PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA E EXTENSÃO

ÁREA DE CIÊNCIAS TECNOLÓGICAS

Curso de Mestrado em Nanociências

HELENO CARMO BORGES CABRAL

A DEFINIÇÃO DE UMA ONTOLOGIA PARA INTEGRAR DADOS DE INTERATOMA E TRANSCRIPTOMA DE CÂNCER

Santa Maria, RS

2010

HELENO CARMO BORGES CABRAL

A DEFINIÇÃO DE UMA ONTOLOGIA PARA INTEGRAR DADOS DE INTERATOMA E TRANSCRIPTOMA DE CÂNCER

Dissertação apresentada ao Curso de Mestrado em Nanociências do Centro Universitário Franciscano de Santa Maria como requisito parcial para obtenção do título de Mestre em Nanociências.

Orientador: Prof° Dr° **GIOVANI RUBERT LIBRELOTTO**

Santa Maria, RS

2010

UNIFRA
CENTRO UNIVERSITÁRIO FRANCISCANO

PRPGPE
Pró-Reitoria de Pós Graduação, Pesquisa e Extensão

ÁREA DE CIÊNCIAS TECNOLÓGICAS

Mestrado em Nanociências

A COMISSÃO EXAMINADORA, ABAIXO-ASSINADA, APROVA A DISSERTAÇÃO:

A DEFINIÇÃO DE UMA ONTOLOGIA PARA INTEGRAR DADOS DE INTERATOMA E TRANSCRIPTOMA DE CÂNCER

Elaborada por

HELENO CARMO BORGES CABRAL

COMISSÃO EXAMINADORA

Prof. Dr. Giovani Rubert Librelotto – UFSM

Prof. Dr. Eugênio Simonetto - UDESC

Profª. Drª. Marta Palma Alves - UNIFRA

Santa Maria, 23 de junho de 2010.

Ficha Catalográfica

C117d
Cabral, Heleno Carmo Borges
A definição de uma ontologia para integrar dados de interatoma e transcriptoma de câncer / Heleno Carmo Borges Cabral; orientação Giovani Rubert Librelotto. – Santa Maria, 2010.
78 f.: il.
Dissertação (Mestrado em Nanociências) – Centro Universitário Franciscano, 2010.

1. Bioinformática. 2. Câncer. 3. Ontocancro - ontologia. I. Librelotto, Giovani Rubert. II. Título.

CDU 004:573:616-006

Elaborada pela Bibliotecária Zeneida Mello Britto CRB10/1374

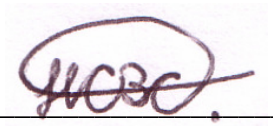
Biblioteca
Termo de autorização

Programa de Pós-Graduação *Stricto Sensu* Mestrado Acadêmico em Nanociências.

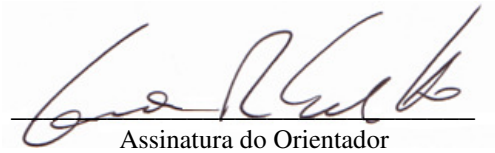
Título da Dissertação: **A DEFINIÇÃO DE UMA ONTOLOGIA PARA INTEGRAR DADOS DE INTERATOMA E TRANSCRIPTOMA DE CÂNCER.**

Orientador: Dr. **GIOVANI RUBERT LIBRELOTTO.**

Eu, **HELENO CARMO BORGES CABRAL**, R. G. n° **5065001348**, autor da Dissertação acima citada, autorizo ao Centro Universitário Franciscano – UNIFRA - a disponibilizar, gratuitamente, sem ressarcimento dos direitos autorais, o documento, em meio eletrônico, na Rede Mundial de Computadores (Internet) para fins de leitura e/ou impressão e para divulgação da produção científica gerada na Instituição, a partir desta data.



Assinatura do Autor



Assinatura do Orientador

Santa Maria/RS
Local

23 / 06 / 2010.
Data

Mesmo que tu já tenhas feito uma longa caminhada, há sempre um caminho a fazer.

(Sto. Agostinho).

AGRADECIMENTOS

Dedico meus agradecimentos a todos que de alguma forma colaboraram com a realização deste trabalho:

– a Deus, sou grato por todas as oportunidades que me foram dadas;

– aos colegas de pós-graduação da UNIFRA;

– aos demais Professores do mestrado em Nanociências, que colaboraram na minha formação acadêmica;

– aos meus familiares, pelo apoio em minhas decisões, especialmente a minha esposa Maíra pelo apoio incondicional nas horas de incertezas;

– à Profª. Drª. Juliana Vizzotto, pela co-orientação neste trabalho;

– ao Profº. Dr. Giovani Rubert Librelotto, pela orientação, tolerância, paciência, amizade e pelos ensinamentos.

- aos colegas e amigos André Seixas, Elenice Martins, Fabrício Dutra, José Carlos Mombach, Marialva Sinigaglia, Karlise Soares, Éder Simão, Jonas Gassen, Jonathan Alves, Pedro Rocha e Marcelo Roza pela constante ajuda e participação neste trabalho.

RESUMO

A Ontocancro é uma ontologia armazenada em um banco de dados de conhecimento projetada para ser a fonte de informação referente a integração de dados de interatoma e transcriptoma envolvidos em vias metabólicas de mecanismo de manutenção do genoma humano (GMM). Esse mecanismo de manutenção são críticos para homeostase celular desde o seu mau funcionamento, o que pode causar câncer. O reparo, a apoptose e as vias de estabilidade cromossômicas compreendem o cerne do GMM. A informação sobre essas vias metabólicas são disseminadas em vários bancos de dados, como o *NCI-Nature*, o *BioCarta*, o *KEGG*, o *Reactome*, o *Prosite* e o *GO*, entre outros. A ontologia Ontocancro foi criada com a intenção de integrar a informação sobre os genes envolvidos em GMM a partir de diversos bancos de dados curados. Essa integração de dados é complexa pela falta de um vocabulário sobre os dados biológicos e a necessidade constante de atualização destes dados. Para sanar essas duas dificuldades, a Ontocancro foi criada. Adicionalmente, ela permite a integração de dados oriundos de transcriptoma obtidos a partir da plataforma *Affymetrix* com os dados de interatoma obtidos a partir do banco de dados chamado *STRING*, o qual possui informação sobre as interações entre as proteínas. Portanto, este trabalho apresenta a integração de dados obtidos de sistemas de informação biológicos usando o paradigma ontológico, de forma a integrar os dados envolvidos em interatoma e transcriptoma em vias metabólicas de estabilidade do genoma.

Palavras-chave: Bioinformática, Câncer, Ontologias.

ABSTRACT

Ontocancro is an ontology stored in a knowledge database designed to be a source of information to integrate transcriptomics and interatomics data involved in gene pathways of genome maintenance/stability mechanisms (GMM). Genome maintenance mechanisms are shown to be critical for cell homeostasis since their malfunctioning can predispose to cancer. Repair, apoptosis and chromosome stability pathways comprise the cornerstone of GMM. The information about these pathways are disseminated in various databases as NCI-Nature, BioCarta, KEGG, Reactome, Prosite, GO and others. Ontocancro was created with the intention of integratin the information of genes involved in GMM from several curated databases. This data integration is difficult for biological data lack a unified vocabulary and need constant update what is provided by Ontocancro. Additionally, it allows the integration of transcriptome data provided by some Affymetrix microarrays platforms with interactome data from the STRING database, which has information about protein interactions. So, this work shows the integration of data from biological information systems using the ontology paradigm, in order to integrate transcriptomics and interatomics data involved in gene pathways of genome stability.

Keywords: *Bioinformatics, Cancer, Ontologies.*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Um dos desafios mais importantes para a biologia da era pós-genômica é a compreensão da estrutura e do comportamento de redes complexas de interações moleculares que controlam o comportamento das células [BARABÁSI, 2004]. O tamanho e a complexidade dos dados biológicos coletados durante os últimos anos incluem informações que requerem uma abordagem integradora [UETZ et al., 2002]. Isto impõe, aos cientistas da computação e biólogos, a procura por métodos inovadores para tratar estes dados para que se aumente a compreensão dos processos biológicos que operam dentro da célula.

Contudo, esta tarefa de integração é difícil, pois os dados biológicos estão disseminados em diversos bancos de dados. Estes bancos possuem diferentes sistemas de gerenciamento, formatos e formas de representar os dados armazenados. A maioria destes está acessível por arquivos de texto ou por interfaces Web que permitem alguns tipos de consultas pré-definidas. Os dois maiores problemas envolvidos aqui são a dificuldade em processar os dados quando se está trabalhando com formatos heterogêneos e com inconsistências, devido à ausência de um vocabulário unificado.

Em bioinformática, as ontologias são cruciais para a manutenção da coerência dos dados em uma coleção de conceitos complexos e seus relacionamentos. Uma ontologia é uma especificação explícita de uma conceitualização [GRUBER, 1993]. Enquanto vocabulários controlados somente restringem as palavras a serem utilizadas em um determinado domínio, as ontologias estendem esta característica simples dos vocabulários controlados e permitem uma especificação formal de termos e seus relacionamentos. Isto torna possível compartilhar e reutilizar o conhecimento. Elas suportam a interoperabilidade entre os sistemas e também permitem inferências sobre o conhecimento representado [GOBLE et al., 2001].

O mapeamento dos genes de um organismo traz respostas a diversas questões que há anos foram formuladas por cientistas ou indivíduos não leigos. Essas questões podem ser desde a curiosidade sobre do que os organismos são formados, até a descoberta das causas de uma doença congênita. Com a bioinformática, foram criadas técnicas para o aprimoramento do mapeamento desses genes, assim como das proteínas que o código genético é capaz de produzir.

A partir da necessidade de integrar diversas informações referentes aos genes ligados ao câncer, este trabalho apresenta a ontologia Ontocancro. Ela visa fornecer dados

centralizados que permitam uma análise consistente de informações extraídas de outros bancos de dados públicos, tornando possível o compartilhamento e a reutilização deste domínio de conhecimento.

1.1. MOTIVAÇÃO

Este projeto propõe-se a investigar o funcionamento de redes biológicas de genes envolvidos em câncer, utilizando-se uma ontologia para comunicação entre sistemas, pessoas e organizações, de forma a suportar projetos e desenvolvimentos de sistemas baseados em conhecimento genético. Segundo Gómes-Péres, as ontologias são desenvolvidas para aplicações científicas, sem a preocupação de reutilização e compartilhamento; estão dispersas por diversos servidores; a formalização difere dependendo do servidor na qual a ontologia está armazenada; são usualmente descritas com diferentes níveis de detalhamento e não há um formato comum para a representação de informações relevantes, sendo que os usuários definem a sua maneira quais as ontologias são melhores aplicadas ao seu propósito.

Neste caso, o projeto abordará uma ontologia para a integração de dados de genes potenciais ao câncer, sendo a meta principal buscar uma padronização para esta ontologia, integrando diversos bancos de dados, homogeneizando-os com a ontologia *MONET* (*Molecular Network*) [SILVA *et al.,*2006], na tentativa de obterem-se redes biológicas mais complexas.

1.1.1. Objetivo Geral

Criar de uma ontologia que permita a representação do conhecimento de redes moleculares e sua expressão agregada a um sistema de informação de forma a facilitar a integração de dados originários de bancos de dados públicos diferentes em um banco de dados único.

1.1.2. Objetivos Específicos

- Identificar as ontologias do domínio de interatoma e transcriptoma existentes;

- Identificar os níveis de compatibilização, codificação e identificação dos genes;
- Desenvolver uma ontologia para integrar esses dados de interatoma e transcriptoma de câncer;
- Criar uma base de dados local para a integração dos dados dos bancos de bioinformática;
- Realizar análise topológica da rede dos genes envolvidos em câncer;
- Criar uma ferramenta gráfica para a análise gênica das vias metabólicas.

1.2. ESTRUTURA DA DISSERTAÇÃO

No primeiro capítulo será abordada a motivação e objetivos deste trabalho, o segundo capítulo tratará o tema bioinformática, onde será possível demonstrar uma visão sobre esta nova área que cresce cada vez mais em nosso meio, incluindo definições de genes, proteínas, vias e uma visão geral sobre interatoma e transcriptoma, bem como o que é câncer.

No terceiro capítulo, será mostrado um apanhado de banco de dados, bem como sua função e particularidade de cada um. A ontologia Ontocancro será enfatizada no quarto capítulo, incluindo sua estrutura, navegação no site e resultados obtidos, enquanto no quinto capítulo será abordada a utilização do software Medusa para análise gênica, fazendo-se referência sua interação com a Ontocancro, apresentando no capítulo seguinte alguns trabalhos relacionados.

2. BIOINFORMÁTICA

A bioinformática tem seu início por volta do ano de 1953, quando os cientistas James Dewey Watson e Francis Crick modelaram em uma estrutura de ferro e madeira uma dupla hélice a fim de representar a molécula de DNA [SETUBAL, 2003], ilustrada na Figura 1. A publicação do seu trabalho na revista científica *Nature* foi um dos grandes marcos na história da biologia no século passado.

Figura 1 - Crick e Watson com o modelo original da dupla hélice do DNA, em 1953.

A molécula de DNA denominada ácido desoxirribonucléico em português, contém o código hereditário (genético) de cada ser e pelo modelo proposto por Watson e Crick ela é constituída por duas cadeias paralelas de nucleotídeos unidas em sequência num espaço disposto helicoidal, ou seja, giram em torno de um eixo. Através deste modelo também foi possível entender como ocorrem as mutações celulares, as codificações e a replicação das moléculas. Posteriormente surgiram outros métodos no sequenciamento dos polímeros de DNA, o que permitiu o estudo das formas mais simples que o compõe. Esses polímeros, desde então, passaram a serem os principais objetos de estudos na nova ciência a Biologia Molecular, sendo que mais de 18 milhões dessas sequências já foram produzidas e estão disponíveis em banco de dados públicos [FILHO, 2002].

Juntamente com a evolução da informática na década de 90, começaram a surgir sequenciadores automáticos de DNA, o que aumentou consideravelmente a quantidade de

sequências genéticas a serem armazenadas, exigindo assim, cada vez mais recursos computacionais para seu armazenamento e manipulação. Em outras palavras pode-se afirmar que somente quando os computadores estavam suficientemente munidos com uma estrutura que pudesse processar milhões de sequências puderam-se obter resultados significativos nas pesquisas com o genoma humano.

Segundo Setubal (2003) em sua análise ao observar que se o sequenciamento automático do DNA tivesse chegado com uns 20 anos de antecedência, não haveria computadores com o poder de processamento suficiente para poder manipular e gerenciar os dados gerados. Em uma analogia pode-se remeter a década de 70, onde um computador de grande porte tinha alguns *kilobytes* de memória e não seria capaz de processar sequer o genoma de um vírus que pode chegar a 20 *kilobases* (20 mil bases ou 20KB).

> Com os computadores mais velozes e de menor custo, juntamente com as pesquisas nas áreas da biologia molecular gerando milhares de informações para serem catalogadas e organizadas, nascia a fusão destas ciências, a bioinformática, uma ciência que envolve diversas linhas de conhecimento, como a engenharia de software, a matemática, física, química, estatística, a ciência da computação e a biologia molecular [FILHO 2002, p. 1].

Com o surgimento dessa nova área, os primeiros profissionais eram basicamente compostos por médicos, físicos, biólogos e outros profissionais basicamente de áreas diferentes da biologia, porém com certos conhecimentos na área da informática. Segundo Filho [FILHO, 2002] havia uma grande dificuldade entre a comunicação de biólogos com cientistas da computação, já que os biólogos levam em consideração resultados como incertezas e erros que podem ocorrer na prática, enquanto os cientistas da computação procuram sempre uma solução direta para um problema.

Através desta variante em seus métodos de pesquisa, houve a necessidade de um profissional com o perfil que tivesse o conhecimento suficiente em ambas as áreas, que fosse capaz de pegar um problema biológico real, analisar quais seriam suas opções e métodos de pesquisa, desenvolver uma solução através de uma abordagem computacional para avaliar os problemas identificados, surgindo assim um profissional que englobassem essas duas qualidades, os bioinformatas.

Pode-se citar como um exemplo clássico do trabalho desse profissional quanto se faz referência a um Sistema de Gerenciamento de Banco de Dados (SGBD) que é de extrema importância a fim de suportar uma demanda robusta no gerenciamento dessas informações que muitas vezes geram *terabytes* de dados biológicos e necessitam de um repositório robusto e seguro.

> No início, o arquivamento de dados nesta área era realizado por grupos de pesquisa individuais, motivado pelo interesse da ciência naqueles dados, com o aumento da demanda por profissionais e equipamentos, juntamente com uma maior ênfase em computação, os arquivamentos passaram a ser de responsabilidade de projetos de grande escala [LESK, 2008, p.141].

De acordo com o grande sucesso do projeto Genoma Humano [NHGRI, 2009] que inicialmente sequenciou um pouco mais de 20% (25.000 genes) do total do material genético humano, a bioinformática mesmo enfrentando problemas por razões de limitações das tecnologias atuais ainda não tenha conseguido sequenciar aproximadamente 1% do genoma é uma área muito influente, considerando dessa forma a grande quantidade de benefícios que trouxe à biologia a fim de facilitar suas pesquisas, desde uma simples planilha eletrônica para a organização de seus dados obtidos até um potente simulador que pode representar os efeitos de uma doença, bem como o teste de possíveis curas.

2.1. GENE, DNA E RNA

O gene é um código distinto, um seguimento de um cromossomo. Segundo o projeto Genoma [MAGATÃO, 2008] um gene pode ser definido como uma unidade fundamental, física e funcional da hereditariedade, ou seja, uma informação capaz de produzir uma determinada proteína ou atuar no controle de uma característica, como por exemplo, a cor dos olhos.

Cada gene tem em sua formação, uma sequência específica de ácidos nucléicos, onde encontramos as informações genéticas. Como se pode analisar na Figura 2, existem dois tipos de ácidos nucléicos, o DNA (ácido desoxirribonucléico) e o RNA (ácido ribonucléico). Em uma análise mais detalhada, pode-se perceber que o DNA e RNA são longos polímeros de

unidades simples, os monômeros do nucleotídeo, porém o RNA é formado por moléculas muito inferiores das que compõem o DNA.

O DNA do ponto de vista biológico é composto de moléculas que contém as instruções genéticas que coordenam o funcionamento de todos os seres vivos, enquanto o RNA é definido como responsável pela síntese de proteínas na célula. Essas proteínas, nada mais são que compostos orgânicos responsáveis por determinadas funções na célula, como por exemplo, a queratina, que é a principal proteína do cabelo humano, garantindo-lhe rigidez, consistência e elasticidade, pode-se ainda citar como proteínas, os anticorpos responsáveis pela defesa do organismo.

