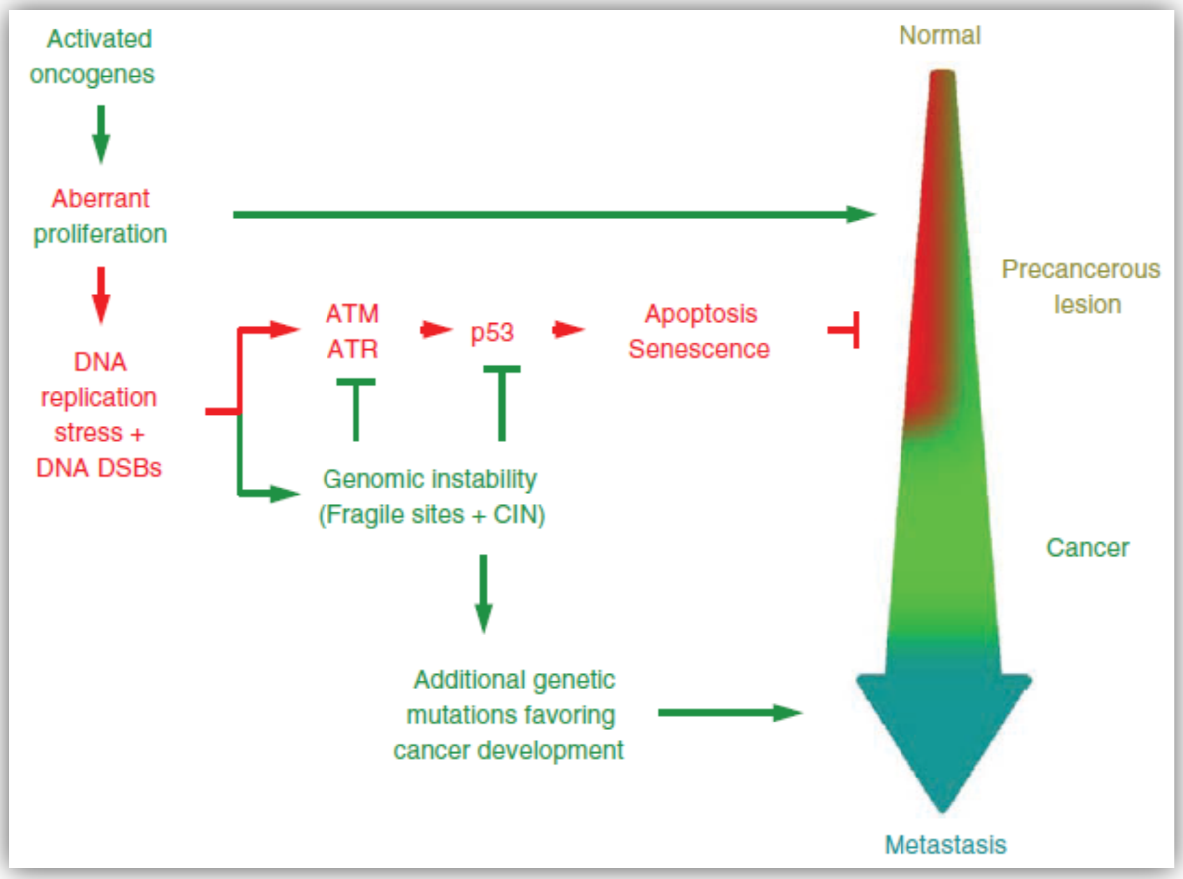


Figura 10 – Modelo de danos do DNA provocado por oncogenes. (HALAZONETIS; GORGOULIS; BARTEK, 2008).

Segundo o modelo, as principais proteínas que irão ativar a resposta aos danos serão a ATM e ATR que irão acionar a p53.

Em pré-câncer, a p53 ativa a apoptose e a senescência para conter os defeitos inerentes e com isso barrar um possível estresse na replicação que poderá levar a célula a uma proliferação aberrante.

Na Figura 10 é possível verificar ainda que a instabilidade genômica é seguida pela instabilidade cromossômica (CIN) em câncer provocando uma adição de genes mutantes o que facilita o desenvolvimento do câncer e a formação de metástase (HALAZONETIS; GORGOULIS; BARTEK, 2008).

2.5. Bancos de dados biológicos, técnicas de análise da expressão gênica e a Ontocancro 1.0

O uso da informática nas pesquisas da área biológica tornou-se relevante a partir de 1977 com a descoberta de um mecanismo de sequenciamento dos pares de bases do DNA pelo cientista inglês Frederick Sanger. Desde então, surgiram áreas como a Biotecnologia, com o desenvolvimento de novos instrumentos capazes de facilitar o sequenciamento das bases, até então, realizado manualmente pelos profissionais da área (GIBAS; JAMBECK, 2001).

A Bioinformática surge como uma ciência por volta de 1995, sendo capaz de utilizar a capacidade de processamento de grandes informações dos computadores no campo das ciências biológicas, tendo ainda seu alicerce nas áreas da física, química e matemática (SETUBAL, 2003).

O gerenciamento da grande demanda de informação produzida nas pesquisas exige a utilização de bancos de dados para o seu armazenamento. A maior dificuldade dos cientistas é a padronização de termos o que causa hoje uma enorme dificuldade em integrar diferentes bancos de dados e informações. Essa dificuldade foi solucionada com a criação de alguns bancos de dados públicos usados para padronizar as descobertas provenientes do Projeto Genoma Humano (WATSON, 1990).

Esta seção apresenta os principais bancos de dados públicos utilizados na obtenção dos dados que compõem a base de dados da Ontocancro (seção 2.5.1), assim como uma explicação sobre o funcionamento das técnicas de análise da expressão gênica (Microarranjos e RNAseq) de onde foram retiradas as informações que alimentam a base de dados da Ontocancro (seção 2.5.2) e a arquitetura original da ontologia denominada Ontocancro 1.0 (seção 2.5.3).

2.5.1. Bancos de Dados Biológicos

Os sistemas biológicos possuem um número grande de componentes (proteínas, genes, compostos químicos) que estão organizados em redes complexas de interação. Atualmente, muitos dados referentes a estes processos encontram-se

facilmente na internet. Através destes processos que foram catalogados, surge uma infinidade de bancos espalhados pela internet, cada um com sua particularidade e alguns interagindo entre si.

2.5.1.1. HUGO Gene Nomenclature Committee - HGNC

O HGNC é um dos principais bancos de dados de nomenclatura de genes e surgiu da necessidade de padronizar os nomes e símbolos de genes, entre outras informações. Essa necessidade surgiu a partir da literatura de diferentes autores que utilizavam uma nomenclatura própria e muitas vezes errônea em trabalhos científicos, assim como também havia nomes diferentes para os mesmos genes, ou genes diferentes com a mesma nomenclatura (SPLENDORE, 2005).

Pensando neste problema e no crescimento dos dados obtidos nas pesquisas, surge assim o banco de dados HGNC[2], gerenciado por uma entidade internacional que possui a responsabilidade de coordenar essas nomenclaturas, aprovando um nome e um símbolo para cada gene catalogado.

Com a utilização deste banco como base central da Ontocancro, foram utilizados do HGNC os seguintes dados: ApprovedSymbol, ApprovedName, PreviousSymbols, Aliases, Chromosome, Status, RefSeqIDs, EntrezGeneIDbyNCBI, Uniprot_ID e Ensembl_ID.

Uma pesquisa realizada no banco HGNC sobre um dos genes conhecidos como promotor do processo carcinogênico, o TP53 (*tumor protein 53*) resulta em dados como a sua localização no cromossomo e diversos links a outros bancos de dados públicos relacionados a esse gene, assim como publicações científicas envolvidas.

2.5.1.2. ArrayExpress - EBI

O *ArrayExpress* é um dos principais bancos de dados sobre expressão gênica. Gerenciado pelo *European Bioinformatics Institute* (EBI), possui resultados

[2] Disponível em www.genenames.org

de experimentos obtidos com técnicas de microarranjos e sequenciamento de DNA, que podem ser acessados através de uma simples consulta no site sobre o tema de interesse.

O *ArrayExpress*[3] está associado ao banco de dados americano GEO (*Gene Expression Omnibus*) da NCBI (*National Center for Biotechnology Information*), que também possui dados de experimentos de técnicas de microarranjos e sequenciamento de DNA, entre outras técnicas de medição.

2.5.1.3. *Gene Expression Omnibus* - GEO

O GEO[4] armazena seus dados com base em três fontes principais: as plataformas (*platforms*), as amostras (*samples*) e as séries (*series*). Os *DataSets* representam o conjunto de estudos curados através das ferramentas de análise, enquanto os *Profiles* demonstram as medições de expressão. Ele ainda disponibiliza aos pesquisadores uma interface *web* bastante amigável e de fácil exploração.

A pesquisa sobre informações contidas nesta base de dados pode ocorrer de duas maneiras, através de termos específicos que são as *queries* ou por uma simples procura dentro dos *DataSets* ou GEO *accessions*. No primeiro campo da *query* é realizada uma pesquisa por termos que localizem determinada série do GEO, pode ser o nome de algum autor específico ou até mesmo palavras-chaves que delimitem um estudo, no segundo campo pode-se entrar com termos que identifiquem a visualização individual de uma expressão genética ou um perfil de interesse, como o nome de um gene e no terceiro campo tem-se uma busca mais otimizada que se dá através da identificação do estudo, que pode ser feito pelo número da série (i.e: GSExxx, onde x é um número que identifica a série), número da plataforma (i.e: GPLxxx, onde x é um número que identifica a plataforma) ou número da amostra (i.e: GSMxxx, onde x é um número que identifica a amostra).

A outra forma de pesquisa é pela navegação propriamente dita através do *Browser* que retorna todas as informações contidas dentro da base de dados do GEO de uma maneira mais geral.

---

[3] Disponível em www.ebi.ac.uk/arrayexpress
[4] Disponível em www.ncbi.nlm.nih.gov/geo

O *ArrayExpress* e o GEO foram utilizados neste trabalho por disponibilizarem dados curados (conferidos) de experimentos para as análises.

.

2.5.2. Técnica de Microarranjos da Affymetrix

Os níveis de expressão dos genes e os genes expressos diferenciam dois grupos celulares morfologicamente distintos. Estes níveis de expressão podem ser medidos através de técnicas experimentais, sendo microarranjo uma delas.

Os microarranjos permitem verificar, de forma simultânea e rápida os níveis de expressão de milhares de genes (transcritos), organizados por sondas, fixadas a uma superfície sólida (*chip*). Devido à incompatibilidade envolvida no tamanho dos nucleotídeos que um gene representa, o número de sondas varia de acordo com o gene.

O protocolo da Affymetrix analisa os níveis de expressão de maneira que são extraídas amostras de mRNA de uma amostra celular, e convertida em cDNA (DNA complementar) em dupla fita para que a amplificação linear da amostra ocorra. Nesta etapa, será criado um modelo de DNA a ser utilizado na síntese de cRNA (RNA complementar). Então este cRNA é purificado e fragmentado para facilitar a hibridização com os oligonucleotídeos que estão presos nas sondas no chip.

O cRNA se ligará com o microarranjo com a ajuda da solução de hibridização e de um marcador fluorescente. Um processo de lavagem irá eliminar os cRNA que não se hibridizaram ao microarranjo. Utilizando um *scanner* a laser os oligonucleotídeos ligados ao cRNA são excitados e seus sinais são captados e convertidos em dados. Estes dados são transferidos para um sistema de arquivos (BECKER; FEIJÓ, 2003).

A Figura 11 apresenta uma visão simplificada do protocolo *Affymetrix.*

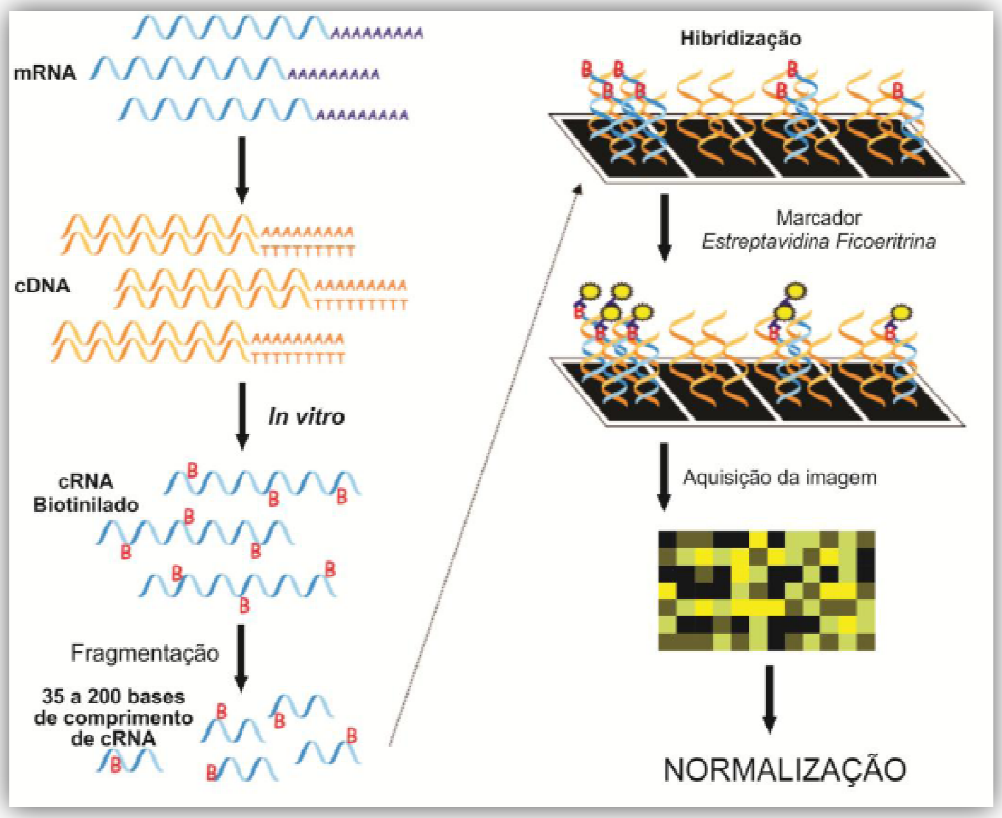


Figura 11 – Protocolo *Affymetrix*. Adaptada de (BECKER; FEIJÓ, 2003)

Os dados gerados durante os processos de hibridização e detecção dos níveis de expressão das amostras de mRNA contêm muitas impurezas e precisam ser normalizados (GOHLMANN; TALLOEN, 2009), ou seja, precisam ser conferidos para uma adequação das informações.

2.5.3. Técnica de Sequenciamento (RNAseq) da Illumina

Considerado como um método revolucionário, o RNAseq, representa o transcriptoma revelado pelo sequenciamento de DNA complementar (cDNA), – o transcriptoma é o conjunto completo de RNA de um organismo (Transcrição –> RNA), grandes variações entre as células dos organismos podem ser percebidas através do estudo do transcriptoma, ou seja, as células de diferentes partes do corpo possuem a mesma cópia de DNA, o que as difere é apenas seus respectivos transcritos, então, o transcriptoma é a análise de dados de expressão gênica – a

técnica de sequenciamento RNAseq possui uma alta sensibilidade e uma ausência quase total de ruídos, uma das suas grandes vantagens (MARIONI *et al.*, 2008).

O processo típico de sequenciamento por RNAseq pode ser observado na Figura 12.

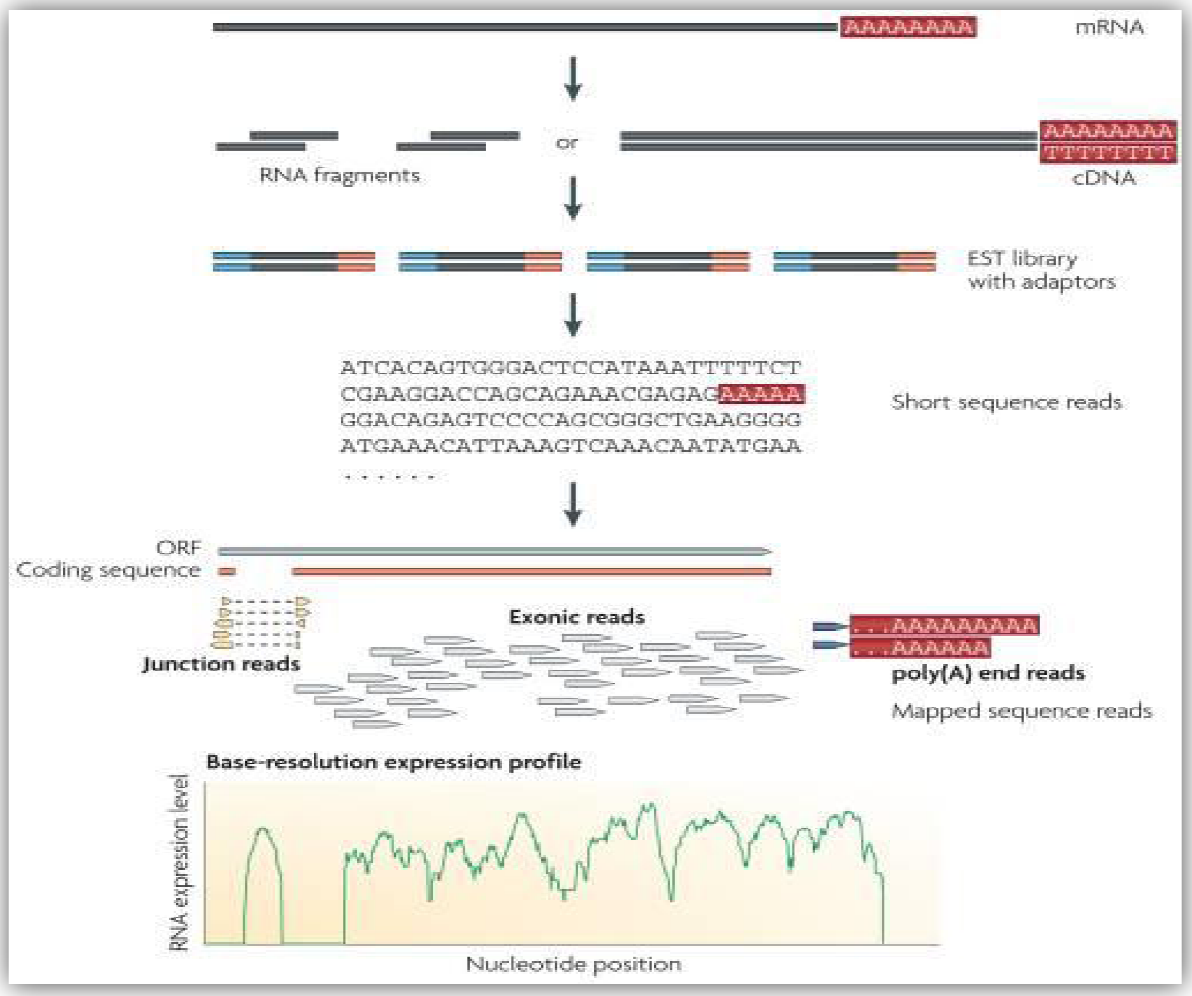


Figura 12 – Técnica de sequenciamento RNAseq (WANG; GERSTEIN; SNYDER, 2009)

Nesta técnica, certa quantidade de RNA é convertida em uma biblioteca de cDNA fragmentada. Cada fragmento que será sequenciado irá receber adaptadores, ou em uma, ou em ambas as extremidades. São geradas sequências curtas, da ordem de 30 a 400 pares de bases. Essas sequências são alinhadas a outro transcriptoma ou a um genoma de referência, ou ainda remontadas sem um genoma de referência, a fim de criar um mapa em escala genômica, sendo este composto

pelo nível de expressão de cada gene individualmente ou pela estrutura transcricional (WANG; GERSTEIN; SNYDER, 2009).