Um gene geralmente codifica uma proteína, mas devido a algumas exceções, existem genes que podem codificar mais de uma proteína, o relacionamento existente é de 1 para 1 entre gene e proteína e muitas vezes o gene tem sua identificação pela proteína que o ativa.

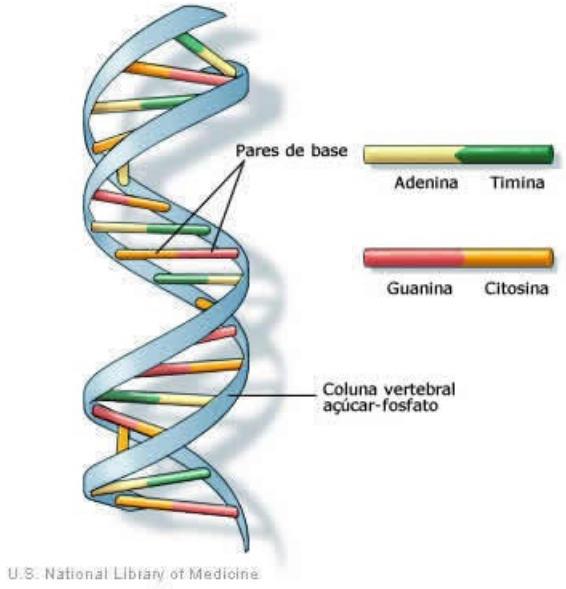


Figura 2 – Estrutura do DNA [NLM, 2009].

Como se percebe, o conhecimento que se tem sobre a informação contida no DNA ainda é muito restrito. Mesmo que nas últimas décadas tenha-se avançado consideravelmente na compreensão dessa informação, ainda há muito sobre o qual não se sabe, por exemplo, sobre a função da grande parte do DNA de eucariotos que não codifica proteínas. Por outro lado, os dados propriamente ditos são de composição simples, consistindo de apenas quatro

tipos de nucleotídeos, que são referenciados pelas bases nitrogenadas que os diferenciam: adenina, citosina, guanina e timina, ou, respectivamente, A, C, G e T [NLM, 2009].

O DNA é geralmente encontrado como uma molécula em forma de dupla hélice, mas geralmente quando se trata dos dados, trata-se apenas de uma das fitas por vez. Uma sequência de DNA é representada por uma sequência de texto contendo as letras que representam as bases dessa fita. Além das quatro letras que codificam os diferentes tipos de bases, por sua vez são usadas mais 11 letras para especificar ambiguidade entre as bases, como se pode notar na Tabela 1.

Tabela 1 – Ambiguidade de nucleotídeos

SÍMBOLO	NUCLEOTÍDEO
R	A ou G
Y	C ou T
W	A ou T
S	G ou C
K	G ou T
M	A ou C
B	C, G ou T
D	A, G ou T
H	A, C ou T
V	A, C ou G
N	A, C, G ou T

Nas estruturas de uma molécula de DNA, cada base em uma das fitas corresponde uma outra base da outra fita, sendo que as bases estão ligadas por pares, especificamente de A-T (adenina e timina) e C-G (citosina e guanina). A estrutura de dupla fita do DNA permite não só que a molécula seja mais estável, mas também funciona como um dispositivo de correção de erro no caso de dano a alguma base, por exemplo, causado pelo excesso de radiação ultravioleta do sol.

Ao longo do DNA, estão codificados os genes, que nada mais são do que unidades hereditárias dos organismos. Apesar de estes poderem ser definidos de diversas maneiras, tratam-se como unidades do DNA que contêm instruções para a codificação de uma proteína, apesar de alguns deles produzirem RNAs que não codificam proteínas.

A principal função do DNA é armazenar toda informação genética de um organismo e, toda vez que uma célula se divide, seu DNA é duplicado num processo chamado

de “replicação” e a produção de proteínas na célula se dá a partir da cópia de uma sequência do DNA para o RNA, em um processo chamado de “transcrição” como se observa na Figura 3.

Esta molécula de RNA é bastante semelhante à do DNA. Entretanto, no RNA encontramos a base U (uracil) no lugar de T (timina). Entre outras diferenças, o RNA é em geral encontrado como uma molécula de fita simples.

Embora o RNA possa ser de diversos tipos e apresentar diversas funções, os RNAs que correspondem a genes que codificam proteínas são denominados de RNA mensageiros, ou mRNA.

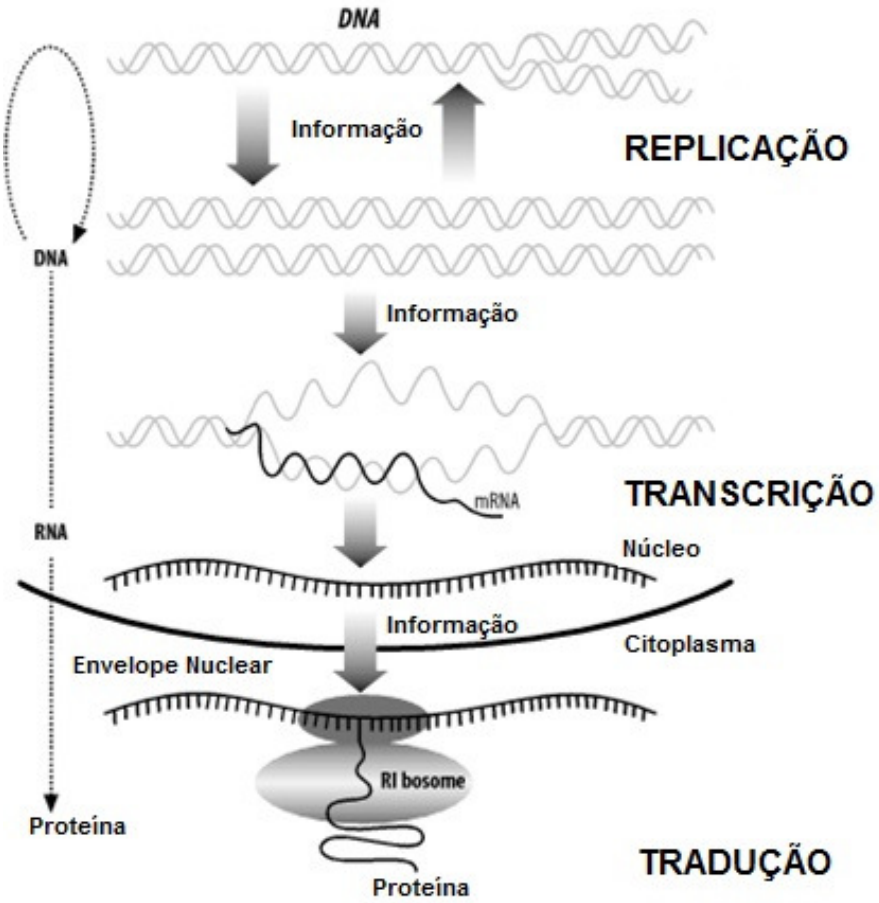


Figura 3 – Fluxo de informação na célula [BEDELL et al, 2003].

2.2. PROTEÍNA

Entre outras funções, as proteínas são constituintes estruturais do “maquinário” da célula. Elas são moléculas que diferem quimicamente do DNA e RNA, pois possuem

aminoácidos na sua composição ao invés de ácidos nucléicos. As proteínas têm a propriedade de se "dobrar" em formas tridimensionais bastante específicas, que dependem de sua sequência de aminoácidos. Deste modo, a sequência de aminoácidos determina a forma de uma proteína e a forma determina sua função, de modo que há proteínas que desempenham as mais diversas funções num organismo. Com isso, percebe-se que, enquanto nesse contexto, o DNA e o RNA são utilizados principalmente para armazenamento e transporte de informações, deste modo, as proteínas são o resultado desse processo, mostrando-se responsáveis por inúmeros processos no organismo.

Por sua vez os aminoácidos que compõem as proteínas, são codificados cada um, por três nucleotídeos. Como são quatro os tipos de nucleotídeos, existem assim 64 combinações possíveis de códons (grupos de três nucleotídeos). Entretanto as 64 combinações, conforme o código genético (Figura 4), codificam apenas 20 aminoácidos (Tabela 2), sendo que um deles, a Metionina indica o início da codificação de uma proteína ("*start*") e outros 3 códons indicam o final dessa codificação ("*stop*"). Portanto, vários desses aminoácidos são codificados com redundância por mais de um tipo de códon.

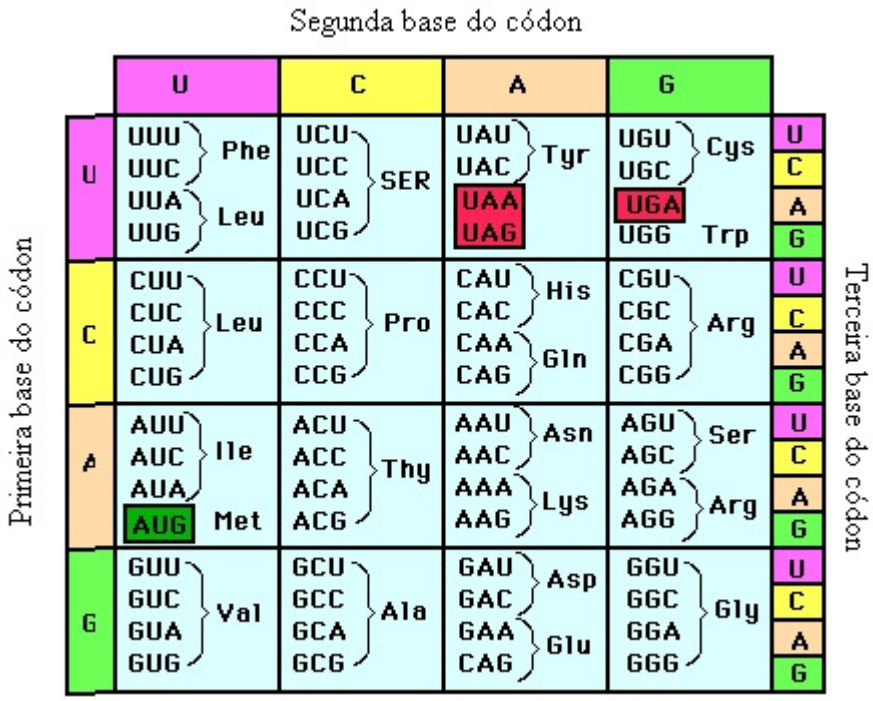


Figura 4 – Código genético padrão do site: www.icb.ufmg.br/prodabi/grupo6/codon1.gif

A Tabela 2 lista os 20 aminoácidos e suas respectivas representações, como se pode perceber a Figura 4 ilustra a conversão de nucleotídeos em aminoácidos usando o código genético padrão.

Tabela 2 – Principais aminoácidos

NOME	REPRESENTAÇÃO
Glicina	Gly, Gli ou G
Alanina	Ala ou A
Leucina	Leu ou L
Valina	Val ou V
Isoleucina	Ile ou L
Prolina	Pro ou P
Fenilalanina	Phe ou Fen
Serina	Ser ou S
Teonina	Thr, The ou T
Cisteina	Cys, Cis ou C
Tirosina	Tyr, Tir ou Y
Asparagina	Asn ou N
Glutamina	Gln ou Q
Aspartato	Asp ou D
Glutamato	Glu ou E
Arginina	Arg ou R
Lisina	Lys, Lis ou K
Histidina	His ou H
Triptofano	Trp, Tri ou W
Metionina	Met ou M

Além das proteínas apresentarem estruturas internas chamadas de domínios, elas podem ser vistas como um trecho da proteína com uma função ou estrutura distinta (ou ambas).

Geralmente, quando se vê uma unidade estrutural específica dentro de uma proteína, essa costuma ter uma função específica associada a ela. Sendo assim, os domínios determinam as características específicas de cada proteína e uma proteína pode conter um ou mais domínios expressos.

2.3. VIAS METABÓLICAS

Uma via metabólica é formada por um grupo de proteínas que são responsáveis por determinadas funções, um exemplo disso é uma via de reparo que ativa as proteínas responsáveis pela função de reparação da célula ou parte do DNA.

O metabolismo de todos os organismos é caracterizado por uma rede complexa de reagentes conectados por reações químicas catalisadas por enzimas. 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. Existem muitos organismos com genomas sequenciados onde as proteínas codificadas estão sendo determinadas.

Barabási e colaboradores [JEONG, 2000] propuseram uma representação gráfica da rede metabólica onde os nós representam os substratos, que estão ligados uns aos outros através de conexões que são as reações metabólicas propriamente ditas. Este trabalho demonstrou que existem princípios universais, e não históricos, na descrição da bioquímica dos seres vivos. Contudo, foram centrados esforços na compreensão da rede metabólica como um todo, não explicitando o papel que as enzimas desempenham, estando essas codificadas no genoma e historicamente construídas pela evolução.

Do ponto de vista biológico, contudo é de importância central compreender o papel que as enzimas desempenham e principalmente determinar como a evolução teceu esta complexa teia de inter-relações.

Em relação ao funcionamento das Redes de Reparo de DNA em câncer, pode-se dizer que a célula possui diferentes mecanismos de reparo para proteger o DNA contra danos, como as quebras de cadeias de DNA ocasionado pela radiação ultravioleta. Os sistemas de reparo se constituem em redes genéticas especializadas nesta proteção, uma vez que impedem que diferentes tipos de danos sejam fixados no material genético. Em células cancerosas, essas redes de proteção não funcionam corretamente, ocasionando uma série de mutações. Sabe-se que os genes de uma das cinco redes de reparo, chamada de Reparo por Excisão de Nucleotídeos (NER), não possui mutações catalogadas causalmente relacionadas a câncer somático, acreditando-se que ela não estaria envolvida no aparecimento de células cancerosas [FUTREAL, 2004].

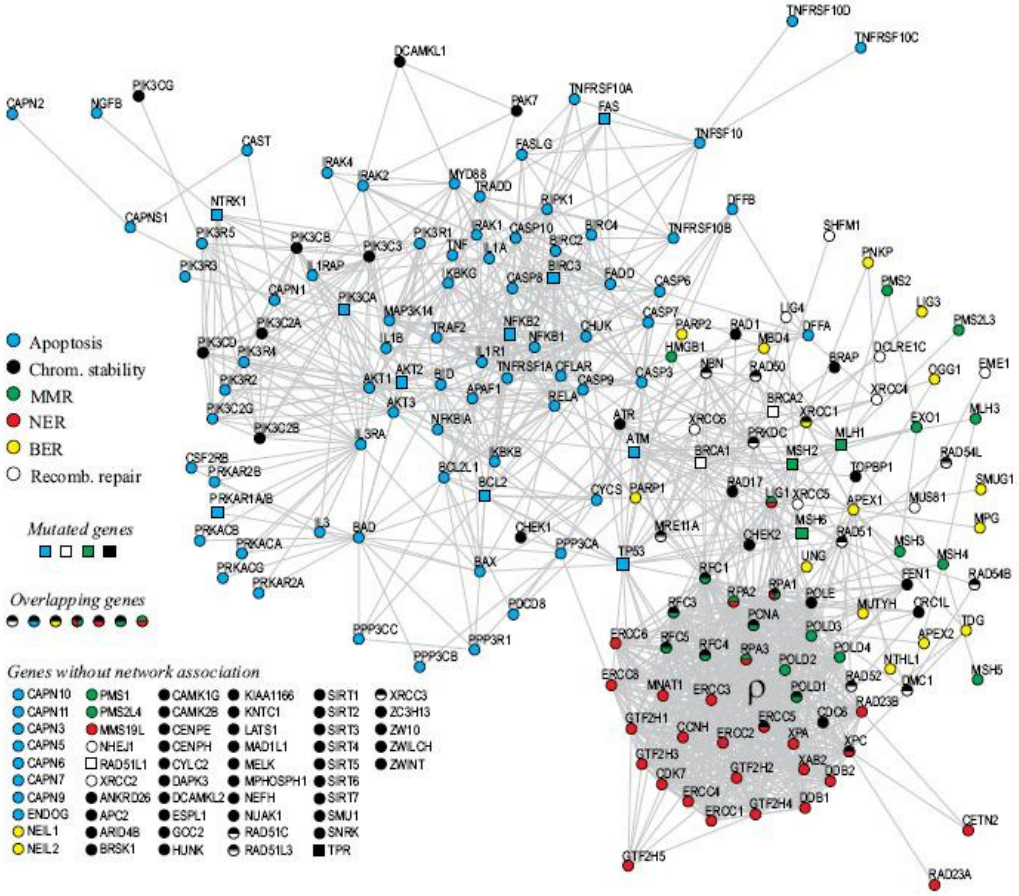


Figura 5 – Redes genéticas envolvidas na apoptose e reparo de DNA. Os genes de apoptose estão representados em azul, os da rede NER em vermelho e os outros genes estão em outras cores conforme a figura. [FUTREAL, 2004].

Investigou-se, então, o funcionamento dos genes desta rede de reparo usando os dados públicos do Projeto Genoma do Câncer Humano disponíveis na Internet [CASTRO, 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.

Construiu-se a rede de interação entre estes genes. Assim, 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 mais comum em câncer, exercendo influência sobre a transcrição, ciclo celular, apoptose e angiogênese. O grafo de interações entre a rede dos genes de reparo e da rede de apoptose pode ser visto na Figura 5.

2.4. INTERATOMA E TRANSCRIPTOMA

Sendo a célula um organismo composto por um número muito grande de genes, e cada gene com uma determinada função e o mesmo podendo fazer parte de mais de uma via ou função segundo Kasahara e colaboradores (2010), dá-se o nome interatoma a todas as interações proteína-proteína de um gene, enquanto o transcriptoma (transcrição do DNA), nada mais é que a síntese dos diferentes tipos de RNA a partir de um molde de DNA, usando as regras da complementaridade. A transcrição de um segmento se inicia quando a RNA polimerase reconhece e liga-se a sequências específicas de nucleotídeos em uma região especial, no início do gene, denominada promotor.

Além destas sequências, o promotor engloba o ponto de início como sendo o primeiro par de bases a ser transcrito em RNA, a partir daí a RNA polimerase move-se ao longo do molde, sintetizando RNA, até alcançar outra sequência específica que sinaliza o término da transcrição, ou seja, a unidade de transcrição estende-se do ponto de início no promotor, até o terminador.

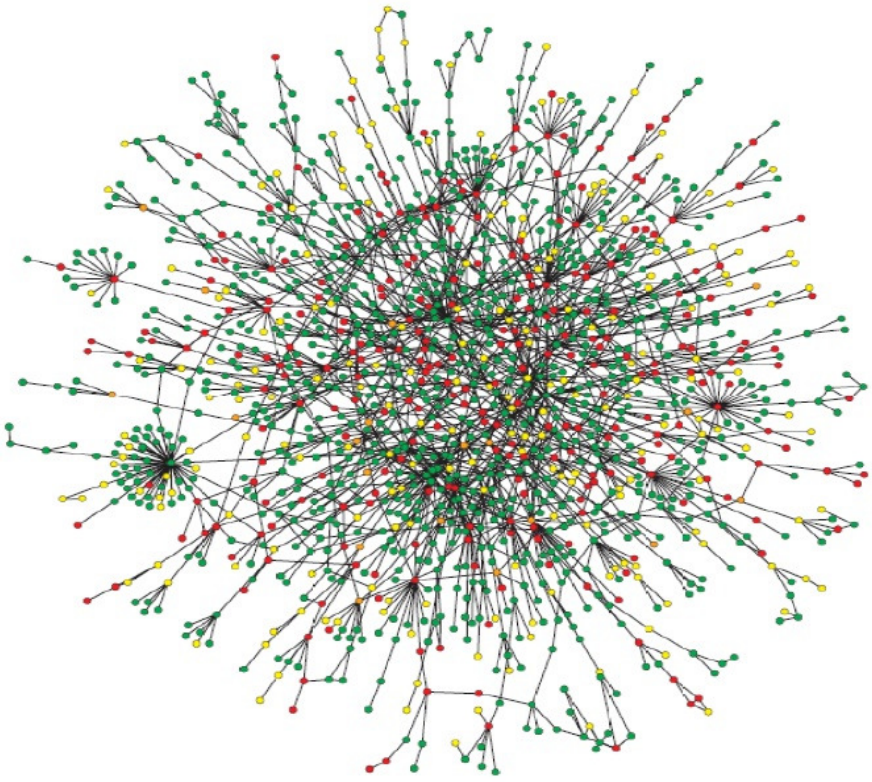


Figura 6 – Redes Protéicas. Em: *http://portal.cbpf.br/themes/pdf/INCT/Workshop/Almeida1-INCT-SC.pdf*

2.5. CÂNCER

Segundo o INCA (Instituto Nacional do Câncer), no site *http://www.inca.gov.br*, define-se câncer, como um tumor maligno, sendo o mesmo um conjunto de mais de duzentas patologias e não simplesmente como uma doença única. Caracteriza-se pelo crescimento desordenado de células anormais que podem invadir outros órgãos e tecidos adjacentes. As variações do câncer estão atreladas aos diferentes tipos de células do corpo, bem como a velocidade de multiplicação das células e a capacidade de invasão dos tecidos ou órgãos vizinhos e/ou distantes.

O câncer pode ser desenvolvido por qualquer indivíduo, embora alguns tenham uma maior predisposição como casos na família (hereditariedade), uso de drogas e defensivos, ou ainda que tiveram doenças congênitas como a síndrome de Down, ataxia telangectásica e imunodeficiências.

Para um melhor entendimento, ressalta-se que a célula sendo composta por três partes: membrana, citoplasma e núcleo, sendo a membrana a parte mais externa, o citoplasma a parte que constitui seu corpo e é no núcleo que estão os cromossomos que são compostos de genes. Os genes são como pequenos arquivos responsáveis por guardar e fornecer instruções para organizar as estruturas, as formas e quais vão ser as atividades das células no organismo.

Toda essa informação é armazenada no DNA, como se fosse escrita numa memória, no ácido desoxirribonucléico, onde os cromossomos passam as informações para o bom funcionamento das células. Por sua vez, uma célula normal pode sofrer alterações no seu DNA, o que se denomina mutação gênica, sofrendo assim uma perda de sua função e começam a multiplicarem-se de forma desordenada, invadindo assim as células sãs, criando novos vasos sanguíneos que irão nutrir e manter as atividades deste crescimento desordenado, formado assim com este grande acúmulo de células os tumores malignos, que por sua vez através dos vasos sanguíneos e linfáticos disseminam-se por órgãos distantes do local onde o tumor teve seu início, formando-se assim as metástases.

Um processo carcinogênico geralmente ocorre em um processo lento, isto em razão do sistema de defesa do organismo podendo levar vários anos essa proliferação de células malignas que irão gerar um tumor visível, sendo este processo passando por vários estágios como: estágio inicial, de promoção e de progressão.

No primeiro estágio, as células sofrem os efeitos dos agentes causadores provocando assim modificações em alguns genes, embora aconteçam estas alterações, ainda não é possível um diagnóstico neste estágio. No estágio de promoção, os genes que sofreram modificação começam gradualmente a transformação de uma célula sadia em uma célula maligna, sendo necessário um longo e continuado contato com o agente promotor cancerígeno. Caso seja interrompido esse contato, cessa todo o processo carcinogênico, muitas vezes o processo é interrompido neste estágio.

No estágio de progressão caracteriza-se pela multiplicação descontrolada e irreversível das células malignas, sendo apenas neste estágio que se pode clinicamente comprovar a presença de câncer, pois neste estágio ele já estará instalado e manifestará os primeiros sintomas da doença.

Como se sabe, estes estágios só ocorrem em razão de alguma falha no organismo, pois em nosso corpo, existem mecanismos de defesas naturais que impedem a proliferação dessas células doentes através de um rede complexa de órgãos, como o fígado, os gânglios linfáticos, o baço, o timo e a medula óssea, também denominados de órgãos linfóides, estando os mesmos relacionados ao crescimento, desenvolvimento e distribuição de células especializadas na defesa de nosso corpo, desenvolvendo um papel muito importante nas atividades do nosso sistema, relacionadas à proteção e defesa deste processo carcinogênico.

As células cancerígenas, por sua vez são menos especializadas nas suas funções do que as células normais, ou seja, conforme as células doentes vão evoluindo sobre as normais, os tecidos invadidos por estas células vão gradualmente perdendo suas funções.

3. BANCO DE DADOS BIOLÓGICOS

Sabe-se que os sistemas biológicos possuem um número grande de agentes (proteínas, genes, compostos químicos) que estão organizados em redes complexas de interação que são o produto de um longo processo evolutivo. Assim sendo, qualquer espécie é o resultado de uma infinidade de acidentes históricos e evolutivos que definiram suas particularidades. Atualmente, muitos dados referentes a estes processos encontram-se facilmente na internet.

As reações químicas catalisadas por enzimas são organizadas em módulos chamados metabólicos que realizam algumas funções específicas, sendo que o metabolismo de todos os organismos é caracterizado por uma rede complexa de reagentes conectados por estas reações.

Através destes processos que foram catalogados, surge uma infinidade de bancos espalhados pela internet e cada um com sua particularidade e alguns interagindo entre si. Neste trabalho foi definido que seria utilizada a totalidade dos dados de alguns bancos, enquanto que outros serviriam para a conexão entre eles. Dentre todos os bancos, utilizou-se como base principal da nomenclatura utilizada na ontologia proposta o banco *"HUGO Gene Nomenclature Committee"*, também chamado de *HGNC*.

3.1. HUGO

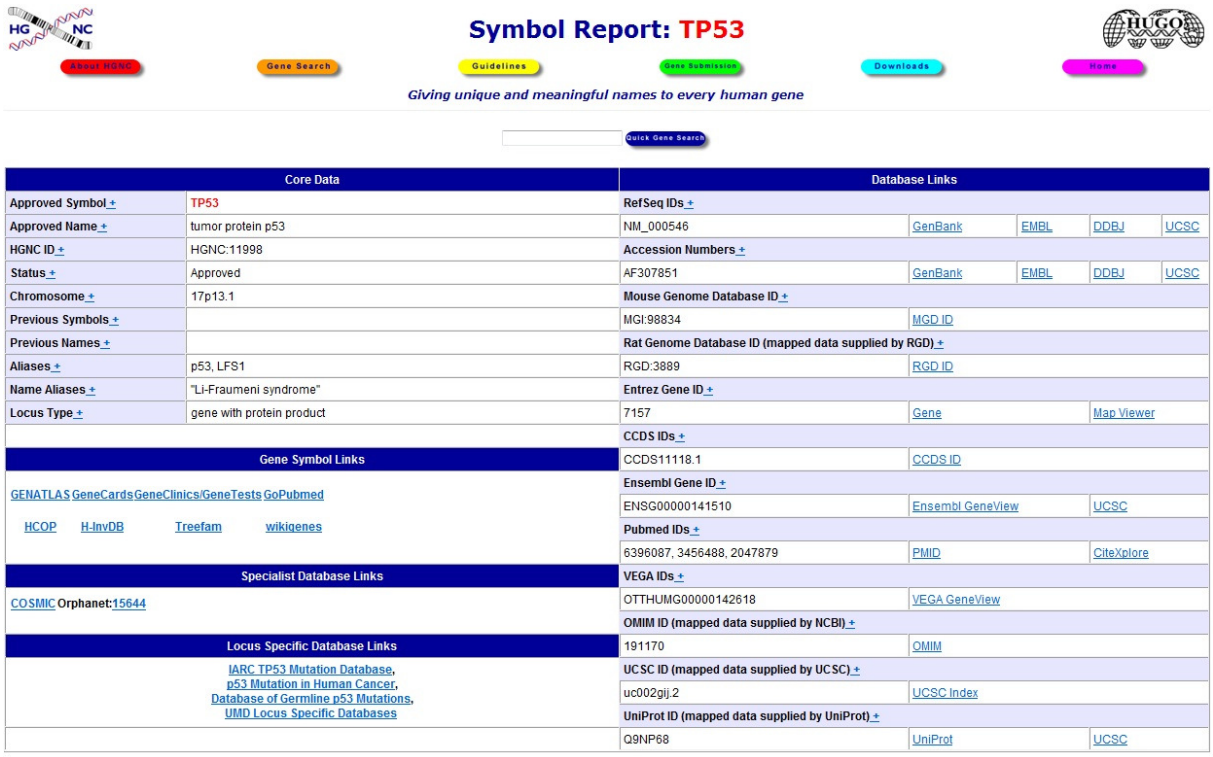
Segundo Splendore (2005), o *HUGO* é um dos principais bancos de dados de nomenclatura de genes e surgiu certamente da necessidade dos problemas mais frequentes relacionados a trabalhos científicos, onde diferentes autores utilizavam uma nomenclatura própria e muitas vezes errônea em determinados genes, bem como também havia nomes diferentes para os mesmos genes, ou genes diferentes com a mesma nomenclatura.

Pensando neste problema, na última década já com o avanço das técnicas de mapeamento e clonagem, bem como no desenvolvimento do projeto do genoma humano, houve a necessidade de uma padronização neste sentido, pois os dados de conhecimento cresciam exponencialmente de tamanho e isto causava um grande transtorno em sua

catalogação, surgindo assim um órgão internacional para coordenar essas nomenclaturas, aprovando um nome e um símbolo para cada gene catalogado, eis que surge o *HUGO*.

Essa entidade além de catalogar e padronizar estes símbolos e nomes utiliza-se de certas regras a fim de facilitar a comunicação entre pesquisadores e a base de dados armazenados eletronicamente, bem como a sua manutenção.

Com a utilização deste banco como base central para esta ontologia, foram utilizados do *HUGO* os seguintes dados: *ApprovedSymbol, ApprovedName, PreviousSymbols, Aliases, Chromosome, Status, RefSeqIDs, EntrezGeneIDbyNCBI, Uniprot_ID, Ensembl_ID, e AccessionNumbers*, como pode-se observar na figura abaixo:



Symbol Report: TP53

About HGNC | Gene Search | Guidelines | Gene Submission | Downloads | Home

Giving unique and meaningful names to every human gene

Quick Gene Search

Core Data	
Approved Symbol +	TP53
Approved Name +	tumor protein p53
HGNC ID +	HGNC:11998
Status +	Approved
Chromosome +	17p13.1
Previous Symbols +	
Previous Names +	
Aliases +	p53, LFS1
Name Aliases +	"Li-Fraumeni syndrome"
Locus Type +	gene with protein product

Gene Symbol Links
GENATLAS GeneCards GeneClinics/GeneTests GoPubmed HCOP H-InvDB Treefam wikigenes

Specialist Database Links
COSMIC Orphanet:15644

Locus Specific Database Links
IARC TP53 Mutation Database, p53 Mutation in Human Cancer, Database of Germline p53 Mutations, UMD Locus Specific Databases

Database Links				
RefSeq IDs +				
NM_000546	GenBank	EMBL	DDBJ	UCSC
Accession Numbers +				
AF307851	GenBank	EMBL	DDBJ	UCSC
Mouse Genome Database ID +				
MGI:98834	MGD ID			
Rat Genome Database ID (mapped data supplied by RGD) +				
RGD:3889	RGD ID			
Entrez Gene ID +				
7157	Gene		Map Viewer	
CCDS IDs +				
CCDS11118.1	CCDS ID			
Ensembl Gene ID +				
ENSG00000141510	Ensembl GeneView		UCSC	
Pubmed IDs +				
6396087, 3456488, 2047879	PMID		CiteXplore	
VEGA IDs +				
OTTHUMG00000142618	VEGA GeneView			
OMIM ID (mapped data supplied by NCBI) +				
191170	OMIM			
UCSC ID (mapped data supplied by UCSC) +				
uc002gij.2	UCSC Index			
UniProt ID (mapped data supplied by UniProt) +				
Q9NP68	UniProt		UCSC	

Figura 7 – Pesquisa realizada no *HUGO* sobre TP53.

3.2. BIOCARTA

A *BioCarta* é conhecida no meio científico [TÔRRES, 2006], como uma empresa web fornecedora de ferramentas didáticas e reagentes para experimentos. Além dos produtos disponibilizados, a *BioCarta* também disponibiliza informações sobre algumas vias bioquímicas de um forma muito organizada. A associação do *CGAP* (*Cancer Genome*

Anatomy Project) a esta empresa agregou significado biológico às sequências existentes neste projeto. As informações que estão disponíveis na *BioCarta* referem-se principalmente a processos celulares como apoptose e transdução de sinal, ou seja, através do qual uma célula converte um tipo de sinal ou estímulo em outro. Neste arranjo podem ser encontradas atualmente onze mil cento dezessete sequências subdivididas/agrupadas em trezentos e três vias.

Sendo assim, esse banco é capaz de propagar informações indicativas da via bioquímica da qual a sequência faz parte no processo de anotação, sendo que foram extraídas as principais vias curadas para a Ontocancro.

Da *BioCarta* foram utilizados os seguintes dados: *ID, Symbol, Name, SequenceID, Hs, e EntrezGene.*

3.3. GENE ONTOLOGY

De acordo com Kellen [PINAGÉ, 2005], a *Gene Ontology* trata-se de um consórcio, cujo projeto é envolver vários centros de pesquisas na área do genoma. Este projeto (ontologia) visa fornecer um domínio biológico específico através de vocabulários estruturados para descrever os genes em um organismo. A padronização dos termos dessa ontologia, permite assim, facilitar a organização das bases de dados das diversas instituições envolvidas neste projeto. Esses termos vão associados através das anotações dos genes das diferentes bases de seus colaboradores, tendo uma lista de termos associadas entre eles.

A grande importância da *GO* é estar atrelada a ela, uma base de dados de mais de cento e vinte mil produtos de genes de uma média de vinte organismos experimentais, sobressaindo-se as plantas, bactérias, vírus, etc., onde as proteínas são identificadas com os identificadores da própria *Gene Ontology*.

Com este trabalho, é possível uma fácil identificação das proteínas, associadas a um termo específico da *GO*, bem como todos os seus termos, associados a uma proteína. Para isso dispõe-se de uma ferramenta web chamada *AmigoGO*, onde para cada gene pesquisado, o usuário é direcionado para uma base de dados onde as informações são mostradas de forma detalhada sobre o gene pesquisado. Além desta ferramenta web o usuário que preferir pode fazer o *download* de outras como o *DAG-Edit*, *COBrA* e o *GoAnnotator* no próprio site da

GO. Outro fator importante que vale a pena ressaltar é que as ontologias podem ser escritas em *frames* (vários arquivos) ou em *flat files* (arquivo único), sendo a GO escrita neste último.

Da *GO* foram extraídos os dados: *GOSymbol, GO_ID, ReferencePubMed, Evidence, HGNC_ID, EntrezGene e fullName.*

3.4. KEGG

O banco *Kegg*, segundo Anna [BAUER-MEHREN, 2009] é um banco subdividido em diversas vias como processos metabólicos, processos celulares, doenças humanas, dentre outros. É um banco que possui um grande número de vias e organismos, curados manualmente. Além de seus dados poderem ser extraídos em arquivos com o formato XML, ainda possui um *software* denominado de *KegArray*, no qual permite uma análise bem detalhada dos dados disponíveis em sua base.

Do *Kegg* foram utilizados os seguintes dados: *Symbol, pathwayName e Name.*

O *Kegg* (*Kyoto Encyclopedia of Genes and Genomes)* é um projeto que visa criar uma base de conhecimento de informações genéticas, ligando funções de genes conhecidos com informações funcionais de mais alto nível [KANEHISA et al., 2004].

Este projeto foi iniciado em conjunto com o programa de genoma humano japonês e provê um serviço de comparação de sequências entradas pelo usuário contra as sequências do projeto. Quando é encontrado algum *hit*, é fornecida informação relativa e ele, indicando a classificação e sua respectiva via. Esse serviço funciona "*online*" e aceita tanto sequências de nucleotídeos quanto de aminoácidos

3.5. NCI

O *NCI-Nature* é um banco de dados com uma grande coleção de vias e suas interações, possui em sua base informações curadas sobre interações biomoleculares e índices de processos celulares montados em vias de sinalização. É um projeto de colaboração entre o *National Cancer Institute (NCI)* e *Nature Publishing Group (NPG).* O *NCI-Nature Pathway* é um banco de interação financiado pelo governo americano, sendo de utilização pública e sem

restrições. Do *NCI* utilizou-se o *pathway, preferredSymbol, Organism, HugoSymbol* e *EntrezGeneID.*

3.6. PROSITE

O *Prosite* é um banco de dados de proteínas e seus domínios baseado no fato de existir um número enorme de proteínas diferentes, sendo que a maioria delas podem ser agrupadas de acordo com as similaridades em suas sequências em um número limitado de famílias, ou seja, proteínas ou domínios de proteínas que pertencem a uma determinada família geralmente compartilham os atributos funcionais e são derivadas de um ancestral comum.