2.5.4. Construção da ontologia Ontocancro 1.0 e sua base de dados

A falta de um vocabulário unificado para genes, subvias e vias de manutenção do genoma e estabilidade genômica foi um dos motivos para o desenvolvimento da ontologia Ontocancro (LIBRELOTTO *et al.*, 2009).

Por ontologia, entende-se uma especificação explícita de uma conceitualização, hierarquizada através de termos relacionados entre si, que descrevem um determinado conhecimento (GRUBER, 1993).

A ontologia Ontocancro propõe-se, portanto, a auxiliar na investigação do funcionamento de redes biológicas de genes envolvidos em câncer. Permitindo a representação do conhecimento de redes moleculares e sua atividade (expressão).

O grafo da ontologia pode ser visualizado na Figura 13. Os dois principais elementos da ontologia Ontocancro 1.0 são as entidades *pathways* e *genes.*

A entidade *pathways* representa as redes moleculares (ou vias metabólicas) deste estudo. A entidade *genes* representa os genes mapeados na Ontocancro, que compõem cada *pathway*, sendo todos de seres humanos, portanto essas relações referem-se ao organismo de *Homo sapiens*.

As interações existentes entre os *pathways* da Ontocancro são representadas na entidade *Interaction*. Estas três entidades são instâncias da classe *Entity* (PEREIRA, 2011).

As entidades *Provenance*, *Evidence* e *Xref* definem metadados para cada uma das demais entidades, necessários para a definição da relevância em uma interação entre dois ou mais *pathways* (PEREIRA, 2011).

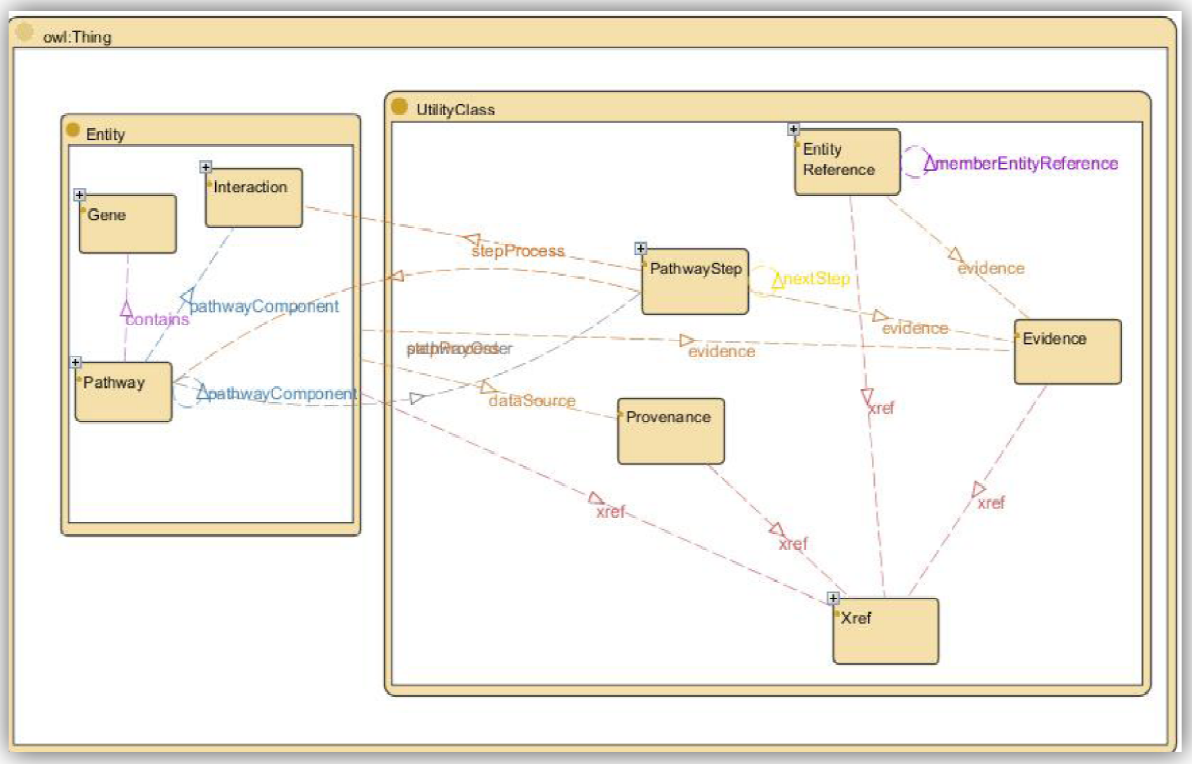


Figura 13 – Grafo da ontologia Ontocancro 1.0

Após a definição da ontologia foram selecionadas as fontes dos dados de onde seriam obtidas as informações sobre os *pathways* e os *genes*. Estas fontes são os bancos de dados públicos, mencionados na seção 3.1, que fornecem repositórios XML, documentos de texto, bancos de dados relacionais e páginas *Web*. Para cada fonte foram selecionados os campos que seriam relevantes para a Ontocancro. Como na maioria dos casos, os dados estão dispostos em formato XML, para a criação da base de dados da Ontocancro optou-se pelo armazenamento e manipulação dos mesmos em seu formato nativo, utilizando o sistema de gerenciamento de banco de dados XML eXist (MEIER, 2000).

Para obter a base de dados que gerencia a ontologia, três etapas foram realizadas: a primeira etapa refere-se à obtenção dos dados, através de convênios firmados entre os mantenedores dos bancos públicos e dos membros do grupo de pesquisa, possibilitando o acesso ao seu conteúdo. A segunda etapa é a de normalização e integração dos dados, nesta etapa *parsers* manipulam os dados que foram obtidos e transformam para o formato de ontologia para serem armazenados neste novo repositório local. E a terceira etapa, refere-se a conferência dos dados,

que trata-se do processo de curagem para corrigir eventuais erros. Nesta última etapa, a presença de um especialista da área de biologia molecular se faz imprescindível, para comparar os dados das diferentes bases com a literatura técnica. Ao final destas etapas, obteve-se um banco de dados unificado contendo informações que permitam a integração de redes de interação molecular do câncer com dados de expressão de genes envolvidos nesta doença (PEREIRA, 2011).

O banco de dados que gerencia a Ontocancro busca organizar e integrar as informações de interatomas e transcriptomas disponíveis a partir dos bancos de dados já citados. Entre estas informações buscadas destaca-se a descrição do gene extraída do banco de dados *String*, o símbolo aprovado pelo HGNC e o EntrezGene, que é um identificador numérico do gene (SIMÃO *et al.*, 2010). A arquitetura da Ontocancro está representada na Figura 14.

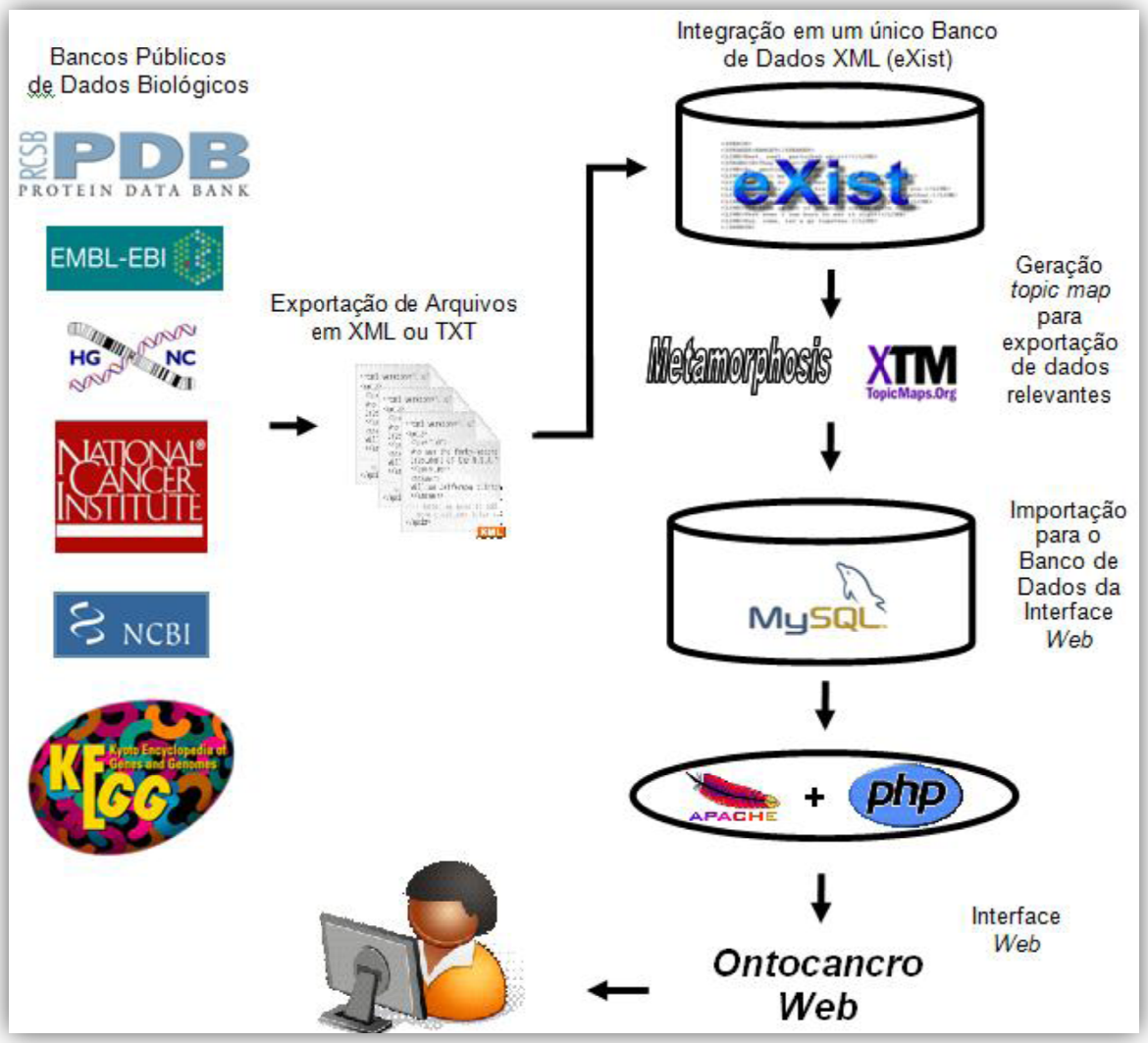


Figura 14 - Arquitetura da Ontocancro 1.0 (PEREIRA, 2011).

A partir do conhecimento representado na ontologia em questão, tornou-se possível efetuar testes experimentais do comportamento celular, permitindo o entendimento de como as redes celulares complexas são afetadas pelas mutações em genes envolvidos no processo canceroso.

Um banco de dados relacional foi criado paralelamente ao banco eXist para ser utilizado na interface *Web* que foi desenvolvida para servir de acesso às informações da ontologia pelo grupo de pesquisadores da Ontocancro. Este banco relacional, da Ontocancro 1.0 é composto por quatro tabelas descritas a seguir.

A tabela *Genes* contém os dados extraídos do banco de dados HGNC relativos aos genes pertencentes às vias metabólicas selecionadas para esta pesquisa. Sua estrutura pode ser visualizada na Tabela 1.

Tabela 1 – Tabela *Genes* do Banco de Dados Relacional.

Tabela Genes	Composta pelos genes da Ontocancro. Nesta tabela são encontradas as informações selecionadas referentes a estes genes.	
Atributos	**Tipo**	**Descrição**
ID	Varchar(30)	Serve para identificar os genes.
ApprovedSymbol	Varchar(30)	Símbolo aprovado pelo HGNC.
EntrezGene	Varchar(30)	Identificador do EntrezGene
HGNC	Varchar(30)	Identificador no banco de dados HGNC
UniGeneID	Varchar(30)	Identificador no banco de dados UniGene
HugoName	Varchar(255)	Nome do gene aprovado pelo HGNC
UniGeneName	Varchar(255)	Nome do UniGene
GeneOntologyID	Varchar(255)	Identificador no banco de dados GeneOntology
ENSG	Varchar(255)	Identificador do Gene na base Ensembl
ENSP	Varchar(255)	Identificador da Proteína na base Ensembl
KeggID	Varchar(255)	Identificador no banco de dados Kegg
ApprovedSymbol2	Varchar(255)	ApprovedSymbol conferido
PreviousSymbols	Varchar(255)	Símbolos previamente aprovados para o gene
NCIName	Varchar(255)	Nome do gene no NCI
Evidence	Varchar(255)	Códigos de evidência da base GeneOntology
Chromosome	Varchar(255)	Localização do gene no cromossomo.
Status	Varchar(255)	Situação da pesquisa e definição de nome do gene
ReactomeID	Varchar(255)	Identificador do Reactome
Aliases	Varchar(255)	Outros nomes do gene (sinônimo)
StringSymbol	Varchar(255)	Identificador no banco de dados String
StringName	Text	Nome no banco de dados String
RefSeq_IDs	Varchar(255)	Referência de Sequencia da base NCBI

A tabela *Pathways* contém os dados das vias metabólicas selecionadas para esta pesquisa. Sua estrutura pode ser visualizada na Tabela 2.

Tabela 2 – Tabela *Pathways* do Banco de Dados Relacional.

Tabela Pathways	Corresponde às vias presentes na Ontocancro.	
Atributos	**Tipo**	**Descrição**
Id_pathway	Int(2)	Identificador das vias na base Ontocancro
Pathway_name	Varchar(71)	Nome da via metabólica
Pathway_type	Varchar(19)	Tipo da via: Cell Cycle, DNA Damage Response ou Apoptosis
db_name	Varchar(50)	Fonte de onde a via foi obtida.
Source	Varchar(229)	Link de referência da fonte.
References	Varchar(255)	2º Link de referência da fonte.
Pathway_apelido	Varchar(20)	Nome abreviado da via na base Ontocancro

A tabela *Affymetrics* contém o identificador da sonda obtida da técnica de microarranjos, explicada na seção 3.2, e faz a relação com a tabela genes, desta forma é possível reconhecer as sondas de um determinado gene. Sua estrutura pode ser visualizada na Tabela 3.

Tabela 3 - Tabela *Affymetrics* do Banco de Dados Relacional.

Tabela Affymetrics	Corresponde às sondas obtidas por experimentos da técnica de microarranjos da empresa Affymetrix.	
Atributos	**Tipo**	**Descrição**
Id_gene	Varchar(255)	Identificador do gene.
affymetrics	Varchar(255)	Identificador da sonda.

Além das três tabelas citadas, a base de dados relacional possui a tabela *pathways_genes* que serve para criar a relação entre a tabela *Pathways* e a tabela *Genes*. O diagrama entidade-relacionamento pode ser visto na Figura 15.

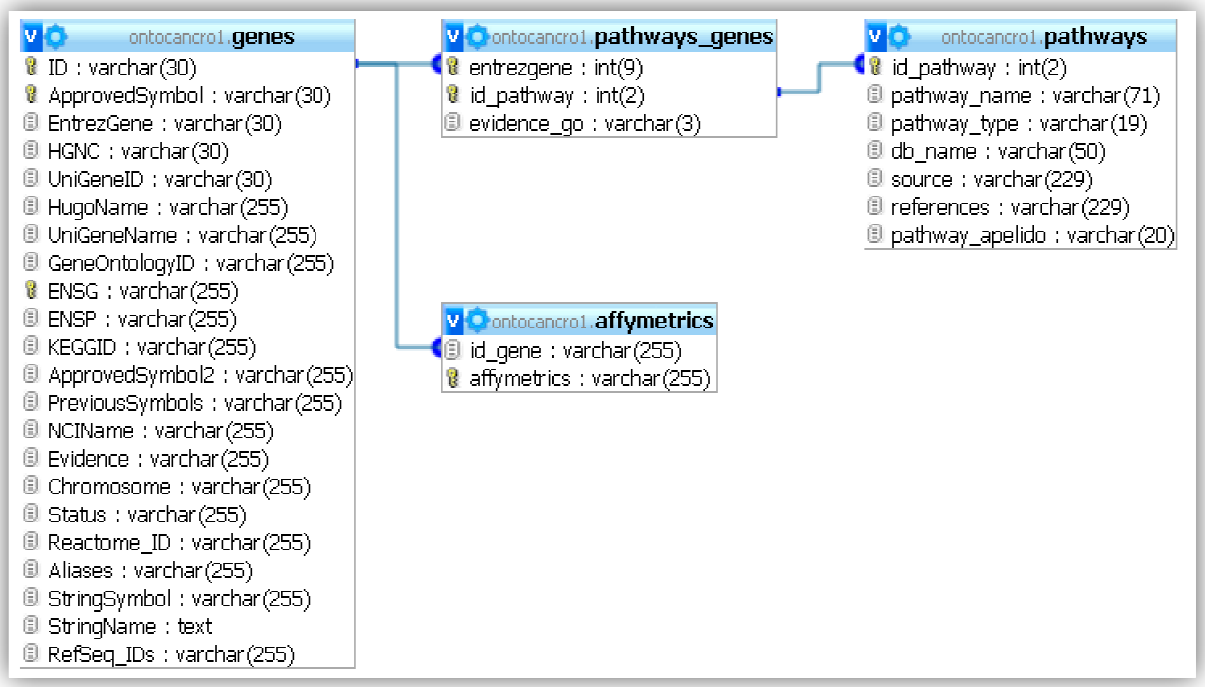


Figura 15 – Esquema da base de dados relacional da Ontocancro 1.0.