Ao serem analisadas as propriedades constantes e variáveis desses grupos de sequências similares, é possível obter uma identificação (*ID*) para uma família de proteínas ou de um domínio, que distingue seus membros de todas as outras proteínas não relacionadas. Como analogia pode-se citar a identificação de impressão digital pela polícia para fins de identificação de algum indivíduo, por exemplo. Da mesma forma, a *ID* da proteína pode ser usada para atribuir uma proteína recentemente sequenciada a uma determinada família de proteínas e, portanto, formular hipóteses sobre sua função. O *Prosite* atualmente contém padrões e perfis específicos para mais de mil famílias de proteínas ou domínios. Cada uma destas *ID´*s fornece informações básicas sobre a estrutura e função dessas proteínas. Deste banco foram utilizados os campos*: Symbol, HGNC_ID, EntrezGene e Name* para esta ontologia.

3.7. REACTOME

Segundo Bauer-Mehren (2009), o *Reactome* é um dos bancos mais completos e bem mais curado, sendo seus dados dispostos de maneira hierárquica, isto é, o nível mais baixo representa uma única reação e os níveis superiores a uma via completa, este banco inicialmente foi desenvolvido para conter unicamente dados humanos, mas devido a sua complexidade e organização, acabou dando espaço para outros organismos. Seus dados provem não só de informações experimentais como também baseadas na literatura, sendo seus dados curados manualmente por seus colaboradores, realizando um trabalho em conjunto com

diversos pesquisadores. Graças a integração dos dados a outras plataformas é possível obter os dados em versões padrões como o *SBML* (*System Biology Markup Language*) e o *BioPax* (*Biological Pathway Exchange*) para representar suas vias. Sendo o formato *BioPax,* desenvolvido para demonstrar uma representação detalhada de processos biológicos e suas interações, enquanto o formato *SBML*, visa detalhar as informações quanto ao armazenamento e compartilhamento desses modelos computacionais de suas redes de interações biológicas.

Na Ontocancro foram utilizados os dados de *Symbol* e *Name* do Reactome.

3.8. STRING

O *STRING* é um banco de dados que contém as interações proteína prevista, segundo o site *http://string.embl.de/.* Essas interações incluem as associações diretas (físico) e indiretas (funcional), provenientes de quatro fontes, sendo elas, o contexto genômico, experimentos de alta capacidade, co-expressão e conhecimento prévio.

O *STRING* quantitativamente integra dados de interação dessas fontes para um grande número de organismos, e as transferências de informações entre estes organismos, quando aplicável. Atualmente o banco possui mais de dois milhões e meio de proteínas de seiscentos e trinta organismos em sua base. Do *STRING* referenciou-se os dados de *ENSG, ENSP, StringSymbol e StringName.*

3.9. UNIPROT

Com a união de esforços entre os grupos responsáveis (*PIR, Swiss-Prot* e *TrEMBL*), surgiu a base *UniProt* [APWEILER et al., 2004]. Esse consórcio tem como objetivo compreender em uma base única, todas as proteínas sequenciadas até o momento no mundo. Além disso, existe a uma preocupação constante com a anotação funcional das sequências depositadas, resultando em uma base pública rica, coerente e com posicionamento biológico.

No site do *UniProt*, é possível fazer a realização de comparações de similaridade usando *BLAST* através de uma interface na sua base de sequências. Para os *hits* encontrados, ainda é possível ver extensas informações relativas à sequência, como dados de ontologia

fornecidos pelo *GO* e referências bibliográficas que tratam desta sequência. Deste importante banco, foi utilizado o dado *fullName*, referenciado ao Reactome.

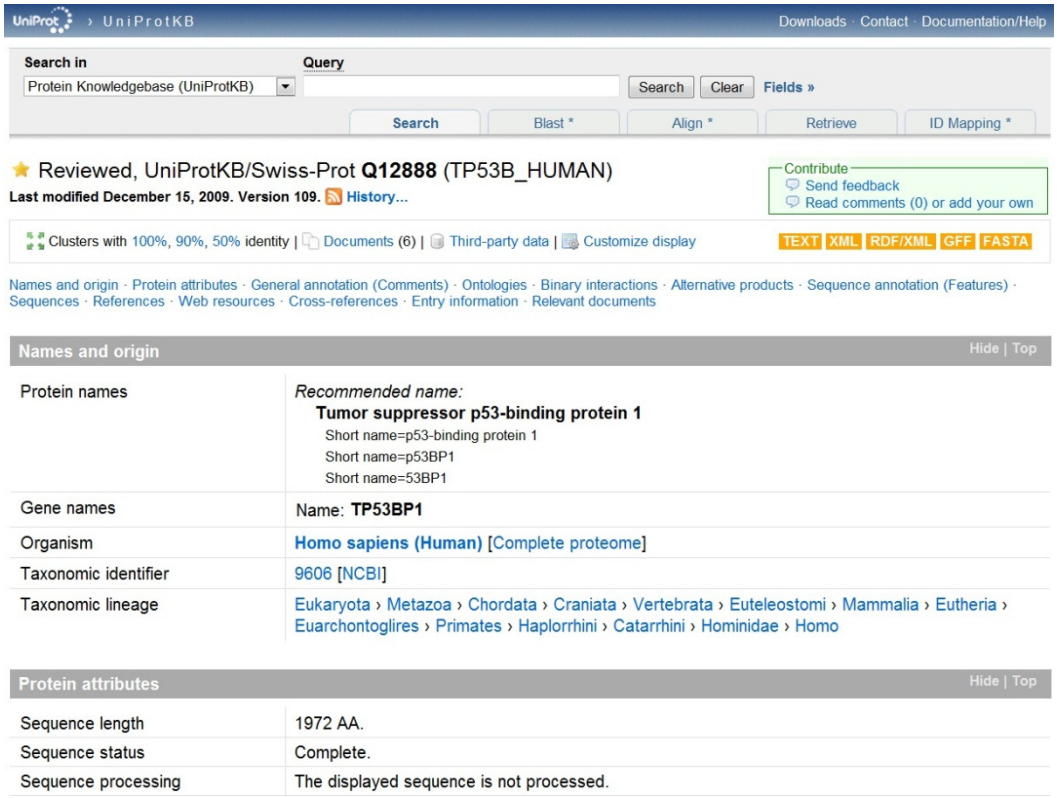


Figura 8 – Informação do banco *Uniprot* sobre *TP53B_HUMAN*

3.10. UNIGENE

De acordo com Schuler [SHULER, 1997], o *Unigene* é uma coleção de sequências de dados similares entre si, de acordo com parâmetros obtidos de uma análise de sequências automatizadas que engloba sequências bem caracterizadas e *ESTs*. Deste banco foram utilizados os dados de *ID, Title, Gene e Express*. Sendo o *Unigene*, uma iniciativa do *National Center for Biotechnology Information* (*NCBI*), promove uma visão organizada do transcriptoma através deste seu sistema analítico automatizado, agrupando assim as sequências bem caracterizadas de genes, sendo utilizado como repositório para inúmeros projetos de mapeamento e projetos de expressão gênica em larga escala.

O *Unigene* também dispõe de uma ferramenta chamada *Digital Differencial Display* (*DDD*), que como o próprio nome diz, serve para fazer uma análise diferencial de

expressão gênica [MURRAY et al., 2007] que emprega um método estatístico de Fisher [FISHER, 1992] a fim de avaliar e determinar a significância da diferença de abundância de *ESTs* de um mesmo *cluster* entre bibliotecas [PINHEIRO, 2009].

3.11. AFFYMETRIX

A *Affymetrix*, segundo fontes do site *http://www.affymetrix.com*, foi a pioneira na exploração da tecnologia utilizando *microarrays*, publicando mais de vinte mil artigos científicos, a *Affymetrix* em sua constante evolução desenvolve inovadoras ferramentas de análise genômica e reagentes para a descoberta, exploração, validação e testes genéticos, contribuindo assim com uma solução completa para análise de estudos do genoma e uma gama de produtos para os estudos de análise celular e de proteínas.

Sendo assim, pesquisadores do mundo todo, utilizam suas ferramentas para melhor compreender o papel que os genes desempenham na doença, eficácia e segurança de terapias, e muitos outros fatores biológicos que afetam o bem-estar humano. A sua tecnologia de *microarrays* tem sido usada no auxílio à descobertas inovadoras para ajudar as comunidades científica e médica a investigar causas como: infertilidade, *HIV*, mal de *Parkinson*, diabetes, malária, e o câncer, objeto desse estudo. Deste banco foram utilizados os dados de *ID; GeneSymbol* e *GeneTitle.*

4. ONTOCANCRO

A crescente quantidade de informações obtidas através das pesquisas científicas da era pós-genômica revela a necessidade do desenvolvimento de ferramentas eficazes no auxílio à organização e compreensão destes dados. Um dos desafios mais importantes na luta contra o câncer, por exemplo, é o entendimento do funcionamento das complexas redes de interações genéticas que controlam as células. No entanto, existe uma grande dificuldade em integrar dados biológicos que se encontram disseminados em diferentes sistemas de informação, como os bancos de dados públicos, os quais armazenam os dados coletados de diversas formas e formatos.

A integração de dados biológicos é uma tarefa complexa, pois exige que o pesquisador busque informações em diversos locais. Sendo que ainda não existe um padrão para referir-se que caracterize alguma informação, o que provoca transtornos quando há uma tentativa de unificar tais dados, é aí que a Ontocancro vem como um diferencial, a fim de auxiliar os pesquisadores e usuários neste árduo processo de pesquisa de genes e vias.

4.1. ONTOLOGIAS E TOPIC MAPS

Uma ontologia é uma especificação explícita de uma conceitualização [GRUBER, 1993]. Enquanto vocabulários controlados somente restringem as palavras que serão utilizadas em um determinado domínio, as ontologias estendem esta característica simples dos vocabulários controlados e permitem uma especificação formal de termos e seus relacionamentos, tornando possível compartilhar e reutilizar o conhecimento, suportando a interoperabilidade entre os sistemas e também permitem inferências sobre o conhecimento representado.

Topic Maps, de acordo com a ISO (*International Organization for Standardization) e a IEC (International Electrotechnical Commission*), é uma norma internacional unificada para descrever estruturas de conhecimento e formalizar a sua associação com recursos de informações [BIEZUNSKY; BRYAN; NEWCOMB, 1999].

Nestes termos, referencia-se *Topic Maps* como uma tecnologia habilitada para a representação e manipulação do conhecimento, proporcionando também uma poderosa forma de navegação sobre recursos de informação. Seu conceito pode ser definido como uma

descrição de um ponto de vista sobre uma coleção de recursos, organizado formalmente por tópicos e pela ligação de algumas partes relevantes do conjunto de informação aos tópicos apropriados.

Um mapa de tópicos pode expressar a opinião de alguém sobre o que os tópicos apresentam, e quais as partes do conjunto de informação que são relevantes para cada tópico. No momento que se fala sobre *Topic Maps* é o mesmo que falar sobre estruturar o conhecimento.

Pode-se definir como os principais objetivos de *Topic Maps* a estruturação de recursos de informação não estruturados, com mecanismos externos aos recursos; permitir buscas que recuperem a informação requisitada; e ainda criar visões diferentes para usuários ou finalidades específicas, filtrando a informação desejada.

Um *Topic Map* em sua representação formal pode ser visto como um conjunto organizado de tópicos, contendo uma estrutura hierárquica de tópicos (definido pelas relações é-um ou contém); vários nomes para cada tópico (ou tema de um índice); ponteiros (ocorrências) entre tópicos e documentos externos (conectando os temas aos recursos de informação); e relacionamentos semânticos (associações) entre tópicos. Um *Topic Map* é, portanto, composto por tópicos associados que dão origem a uma rede semântica estruturada que agrupa informações relacionadas a um domínio específico.

4.2. A ONTOLOGIA ONTOCANCRO

A descrição de uma rede molecular complexa responsável pelo comportamento da célula requer que novas ferramentas sejam desenvolvidas para integrar as enormes quantidades de dados experimentais existentes em sistemas de informações biológicas. Essas ferramentas poderiam, portanto, ser utilizadas na caracterização destas redes e na formulação de hipóteses biológicas relevantes.

A ontologia Ontocancro propõe-se, portanto, a auxiliar na investigação do funcionamento (expressão gênica) de redes biológicas de genes envolvidos em câncer. Ela permite a representação do conhecimento de redes moleculares e sua atividade (expressão). Está agregada a um sistema de informação, de forma a facilitar a integração de dados originários de bancos de dados públicos diferentes em um único banco. A visão gráfica da ontologia proposta pode ser vista na Figura 9.

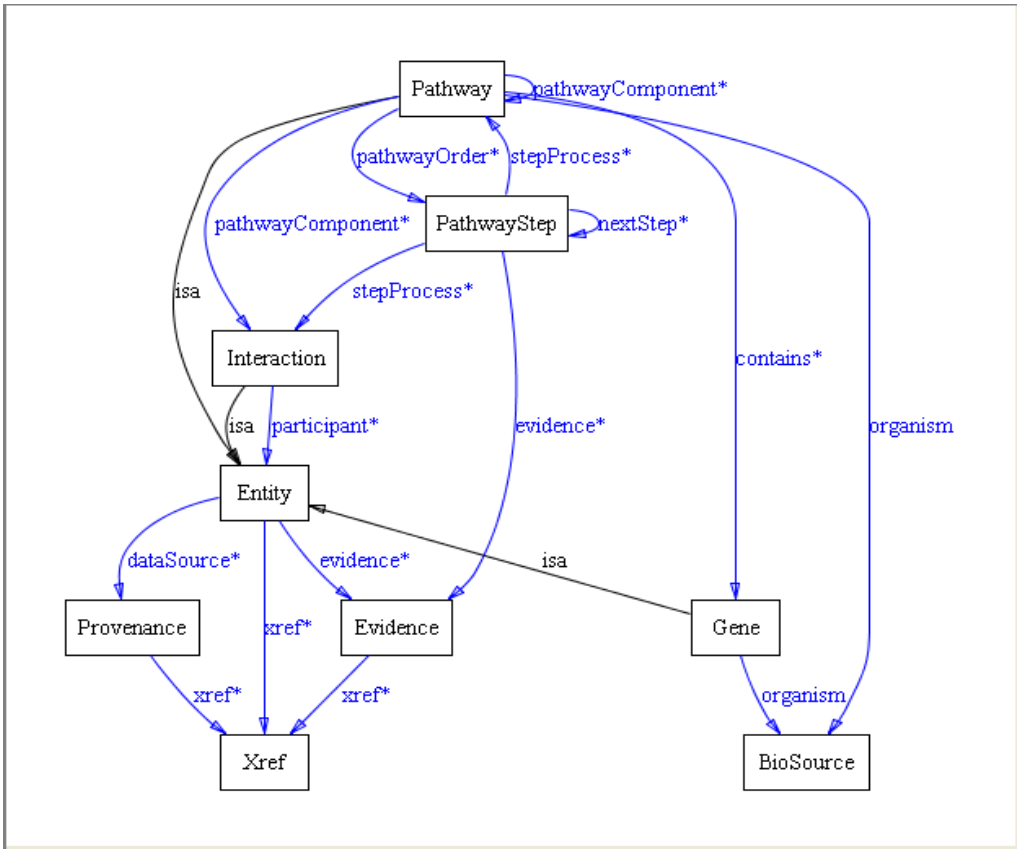


Figura 9 – A ontologia Ontocancro.

Conforme apresentado na Figura 9, os dois principais elementos da ontologia Ontocancro são os *pathways* e os *genes* que compõe cada *pathway*. Os *pathways* ainda estão organizados de acordo com a sua ordem obtida a partir dos bancos que deram origem a tais vias. Esta informação fica armazenada na entidade *PathwayStep*. As interações existentes entre os *pathways* da Ontocancro são representadas na entidade *Interaction*. Estas três entidades são instâncias da classe *Entity*.

Tanto os genes, quanto os pathways, possuem relações com a classe *BioSource*. Entretanto, é importante lembrar que os genes que estão mapeados na Ontocancro são todos de seres humanos, portanto essas relações referem-se todas ao *Homo sapiens*. 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*.

4.3. A ARQUITETURA DO SISTEMA PARA PROCESSAR A ONTOCANCRO

A arquitetura proposta é baseada na abordagem de transformação de dados, integrando dados de diferentes bases em um único repositório de informação que permite a

manipulação destes em seu estudo. Como a maioria dos bancos disponibiliza os dados em formato XML, tornou-se interessante o armazenamento e manipulação dos mesmos em seu formato nativo.

A integração dos dados foi dividida em etapas e ao final, obteve-se um repositório de dados unificados contendo as informações que permitem a integração de redes de interação molecular de câncer com dados de expressão de genes envolvidos em câncer. A partir deste conhecimento, foi 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.

A arquitetura do sistema que proporcionará a interoperabilidade dos dados contidos nas diversas bases de dados disponíveis deverá ser composta pelas camadas de fontes (*source layer*), *wrapper* e mediação (*mediating layer*). A camada de fontes corresponde às fontes de dados estruturados, ou seja, os bancos de dados biológicos, os quais são compostos por repositórios XML, documentos de texto, bancos de dados relacionais e páginas Web. A camada de *wrapper* conterá *wrappers* para cada fonte de dados a fim de selecionar os campos que são importantes para a ontologia desejada. Cada *wrapper* criará uma ontologia representando cada fonte e seu conteúdo para a Ontocancro.

A listagem completa dos bancos de dados biológicos que são consultados para a composição da Ontocancro são os seguintes:

- *KEGG - http://www.genome.jp/kegg/*
- *NCBI - http://www.ncbi.nlm.nih.gov/*
- *The NCI-Nature Pathway Interaction Database - http://pid.nci.nih.gov/*
- *GeneOntology - http://www.geneontology.org/*
- *BioCarta Pathways – http://www.biocarta.com/*
- *Reactome - http://www.reactome.org/*
- *HGNC: Hugo Gene Nomenclature- http://www.genenames.org/*
- *Prosite - http://ca.expasy.org/prosite/*
- *String - http://string.embl.de/*
- *UniGene - http://www.ncbi.nlm.nih.gov/unigene/*
- *UniProt - http://www.uniprot.org/*
- *Affymetrix - http://www.affymetrix.com/*

A camada de mediação contém o mediador que possibilita interoperabilidade entre as fontes locais. Uma de suas principais funções é integrar ontologias locais, para

garantir acesso global às fontes. Ele contém uma máquina de inferência que lida com as ontologias e os mapeamentos, e um processador de consultas.