Com a estrutura relacional definida, os dados da Ontocancro 1.0 formam um conjunto de 1434 genes dispostos em 131 vias retiradas dos principais bancos de dados biológicos públicos.

No entanto, a necessidade de integração das vias metabólicas com doenças relacionadas com o câncer e a análise da expressão gênica significativa exigiu uma atualização da estrutura da ontologia e da base de dados que a gerencia. Esta nova estrutura será apresentada no Capítulo 4.

2.6. Sumário do Capítulo

No presente capítulo foram abordados os fundamentos da biologia molecular, como o dogma central, que descrevem como a informação genética é transmitida para os seus descendentes (seção 2.1). A seção 2.2 apresentou o funcionamento das redes genéticas, abordando os mecanismos de manutenção do genoma (que são os responsáveis por garantir que o DNA se replique dentro de uma célula sem ocorrer danos ao mesmo) e o funcionamento das redes de reparo de DNA, redes do ciclo celular e as redes de apoptose. A seção 2.3 descreveu como o câncer pode resultar de algum processo de replicação do DNA mal sucedido. A seção 2.4

apresentou a barreira anticâncer proposta por Halazonetis e colaboradores (2008), responsável por impedir a transformação tumoral através das redes de manutenção do genoma.

A seção 2.5 teve o objetivo de apresentar os principais bancos de dados biológicos que serviram como fonte para a Ontocancro e as técnicas utilizadas para análise da expressão de genes. Também foi apresentado o desenvolvimento da Ontocancro 1.0, as fontes de origem dos dados importados e o armazenamento de suas informações em uma base de dados relacional. Por fim, explicou-se os motivos que levaram a atualização da Ontocancro.

O capítulo 3 tratará da reestruturação da ontologia e da sua base de dados, foco desta dissertação.

3. DESENVOLVIMENTO DA ONTOCANCRO 2.0

A partir de estudos realizados sobre os dados contidos na Ontocancro 1.0, surgiu a necessidade de uma atualização em sua estrutura e inclusão de novas informações em sua base de dados. Isso porque, nas vias metabólicas contidas na Ontocancro 1.0, a grande quantidade de genes dificultava a análise da expressão dos mesmos, resultando em distorções dos dados analisados, isto devido à falta de curagem em algumas vias nas bases de dados que as mantinham.

Para resolver o problema da grande quantidade de genes presentes em uma mesma via, a estratégia encontrada foi a separação de vias em subvias de forma a agregar genes que possuam uma maior proximidade dentro de uma mesma via. Decidiu-se por incluir também na base de dados somente vias metabólicas que passaram por uma curagem (conferência através de especialista da área), na base de dados original, aumentando assim a fidelidade da análise do perfil das mesmas. Com isso, diminuiu-se bastante o número de genes a serem analisados, o que resulta numa análise estatística mais confiável e mais significativa (PEREIRA, 2011).

Para garantir que todas as vias metabólicas utilizadas na Ontocancro houvessem passado por um processo de curagem em sua fonte, quase todos os dados foram retirados da base de dados *NCI-Nature* (*National Cancer Institute*), pois seus dados são curados e nele encontram-se informações sobre vias metabólicas, interações biomoleculares e processos celulares montados de maneira mais confiável.

Na atualização da Ontocancro foram levadas em consideração vias provenientes dos bancos públicos: *NCI-Nature* (*BioCarta* e *Reactome*), Ontocancro, KEGG e *Gene Ontology*.

Outro fator importante para a atualização da Ontocancro foi a necessidade de integração de suas vias metabólicas com doenças relacionadas com o câncer, de forma a apresentar os genes relativamente expressos em processos cancerígenos.

Tal reestruturação pode ser dividida em três momentos. O primeiro momento iniciou com a inclusão dos dados oriundos da técnica de microarranjos da empresa Affymetrix. O segundo momento ocorreu com a inclusão do algoritmo para cálculo da atividade e diversidade relativa e o terceiro momento ocorreu com a inclusão das

informações da análise das expressões gênicas obtidas pela técnica de sequenciamento RNAseq da empresa Illumina.

Este capítulo descreve as alterações realizadas na estrutura da ontologia (seção 3.1), a versão atual da estrutura do banco de dados (seção 3.2), os estudos selecionados para análise da expressão gênica pelo método de microarranjos (seção 3.3), os estudos selecionados para análise da expressão gênica pelo método RNASeq (seção 3.4) e a importação das informações para a base de dados da Ontocancro (seção 3.5).

3.1. Reestruturação da nova Ontocancro

O objetivo principal da Ontocancro 1.0 era desenvolver uma ontologia capaz de unificar o vocabulário para genes e vias de manutenção do genoma e estabilidade genômica. Ou seja, sua função era organizar e integrar as informações de interatomas (vias) e transcriptomas (doenças) disponíveis a partir dos principais bancos de dados que fornecem dados relacionados à pesquisa. Inicialmente, foram inseridas informações sobre as vias metabólicas e os genes que as compõem, formando um total de 1434 genes distribuídos em 131 vias de manutenção do genoma envolvidas no ciclo celular, resposta ao dano do DNA, apoptose e senescência.

Porém, a quantidade de vias e genes não permitia uma análise da expressão gênica mais precisa, e para resolver este problema, decidiu-se por não fazer a análise em uma via inteira, e sim, segmentar a via em subvias. Isso proporcionaria uma análise mais confiável.

Posteriormente, estudos realizados com a técnica de microarranjos desenvolvida pela empresa Affymetrix para análise da expressão gênica foram incluídos no contexto da pesquisa, para que fosse possível analisar os valores de expressão. Desta forma, novos conceitos foram criados para a ontologia.

Estes estudos referem-se à experimentos publicados por cientistas em bancos de dados públicos, com base em amostras de tecidos normais e tecidos afetados por diversas doenças em diferentes organismos. No banco de dados GEO é possível buscar informações relacionadas ao câncer e obter arquivos com os valores de expressão obtidos nas pesquisas. O banco GEO disponibiliza atualmente,

40.662 séries, que correspondem aos estudos realizados de diversas doenças. Para poder realizar a análise da técnica de microarranjos com o uso da Ontocancro, foram definidas 4 séries de microarranjos. Desta forma, criou-se uma nova entidade na ontologia, denominada *Series*, que define a doença estudada e o organismo afetado, assim como inclui informações sobre os autores da pesquisa e a data de publicação da mesma.

São quatro estudos de microarranjos envolvidos em câncer extraídos do banco de dados GEO que usam a plataforma Affymetrix: GSE10927 (adrenocortical), GSE27155 (tireoide), GSE19650 (pancreático) e GSE4183 (colón retal). Estes estudos serão detalhados na seção 4.3. Cada série possui um conjunto de amostras *(samples)* que representam os registros das expressões obtidas nos experimentos. O banco GEO fornece atualmente 978.663 arquivos das amostras. Assim, essas informações foram inseridas na ontologia através da entidade *Samples*.

A nova estrutura da ontologia está representada no grafo da Figura 16.

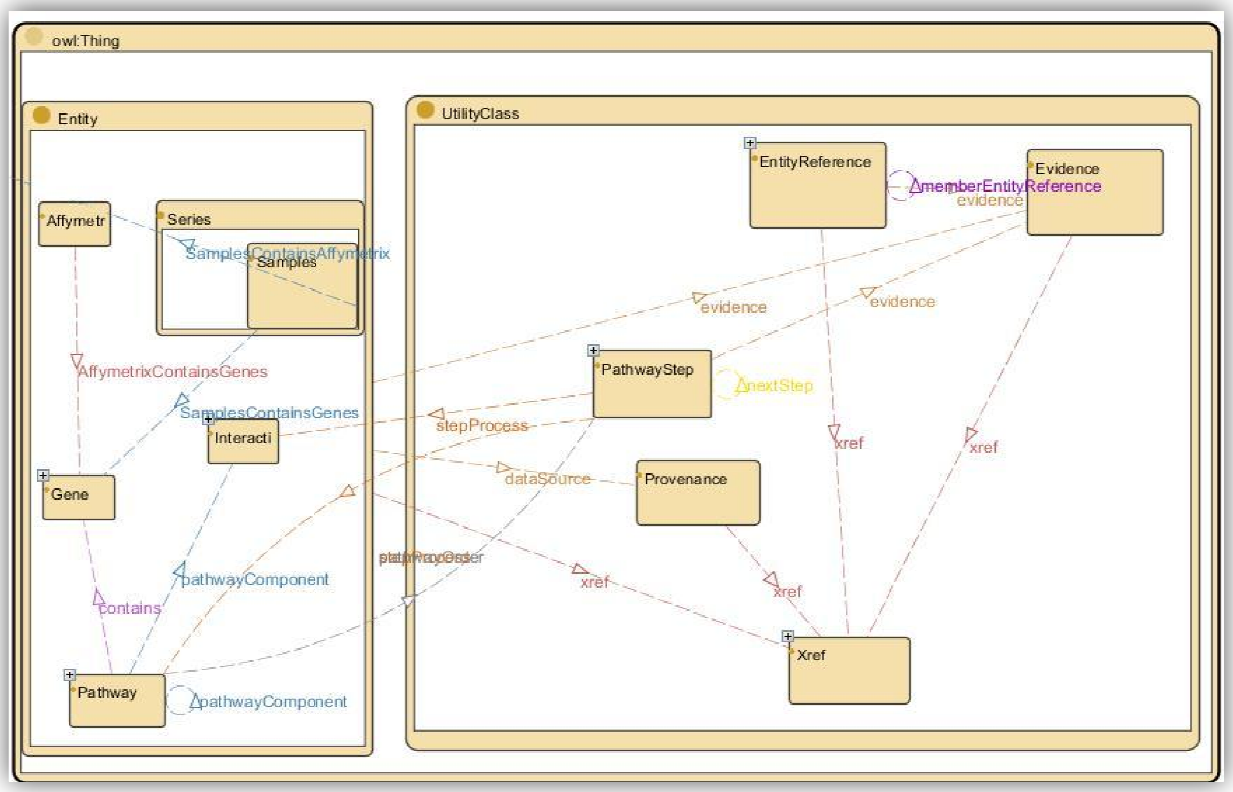


Figura 16 - Grafo da ontologia Ontocancro 2.0.

Outra entidade adicionada à ontologia Ontocancro refere-se às Plataformas, que definem os métodos utilizados para detectar e analisar os dados obtidos com a técnica de microarranjos. Desta forma, qualquer experimento possui associado a ele as especificações que definem o método em que foi realizado, em sua plataforma. Para este conceito, foi criada a entidade *Platforms*. O banco GEO possui 11925 plataformas registradas.

3.2. Reestruturação da base de dados da Ontologia Ontocancro

Após a definição dos novos conceitos necessários para análise da expressão gênica, o banco de dados relacional que gerencia as informações da ontologia foi reestruturado. Os campos utilizados em cada nova tabela foram originados diretamente dos arquivos obtidos no GEO, com a intenção de manter todas as informações, conforme a sua disponibilização.

A relação com a estrutura da Ontocancro 1.0 foi feita através do identificador *affymetrics*, da tabela *Affymetrics*, que encontra-se disponível nas amostras dos estudos. Assim, cada gene pode possuir uma ou mais sondas para um determinado estudo.

Para conter os 9.696.492 (nove milhões, seiscentos e noventa e seis mil, quatrocentos e noventa e dois) registros das expressões gênicas dos estudos selecionados, a tabela *Samples* foi adicionada no banco de dados e sua estrutura está representada na Tabela 4.

Tabela 4 – Tabela *Samples* do Banco de Dados Relacional.

Tabela *Samples*	Composta pelos dados obtidos dos estudos de microarranjos envolvidos em câncer.	
Atributos	**Tipo**	**Descrição**
Num_serie	Varchar(30)	Série que pertence a amostra. Ex. GSE19650
Id_ref	Varchar(255)	Identificador Affymetrix
Abs_call	double(50,0)	Estado de expressão
Status_Abs_Call	char(1)	Estado da amostra: Presente, Ausente Marginal.
Detection_P_Value	text	Valor de significância do gene na amostra
Num_sample	Varchar(255)	Identificação da amostra. Ex. GSM490138
Status	Varchar(100)	Descrição da amostra.
Type	Varchar(8)	Tipo do tecido: controle (saudável) ou câncer

A tabela *Series* também foi adicionada para manter as informações do estudo a que os valores pertencem, de cada amostra (*samples*), sua estrutura está representada na Tabela 5

Tabela 5 – Tabela *Series* do Banco de Dados Relacional

Tabela Series	Composta pelas informações sobre os estudos de microarranjos selecionados	
Atributos	**Tipo**	**Descrição**
Id_Serie	Varchar(255)	Identificador da Serie.
Title	double(50,0)	Nome do estudo.
Status	char(1)	Status e data da publicação
Organism	text	Organismo que se refere o estudo
Pmid	Varchar(255)	Identificação do artigo no banco de dados PubMed.

E para conter as informações sobre a plataforma em que o estudo foi trabalhado, a tabela *Platforms* foi adicionada. A estrutura da tabela *Platforms* está representada na Tabela 6.

Tabela 6 - Tabela *Platforms* do Banco de Dados Relacional

Tabela Platforms	Composta pelas informações sobre as plataformas dos estudos	
Atributos	**Tipo**	Descrição
Platform	Varchar(100)	Identificador da Plataforma
ID	Varchar(1000)	Identificador Affymetrix
GB_ACC	Varchar(1000)	Identificador GenBank
SPOT_ID	Varchar(1000)	Identificador alternativo
SpeciesScientificName	Varchar(1000)	Organismo que foi pesquisado
AnnotationDate	Varchar(1000)	Data da realização do estudo
SequenceType	Varchar(1000)	Tipo de Sequência
SequenceSource	Varchar(1000)	Fonte da Sequência
TargetDescription	Mediumtext	Anotações GenBank
RepresentativePublicID	Varchar(1000)	Número de acesso da Sequência
GeneTitle	Mediumtext	Nome do gene
GeneSymbol	Varchar(1000)	Símbolo aprovado do gene
ENTREZ_GENE_ID	Varchar(1000)	Identificador Entrezgene
RefSeqTranscriptID	Mediumtext	Conjunto de identificadores RefSeq
GeneOntologyBiologicalProcess	Longtext	Anotações sobre Processo Biológico da base do GeneOntology
GeneOntologyCellularComponent	Longtext	Anotações sobre Componente Celular da base do GeneOntology
GeneOntologyMolecularFunction	Lontext	Anotações sobre Função Molecular da base do GeneOntology

Conforme mencionado no início deste capítulo, a reestruturação da Ontocancro ocorreu em três momentos principais, sendo o primeiro momento marcado pela inclusão dos dados oriundos da técnica de microarranjos. O segundo momento ocorreu com a inclusão do algoritmo para cálculo da atividade e diversidade relativa desenvolvido em Pereira (2011) para traçar o perfil das vias de estabilidade genômica e, assim, caracterizar quantitativamente os resultados obtidos pelas pesquisas de Halazonetis (2008). O terceiro momento ocorreu com a inclusão das informações da análise das expressões gênicas obtidas pela técnica de sequenciamento RNAseq.

Para a inclusão dos dados obtidos pela técnica de sequenciamento na Ontocancro 2.0 foi necessário importar as sequencias *Probe* (fragmentos de DNA ou RNA) de RNAseq obtidas através dos arquivos gerados e disponibilizados pelo banco de dados GEO. A partir dos dados contidos desses arquivos, criou-se a tabela *ilmn_tags*, que contém as sequências *Probe*, o identificador ENSG e índices relacionados à localização da sequência. A estrutura está representada na Tabela 7**.**

Tabela 7 – Tabela *ilmn*_tags do Banco de Dados Relacional.

Tabela ilmn_tags	Composta pelos dados obtidos dos arquivos de RNAseq do GEO.	
Atributos	**Tipo**	**Descrição**
ID_ENSG	Varchar(255)	Identificador correspondente a tabela genes. Ex. ENSG00000000003
ENSG	Varchar(255)	Identificador ENSG contido nos arquivos GEO. Ex. ENSG00000000003_at
Probe_X	Varchar(255)	Localização X do Probe contido nos arquivos GEO. Ex 302
Probe_Y	Varchar(255)	Localização Y do Probe contido nos arquivos GEO. Ex 585
Illum_tg_GAllx	Varchar(255)	Sequencia Probe contida nos arquivos GEO. Ex TCTTAGTTTTCGTTTGTGCCTTTGA

O identificador *ENSG* foi usado como chave estrangeira para fazer a relação entre a tabela *ilmn_tags* com a tabela *genes* da Ontocancro 1.0. Esta relação é do tipo um-para-muitos e por isso, foi criada a tabela *ensg_ilnm*, com sua estrutura representada na Tabela 8.

Tabela 8 – Tabela *ensg_ilmn* do Banco de Dados Relacional.

Tabela *ensg_ilmn*	Composta pelos dados obtidos dos arquivos de RNAseq do GEO.	
Atributos	**Tipo**	**Descrição**
IDONTO_ENSG_ILMN	Varchar(255)	Identificador correspondente a tabela genes. Ex. ENSG00000000003
Probe_Set_Name	Varchar(255)	Identificador ENSG contido nos arquivos GEO. Ex. ENSG00000000003_at
Description	Varchar(255)	Descrição contida nos arquivos GEO. Ex tetraspanin 6 [Source:HGNC Symbol;Acc:11858]

Com a importação dos arquivos do GEO foi possível fazer o mapeamento entre as sondas de microarranjo e o identificador ENSG das *Probes* de RNAseq com o símbolo aprovado do HGNC, uma vez que na arquitetura da Ontologia Ontocancro já existia esse identificador armazenado.

A Figura 17 ilustra a versão atualizada do Diagrama Entidade-Relacionamento da base de dados da Ontocancro após as alterações realizadas.

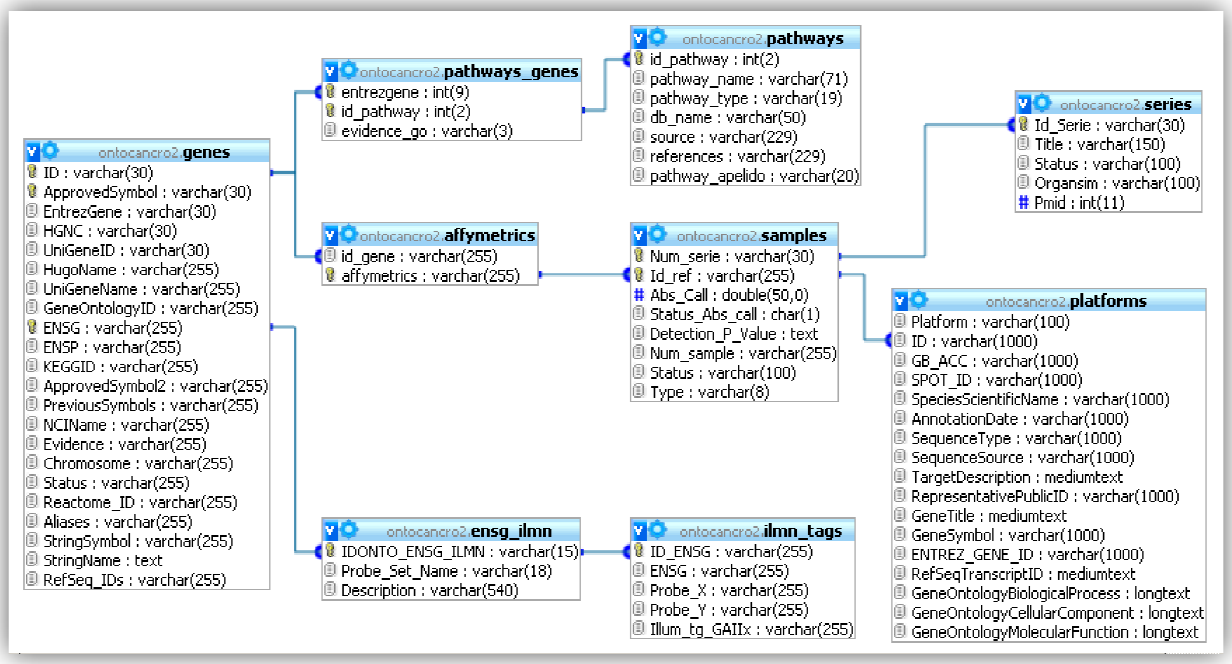


Figura 17 – Esquema da base de dados relacional da Ontocancro 2.0.

Na versão atual, existem 896 genes na na base de dados da Ontocancro 2.0, distribuídos em 40 subvias. A tabela "*ensg_ilmn*" possui a descrição de cada identificador ENSG encontrado no estudo selecionado no GEO, enquanto a tabela "*ilmn_tags*" possui as sequencias *Probe* obtidas no estudo para cada identificador ENSG. Para um novo estudo, basta atualizar as duas tabelas citadas com as novas informações.

3.3. Seleção dos estudos envolvidos em câncer extraídos da técnica de microarranjos da Affymetrix

Os quatro estudos de microarranjos envolvidos em câncer que foram extraídos do banco de dados GEO são:

1. GSE10927 (*Human adrenocortical carcinomas, adenomas, and normal*): Contém 33 amostras de carcinomas (tecidos cancerosos), 22 amostras de adenomas (tecidos pré-cancerosos), e 10 amostras de tecido normal (tecidos saudáveis), de diferentes pacientes. Contendo ensaios de mRNA utilizando o microarranjo *Affymetrix* HG_U133_plus_2 *arrays*, formado por 54.675 conjuntos de sondas (GIORDANO *et al.*, 2009). Totalizando 3.553.875 registros[5].
2. GSE27155 (*Human thyroid carcinomas, adenomas and normal*): Contém 4 amostras de tecido normal, 17 amostras de adenoma da tireoide, 78 amostras de carcinomas. Estas 99 amostras contêm ensaios de mRNA utilizando o microarranjo *Affymetrix* HG_U133A, formado por 22.283 conjuntos de sondas (GIORDANO *et al.*, 2006). Totalizando 2.206.017 registros[6].
3. GSE19650 (*Human pancreatic carcinomas, adenomas, and normal*): Contém 6 amostras de carcinomas, 6 amostras de adenomas, e 7 amostras de tecido normal, de diferentes pacientes. Contendo ensaios de mRNA utilizando o microarranjo *Affymetrix* HG_U133_plus_2 *arrays*, formado por 54.675 conjuntos de sondas (HIRAOKA *et al.*, 2011). Totalizando 1.038.825 registros[7].

---

[5] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10927
[6] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27155
[7] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19650

4. GSE4183 (*Human colon rectal carcinomas, adenomas, inflammations and normal*): Contém 8 amostras de tecido normal, 15 amostras de adenomas, 15 amostras de carcinoma e 15 com inflamação. Estas amostras contêm ensaios de mRNA utilizando o microarranjo *Affymetrix* HG_U133A, formado por 54.675 conjuntos de sondas (GALAMB *et al.*, 2010). Totalizando 2.897.775 registros[8].

O banco de dados GEO possui mais de trinta mil estudos (series) de diversas espécies, áreas de pesquisa e técnicas de análise. Para escolher os estudos deste trabalho foram consideradas as características necessárias para a pesquisa relacionada ao câncer em humanos, com dados produzidos pela técnica de microarranjos da empresa Affymetrix. Além disso, para que as comparações pudessem ser feitas, era de suma importância que as séries tivessem amostras (*samples*) de tecidos com câncer (adenomas e carcinomas) e de tecidos normais.

3.4. Seleção dos estudos envolvidos em câncer extraídos da técnica de sequenciamento (RNAseq) da Illumina

Para análise de dados obtidos da técnica de sequenciamento RNAseq foi selecionada a série GSE29007 do banco de dados GEO que também possui amostras de microarranjo, possibilitando desta forma, a comparação entre as duas técnicas.

O estudo GSE29007 possui duas plataformas:

GPL10999 *Illumina Genome Analyzer IIx (Homo sapiens);*

GPL13447 Affymetrix Gene Chip Human Genome U133A 2.0 Array [CDF: Ensembl v58, Hs133Av2_Hs_ENSG].

Esta série inclui para RNASeq as amostras de tecido de indivíduos classificados como: Saudável não Fumante; Saudável Fumante; Fumante sem Câncer de Pulmão; Fumante com Câncer de Pulmão. E para Microarranjo as amostras Ex-Fumante sem Câncer de Pulmão, Ex-Fumante com Câncer de Pulmão, Fumante com Câncer de Pulmão.

[8] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4183

3.5. Importação dos dados

A importação dos dados selecionados ocorreu de forma semelhante, pois ambos foram retirados do banco GEO, que disponibiliza aos pesquisadores, ferramentas e informações dos experimentos realizados com diversas técnicas, como microarranjos e RNAseq. Através da interface é possível configurar consultas de acordo com os campos e características disponíveis. A Figura 18 mostra o resultado da consulta de uma série, onde é possível ver os arquivos disponíveis.

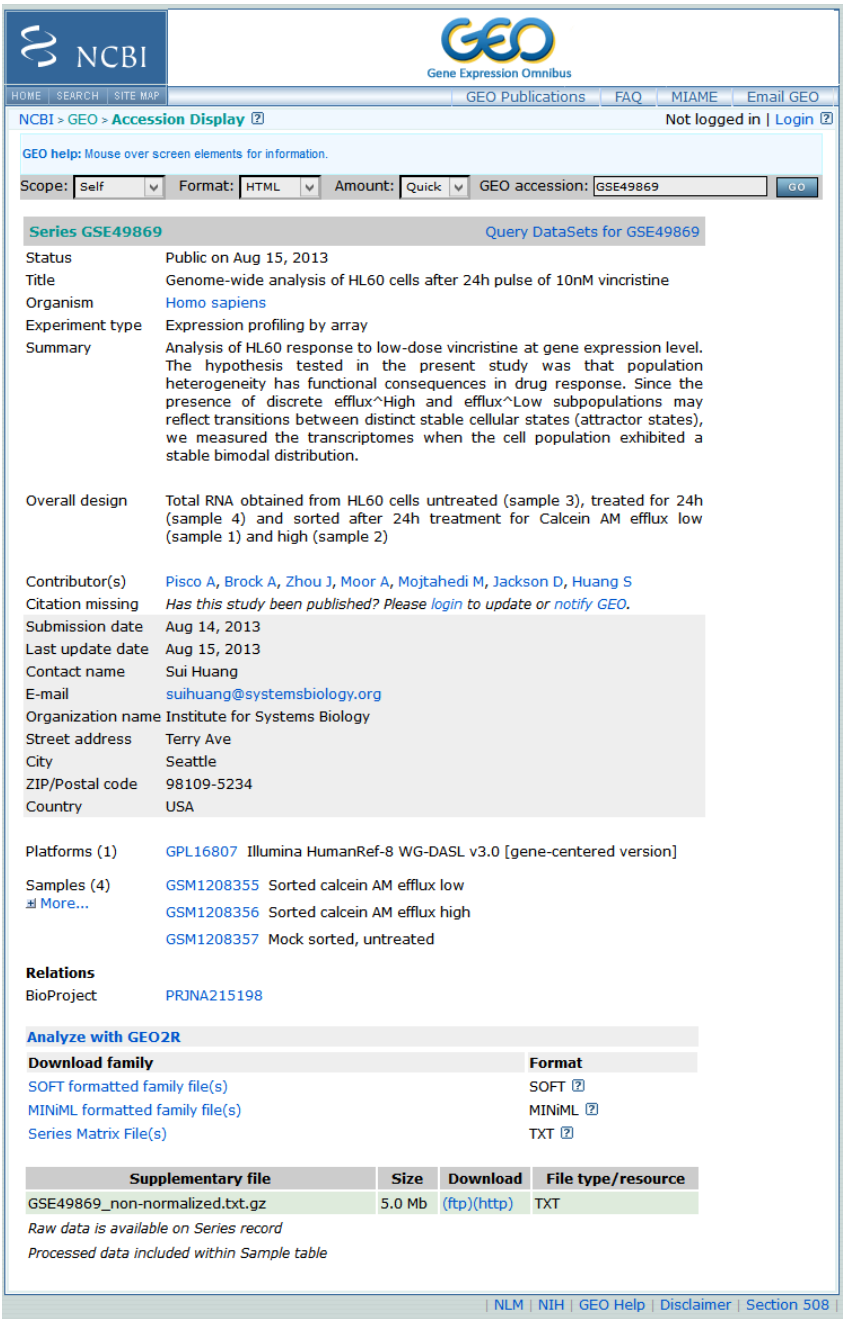


Figura 18 - Interface de consulta do banco GEO

Após a seleção do estudo que será analisado, tem-se à disposição as informações sobre o desenvolvimento do estudo e os arquivos para *download*. Os formatos de arquivos disponíveis são variados, mas os estudos selecionados para a Ontocancro eram do tipo texto (txt).

Os arquivos texto disponíveis das amostras de microarranjos, têm seus dados tabulados, como mostra parcialmente a Tabela 9, e contém as colunas: *id_ref*, que identifica a sonda do estudo; *value*, que guarda o valor da expressão sobre esta sonda; *abs_call*, que indica se o transcrito está presente, ausente ou marginal; e, o *p_value*, que indica o nível de significância da amostra.

Tabela 9 - Estrutura dos arquivos das amostras de microarranjo

ID_REF	*VALUE*	*ABS_CALL*	*P_VALUE*
AFFX-BioB-5_at	374.5	P	0.00034
AFFX-BioB-M_at	569.6	P	0.000044
AFFX-BioB-3_at	442.2	P	0.000052
AFFX-DapX-3_at	2.9	A	0.876428

Para cada amostra, uma instrução de inserção SQL (*Structured Query Language*) (MILANI, 2007) foi criada para adicionar os dados dos arquivos e as demais informações na tabela *samples*, tais como: *Num_serie*, que identifica a série a que pertence a amostra; *Num_sample*, que identifica a amostra; *Status*, que contém a descrição do nome da amostra; e, *Type* que identifica se a amostra pertence a um tecido normal, também chamado de controle, ou se pertence a um tecido afetado com câncer.

De forma semelhante, os arquivos texto disponíveis das amostras de RNAseq possuem seus dados tabulados, conforme mostra a Tabela 10. Porém nestes casos, os arquivos foram tratados para separar a descrição textual do identificador ENSG, que foi importado na tabela *ensg_ilmn* afim de evitar redundâncias.

Tabela 10 - Estrutura dos arquivos das amostras de RNAseq

ENSG	Probe_X	Probe_Y	Illum_tg_GAllx
ENSG00000000003_at	302	585	TCTTAGTTTTCGTTTGTGCCTTTGA
ENSG00000000003_at	513	665	TTATGATGTTCATACTTTCCCTCTT
ENSG00000000003_at	293	109	ACACAACTTACATTTCTTTGCCTCC
ENSG00000000003_at	473	631	TACATCAGGTTTCAGCACACAACTT

As demais tabelas foram preenchidas manualmente, pois possuíam poucos registros, como a tabela *pathways* e *series*.

3.6. Sumário do Capítulo

No presente capítulo foi mostrado como a ontologia Ontocancro foi reestruturada e seus novos conceitos (seção 3.1), como ficou a versão atual da estrutura do banco de dados (seção 3.2), quais dados foram selecionados pela técnica de microarranjos da Affymetrix (seção 3.3) e quais dados foram obtidos pela técnica de sequenciamento RNASeq (seção 3.4). A seção 4.5 descreveu como os dados foram importados para o banco de dados da Ontocancro.

No próximo capítulo será descrita o procedimento de acesso aos dados para comprovação dos resultados obtidos de Halazonetis (2008), através da descrição dos cálculos da atividade e diversidade relativa e como foram implementados na página que armazena as informações da Ontocancro.

4. IMPLEMENTAÇÃO DOS CÁLCULOS DE ATIVIDADE RELATIVA E DIVERSIDADE RELATIVA NA ONTOCANCRO 2.0

Diversas metodologias podem ser utilizadas para análise das interações envolvendo vias e subvias em microarranjos, inclusive metodologias destinadas a estudos específicos. Isso devido à grande quantidade de dados de microarranjos e da grande variabilidade na expressão de genes isolados (SIMÃO, 2012). Na metodologia de análise por representação excedente, por exemplo, há uma busca por vias que contenham genes expressos significativamente. A partir de uma lista de genes é aplicado um teste hipergeométrico que avaliará a expressão do conjunto e calculará a probabilidade de conter genes diferencialmente mais expressos do que seria esperado por acaso (GOHLMANN; TALLOEN, 2009). O teste hipergeométrico é uma técnica para análise de perfis moleculares dos dados de expressão gênica, que classifica genes em uma determinada categoria.

Outro método que se destaca é a contagem funcional de classes que é usado em microarranjos que apresentam baixos valores p (*p-value*) em suas amostras. O valor de p é a probabilidade de obter-se uma estatística de teste no mínimo tão extrema como o que foi realmente observado, muitas vezes o valor de p é menor que o nível de significância que geralmente varia entre 0,05 e 0,1. Esse método é usado para calcular uma pontuação que define as alterações significativas de todos os genes incluídos na análise. Um terceiro método analisa os conjuntos de genes diretamente ligados a conjuntos pré-definidos com o objetivo de identificar as mudanças de expressão entre essas vias (GOHLMANN; TALLOEN, 2009).

Este capítulo compreende a descrição dos cálculos de Atividade Relativa e Diversidade Relativa (seção 5.1) e apresenta o algoritmo que implementa estes cálculos (seção 5.2).

4.1. Descrição dos cálculos da Atividade Relativa e Diversidade Relativa

Castro e colaboradores (2007) introduziram outro método de análise direto de expressão de vias e subvias. Os cálculos da atividade relativa e da diversidade

relativa são usados para avaliar os níveis de expressão de conjuntos de genes e definir os padrões de expressão entre os genes das vias (CASTRO, MAURO A. A. *et al.*, 2007).

Para calcular a atividade relativa de uma dada via α com um número de genes M_α, deve-se somar a expressão dos genes em dois grupos de vias: o primeiro grupo representa as amostras de tecidos alterados ou experimentais N_α^e e o segundo grupo é composto pelas amostras de tecidos normais ou controle N_α^γ. Então, a atividade relativa n_α da via α será dada por:

$$n_\alpha = \frac{N_\alpha^e}{N_\alpha^e + N_\alpha^\gamma} \quad (1)$$

O valor de n_α varia entre $0 \leq n_\alpha \leq 1$, se $n_\alpha < 0{,}5$ a atividade da via com amostra alterada é menor que a atividade do controle, enquanto que $n_\alpha > 0{,}5$ representa o caso inverso (CASTRO, MAURO A. A. *et al.*, 2007).

Para caracterizar de forma quantitativa a diversidade para uma via α é utilizada a entropia de Shannon que é descrita como:

$$H_\alpha = -\frac{1}{\ln(M_\alpha)} \sum_i^{M_\alpha} p(i,\alpha) \ln p(i,\alpha) \quad (2)$$

onde $M\alpha$ é o número de genes na via e p(i,α) é a frequência da diversidade do gene i, dada por:

$$p(i,\alpha) = \frac{s(i,\alpha)}{N_\alpha} \quad (3)$$

com s(i,α) sendo a atividade do gene (i) e N_α a soma da expressão dos genes na via (α). O termo $\ln(M_\alpha)$ é um fator de normalização que garante $0 \leq H_\alpha \leq 1$, desta forma, pode-se comparar as vias com diferentes quantidades de genes. Tendo como referência o sinal controle da amostra, pode-se definir a diversidade relativa (h_α) como:

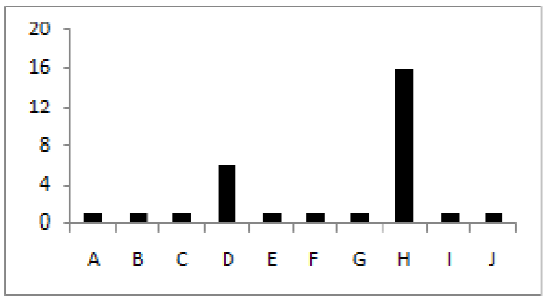
$$h_\alpha = \frac{H_\alpha^e}{H_\alpha^e + H_\alpha^\gamma} \quad (4)$$

onde H_{α}^{e} e H_{α}^{γ} são as diversidades das amostras com alteração (experimento) e o controle, respectivamente. O valor de h_{α} varia entre $0 \leq h_{\alpha} \leq 1$, se $h_{\alpha} < 0,5$ implica que $H_{\alpha}^{e} < H_{\alpha}^{\gamma}$, isto é, a diversidade dos valores de expressão dos genes na via é menor para a amostra alterada do que para o controle, enquanto que $h_{\alpha} > 0,5$ representa o caso inverso (CASTRO, MAURO A. A. *et al.*, 2007). A Figura 19 ilustra a diversidade relativa para um conjunto de 10 genes alterados com seus respectivos genes de controle pertencentes a uma via qualquer.