Para a tarefa de geração dos *wrappers* e integração das fontes, utiliza-se o *Metamorphosis* [LIBRELOTTO et al., 2006], o qual permite obter a interoperabilidade semântica entre sistemas heterogêneos de informação porque os dados relevantes são extraídos e armazenados de acordo com uma ontologia expressa em *Topic Maps* [PARK, 2003]. O ambiente valida a ontologia gerado de acordo com um conjunto de regras definido numa linguagem para descrição de restrições. Esta ontologia fornece fragmentos de informação (as instâncias das classes definidas na ontologia) conectados por relações específicas para outros conceitos, em diferentes níveis de abstração. A navegação sobre a ontologia será realizada seguindo a idéia de uma rede semântica, a qual proporcionará uma visão homogênea sobre os recursos.

A arquitetura proposta é baseada na abordagem de transformação de dados. Como na maioria dos casos, os dados estão dispostos em formato XML, optou-se pelo armazenamento e manipulação dos mesmos em seu formato nativo, utilizando o sistema de gerenciamento de banco de dados XML *eXist* [MEIER, 2003].

Após a decisão de que modelagem seria utilizada, e a correspondente decisão de que sistema gerenciador será escolhido, partiu-se para a integração dos dados. Esta fase estará dividida, então, em três etapas:

- **Aquisição dos dados:** o acesso aos dados dos bancos públicos se dá através de convênios firmados entre os mantenedores do banco e dos membros do grupo de pesquisa, possibilitando o acesso ao seu conteúdo, geralmente em formato XML.
- **Normalização e integração dos dados:** nesta etapa, criou-se *parsers* para manipular os dados que são adquiridos no estágio anterior e traduzi-los para o formato de ontologia e armazená-los neste novo repositório local.
- **Limpeza dos dados:** este processo corrige os dados incorretos. Nesta fase, é imprescindível a presença de um especialista na área de biologia molecular, de forma a comparar os dados das diferentes bases e da literatura técnica.

Ao final destas etapas, obteve-se um repositório de dados unificados contendo os dados que permitem a integração de redes de interação molecular de câncer com dados de expressão de genes envolvidos em câncer. 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. A arquitetura da ontologia e da interface está representada na Figura 10.

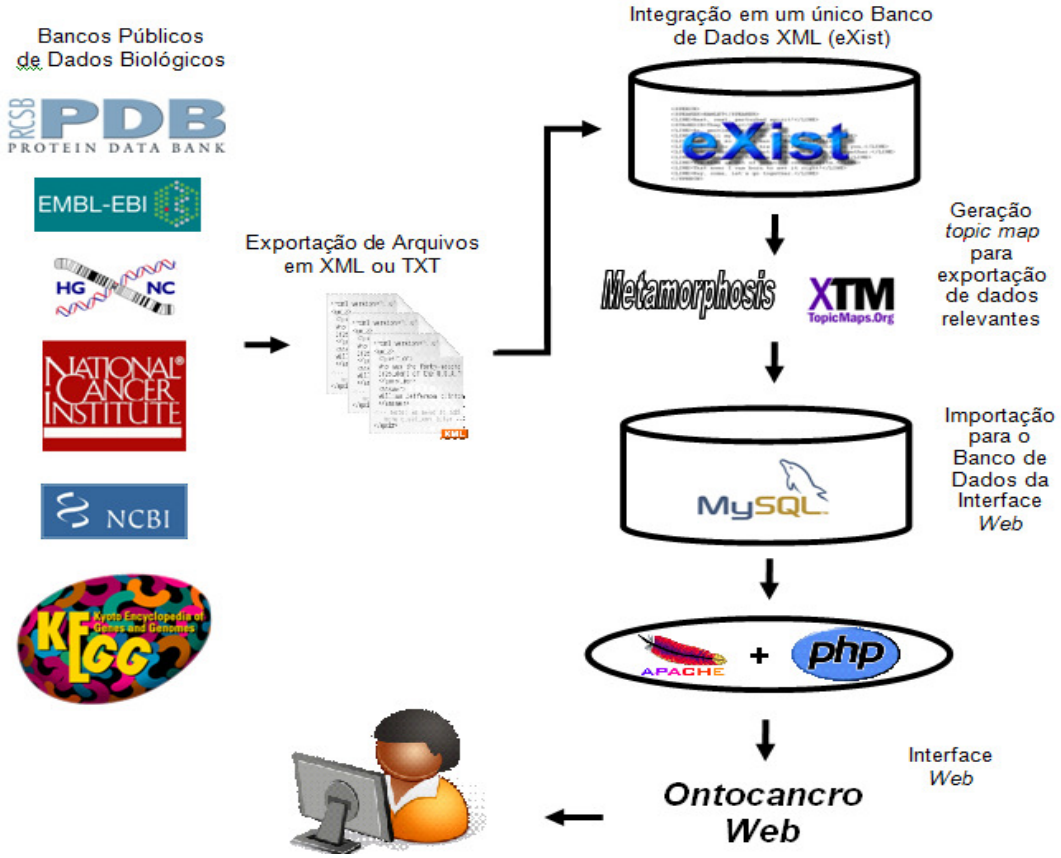


Figura 10 – Arquitetura do sistema para o processamento da Ontocancro.

4.4. INTEGRAÇÃO DOS DADOS DOS BANCOS DE DADOS BIOLÓGICOS

A Ontocancro é composta de informações de bancos de dados biológicos. Como se percebe na Figura 10, a partir dos acordos com todos os mantenedores dos bancos de dados biológicos utilizados como fonte da Ontocancro, foram obtidos os arquivos texto e XML referente a cada banco. Os arquivos de texto são posteriormente tratados, para que seus dados estejam em formato XML e, desta forma, possam ser lançados no banco de dados XML da ontologia.

O banco XML da ontologia está armazenado em um sistema gerenciador de banco de dados XML nativo *eXist*. Neste banco, os arquivos XML obtidos de cada fonte é mantido em sua forma original.

Entretanto, a partir dos bancos de dados biológicos consultados se obtém um arquivo para cada uma de suas vias relacionadas com interatoma e transcriptoma. Assim, o banco de dados *eXist* é, neste momento, composto por mais de centro trinta arquivos XML. Dentro deste montante, destacam-se trinta e dois arquivos oriundos da *Biocarta* e sessenta cinco arquivos obtidos do *GeneOntology*; estes são os bancos biológicos que mais contribuem com vias metabólicas à Ontocancro.

Os arquivos XML obtidos a partir do mesmo banco biológico são estruturados de acordo com o mesmo esquema XML. Entretanto, os arquivos de bancos distintos possuem esquemas diferentes entre si. Para possibilitar a integração destes documentos com esquemas distintos, usa-se o *Metamorphosis*. Esta ferramenta cria um *topic map* para cada banco de dado biológico, composto pelos dados que são importantes para a Ontocancro.

Para possibilitar uma integração coerente entre os genes e vias metabólicas, encontrados nos diversos bancos, são utilizados principalmente dados como:

- código *EntrezGene* de cada gene, gerenciado pelo *NCBI*;
- código e o nome de acordo com o banco *HGNC*;
- identificador no banco *NCI*.

A partir desta integração, o *Metamorphosis* gera um único *topic map* que contém todos os dados oriundos dos diversos bancos de dados biológicos consultados. Este *topic map* único contém a ontologia chamada Ontocancro.

A necessidade de ter-se criado um banco de dados relacional *MySQL* paralelo ao *eXist* deve-se ao fato de que a atualização dos bancos de dados biológicos se dá, geralmente, a partir de documentos XML. Desta forma, mantém-se o banco de dados XML para permitir uma atualização permanente da Ontocancro, enquanto que o banco de dados relacional é utilizado para a geração das páginas Web e para o seu motor de busca, em sua versão mais atualizada.

4.5. NAVEGAÇÃO NO SITE DA ONTOCANCRO

Para que fosse possível uma interação ágil e simples, capaz de ser entendida por qualquer usuário, partiu-se para o desenvolvimento de uma interface Web, para disponibilizar

o acesso aos dados da Ontocancro. Desejando-se a disponibilização *online* das informações extraídas pela ontologia, o banco de dados relacional *MySQL* com a ontologia serviu de base para o acesso Web. De acordo com os requisitos para a construção do site, foram construídas as seguintes funcionalidades:

- listagem das vias metabólicas disponíveis;
- visualização dos genes de uma via metabólica e exportação dos seus dados para arquivo de texto ou em planilha;
- listagem de todos os genes já cadastrados com seu respectivo detalhamento;
- relacionamento das vias metabólicas pertencentes a um gene;
- disponibilização do motor de busca para genes e vias metabólicas;
- desenvolvimento de um módulo administrativo com autenticação por usuário, para manutenção do site;
- importação dos arquivos gerados pela ontologia: com tratamento para atualização de via metabólica.

Na ontologia Ontocancro, a navegação é dada de forma simples e sucinta, acessando o site *http://www.ontocranco.org* tem-se a tela apresentada na Figura 11.

Figura 11 – Página inicial do site da Ontocancro.

Para iniciar-se uma busca, deve se colocar o nome do gene ou via que se deseja a informação e clicar no botão *search* para a obtenção do gene/via como mostra a Figura 11.

O resultado da busca aparece de acordo com o gene/via pesquisado listado de forma ordenada e o usuário logo após escolher o que deseja e clicar no *link*, será levado para a página onde estão todas as informações desejadas, como se pode notar nas Figuras 12 e 13.

Na tela de importação têm-se os dados da via metabólica, surgindo formulários para cada gene importado, onde poderão aparecer somente os dados do arquivo quando o gene for novo no banco. Disponibiliza-se também a quantidade de genes importados e as informações são enviados para a base de dados uma única vez.

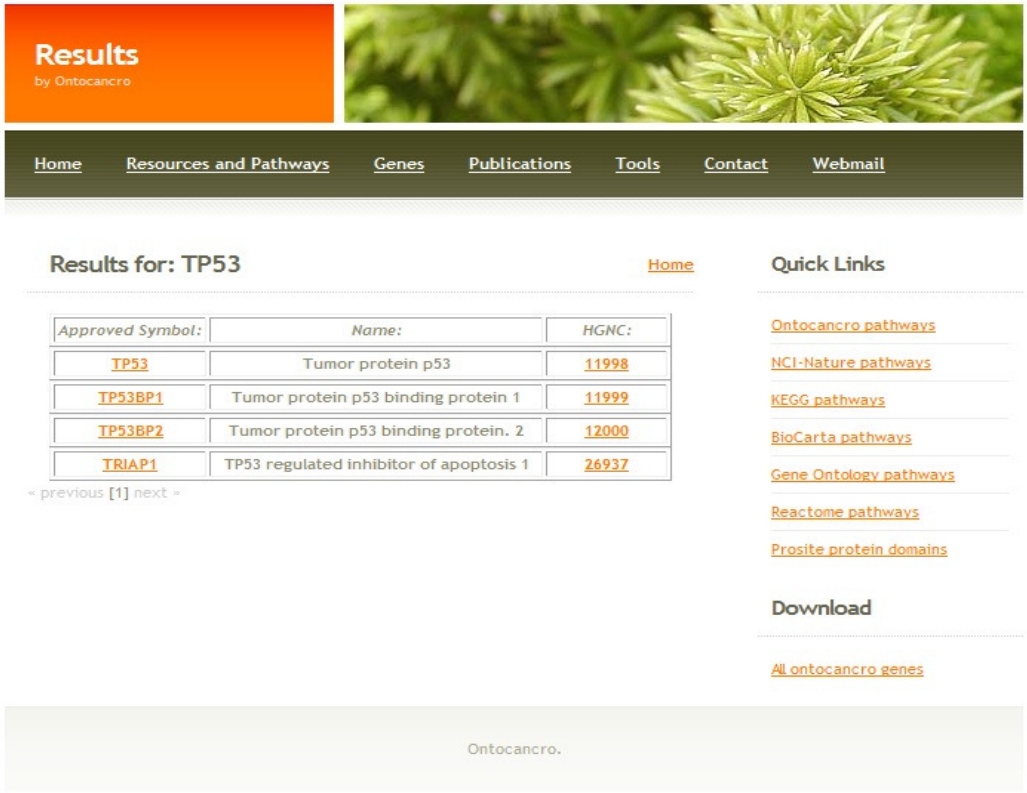
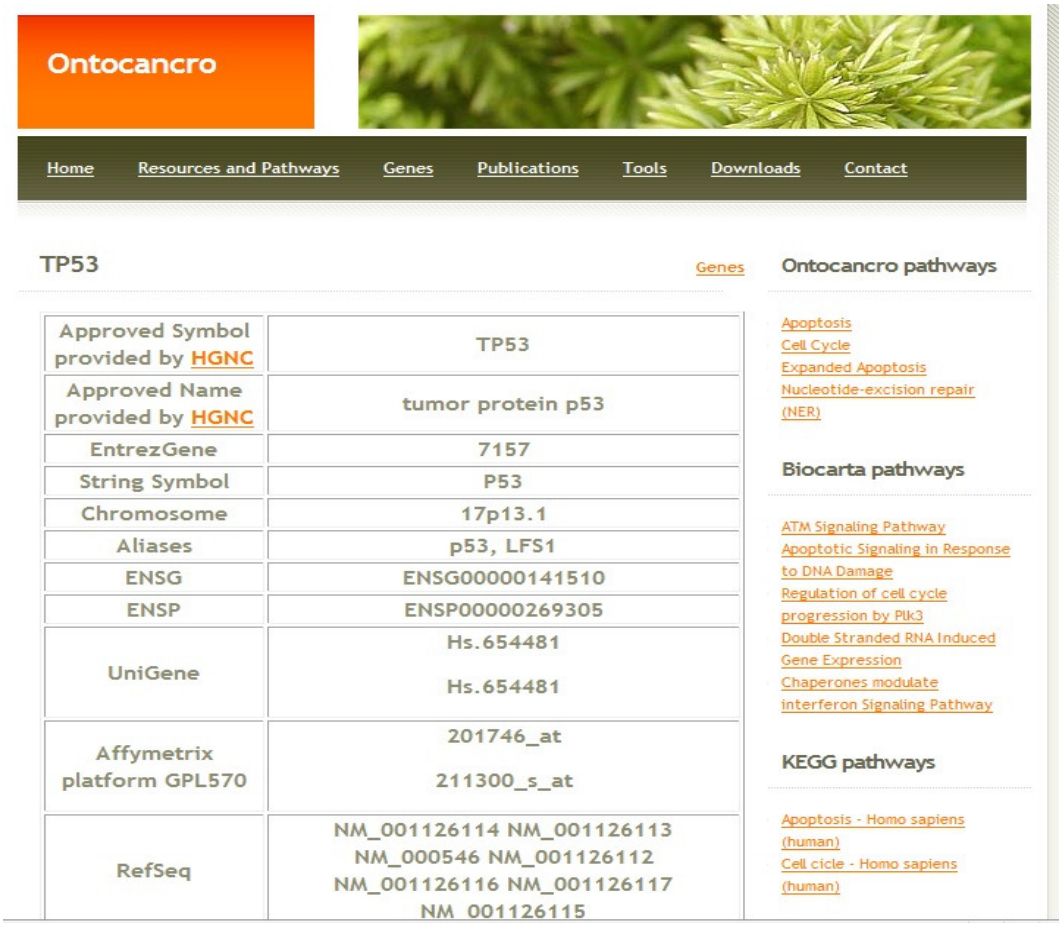


Figura 12 – Resultados obtidos na busca.

Nas Figuras 13 e 14, tem-se a visualização da tela referente ao TP53, apresentando a sua descrição, as vias metabólicas que estão inseridas e os genes com o qual o mesmo está relacionado nestas vias.



Approved Symbol provided by HGNC	TP53
Approved Name provided by HGNC	tumor protein p53
EntrezGene	7157
String Symbol	P53
Chromosome	17p13.1
Aliases	p53, LFS1
ENSG	ENSG00000141510
ENSP	ENSP00000269305
UniGene	Hs.654481 Hs.654481
Affymetrix platform GPL570	201746_at 211300_s_at
RefSeq	NM_001126114 NM_001126113 NM_000546 NM_001126112 NM_001126116 NM_001126117 NM_001126115

Figura 13 – Imagem parcial com dados da busca sobre TP53.

	NM_001126115
Status	Approved
Organism	Homo sapiens

Interacting partners:

Gene Symbol	ENSP
SOD1	ENSP00000270142
PRKCA	ENSP00000284384
CSNK1D	ENSP00000324464
HSP90AA1	ENSP00000335153
UBC	ENSP00000344818
CDKN2A	ENSP00000355153
PARP1	ENSP00000355759
TFAP2A	ENSP00000368933
HTT	ENSP00000374402
S100B	ENSP00000291700

Figura 14 – Continuação da imagem parcial com dados da busca sobre TP53.

4.6. RESULTADOS OBTIDOS

A Ontocancro consiste em uma base de dados que reúne informações de genes e vias envolvidas no processo carcinogênico, onde foram filtrados e catalogados aproximadamente 1428 genes distribuídos em 130 vias. Todos estes dados foram extraídos dos principais bancos de dados públicos de genes: *NCI-Nature, BioCarta, KEGG, Reactome, Prosite, GO e STRING*, além dos demais citados anteriormente.

Devido à falta de consenso na definição dos conjuntos de genes e das vias de cada base de dados, o projeto apresenta um padrão próprio para as vias Ontocancro, que estão relacionadas da seguinte forma:

- Apoptose – 491 genes
- Reparação por Excisão de Base (BER) – 44 genes
- Ciclo Celular (CC) – 286 genes
- Estabilidade Cromossômica (CS) – 76 genes
- Apoptose expandida – 955 genes
- Recombinação homologa (HR) – 34 genes
- Reparação por mau pareamento (MMR) – 28 genes
- Ligação de extremidades não coesivas – 14 genes
- Reparação por Excisão de Nucleotídeos (NER) – 51 genes

A disfunção destas vias pode provocar inúmeros tipos de anomalias celulares, uma delas (e que se estuda neste trabalho) é o câncer. As vias se dividem de acordo com a sua função. Por exemplo, na via de reparação por excisão de base (BER) encontra-se num conjunto de genes envolvidos diretamente no reparo das bases do DNA, assim como as outras vias de reparação, tais como, reparação por excisão de nucleotídeos (NER), reparação por mau pareamento (MMR) e recombinação homologa (HR), entre outras, também auxiliam na manutenção e no reparo do DNA.

Em outros casos, aparecem as vias de ciclo celular (CC), responsáveis pelo processo de duplicação do DNA. Outra via importante é a via de estabilidade cromossômica (CS), que coordena todas as proteínas responsáveis pela estabilidade dos cromossomos. Se algumas destas vias não são capazes de corrigir ou atuar de forma adequada, entra em função a via de apoptose, que tem como função destruir a célula ou então o DNA. É ela a responsável pelo suicídio celular, também conhecido por morte celular programada.