Experimento (e)

Genes	s(i,α)	p(i,α)	ln(p(i,α)).p(i,α)
A	1	0.03	-0.11
B	1	0.03	-0.11
C	1	0.03	-0.11
D	6	0.20	-0.32
E	1	0.03	-0.11
F	1	0.03	-0.11
G	1	0.03	-0.11
H	16	0.53	-0.34
I	1	0.03	-0.11
J	1	0.03	-0.11
Σ	**30**	**1.00**	**-1.56**

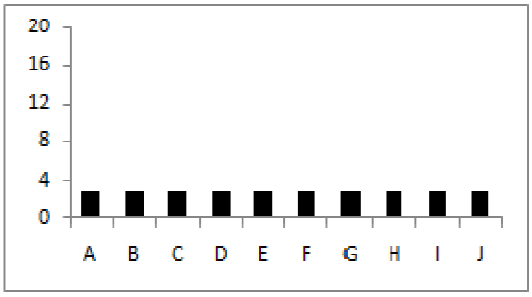
M_{α}= 10
N_{α}= 30
H_{α}= 0.68



Controle (γ)

Genes	s(i,α)	p(i,α)	ln(p(i,α)).p(i,α)
A	3	0.10	-0.23
B	3	0.10	-0.23
C	3	0.10	-0.23
D	3	0.10	-0.23
E	3	0.10	-0.23
F	3	0.10	-0.23
G	3	0.10	-0.23
H	3	0.10	-0.23
I	3	0.10	-0.23
J	3	0.10	-0.23
Σ	**30**	**1.00**	**-2.30**

M_{α}= 10
N_{α}= 30
H_{α}= 1.00



h_{α}= 0.405

Figura 19 – Exemplo mostrando os níveis de expressão de 10 genes quaisquer, com diminuição da diversidade do experimento em relação ao controle ($\mathbf{H_{\alpha}^{e} < H_{\alpha}^{\gamma}}$). Para este caso a diversidade dos genes de controle é maior do que o experimento, desta forma a diversidade relativa é menor que 0,5 deste modo hα=0,405. (SIMÃO, 2012)

Os genes das amostras de controle apresentam níveis de expressão com valores mais próximos (constantes ou sem dispersão) do que os níveis de expressão

dos genes do experimento, isso indica que $H_{\alpha}^{e}<H_{\alpha}^{\gamma}$ e $h_{\alpha}<0,5$. Ou seja, há uma diminuição da diversidade nos níveis de expressão dos genes de experimento em relação ao controle. Para $H_{\alpha}^{e}>H_{\alpha}^{\gamma}$ tem-se o caso contrário e haverá aumento da diversidade relativa $h_{\alpha}>0,5$ do experimento *versus* o controle. Na reprogramação da expressão dos genes poderão ocorrer níveis de aumento ou diminuição da diversidade relativa entre os genes de uma via.

Para determinar se uma alteração em uma via (atividade) ou em um conjunto de genes (diversidade) é estatisticamente significativa em um determinado estudo (microarranjo), aplica-se o método de *bootstrap.* Esse método é usado para calcular a distribuição amostral de h_{α} e n_{α} através de uma análise de reamostragem aleatória que cobre todos os genes do estudo com repetições que variam de 100 a 100.000 com as mesmas quantidades de genes das vias de interesse para investigar a convergência do valor de p (*p-value*) da amostra sobre uma distribuição de probabilidades normal.

O nível de significância em uma distribuição de probabilidades é comparado com o valor de p e serve para delimitar se os resultados da atividade ou diversidade relativa de um determinado conjunto de genes são ou não significativos a um determinado nível. Desta forma, se o nível de significância for fixado em 0,05 os valores de p menores que 0,05 ($p < 0,05$) apresentarão aumento ou diminuição significativa de expressão da via em relação ao estudo. Dependendo em qual dos lados da distribuição de probabilidade unicaudal cair o valor de p poderá ter aumento ou diminuição de expressão.

As análises mostradas acima podem ser calculadas usando o *software ViaComplex. O software* é um aplicativo usado para construir mapas funcionais de expressão gênica e utiliza a entropia de Shannon para obter um parâmetro quantitativo usado para caracterizar a atividade e a diversidade relativa de vias (CASTRO, MAURO A A *et al.*, 2009). O *software* é utilizado para calcular significâncias de vias pelo método de *bootstrap* e pela correção de falsos positivos. Os falsos positivos são usados como método de controle estatístico quando se faz os testes de múltiplas hipóteses (*bootstrap*) para efetuar uma correção nas múltiplas comparações (entre as vias). O código fonte do *software ViaComplex* contendo todas as equações apresentadas acima está disponível em ambiente virtual (CASTRO, MAURO A A *et al.*, 2009).

4.2. Algoritmo para o cálculo da Atividade Relativa e Diversidade Relativa

Uma vez que a base de dados da Ontocancro possui todas as informações relacionadas às doenças a serem estudadas e às vias de manutenção do genoma (Reparo ao Dano no DNA, Apoptose e Ciclo Celular), está pronta para a realização do estudo estatístico que integre estes dois tipos de informação de forma que seja possível traçar o perfil destas vias.

Para isso, inicialmente, calcula-se o valor da atividade relativa através da escolha do estudo que será analisado dentre as quatro doenças (transcriptomas) selecionadas pela técnica de microarranjos, além de escolher a sub-via metabólica dentre as sub-vias (interatomas) disponibilizadas pela Ontocancro. O resultado será a lista de *samples*, que é composta por amostras de tecido saudável (normal) e de tecido afetado (com câncer).

Esta análise se fará através da entropia de Shannon (SHANNON, 1948), que permite obter um parâmetro quantitativo usado para caracterizar a diversidade e a atividade relativa de vias metabólicas (CASTRO, M. A. A. *et al.*, 2007). Para isso, é essencial realizar um processo de filtragem nas amostras a serem avaliadas, pois nem todas apresentam um nível de relevância considerável em seu sinal de expressão dentro do microarranjo. Além disso, deve-se levar em consideração também que os genes podem estar repetidos dentro de uma mesma amostra; neste caso faz-se a média entre os diferentes valores de expressão do mesmo gene.

Para a filtragem dos valores de expressão das amostras é importante definir o grau de significância para todos os genes expressos dentro de um microarranjo. Em cada amostra o gene apresenta um valor de significância (*p_value*) ou um estado de expressão (*abs_call*) que determinam a relevância de sua expressão para o estudo; se este valor de significância estiver entre 0.00 e 0.04 indica que tal gene está expresso na sonda; caso contrário este gene terá o seu valor de expressão igualado a zero. Para o caso dos genes que não possuem valor de significância, considera-se o seu estado de expressão: ausente (A), marginal (M) ou presente (P). Para os genes que apresentam seu estado de expressão presente ou marginal, seus valores de expressão são mantidos. Para genes que apresentarem estado de expressão ausente, seus valores de expressão serão igualados a zero. Esta etapa serve para

selecionar amostras que contenham um nível de expressão muito baixo fazendo com que elas não tenham significância na análise do perfil das vias,

Dessa forma, deve ser escolhida pelo menos uma amostra de tecido saudável que será usado como controle para cálculo da expressão, e uma amostra de tecido afetado. Além dessas seleções, o valor de significância deve ser informado, caso haja.

Para relacionar os genes contidos na Ontocancro com os dados das doenças importadas do GEO, necessita-se um cruzamento com as informações das sondas *Affymetrix*. Na tabela *samples* há um atributo chamado *Id_ref*, nele está contido o identificador de todas as sondas que foram utilizadas para realizar o experimento. Para relacionar aos genes que compõem as amostras, foi criada a tabela *affymetrix* contendo o identificador das sondas e os genes que fazem parte dela e criou-se o relacionamento com a tabela *genes*. A partir do cruzamento destas informações pode-se obter quais genes de uma via estão presentes na doença analisada.

Para realizar os cálculos foi criada uma tabela temporária chamada *media_final* que recebe todos os dados das amostras e os genes encontrados nas vias, para facilitar os cálculos. Sua estrutura está representada na Tabela 11.

Tabela 11 – Tabela *media_final* do Banco de Dados Relacional.

Tabela *media_final*	Composta pelos dados necessários aos cálculos	
Atributos	**Tipo**	**Descrição**
Serie	Varchar(8)	Identificador da Série ou Doença. Ex. GSE19650
entrezgene	int(9)	Identificador do gene. Ex. 5982
Abs_call	double(50,0)	Valor da expressão do gene em determinada sonda. Ex. 16
Status_abs_call	char(1)	Estado da amostra, se está presente (P), ausente (A) ou marginal (M)
Detection_P_Value	Text	Grau de significância da amostra
Status	Varchar(100)	Nome dado a amostra
Id_ref	Varchar(255)	Identificador da sonda que a amostra pertence
Type	Varchar(8)	Tipo de amostra: cancer ou controle

Com os dados armazenados em uma única tabela, foram desenvolvidas rotinas SQL que possibilitam a execução dos cálculos da atividade, do desvio padrão e da diversidade relativa.

O cálculo do Desvio Padrão é utilizado para calcular as medidas de variação entre o grupo de expressões. Ele permite uma interpretação direta do conjunto de dados. Ele é calculado através da fórmula:

$$STD = \frac{\sqrt{\sum_{i=0}^{n}(x_i - \bar{X})^2}}{n-1} \quad (5)$$

Onde, n é o número total de genes encontrados na via, x_i é o valor da expressão referente a cada gene e $\bar{X}$ é a média dos valores de expressão encontrados em cada amostra.

A própria linguagem SQL possibilita o uso de funções nativas, como a média e o desvio padrão, que foram inseridos na mesma rotina, uma vez que os genes já estão agrupados por amostras, basta utilizarmos os seguintes parâmetros antes de cada seleção: AVG (*Average*) para o cálculo da média e STD (*Standard Derivation*) para calcular o desvio padrão.

A média então resultará no valor da atividade. Para o cálculo da atividade relativa necessita-se dividir as amostras em dois tipos: controle e câncer. Para isso utiliza-se a mesma consulta que calcula a média e o desvio padrão, porém agregando uma instrução condicional que armazena em duas variáveis distintas os tipos de amostras encontradas.

Na Figura 20 é possível ver parcialmente o código da rotina criada. Os cálculos do desvio padrão e da média podem ser observado na linha 123. O cálculo da atividade relativa encontra-se na linha 145, enquanto o cálculo da diversidade relativa encontra-se na linha 158. Também é possível verificar a separação das amostras por tipo controle e câncer, nas linhas 138 e 141.

Para o cálculo da diversidade relativa outro fator precisou ser observado: o SQL não consegue tratar o fato da fórmula ser composta por vários tipos de operadores. Para solucionar este problema desenvolveu-se uma função determinística em SQL (chamada *func_Atividade7*, linha 150 da Figura 20).

```
115 function calc_media_amostras() {
116     $sql_1 = "SELECT Distinct substr(Status,1,LENGTH(Status)-2)
117             FROM `media_final`";
118
119     $result_1 = mysql_query($sql_1) or die (mysql_error());
120       while($reg_1 = mysql_fetch_array($result_1)){
121        mysql_query("UPDATE media_final set Status = '$reg_1[0]'
122        WHERE Status like '$reg_1[0]%'") or die (mysql_error()); }
123       $sql = "Select STD(Abs_Call) as desvio, AVG(Abs_Call)
124          as media,Status, Type
125          from media_final GROUP BY Status";
126
127     $result = mysql_query($sql) or die (mysql_error());
128       if(!$result){ die ("Erro da sintaxe SQL :<hr>$sql"); }
129
130      $atr[]=array();
131
132       while($reg = mysql_fetch_array($result)){
133        echo "</br>Status: ".$reg['Status'];
134        echo "</br>Value of activity (Average): ".$reg['media'];
135        echo "</br>Standard deviation: ".$reg['desvio'];
136        echo "</br>Type: ".$reg['Type'];
137
138            if($reg['Type'] == 'cancer'){
139                  $atr[0] = $reg['media'];
140            }
141            else if($reg['Type'] == 'controle'){
142                  $atr[1] = $reg['media'];
143            }
144       }
145          $atr_final=$atr[0]/($atr[0]+$atr[1]);
146
147            echo "Relative activity: "." ". $atr_final ."</br>";
148
149            $sql_divrs = "SELECT (-1/log(count(entrezgene)))*
150                        SUM(`Abs_Call`/func_Atividade7()
151                        *log(`Abs_Call`/func_Atividade7()))
152                        as valor from `media_final`
153                    GROUP BY Type";
154
155      $result_divrs = mysql_query($sql_divrs);
156            $reg_divrs_c=mysql_result($result_divrs,0,"valor");
157            $reg_divrs_n=mysql_result($result_divrs,1,"valor");
158            $reg_divrs_t=$reg_divrs_c/($reg_divrs_c+$reg_divrs_n);
159            echo "Relative diversity: ".$reg_divrs_t."<br>";
160 }
```

Figura 20 - Rotina para Cálculo da Atividade e Diversidade Relativa

Funções determinísticas sempre retornam o mesmo tipo de valor quando são chamadas por um conjunto específico de valores de entrada e, dado o mesmo estado da base de dados. Existem várias propriedades de funções que podem ser determinadas pelo usuário para indexar os resultados da função (MICROSOFT, [S.d.]). Esta função realiza o somatório de todas as médias do valor de expressão das amostras de cada gene contido nas vias. O resultado retornado por essa função diz respeito ao elemento N_a referenciado na fórmula (3) conhecido pelo valor da atividade da via. Assim, tendo o número total de genes na via, o valor de expressão de cada gene e atividade relativa, aplicou-se a fórmula da entropia de *Shannon* (2) e obteve-se como resultado a diversidade da via para cada tipo de amostra. Para o cálculo da diversidade relativa fez-se o uso da fórmula (4) obtendo os valores referentes a mesma (PEREIRA, 2011). A função criada pode ser observada na Figura 21.

```
1 CREATE FUNCTION `func_Atividade7`() RETURNS double
2  DETERMINISTIC
3
4 begin
5 declare c,d double;
6 declare done int default 0;
7 declare temp, soma double;
8 declare cur1 cursor for
9 select AVG(abs_call)  from media_final GROUP BY entrezgene;
10 declare continue handler for not found set done=1;
11 open cur1;
12 set soma = 0;
13 atividade_loop:LOOP
14 fetch cur1 INTO temp;
15 set soma = soma + temp;
16 if done=1 then
17 leave atividade_loop;
18 end if;
19 end loop atividade_loop;
20 close cur1;
21
22 return soma;
23 end;
```

Figura 21 - Função Determinística para Cálculo da Diversidade Relativa

A Figura 22 apresenta um fluxograma dos passos descritos, que começam com a escolha da subvia (interatoma) e da doença (transcriptoma), a partir de então é feita uma consulta na base de dados da Ontocancro e são extraídas as informações referentes à subvia selecionada e a doença.

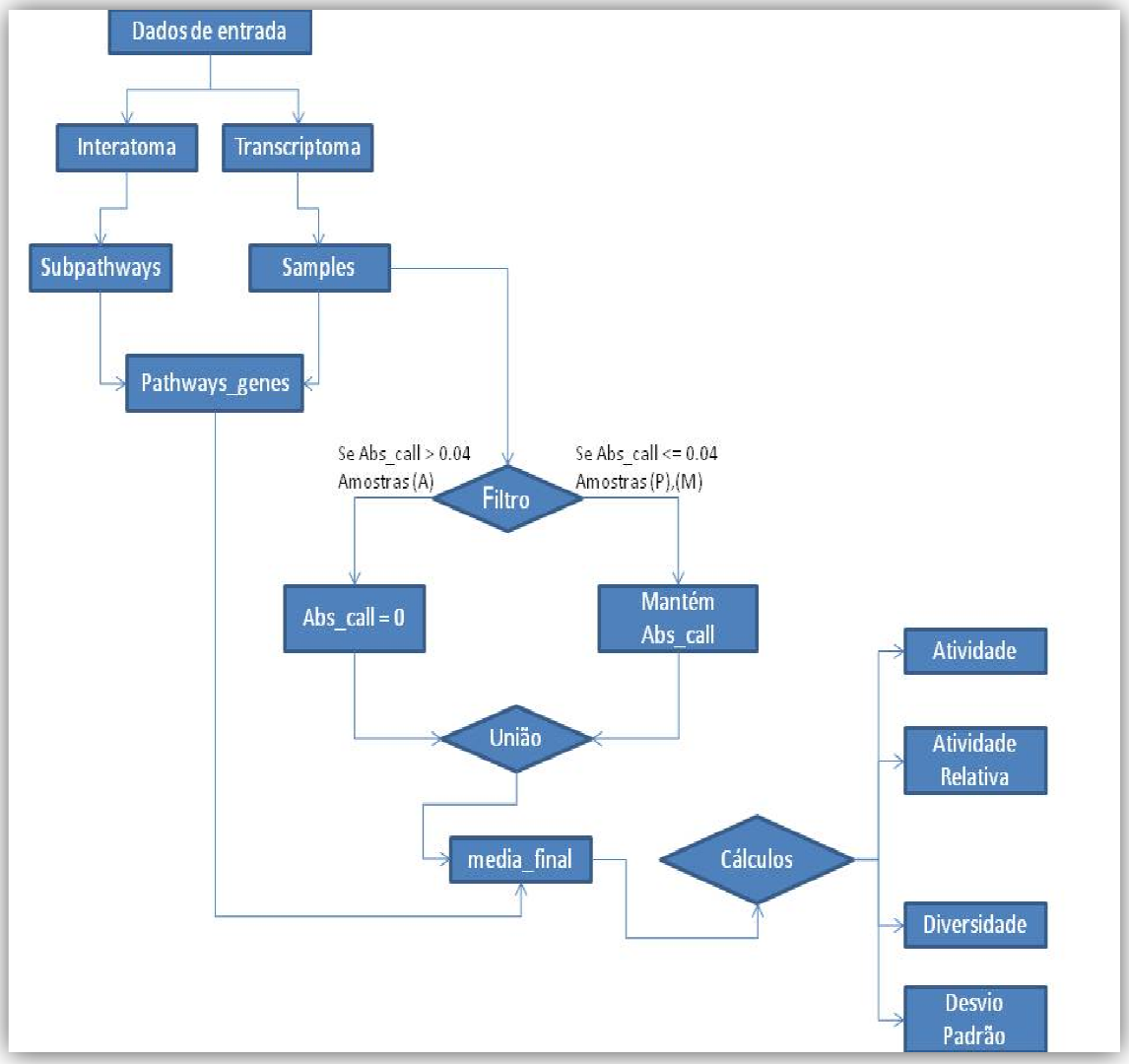


Figura 22 - Fluxograma representando a metodologia utilizada (Pereira, 2011).

A tabela *pathways_genes* é usada para relacionar os genes encontrados nas vias com os genes encontrados nas amostras. Os dados provenientes das doenças passam por uma filtragem, onde amostras com nível de significância maior que 0.04 tem seus valores de expressão zerados e os valores das amostras com significância menor que 0.04 são mantidos.

Então é criada a tabela temporária chamada *media_final* que reúne todas essas informações e possibilita a realização dos cálculos.

4.3. Principais telas de acesso as funcionalidades da Ontocancro 2.0

Conforme apresentado na Figura 23, inicia-se o acesso aos cálculos através da seleção da subvia e da doença, que serão analisadas.

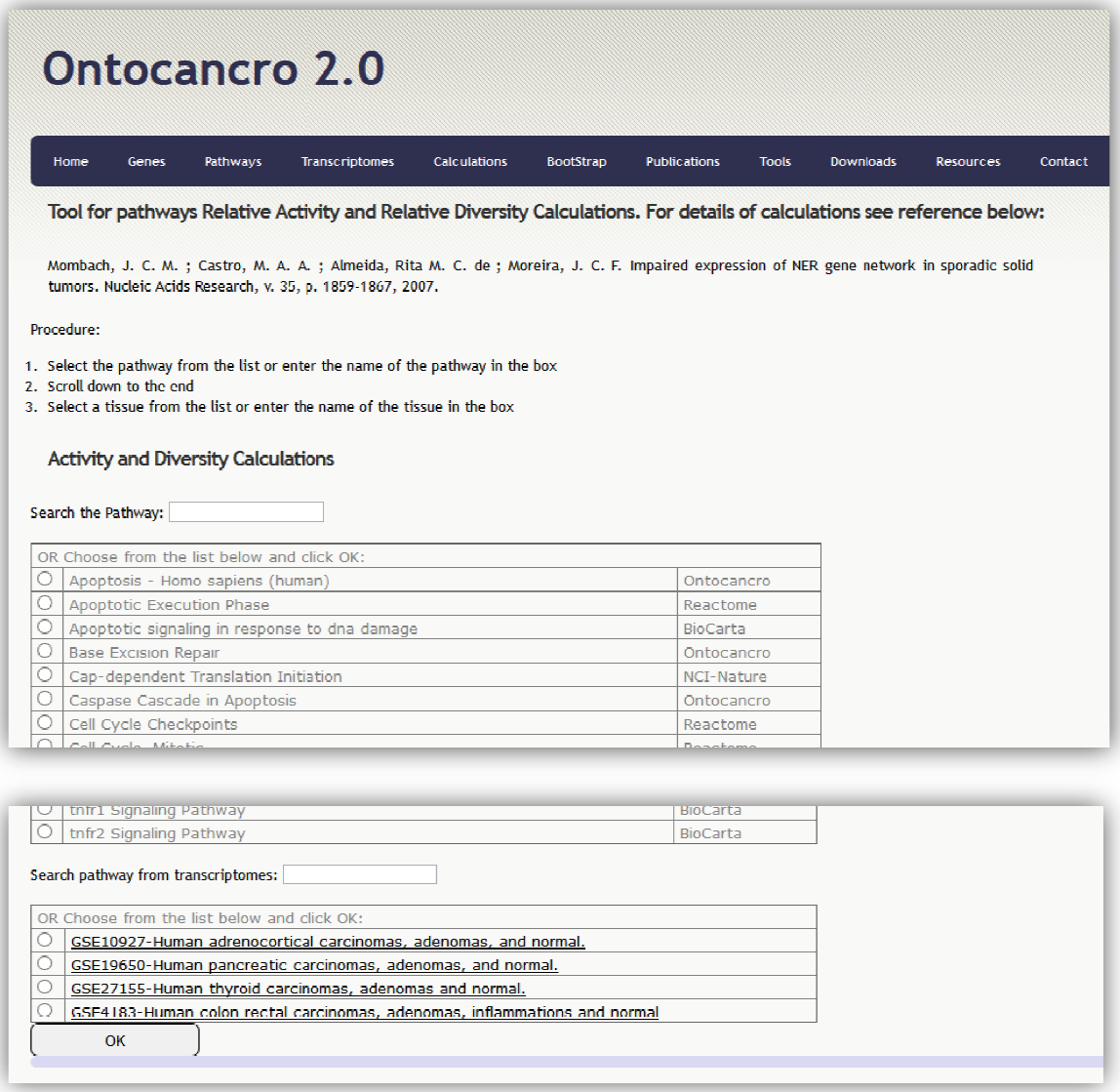


Figura 23 - Tela de seleção da subvia e da doença

Em seguida, o sistema relacionará as amostras com os genes encontrados nas vias. O pesquisador deverá preencher o valor de significância e selecionar amostras do tipo controle (tecido normal) e amostras afetadas com câncer, como mostrado na Figura 24.

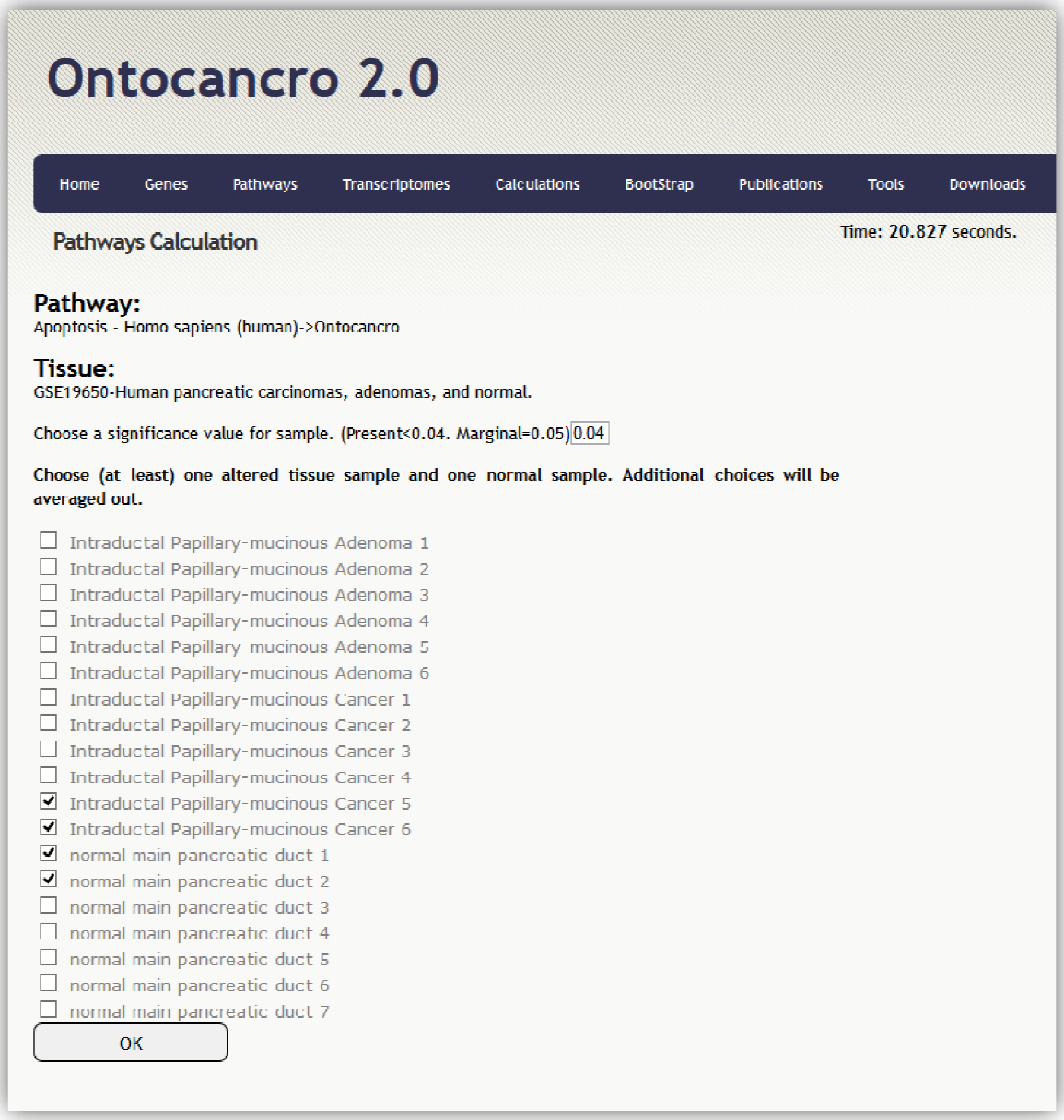


Figura 24 - Tela de seleção das amostras e definição do valor de significância

A Ontocancro realiza então os cálculos e apresenta o relatório com os valores encontrados como pode ser observado na Figura 25.

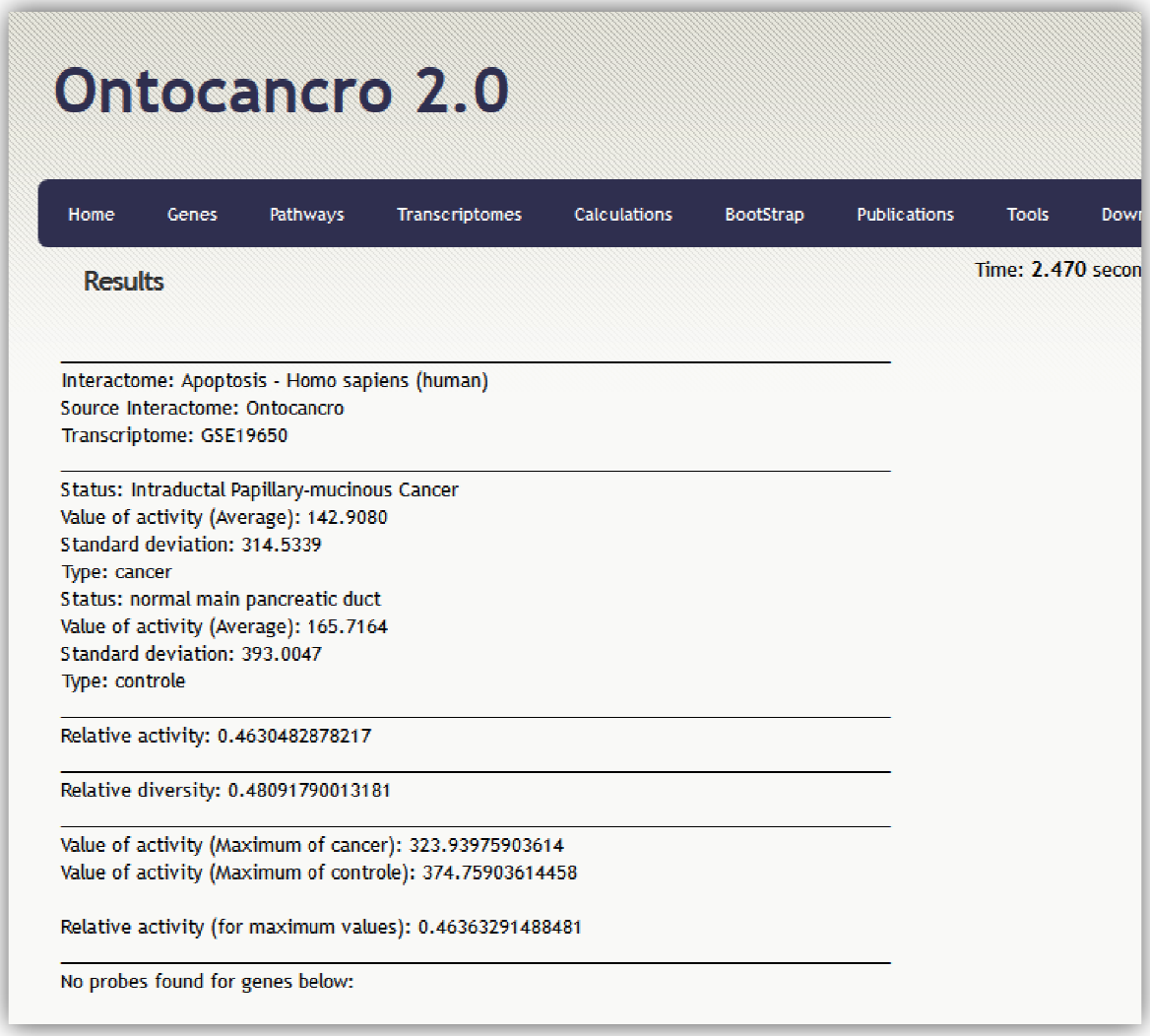


Figura 25 - Tela de apresentados dos valores de expressão encontrados.

Além disso, os menus Genes e Pathways apresentam informações sobre os genes e subvias armazenados na Ontocancro e disponibiliza um módulo de exportação de dados que podem ser filtrados pelo pesquisador. A Ontocancro 2.0 está disponível em www.ontocancro.org.

4.4. Sumário do Capítulo

Este capítulo apresentou os cálculos da Atividade Relativa e Diversidade Relativa e a criação do algoritmo para implementação das suas respectivas fórmulas.

Esta implementação permite traçar o perfil das vias metabólicas envolvidas no processo carcinogênico, com base em dados da expressão gênica obtidas com a técnica de microarranjos.

Também foram apresentadas as principais telas da Ontocancro 2.0, que permitem selecionar os dados para os cálculos citados.

5. RESULTADOS

Um dos objetivos desta dissertação é comprovar quantitativamente a proposta de Halazonetis e colaboradores (2008), que prevê a existência de uma barreira anticâncer em tecidos pré-cancerosos. Desta forma, após a implementação dos cálculos da atividade relativa e diversidade relativa explicados na seção 5.1, através dos algoritmos apresentados na seção 4.2, elaborou-se um estudo de caso para análise dos dados, que será detalhado na seção 5.1.

Outro objetivo foi importar dados oriundos da técnica de sequenciamento RNAseq para comparar com a abordagem de microarranjos, o qual será detalhado na seção 5.2.

5.1. Estudo de Caso comparando a Ontologia Ontocancro com o Software ViaComplex

Como dito na seção 5.1, o *software ViaComplex* é uma ferramenta de código fonte aberto que possibilita a construção de mapas metabólicos a partir de expressões gênicas, e que também fornece uma ferramenta generalizada para avaliar estas redes baseada na entropia de Shannon.

Para este estudo de caso, foram selecionadas amostras de tecidos da tireoide (série GSE27155) que foram usados para determinar a presença da barreira anticâncer em tecidos pré-cancerosos (adenomas) apresentada na seção 2.4. O conjunto de vias GMM extraídas da base de dados da Ontocancro é formado por duas vias de apoptose (*Death Receptor Signalling e Apoptotic Signaling in Response to DNA Damage*) e duas vias de ciclo celular (*RB Tumor Suppressor/Checkpoint Signaling in Response to DNA Damage e G2/M Checkpoints*).

Nos estudos realizados por Halazonetis e colaboradores (2008) é possível constatar, em tecidos pré-cancerosos (adenomas), um aumento na atividade das vias de apoptose e uma diminuição da atividade das vias de ciclo celular. Aplicando a abordagem da Ontocancro às amostras de tecidos pré-cancerosos (adenomas) das vias mencionadas, a mesma constatação foi verificada, como indicado no

gráfico da Figura 26, onde os maiores valores, das categorias que representam as subvias selecionadas, correspondem a atividade da via de Apoptose. Isso indica uma diminuição da proliferação celular, que poderia conter os defeitos em células aberrantes.

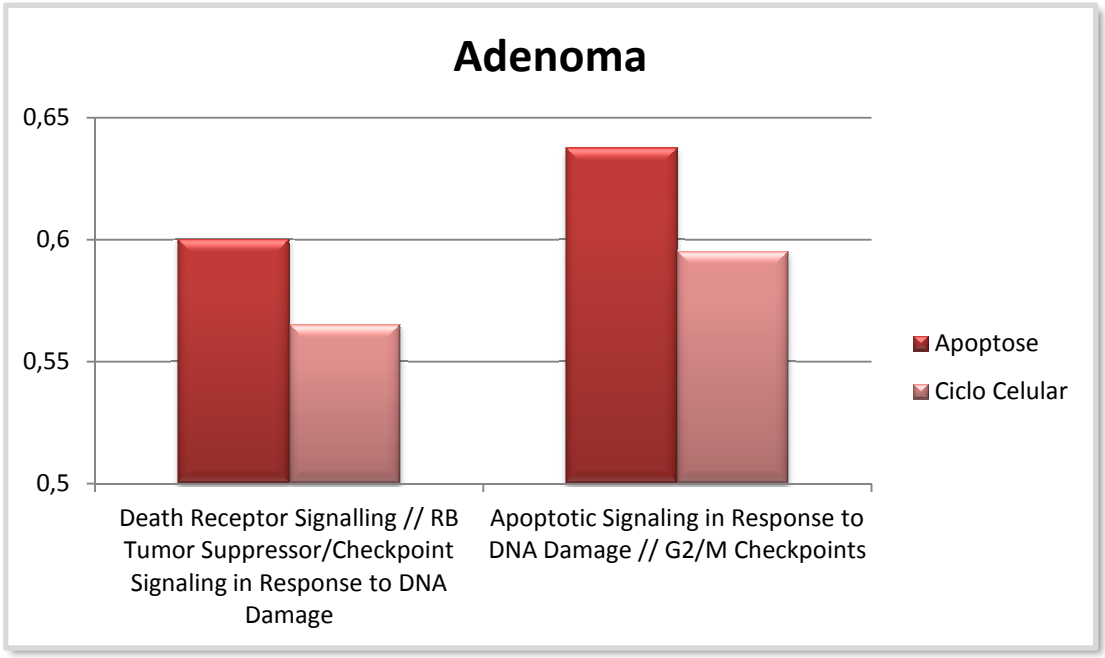


Figura 26 - Gráfico resultante da análise de tecidos pré-cancerosos (adenomas).