O acesso a base de dados da Ontocancro é feita a partir do endereço eletrônico *http://www.ontocancro.org*, frequentemente utilizado pelos pesquisadores que estudam o comportamento dos genes relacionados ao câncer.

O site também possibilita ao usuário baixar os arquivos textos com todas as informações das vias, que pode ser lido e processado em planilhas eletrônicas ou editores de texto. Isto facilita a extração das informações por outros aplicativos, tornando o acesso aos dados mais dinâmico. Pode-se ainda selecionar um gene pertencente a uma determinada via para obter informações detalhadas sobre o mesmo.

A Ontocancro reúne diversas informações sobre cada gene catalogado. Estas informações estão descritas também no arquivo texto da via, que pode ser visualizado e mais facilmente descrito quando aberto em uma planilha eletrônica.

5. INTEGRANDO NOVAS FUNÇÕES AO SOFTWARE MEDUSA DE ANÁLISE GÊNICA EM VIAS ONTOCANCRO NO COMBATE AO CÂNCER

Um desafio para os pesquisadores é desenvolver técnicas de integração de dados que suporte o crescimento continuo dos bancos de dados biológicos. O grande número de informações necessárias para relacionar às redes genéticas, no combate as doenças, necessita de cada vez mais ferramentas para integrar essas informações que estão distribuídas em vários bancos de dados biológicos. Com a finalidade de integrar proteínas envolvidas no processo carcinogênico desenvolveu-se o projeto Ontocancro (LIBRELOTTO, 2009). Baseado em uma ontologia que organiza o conhecimento da área em questão e integra-o com dados relativos à transcriptoma e interatoma dos genes em vias obtidas a partir dos bancos de dados NCI-Nature, BioCarta, KEGG, Reactome, Prosite, GO e STRING. Para visualização das redes de interação de proteínas, os pesquisadores utilizam um software chamado Medusa. Medusa é uma ferramenta interativa muito utilizada em uma série de projetos científicos da área. Além de visualizar, o pesquisador pode manipular a rede, incluindo ou excluindo nós e arestas (HOOPER, 2005).

Na utilização do software Medusa, os pesquisadores inserem um grupo de proteínas para analisar os genes envolvidos e as interações entre elas. Para realizar essa análise é necessário descobrir e identificar a quais vias pertencem cada gene envolvido no grafo e para isto é necessário pesquisar nos dados do Ontocancro. Atualmente, a Ontocancro conta com nove vias próprias, ou seja, conjuntos de proteínas responsáveis por determinadas funções, cinco vias de reparo, duas de morte celular (apoptose), uma de ciclo celular e uma de estabilidade cromossômica, somando um total de 1428 proteínas, além de outros grupos de genes ligados a morte celular e ao processo carcinogênico.

Outra análise importante na pesquisa de células cancerígenas é descobrir quais genes envolvidos são mais ativos e menos ativos na célula afetada, estes dados são disponibilizados em listas com milhares de genes, o que torna a pesquisa um verdadeiro "garimpo".

Tendo em vista o trabalho exaustivo dos pesquisadores, tem-se uma idéia da necessidade de uma ferramenta que tivesse todas essas funções. Sendo assim neste trabalho foram desenvolvidas novas funções a ferramenta Medusa, agregando assim mais funcionalidades que diminuem o trabalho dos profissionais que utilizam o Medusa para analise gênica e agilizando o processo de busca e manipulação de proteínas e suas interações.

5.1. O SOFTWARE DE ANÁLISE GÊNICA MEDUSA

A ferramenta Medusa foi desenvolvida por Sean Hooper e Peer Bork. Segundo Hooper (2005) trata-se de um aplicativo desenvolvido na linguagem Java, livre para uso acadêmico, e distribuído sob a licença pública GPL. Tem como funcionalidade principal a visualização do grafo de interação de proteínas. Os dados, bem como o código-fonte desta ferramenta, encontram-se disponíveis no endereço eletrônico *http://coot.embl.de/medusa/*.

Para gerar o grafo que representa a via metabólica, o pesquisador necessita buscar no banco de dados *STRING* as informações necessárias. Segundo Jensen (2008) o *STRING* consiste de um banco de dados específico de interação de genes, já está na sua versão 8.0 e abrange cerca de dois milhões e meio de proteínas de seiscentos trinta organismos. Pode ser acessado pelo endereço eletrônico *http://string-db.org*.

O procedimento para abrir um grafo no Medusa funciona da seguinte forma: o pesquisador insere no banco de dados *STRING*, uma amostra de proteínas que deseja analisar. O banco de dados *STRING* gera um arquivo texto com todas os dados sobre as interações daquele grupo de proteínas. Para visualização e manipulação do grafo descrito neste arquivo é necessário que o pesquisador abra ele na Medusa.

As informações contidas no arquivo gerado na base de dados da *STRING* são: posição da proteína no plano, cor, interações e outros dados de cada proteína da amostra como representado na Figura 15. Todos esses dados estão catalogados e disponíveis para o pesquisador no banco de dados *STRING*.

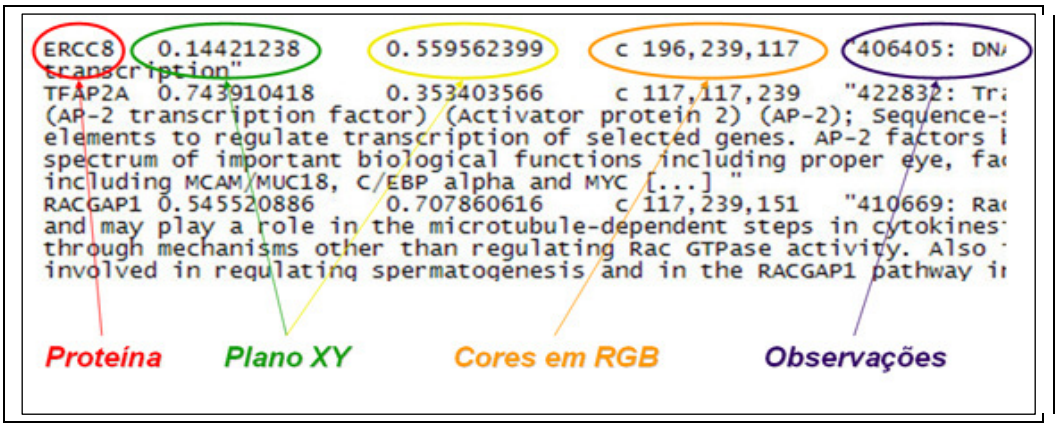


Figura 15 – Formato do arquivo gerado no STRING

Depois de lida as informações necessárias para gerar o grafo, o Medusa permite que o usuário visualize e modifique o grafo aberto, conforme mostra a Figura 16, podendo também salva-lo no arquivo texto para posteriores visualizações.

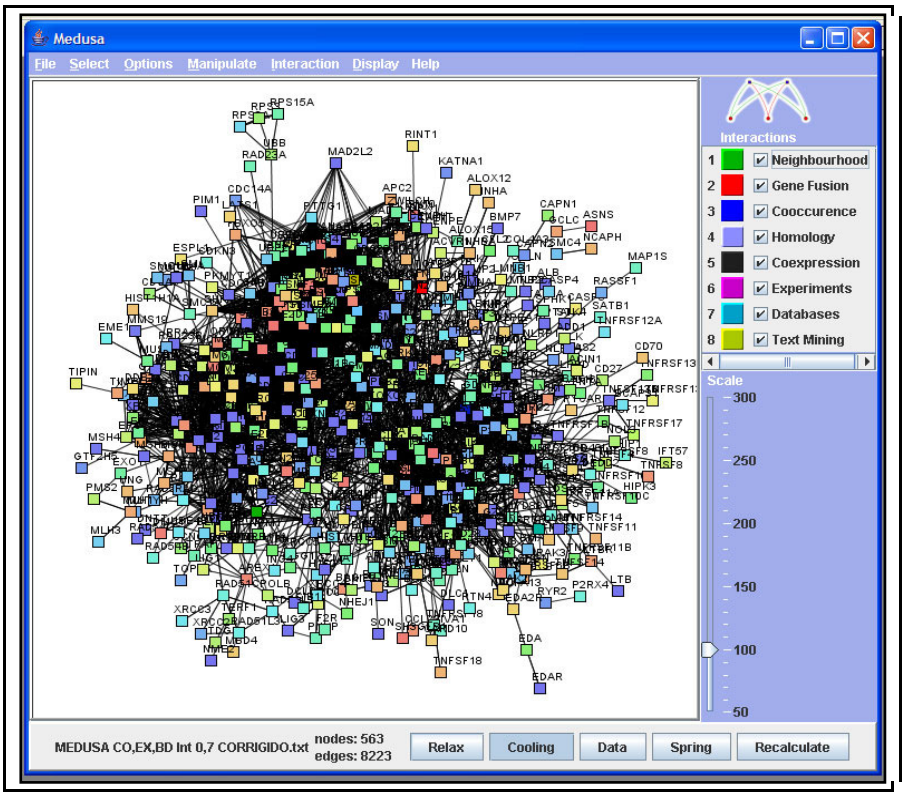


Figura 16 – Captura da interface do Medusa com o grafo gerado.

O Medusa permite manipular o grafo de várias maneiras. No menu *Manipulate* da ferramenta pode-se encontrar opções como:

- *Change node shape*: usado para alterar o formato do nodo para circulo, triangulo, retângulo ou diamante;
- *Delete nodes*: nesta opção é possível apagar os nodos selecionados;
- *Edit Graph*: abre uma tabela, conforme a Figura 17, que permite adicionar novos nodos e/ou alterar dados dos existentes.

Outra funcionalidade que o Medusa possibilita é de visualizar as interações das proteínas de acordo com o tipo de interação. Assim como no *STRING*, o Medusa identifica as interações pela cor do índice no lado direito do grafo.

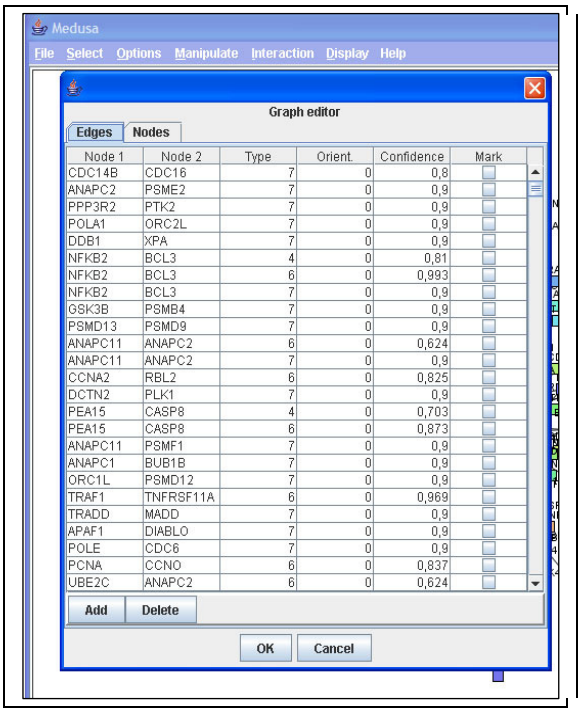


Figura 17 – Tabela para edição do grafo no Medusa.

Para ativar esta função basta ir ao menu *Display*, e depois clicar em *Interactions*. Assim as interações do grafo serão facilmente identificadas pela cor, conforme mostra a Figura 18.

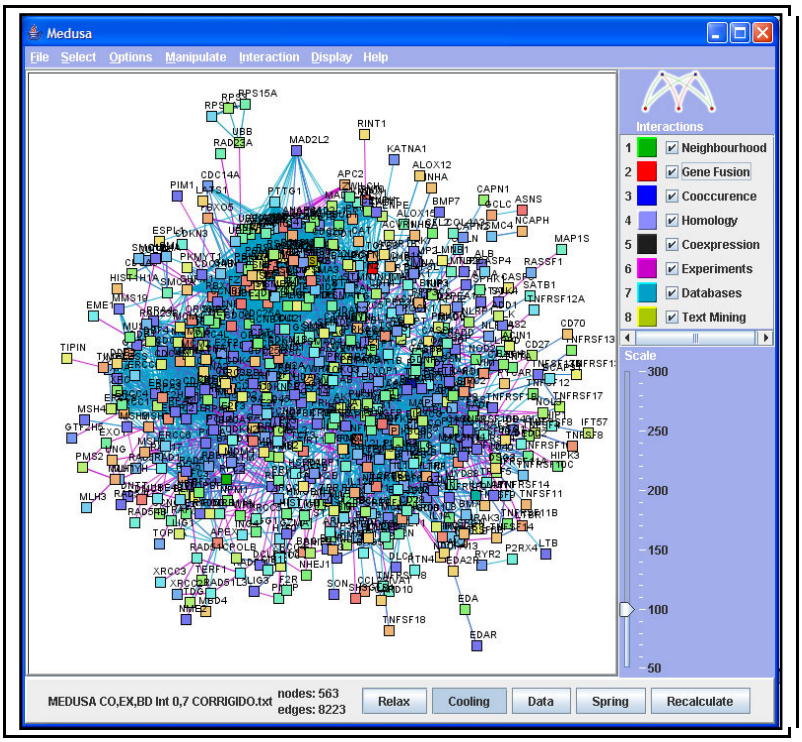


Figura 18 – Medusa com a identificação das interações ativada.

5.2. FUNÇÃO QUE IDENTIFICA AS VIAS DO ONTOCANCRO NO MEDUSA

O grafo que a ferramenta Medusa gera, a partir do arquivo obtido na base de dados do *STRING*, tem a vantagem de ser dinâmico. Porém, as informações obtidas não são suficientes para identificação das vias que as proteínas pertencem. E estas informações demandam uma busca em outras bases de dados. Muitas vezes este trabalho era feito pelo próprio pesquisador, o que demandava muito tempo de pesquisa, pois identificar uma amostra de proteínas em um grafo que contém em média oitocentos nodos exigia dias pesquisando. Para então obter o resultado conforme a Figura 19 [SIMÃO, 2009].

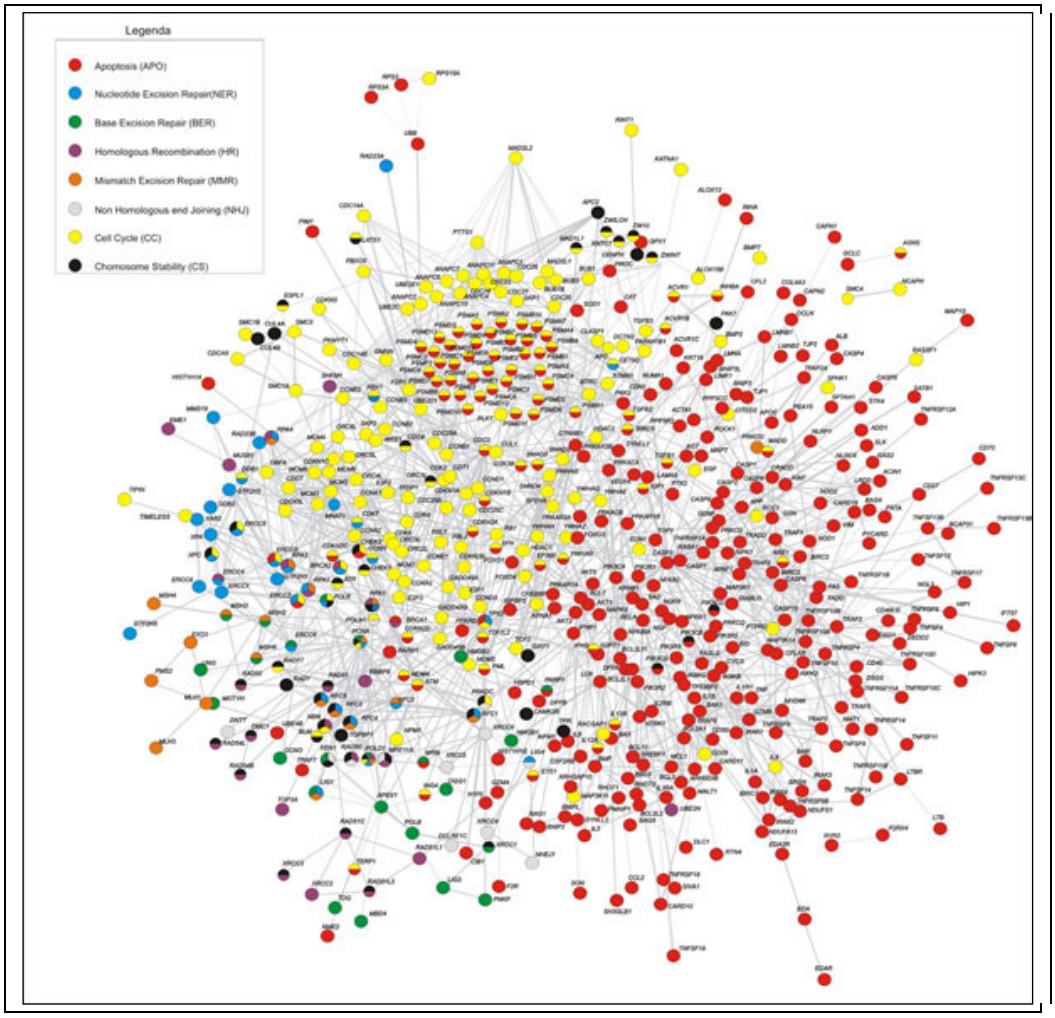


Figura 19 – Rede de interação identificando as vias [SIMÃO, 2009].

Para isto o pesquisador necessitava abrir simultaneamente, no Medusa e em uma planilha eletrônica, o arquivo gerado no *STRING* contendo um número N de proteínas. Na planilha eletrônica eram filtrados os dados necessários de cada proteína, e então era

necessário pesquisar as N proteínas em todas as vias da base de dados do Ontocancro, uma por uma, pois cada proteína pode pertencer a mais de uma via. Depois que todas as vias envolvidas na amostra fossem catalogadas. O grafo aberto no Medusa era redesenhado e pintado em um aplicativo de desenho, exigindo um processo quase manual. E a partir dos dados filtrados na planilha eletrônica criava-se um índice para identificar as vias no grafo.

Como se pode perceber todo este trabalho exigia um grande fluxo de aplicações, conforme ilustra a Figura 20. Para evitar este processo repetitivo e exaustivo ao pesquisador, neste trabalho implementou-se a integração das informações que o Medusa obteve do banco de dados *STRING* com a base de dados das vias mapeadas pelo Ontocancro.