Halazonetis também propõe que em tecidos cancerosos (carcinomas) há um aumento da atividade em vias de ciclo celular e uma diminuição da atividade em vias de apoptose. Este aumento de atividade em vias de ciclo celular dá-se a partir da não ocorrência de apoptose em células alteradas. O organismo então começa a proliferar estas células modificadas aumentando gradativamente o grau de câncer. O que também foi constatado na análise com a ontologia.

No gráfico da Figura 27, referente às amostras de tecidos cancerosos, onde os maiores valores, das categorias que representam as subvias selecionadas, correspondem a atividade da via de Ciclo Celular. Isso indica um aumento da proliferação de células afetadas.

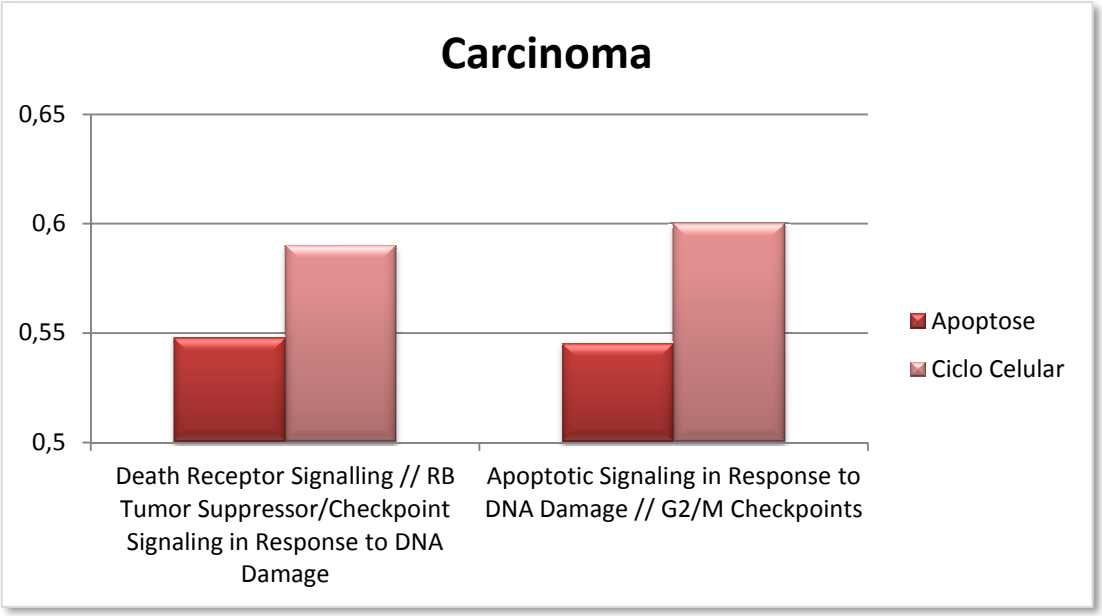


Figura 27 - Gráfico resultante da análise de tecidos cancerosos (carcinomas).

Realizou-se assim, uma comparação de abordagens entre a Ontocancro e a ferramenta *ViaComplex* para inferir o nível de diferenciação entre os métodos utilizados, uma vez que a abordagem desenvolvida pela Ontocancro leva em consideração o identificador *EntrezGene* (código universal que caracteriza o gene em diversas bases de dados de bioinformática), enquanto a ferramenta *ViaComplex* utiliza como identificador o *HugoName*, que pode sofrer alterações a medida que são feitos ajustes no nome do gene de acordo com suas características.

A Tabela 12 descreve na primeira coluna as vias metabólicas associadas aos mecanismos de manutenção do genoma (GMM): NER (*Nucleotide Excision Repair*), BER (*Base Excision Repair*), MMR (*Mis-match Repair*), HR (*Homologous Recombination*), EJ (*Non-homologous End Joining*) e CS (*Chromosome Stability*); na segunda e na terceira coluna os valores de atividade relativa encontrados em cada abordagem e na última coluna a diferença entre esses valores.

Tabela 12 - Diferença entre abordagens: Ontocancro e *Viacomplex*.

VIAS	ADENOMA ONTOCANCRO	ADENOMA VIACOMPLEX	DIFERENÇA
NER	0.503323378	0.5086	-0.005276622
BER	0.501151644	0.5054	-0.004248356
MMR	0.502682944	0.5033	-0.000617056
HR	0.500873387	0.5024	-0.001526613
EJ	0.5	0.5033	-0.0033
CS	0.501154162	0.5033	-0.002145838

É possível notar que a diferença entre as vias relacionadas com os GMM dá-se somente na terceira casa decimal, ou seja, a diferença é pouco relevante o que permite concluir que a abordagem utilizada pela Ontocancro é válida.

5.2. Comparação entre as técnicas de Microarranjo e RNAseq

Para realizar esta análise, buscou-se no banco de dados GEO uma série que contivesse amostras de tecido utilizando as técnicas de microarranjo e RNASeq, simultaneamente. Foi encontrada a série GSE29007, que possui duas plataformas: *GPL10999 Illumina Genome Analyzer IIx (Homo sapiens); GPL13447 Affymetrix Gene Chip Human Genome U133A 2.0 Array.*

Para a plataforma GPL10999, referente ao RNAseq, há as amostras de tecido classificados como "Saudável não Fumante"; "Saudável Fumante"; "Fumante sem Câncer de Pulmão"; "Fumante com Câncer de Pulmão". E para a plataforma GPL13447, referente à microarranjos, existem as amostras classificadas como "Ex-Fumante sem Câncer de Pulmão", "Ex-Fumante com Câncer de Pulmão" e "Fumante com Câncer de Pulmão".

O primeiro passo foi fazer o *download* dos arquivos referentes a série GSE29007, de ambas as técnicas. Em seguida, foram separadas as amostras experimento (de tecido afetado com câncer) e as controle (de tecido saudável). Para cada controle e experimento, montou-se um arquivo texto, com quatro colunas: as sondas, o símbolo aprovado do HGNC, o experimento e o controle. Estes arquivos, juntamente com a lista de vias da Ontologia Ontocancro, foram usados no Software *ViaComplex* (CASTRO, M. A. A. *et al.*, 2009), para obter a atividade relativa e a diversidade relativa, uma vez que não foi realizada a importação dos dados para a Ontocancro realizar os cálculos.

Para melhor compreensão elaborou-se o fluxograma da Figura 28, que demonstra as etapas para obter a análise das técnicas de RNAseq e Microarranjos, utilizando o *software* ViaComplex.

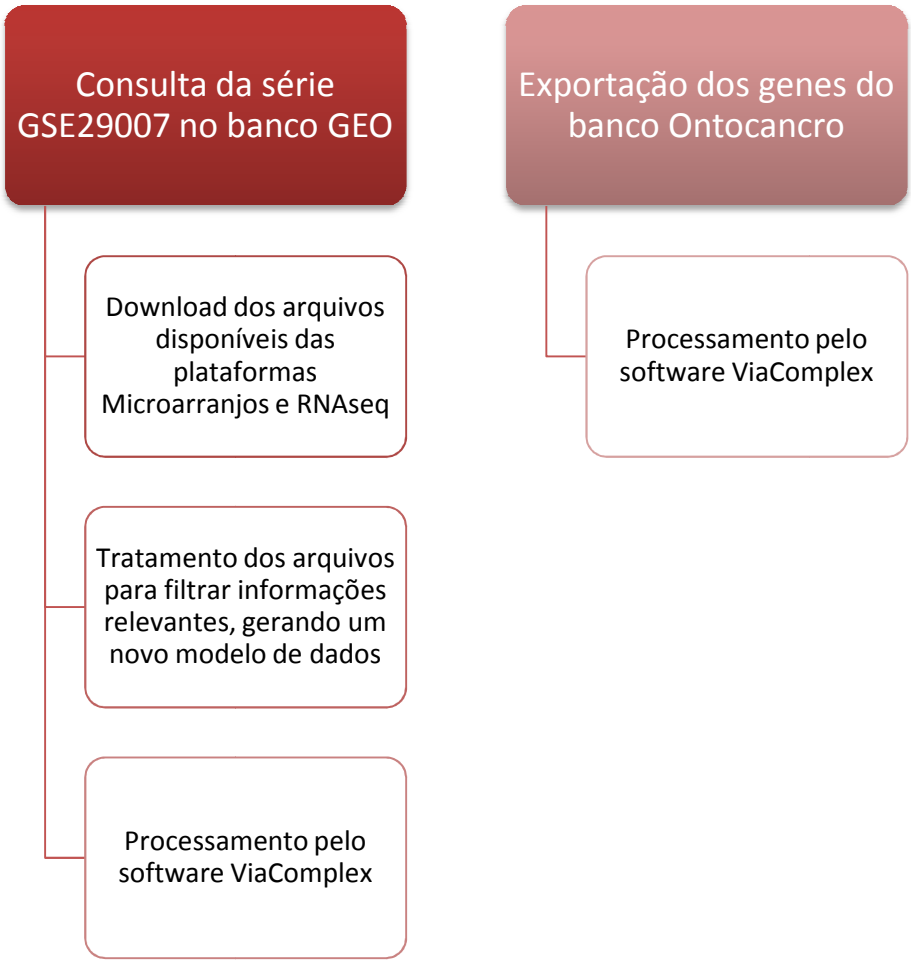


Figura 28 - Fluxograma da análise de RNAseq e Microarranjos pelo software ViaComplex

Os resultados obtidos desta análise não foram satisfatórios, sendo que poucas vias alteradas foram encontradas. Desta forma, concluiu-se que os dados tem baixa sensibilidade e, portanto passou-se a utilizar outra técnica estatística de análise envolvendo diferenciação de expressão (*fold change*) de genes. Para isso, utilizou-se a linguagem estatística de programação R, com os pacotes do *Bioconductor* para fazer a análise da diferenciação de expressão (GENTLEMAN *et al.*, 2004).

A análise de genes, diferencialmente expressos, teve como objetivo identificar genes com diferenças de expressão entre as amostras, ou seja, se entre duas condições de tratamento (amostra de controle *versus* experimento), o nível de expressão é alterado significativamente, o gene é expresso diferencialmente.

Assim, para calcular o *fold change*, foram obtidas as médias dos valores de expressão de todas as amostras de RNAseq e Microarranjo. Então, relacionaram-se os 896 genes da Ontocancro 2.0 com a lista de genes expressos e, com a relação

dos genes encontrados, calculou-se o *fold change* que nada mais é do que a divisão do valor de expressão de um gene, nas amostras experimento, pela média de seu valor nas amostras controle.

Para melhor compreensão elaborou-se o fluxograma da Figura 29, que demonstra as etapas para obter a análise das técnicas de RNAseq e Microarranjos, utilizando o pacote *Bioconductor*.

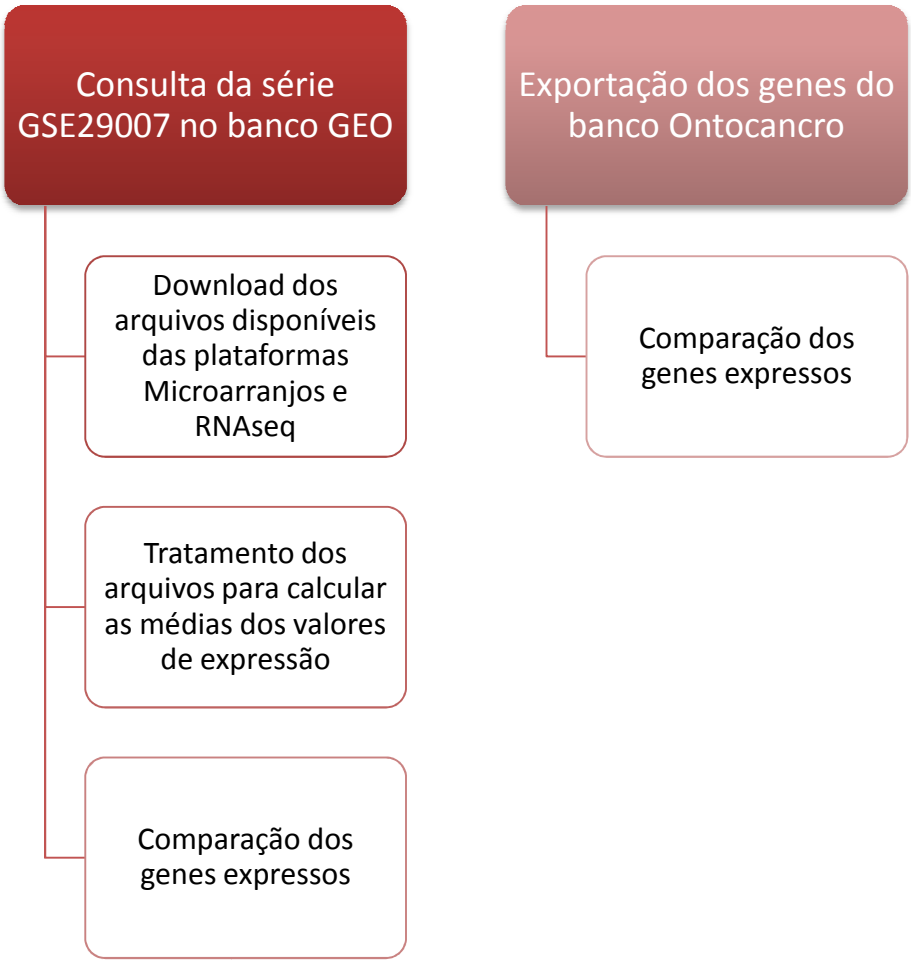


Figura 29 - Fluxograma da análise de RNAseq e Microarranjos pelo *Bioconductor*

Visto que as amostras disponíveis de RNAseq e Microarranjo não são exatamente correspondentes, escolheu-se as mais próximas, isto é, comparamos a amostra Fumante com câncer de pulmão *versus* Saudável não fumante para RNAseq e a de Fumante com câncer de pulmão *versus* ex-fumante sem câncer de

pulmão para Microarranjo. Desta forma, era esperado encontrar algumas diferenças entre os resultados de Microarranjo e RNAseq, mas alguns genes e vias deveriam estar igualmente alterados devido ao mesmo tipo de agravo celular presente nas amostras.

Fez-se a análise dos *top* 100 genes mais alterados listados em ordem de *fold change* de cada um. Comparou-se a lista dos genes repetidos em RNAseq e Microarranjo e foi encontrado um total de vinte e três genes significativamente expressos, entre os 100.

Partiu-se para a comparação dos resultados através da coincidência de vias Ontocancro com a lista dos *top* 100 genes mais alterados. Com base nas três vias da Ontocancro 2.0 (Resposta ao Dano, Apoptose, Ciclo Celular), verificou-se as sub-vias que os Top 100 genes mais alterados pertenciam.

Para RNAseq, observou-se um total de 39 sub-vias alteradas, onde 9 pertenciam a via de Reparo ao Dano (DDR), 15 subvias pertenciam a via de Apoptose e 15 sub-vias pertenciam a Via do Ciclo Celular, conforme representado pelo gráfico da Figura 30.

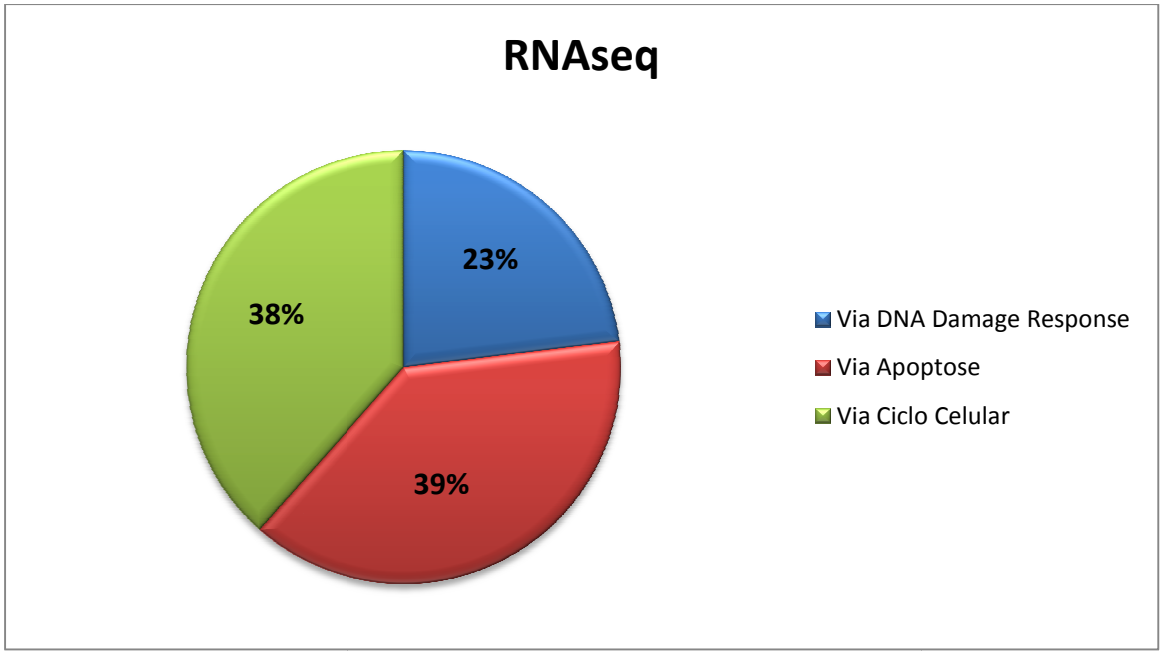


Figura 30 - Gráfico da análise de RNAseq

Para microarranjos, observou-se um total de 35 sub-vias, onde 6 sub-vias pertencem a via de Reparo ao Dano (DDR), 15 sub-vias pertencem a via de Apoptose e 14 sub-vias pertencem a via do Ciclo Celular, conforme representado pelo gráfico da Figura 31.