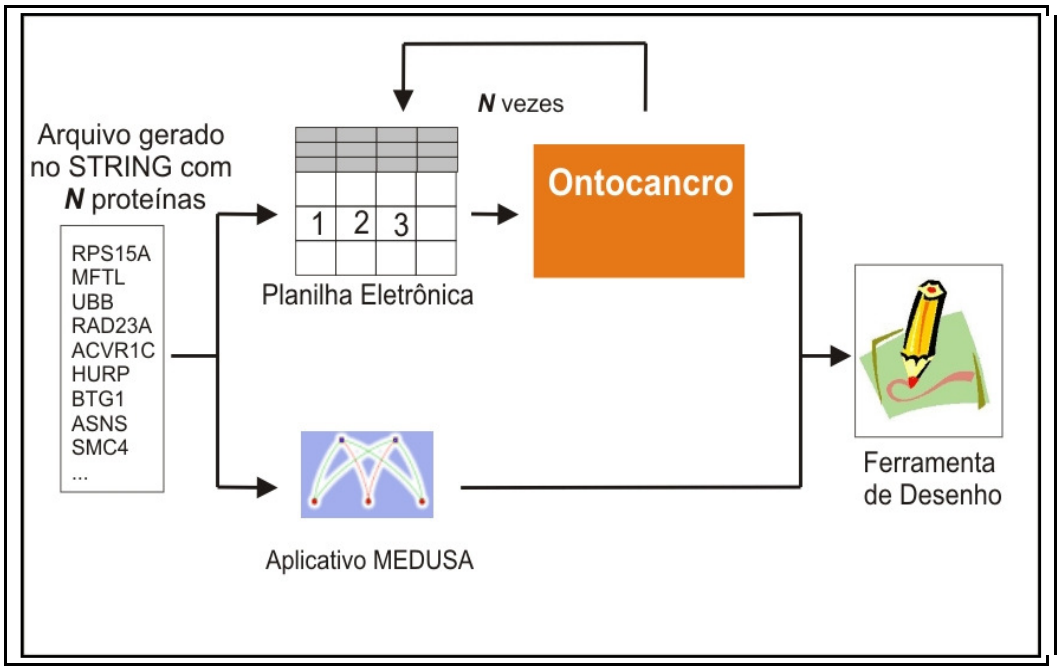


Figura 20 – Fluxo das aplicações utilizadas pelos pesquisadores.

Desta maneira reduziu o número de aplicações necessárias para este processo de pesquisa, ilustrado na Figura 21. Pois com esta nova funcionalidade agregou-se ao Medusa processos que eram feitos nas outras ferramentas. Agora, basta o pesquisador abrir o arquivo gerado no *STRING* juntamente com um arquivo contendo os dados das vias e suas proteínas. Para então obter o resultado, em alguns segundos, com a identificação das proteínas e as suas respectivas vias, como apresentado na Figura 22.

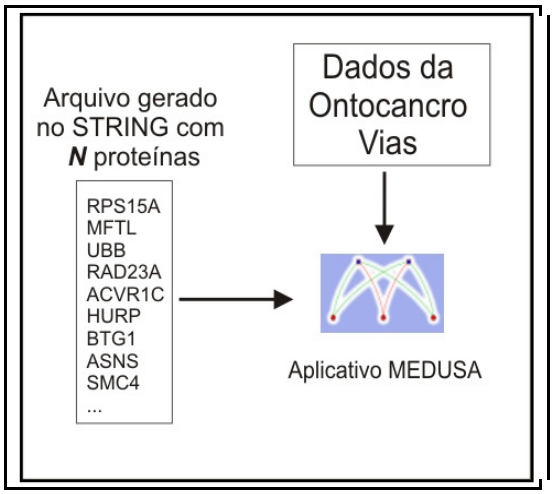


Figura 21 – Redução do fluxo de aplicações com a nova função implementada.

Sendo assim, automatizou-se o trabalho de pesquisa, evitando possíveis erros que podem ocorrer em um trabalho manual. Como por exemplo: o grande número de nodos em um grafo, algum deles pode facilmente passar despercebido aos olhos do pesquisador no meio de tantos envolvidos.

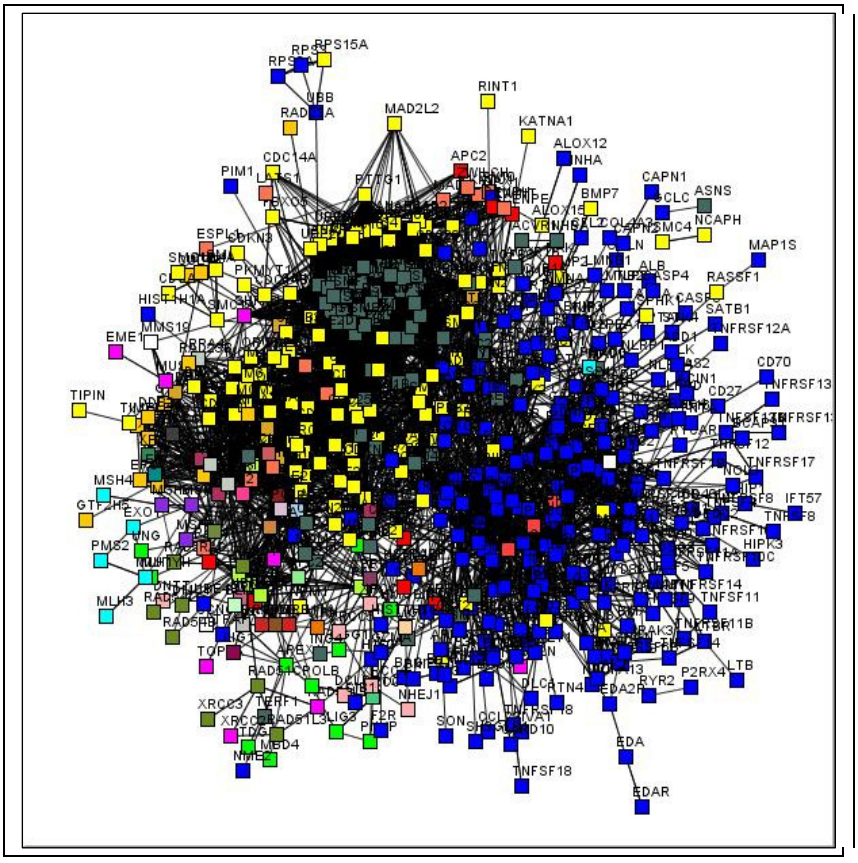


Figura 22 – Resultado obtido com a nova função.

5.2.1 Alterações no Medusa para Receber a nova Função

Para iniciar a implementação desta funcionalidade na ferramenta, necessitou-se a alteração no código de algumas classes do Medusa. A primeira a ser alterada foi a classe *Node* do pacote de desenvolvimento *Medusa.graph*, esta é a classe que determina todos os atributos e métodos dos nós envolvidos no grafo. Nela foi criada a variável *pathway*, que carrega o nome da via a qual o nodo pertence, e os métodos para manipulação da mesma: *setPathway*, que recebe o nome da via, e *getPathway*, que retorna uma variável com o nome da via.

Na classe Principal *MedusaFrame* do pacote Medusa foi criado o menu de acesso a esta nova função. Da seguinte maneira: No menu *Options* foi criado o sub-menu "Ontocancro", para acesso as novas funções implementadas neste trabalho. Neste sub-menu o item "Ontocancro *Pathways*" é o que referencia à função que identifica as vias, conforme mostra a Figura 23.

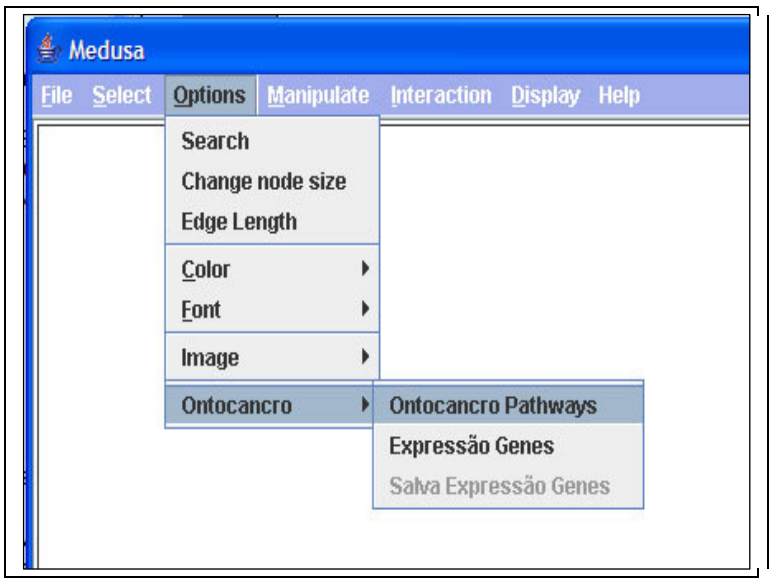


Figura 23 – Menu adicionado à ferramenta.

Outra alteração necessária foi na classe *DisplayPanel* do pacote *Medusa.display*, esta classe é responsável por manipular o grafo aberto no Medusa, onde foi implementada a função *repaintOntoPath*. Esta função é onde consta a maior parte da implementação deste trabalho e que está descrita nos próximos itens.

5.2.2 Carregar os Dados do Ontocancro

A partir dos dados obtidos no Ontocancro, foi gerado um arquivo com o relacionamento das vias e suas proteínas, no formato texto, por ser de fácil manipulação e integração dos dados. Este arquivo disponibiliza um relacionamento por linha, contendo o nome da via e o nome da proteína.

Para ler este arquivo foi desenvolvido um *parser* na função *repaintOntoPath*, onde o primeiro *token* lido no arquivo é o nome da via e o próximo o nome da proteína, conforme Figura 24.

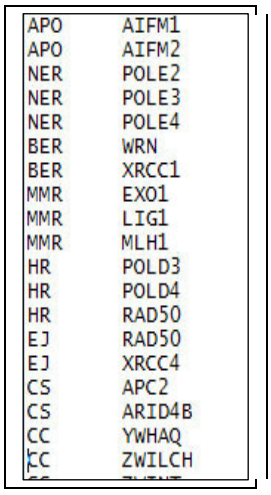


Figura 24 – Arquivo que relaciona via e proteína.

5.2.3 Identificar e Representar as Vias

O fato de algumas proteínas estarem envolvidas em mais de uma via, e o modo como a ferramenta Medusa disponibiliza o grafo, determinou o modo de representação destas vias. A base de dados do Ontocancro Vias envolve um total de nove vias, que foram chamadas de vias principais. Conforme o *parser* lê o arquivo com as vias e suas respectivas proteínas, pode ocorrer a possibilidade de uma proteína que já pertence à determinada via "X" ser encontrada em outra via "Y", é criado então uma nova sub-via que leva o nome de "X-Y", esta sub-via que é tratada como uma nova via, para os casos em que outra proteína venha a pertencer às mesmas vias envolvidas nesta sub-via. Para cada via e sub-via identificada a função determina uma cor diferente, para posteriormente ser usado como um índice.

Para carregar essas informações foi criada a classe *OntocancroPathways* dentro do pacote de desenvolvimento Medusa. Esta classe é um *arraylist*, que recebe todas as proteínas da via, e possui os atributos *pathNome* e *pathColor* e os métodos de manipulação dos atributos: *setPathName, setPathColor, getPathName* e *getPathColor* para identificação da via no grafo.

Depois de carregar todas as vias do arquivo, então é percorrido os nodos do grafo no Medusa fazendo uma pesquisa em todas as proteínas e atribuindo a cor do nodo identificado para a cor da via relacionada.

Quando o grafo foi analisado, a função gera um índice ao lado da ferramenta para que o usuário possa identificar as proteínas expressas no grafo pela cor da via a qual pertencem. Para mostrar este índice na tela foi criada a classe *OntocancroPathwaysDialog*, que tem como base um *JFrame* com a tabela de cores das vias relacionadas no grafo, conforme mostra na Figura 25.

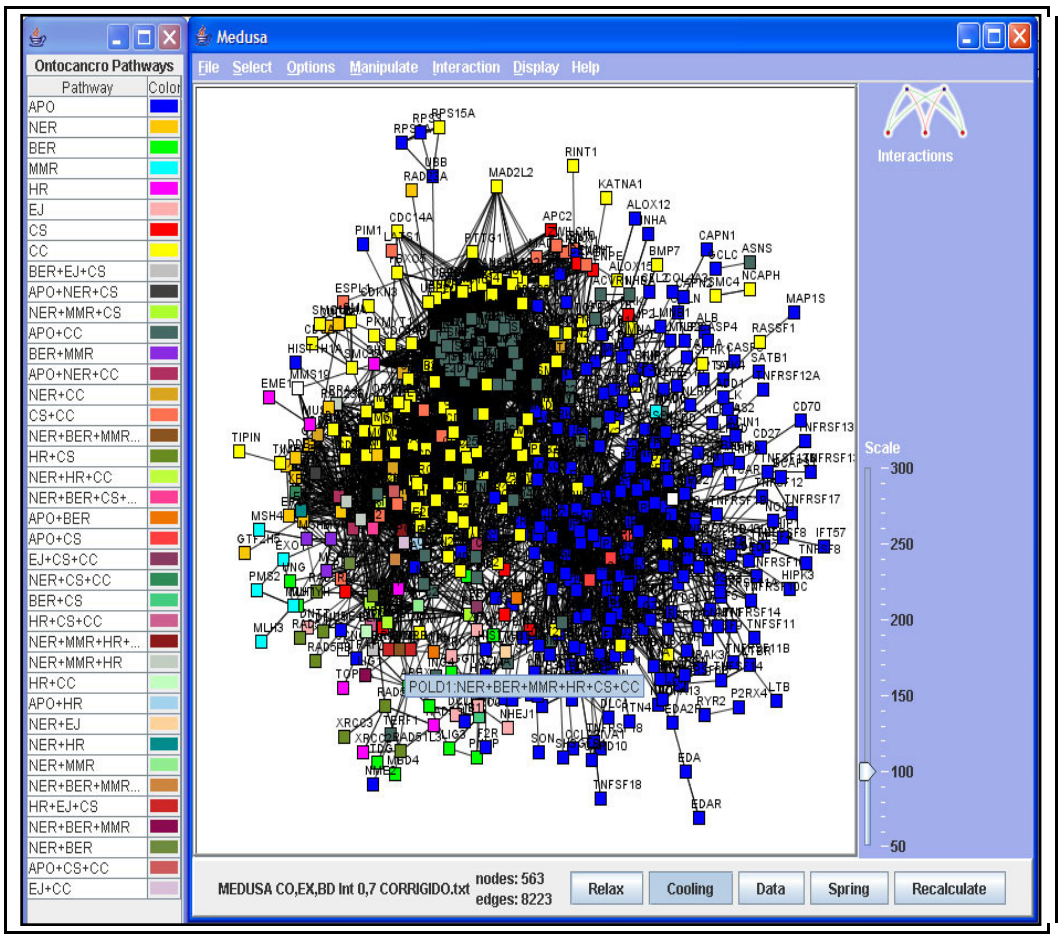


Figura 25 – Grafo com as vias identificadas.

Para que não ocorra confusão na identificação dos índices, assim que gerado o índice das vias, é desabilitado o índice de interações. O qual por padrão já está inserido na interface da ferramenta, podendo ser ativado novamente no menu *Display*, no item *Interactions*.

Muitas vezes o modo como os nodos estão distribuídos no grafo dificultam a identificação dos mesmos, fazendo com que o usuário necessite alterar a escala do grafo no Medusa para identificá-los. Com este objetivo, foi desenvolvida nesta função a possibilidade do pesquisador passar o mouse sobre o nodo para obter informações como o nome da proteína e a via a qual ela pertence.

Para tornar a função ainda mais completa adicionou-se a possibilidade do usuário trocar as cores relacionadas às determinadas vias, caso exista a necessidade.

Por exemplo, pode ocorrer que a cor direcionada para uma via esteja muito parecida com a cor de outra, o que causaria confusão para o usuário identificá-las. Logo, para alterar a cor da via, basta o usuário clicar sobre determinada cor no índice e terá assim uma paleta de cores, um objeto da classe *JColorChooser* que pertence ao pacote *javax.swing*, para selecionar a cor desejada (Figura 26).

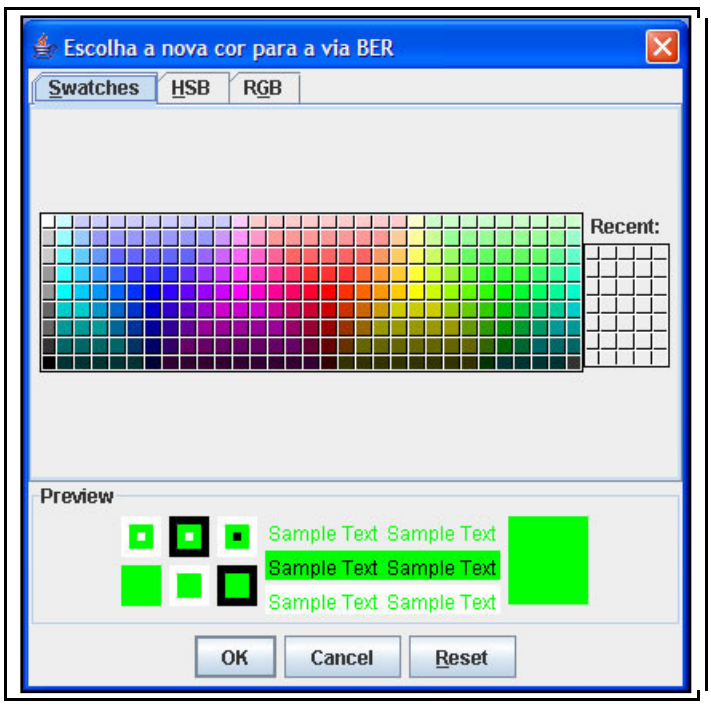


Figura 26 – Paleta para seleção de nova cor à via.

Quando a cor é selecionada, é chamada a função *findChangePathwayColor* que recebe como parâmetros o nome da via selecionada e a nova cor. Esta função, também

implementada neste trabalho, percorre o grafo em busca dos nodos que representam as proteínas da via e, quando encontrados, são alterados para a nova cor.

5.2.4 Selecionar e Apagar Todos os Genes de uma Via

Durante a implementação deste trabalho, notou-se a possibilidade de melhorar a função que identifica as vias Ontocancro no grafo do Medusa. Para isto foi desenvolvida uma função que seleciona todas as proteínas de uma determinada via no grafo gerado, similar ao mostrado na Figura 27.