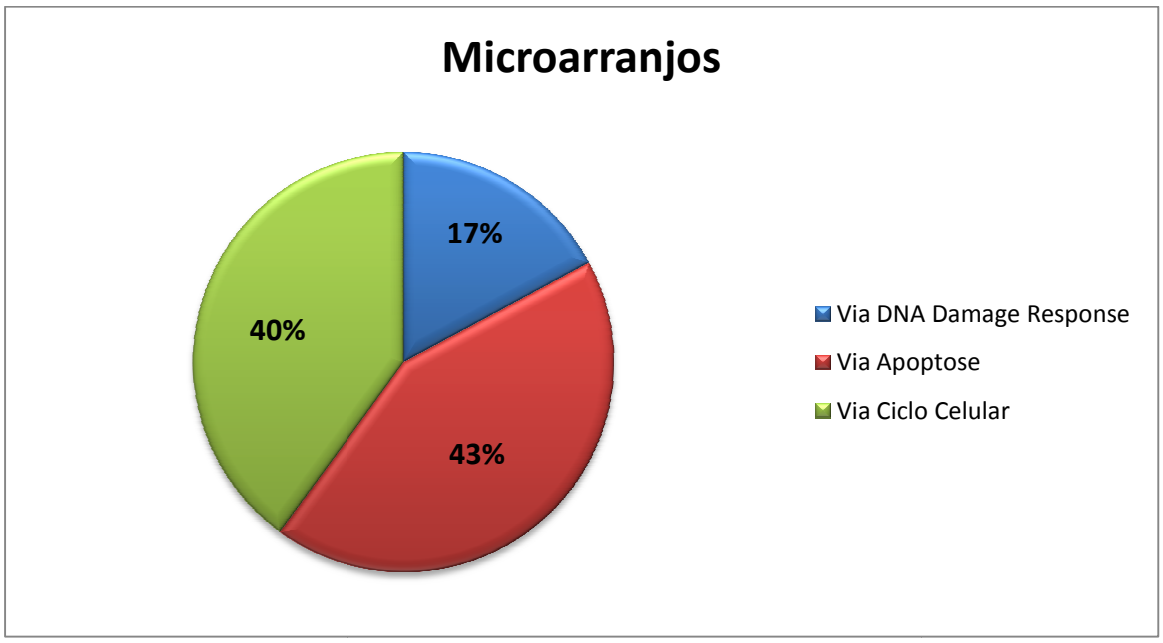


Figura 31 - Gráfico da análise de microarranjos

Os resultados desta análise permitem apenas estabelecer a ativação de vias de manutenção do genoma, geralmente associadas a *stress* induzido por danos ao DNA. No entanto, neste trabalho, não foi possível analisar as vias coincidentes entre as duas técnicas e se sua alteração se dá pelo mesmo tipo de regulação. Sugere-se a continuidade desta análise em trabalhos futuros para embasá-los do ponto de vista biológico e para estabelecer melhor a relação entre as duas técnicas de análise de expressão.

5.3. Sumário do Capítulo

Neste capítulo foram apresentados os resultados obtidos na comparação entre as ferramentas de análise estatística *ViaComplex* e *Ontocancro 2.0* (seção 5.1). E também, a comparação entre as técnicas de microarranjo e RNAseq (seção 5.2).

6. CONCLUSÃO

A ontologia Ontocancro foi projetada a partir da necessidade de integrar informações provenientes de diversos bancos de dados públicos, relativas a pesquisas sobre o desenvolvimento do câncer em seres humanos. Inicialmente, a Ontocancro 1.0 fornecia aos pesquisadores apenas dados não curados sobre genes e vias metabólicas ligadas ao processo carcinogênico, sem possibilitar uma análise precisa da expressão gênica, devido ao grande número de genes e vias registradas.

A partir de uma pesquisa qualitativa realizada por Halazonetis e colaboradores (2008), que propôs a existência de uma barreira de evolução tumoral, que pode ser observada nas alterações em vias e genes ligadas ao processo carcinogênico, buscou-se comprovar quantitativamente através de métodos de análise estatística da expressão de genes e vias, as alterações que ocorrem em tecidos de adenoma que atuam como uma barreira anticâncer.

Para esta comprovação, uma reestruturação do banco de dados foi necessária para conter os dados sobre estudos em tecidos de pacientes com câncer, disponibilizados no banco de dados GEO e produzidos através de técnicas de microarranjos (Affymetrix) e RNAseq (Illumina).

A contribuição desta dissertação consistiu na remodelagem da ontologia e da sua base de dados, obtendo um sistema dinâmico de fácil acesso, que permite o *download* dos dados armazenados e, principalmente, a obtenção dos resultados dos cálculos estatísticos da atividade relativa e diversidade relativa. Estes cálculos permitem caracterizar o perfil das vias metabólicas envolvidas nos mecanismos de manutenção do genoma.

Para a realização destas atividades, um estudo sobre as técnicas de microarranjos e de sequenciamento foi realizado, assim como das informações produzidas por elas. Após a definição dos requisitos necessários, novos conceitos foram introduzidos na ontologia, e posteriormente, as tabelas correspondentes foram adicionadas ao banco de dados. Também foram implementados os cálculos da atividade relativa e da diversidade relativa. Desta forma, através da interface *web* da Ontocancro 2.0, os pesquisadores podem gerar dados que auxiliem na análise do perfil das vias de estabilidade genômica.

Com a análise do perfil das vias de estabilidade genômica foi possível perceber que as mesmas apresentam desempenhos diferenciados em transcriptomas de adenomas e câncer. Nos estudos com os microarranjos de adenoma em tireoide e câncer de tireoide, realizados através dos cálculos de atividade relativa e diversidade, foi possível observar as evidências sobre a barreira de progressão do câncer proposta por Halazonetis e colaboradores (2008).

No entanto, os estudos com os dados da técnica de sequenciamento RNAseq ainda não foram conclusivos, necessitando de uma nova análise, uma vez que esta técnica é mais complexa e exige maior conhecimento sobre o seu funcionamento, porém a alta sensibilidade proposta pela técnica permite dados mais precisos, e por isso, a continuação da pesquisa utilizando esses dados torna-se imprescindível.

A utilização da Ontocancro 2.0 como ferramenta para análise das vias envolvidas com os mecanismos de manutenção do genoma (GMM) mostrou-se eficiente, pois percebeu-se que as vias de estabilidade genômica apresentam diferentes comportamentos em doenças que envolvem alterações genéticas. O estudo do transcriptoma, apresentando amostras com tecidos pré-cancerosos e amostras de tecidos com câncer, revela alterações nas vias de apoptose e ciclo celular. Para identificar os transcriptomas de doenças como o câncer, adenoma e doenças relacionadas à instabilidade do genoma, pode-se considerar diferenças nas vias GMM. Em câncer, as vias de apoptose e ciclo celular deverão aparecer com a sua significância aumentada, enquanto em adenomas é possível observar um aumento do valor de significância em vias de ciclo celular.

Como trabalhos futuros, sugere-se: a implementação de um módulo de importação de estudos, a realização de novas análises usando a técnica de RNAseq e a comparação com outras ferramentas de análise da expressão de genes e vias metabólicas, com o intuito de verificar também, os mecanismos que envolvem a destruição da barreira anticâncer.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBERTS, B. *et al. Biologia Molecular da Célula*. 5. ed. São Paulo: Artmed, 2010.

BARABÁSI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. fev. 2004, [S.l: s.n.], fev. 2004. p. 101–113. Disponível em: <http://dx.doi.org/10.1038/nrg1272>. Acesso em: 27 fev. 2013.

BECKER, J. D.; FEIJÓ, J. A. Profiling Genomes With Oligonucleotide Arrays, Affymetrix Core Facility. 2003, [S.l: s.n.], 2003. p. 24: 2–6.

CALZONE, L. *et al.* A comprehensive modular map of molecular interactions in RB/E2F pathway. *Molecular systems biology*, v. 4, p. 173, jan. 2008. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2290939&tool=pmcentrez&rendertype=abstract>. Acesso em: 9 ago. 2013.

CASTRO, M. A. A. *et al.* Impaired expression of NER gene network in sporadic solid tumors. 2007, [S.l: s.n.], 2007. p. 1859–1867. Disponível em: <http://nar.oxfordjournals.org/content/35/6/1859>.

CASTRO, M. A. A. *et al.* ViaComplex: software for landscape analysis of gene expression networks in genomic context. 1 jun. 2009, [S.l: s.n.], 1 jun. 2009. p. 1468–9. Disponível em: <http://bioinformatics.oxfordjournals.org/cgi/content/long/25/11/1468>. Acesso em: 17 nov. 2012.

CASTRO, MAURO A A *et al.* ViaComplex: software for landscape analysis of gene expression networks in genomic context. *Bioinformatics (Oxford, England)*, v. 25, n. 11, p. 1468–9, 1 jun. 2009. Disponível em: <http://bioinformatics.oxfordjournals.org/cgi/content/long/25/11/1468>. Acesso em: 17 nov. 2012.

CASTRO, MAURO A. A. *et al.* Impaired expression of NER gene network in sporadic solid tumors. *Nucleic Acids Research*, v. 35, p. 1859–1867, 2007. Disponível em: <http://nar.oxfordjournals.org/content/35/6/1859>.

COOPER, G. M.; HAUSMAN, R. E. *A célula: uma abordagem molecular*. [S.l.]: Artmed, 2007. p. 716. Disponível em: <http://books.google.com/books?id=JvpVOgAACAAJ&pgis=1>. Acesso em: 9 ago. 2013.

DE ROBERTIS, E. *Bases da Biologia Celular e Molecular*. 4. ed. Rio de Janeiro: Editora Guanabara Koogan, 2006.

FUTREAL, P. A. *et al.* A census of human cancer genes. *Nature reviews. Cancer*, v. 4, n. 3, p. 177–83, mar. 2004. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2665285&tool=pmcentrez&rendertype=abstract>. Acesso em: 29 out. 2012.

GALAMB, O. *et al.* Reversal of gene expression changes in the colorectal normal-adenoma pathway by NS398 selective COX2 inhibitor. *British journal of cancer*, v.

102, n. 4, p. 765–73, 16 fev. 2010. Disponível em: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4183>. Acesso em: 6 jan. 2013.

GENTLEMAN, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, v. 5, n. 10, p. R80, jan. 2004. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=545600&tool=pmcentrez&rendertype=abstract>. Acesso em: 7 ago. 2013.

GIBAS, C.; JAMBECK, P. *Desenvolvendo Bioinformática: ferramentas de software para aplicações em biologia*. Rio de Janeiro: Campus Elsevier, 2001.

GIORDANO, T. J. *et al.* Delineation, functional validation, and bioinformatic evaluation of gene expression in thyroid follicular carcinomas with the PAX8-PPARG translocation. *Clinical cancer research*, v. 12, n. 7 Pt 1, p. 1983–93, 1 abr. 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27155>. Acesso em: 25 nov. 2012.

GIORDANO, T. J. *et al.* Molecular classification and prognostication of adrenocortical tumors by transcriptome profiling. *Clinical cancer research*, v. 15, n. 2, p. 668–76, 15 jan. 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10927>. Acesso em: 5 dez. 2012.

GOHLMANN, H.; TALLOEN, W. *Gene Expression Studies Using Affymetrix Microarrays (Chapman & Hall/CRC Mathematical & Computational Biology)*. [S.l.]: Chapman and Hall/CRC, 2009. p. 359Disponível em: <http://www.amazon.com/Expression-Affymetrix-Microarrays-Mathematical-Computational/dp/1420065157>. Acesso em: 6 jan. 2013.

GRIFFITHS, A. J. F. *et al. Genética Moderna*. Rio de Janeiro: Guanabara Koogan, 2001.

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing. 1993, Dordrecht, Netherlands: Kluwer Academic, 1993.

HALAZONETIS, T. D.; GORGOULIS, V. G.; BARTEK, J. An oncogene-induced DNA damage model for cancer development. *Science (New York, N.Y.)*, v. 319, n. 5868, p. 1352–55, 7 mar. 2008. Disponível em: <http://www.sciencemag.org/content/319/5868/1352.abstract>. Acesso em: 2 nov. 2012.

HIRAOKA, N. *et al.* CXCL17 and ICAM2 are associated with a potential anti-tumor immune response in early intraepithelial stages of human pancreatic carcinogenesis. *Gastroenterology*, v. 140, n. 1, p. 310–21, jan. 2011. Disponível em: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19650>. Acesso em: 6 jan. 2013.

JEFFORD, C. E.; IRMINGER-FINGER, I. Mechanisms of chromosome instability in cancers. jul. 2006, [S.l: s.n.], jul. 2006. p. 59: 1–14. Disponível em: <http://dx.doi.org/10.1016/j.critrevonc.2006.02.005>. Acesso em: 27 mar. 2013.

KAO, J. *et al.* Cellular response to DNA damage. *Annals of the New York Academy of Sciences*, v. 1066, p. 243–58, dez. 2005. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16533929>. Acesso em: 8 ago. 2013.

LENHARD JR, R. E.; OSTEEN, R. T.; GANSLER, T. *Clinical Oncology (Book with CD-ROM for Windows & Macintosh)*. [S.l.]: Wiley-Blackwell, 2000. p. 800Disponível em: <http://www.amazon.com/Clinical-Oncology-CD-ROM-Windows-Macintosh/dp/0944235158>. Acesso em: 6 jan. 2013.

LIBRELOTTO, G. R. *et al. An Ontology to Integrate Transcriptomics and Interatomics Data Involved in Gene Pathways of Genome Stability*. [S.l.]: Springer Berlin Heidelberg, 2009. v. 5676. p. 164–167. Disponível em: <http://dl.acm.org/citation.cfm?id=1614645.1614668>. Acesso em: 6 jan. 2013.

LODISH, H. *Biologia Celular e Molecular*. [S.l.]: Artmed, 2005. p. 1054

LOEB, L. A.; LOEB, K. R.; ANDERSON, J. P. Multiple mutations and cancer. 4 mar. 2003, [S.l: s.n.], 4 mar. 2003. p. 776–81. Disponível em: <http://www.pnas.org/content/100/3/776.abstract>. Acesso em: 30 mar. 2013.

MARIONI, J. C. *et al.* RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. set. 2008, [S.l: s.n.], set. 2008. p. 1509–17. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2527709&tool=pmcentrez&rendertype=abstract>. Acesso em: 28 fev. 2013.

MEIER, W. eXist: An open source native xml database. 2000, [S.l: s.n.], 2000. p. 2593:169–183.

MICROSOFT. *Microsoft developer network - virtual library*. Disponível em: <http://msdn.microsoft.com/pt-br/library/ms178091.aspx.>.

MILANI, A. *MySQL Guia do Programador*. [S.l.]: Novatec, 2007.

MOMBACH, J. C. M. *et al.* On the absence of mutations in nucleotide excision repair genes in sporadic solid tumors. *Genetics and Molecular Research: GMR*, v. 7, n. 1, p. 152–60, jan. 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/18393219>.

PEREIRA, R. T. *Desenvolvimento de uma ferramenta para a análise de vias de estabilidade genômica*. 2011. Universidade do Minho, Portugal, 2011.

ROCHA, M. *Bioinformática: passado, presente e futuro!* Disponível em: <http://wiki.di.uminho.pt/twiki/pub/Education/MICEI/MatPedSem/seminario-05-19.pdf>. Acesso em: 8 out. 2009.

SABATEL, H. *et al.* Importance of PIKKs in NF-κB activation by genotoxic stress. *Biochemical pharmacology*, v. 82, n. 10, p. 1371–83, 15 nov. 2011. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/21872579>. Acesso em: 9 ago. 2013.

SETUBAL, J. C. *A origem e o sentido da bioinformática*. Disponível em: <http://www.comciencia.br/reportagens/bioinformatica/bio10.shtml>. Acesso em: 8 out. 2009.

SHANNON, C. E. A mathematical theory of communication. 1948, [S.l.]: The Bell System Technical Journal, 1948. p. 27:379–423.

SIMÃO, É. M. *Dinâmica da Transição Pré-Câncer para Câncer: Estudo da Expressão de Vias de Manutenção do Genoma*. 2012. Universidade Federal de Santa Maria, 2012.

SIMÃO, É. M. *et al.* Modeling the Human Genome Maintenance network. *Physica A*, v. 389, n. 19, p. 4188–4194, out. 2010. Disponível em: <http://dx.doi.org/10.1016/j.physa.2010.05.051>. Acesso em: 6 jan. 2013.

SPLENDORE, A. Para que existem as regras de nomenclatura genética? jun. 2005, [S.l: s.n.], jun. 2005. p. 148–152. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-84842005000200020&lng=en&nrm=iso&tlng=pt>. Acesso em: 15 abr. 2013.

STRACKER, T. H.; PETRINI, J. H. J. The MRE11 complex: starting from the ends. fev. 2011, [S.l.]: Nature Publishing Group, fev. 2011. p. 90–103. Disponível em: <http://dx.doi.org/10.1038/nrm3047>. Acesso em: 8 ago. 2013.

UETZ, P.; IDEKER, T.; SCHWIKOWSKI, B. Visualization and Integration of Protein-Protein Interactions. Protein -protein interactions - a molecular cloning manual. *Cold Spring Harbor Laboratory Press*, 2005. Disponível em: <http://130.203.133.150/viewdoc/summary?doi=10.1.1.23.9572>. Acesso em: 11 abr. 2013.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, v. 10, n. 1, p. 57–63, jan. 2009. Disponível em: <http://dx.doi.org/10.1038/nrg2484>. Acesso em: 28 jan. 2013.

WATSON, J. D. The human genome project: past, present, and future. *Science (New York, N.Y.)*, v. 248, n. 4951, p. 44–9, 6 abr. 1990. Disponível em: <http://www.sciencemag.org/content/248/4951/44.long>. Acesso em: 6 jan. 2013.

WATSON, J. D.; CRICK, F. H. C. A Structure for Deoxyribose Nucleic Acid. 1953, [S.l: s.n.], 1953. p. 737–738. Disponível em: <http://www.nature.com/nature/dna50/watsoncrick.pdf>.

WEINBERG, R. A. *The biology of cancer*. [S.l.]: Garland Science, 2007.

ZAHA, A.; FERREIRA, H. B.; PASSAGLIA, L. M. P. *Biologia Molecular Básica*. 4. ed. Porto Alegre: Artmed, 2012.