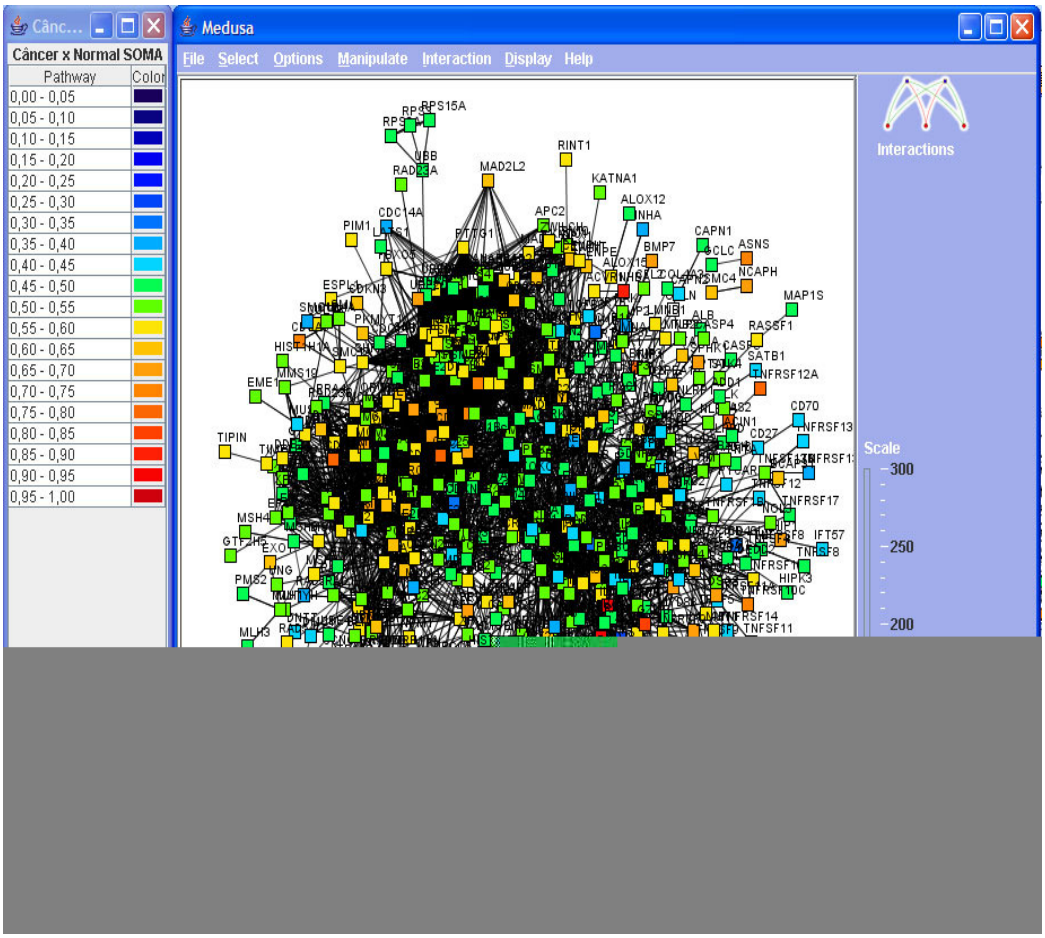


Figura 27 – Interface do Medusa com as proteínas de apoptose selecionadas.

Implementou-se a função *findPathway* que recebe o nome da via selecionada e retorna o valor booleano *true* quando é identificada todas a proteínas da via selecionada. A função inicia percorrendo todo o grafo e para cada objeto da classe *Node* compara o valor retornado no método *getPathway*() com a via selecionada. Quando é verificado que o nodo

pertence à via indicada, é então chamado o método *setFixed*(*true*) para o nodo em questão, sendo que este método já estava implementado pelos autores da ferramenta na classe *Node*. A função *findPathway* foi implementada na classe *DisplayPanel* do pacote *Medusa.display*.

Para o usuário utilizar esta função basta clicar na via desejada, indicada no índice gerado pela função que identifica as via Ontocancro no grafo.

Desta maneira é possível selecionar todos os nós da via e então apagá-los. Possibilitando assim, análises com proteínas pertencentes a determinadas vias. Para isto, o usuário seleciona a via desejada e então utiliza a função *Delete Nodes* no menu *Manipulate* do Medusa, conforme mostra a Figura 28.

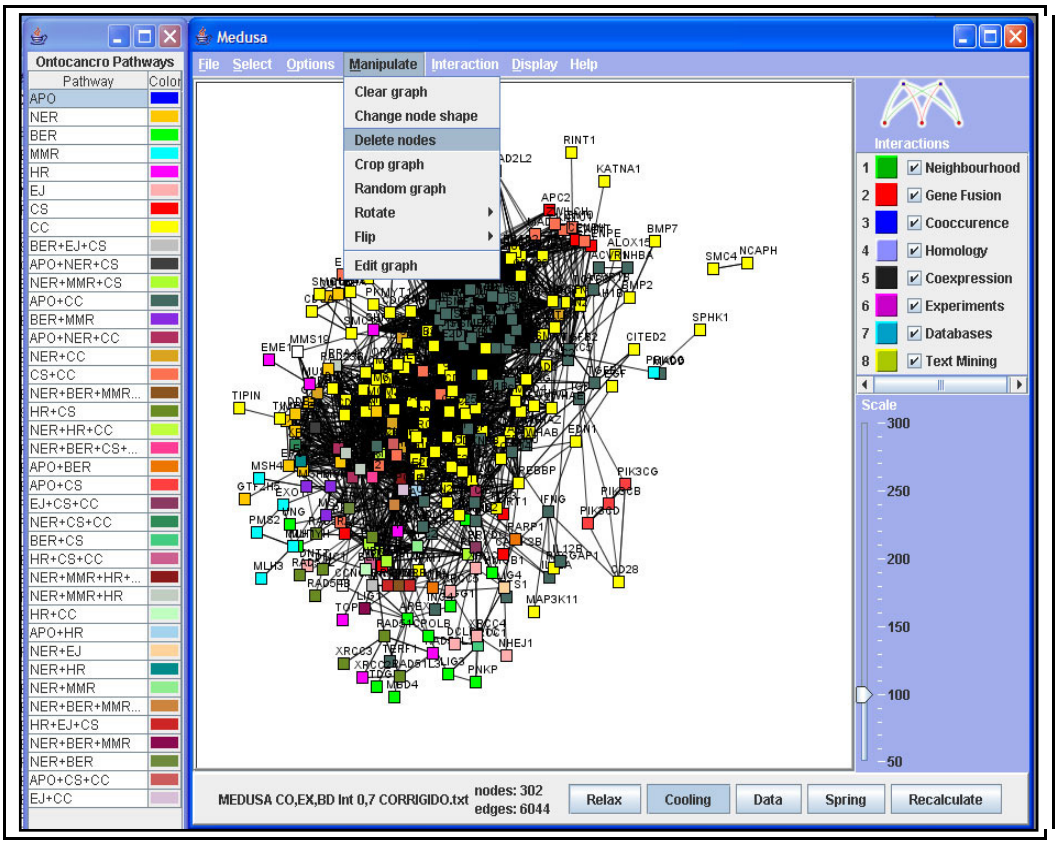


Figura 28 – Grafo com as proteínas de apoptose excluídas.

O Medusa permite que se aplique outro processo aos nodos selecionados, no menu *Manipulate* tem-se ainda a opção de alterar o formado dos nodos selecionados, no item *Change node shape*. Os formatos permitidos são quadrado (padrão), círculo, triângulo e diamante.

5.3. IDENTIFICAR A ATIVIDADE RELATIVA DAS PROTEÍNAS NO MEDUSA

Outra análise necessária para os pesquisadores é identificar o grau de atividade relativa entre as proteínas envolvidas na amostra carregada no Medusa, identificando se uma proteína está mais ativa que a outra para casos particulares. Por exemplo, caso de amostras de células retiradas de tecidos com inflamação, adenoma ou câncer.

Este tipo de pesquisa também demanda o uso de mais ferramentas como planilhas eletrônicas para selecionar as proteínas desejadas e fazer o cálculo da expressão. Neste trabalho implementou-se a função que identifica a atividade relativa entre as proteínas no software Medusa, apresentando um resultado dentro das disponibilidades que o grafo apresentado na ferramenta oferece.

5.3.1. Alterações no Medusa para Receber a nova Função

Assim como na função que identifica as vias no grafo, para esta nova função também foi necessário alterar o código de algumas classes implementadas no Medusa. Na classe *Node* acrescentou-se uma nova variável do tipo *Double* chamada de *atR*, que é carregada com o dado da atividade relativa da proteína que o nodo representa. Também implementou os métodos *setATr* e *getATr* para manipulação desta variável.

Igualmente na classe principal *MedusaFrame* do pacote Medusa foi criado o menu de acesso a esta função, ficando dentro no menu Ontocancro junto com o item Ontocancro *Pathways* e representado pelo item “Expressão Genes”.

Na classe *DisplayPanel* do pacote *Medusa.Display*, foi implementada a função *repaintExpr*, sendo esta a qual determina a funcionalidade implementada à ferramenta. E que está descrita nos itens a seguir.

5.3.2. Carregar os Dados da Atividade Relativa

Estão disponíveis para os pesquisadores, no site da *GEO* os arquivos de expressão contendo resultados de estudos em tecidos, células ou DNA. Estes arquivos contêm dados suficientes para expressar a atividade dos genes envolvidos no grafo. Sendo estes dados, o

identificador da prova, o nome da proteína, o sinal de expressão e o sinal de controle, conforme mostra a Figura 29.

Para ler este arquivo foi implementado um *parser* na função *repaintExpr*. Para receber os dados lidos pelo *parser* desenvolveu-se a classe *GeneExpr* que tem como atributos: o nome, o sinal de expressão e sinal de controle.

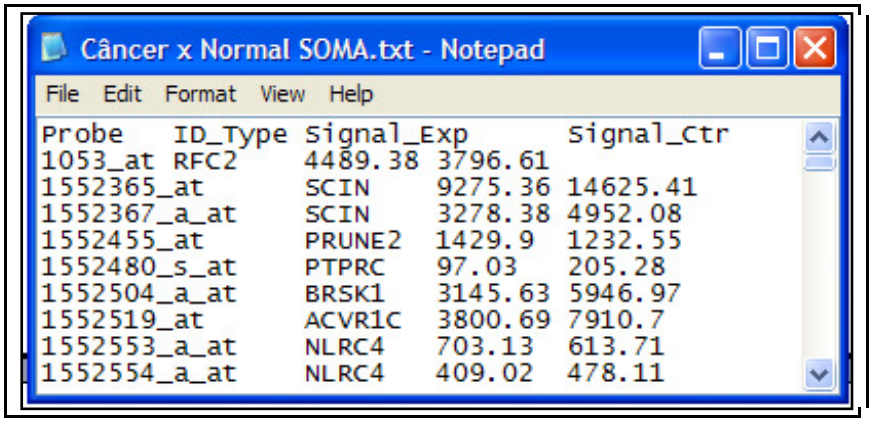


Figura 29 – Arquivo usado para calcular a atividade relativa.

Pode ocorrer de algumas proteínas aparecerem mais de uma vez no arquivo. E para calcular a atividade relativa, todas essas ocorrências devem ser levadas em conta. Para isso os atributos de sinal de expressão e o de sinal de controle da classe são do tipo arraylist de valores *Double*. Podendo assim, ser carregado vários valores nestes campos.

5.3.3. Identificar a Atividade Relativa das Proteínas

Depois de carregado todas as informações contidas no arquivo, é então calculado a atividade relativa para cada proteína. A fórmula da atividade relativa é dada pela média dos sinais de expressão da proteína dividida pela soma da média dos sinais de expressão com a média dos sinais de controle da proteína, de acordo com a seguinte fórmula:

$$AR = \frac{\text{Média(Expressão)}}{\text{Média(Expressão)+Média(Controle)}}$$

Para isto foi necessário criar três métodos na classe que retornam um valor do tipo ponto flutuante:

- ***getMedExp***(): Retorna a soma os valores do *arraylist* de sinais de expressão da proteína dividido pelo tamanho do vetor;

- ***getMedCtr***(): Semelhante ao *getMedExp*(), porém soma os valores do *arraylist* de sinais de controle da proteína e divide pelo tamanho do vetor;
- ***getActR***(): Retorna o valor da atividade relativa da proteína, dividindo o valor de *getMedExp*() pela soma de *getMedExp*() com *getMedCtr*().

O valor retornado pelo *getActR*() é o grau de atividade relativa da proteína. Para isto foi criado uma tabela de forma gradual de cores para que depois de identificado os valores das atividades, os nodos referentes às proteínas sejam coloridos de acordo com essa tabela. A tabela consiste em uma graduação de 0 a 1, para pontos flutuantes, desta maneira é possível representar infinitos valores. O índice identifica os mais ativos com cores mais fortes como vermelho, e os menos ativos com a cor azul, de acordo com a Figura 30. Assim como na função que identifica as vias Ontocancro no Medusa, foi desabilitado o índice de Interações. Para que não ocorra confusão na identificação com os dois índices ao mesmo tempo na interface. Podendo o índice de Interações ser ativado novamente no menu *Display* e habilitando item *Interactions*.

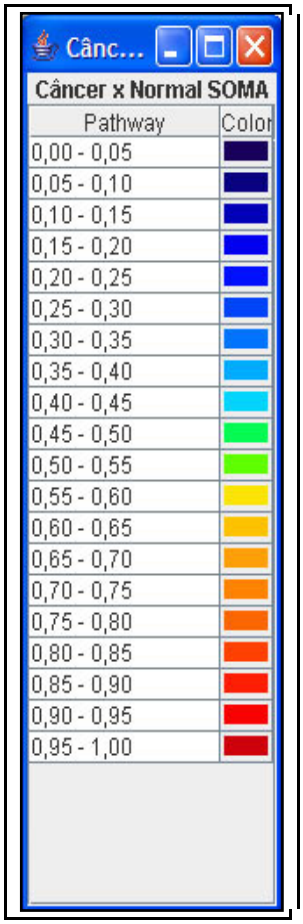


Figura 30 – Graduação de cores para a atividade relativa.

A graduação ocorre dentro de uma margem de 0,05. Esta margem determina as cores que a atividade da proteína será representada pelo valor da sua atividade relativa.

Quando todos os valores são carregados, percorrem-se os nodos do grafo para identificar as proteínas envolvidas. Para cada proteína encontrada é relacionada sua atividade e então é colorida de acordo com a tabela de cores referente ao valor da atividade relativa. A Figura 31, mostra o resultado obtido no Medusa com esta função. Possibilita ainda que o pesquisador passe o mouse sobre o nodo desejado e obtenha na tela o nome da proteína e o valor de sua atividade relativa.

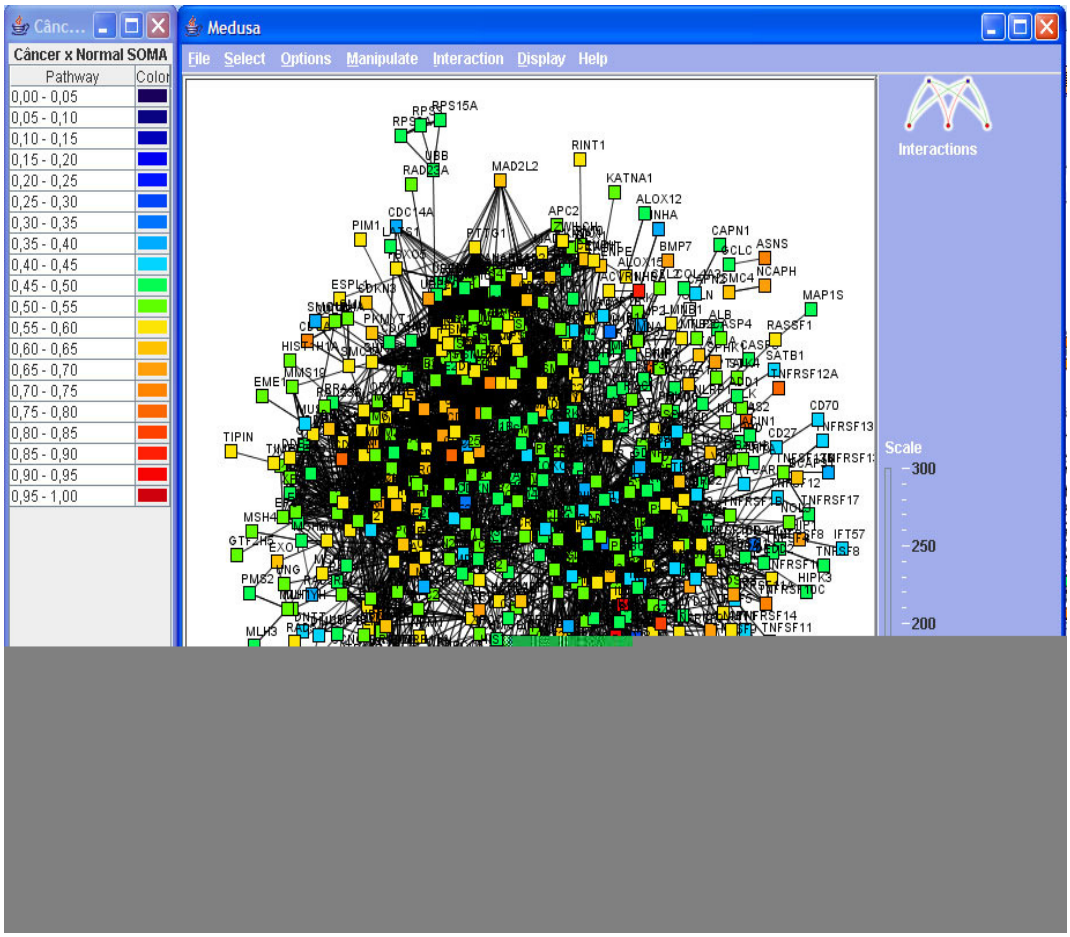


Figura 31 – Atividades relativas expressas no Medusa.

5.3.4. Selecionar e Deletar os Genes de Acordo com a Atividade Relativa

Com a implementação da função que identifica o grau de atividade relativa das proteínas envolvidas no grafo, observou-se a necessidade de selecionar as proteínas inseridas em uma determinada margem do índice de atividades.

Sendo assim, implementou-se a função *findExpr* na classe *DisplayPanel* do pacote *Medusa.display*. A função é chamada quando o usuário clica sobre um item na tabela do índice de atividade relativa. Inicia-se a procura em todos os nodos do grafo, comparando a atividade relativa obtida de cada nodo pelo método *getAtR*, implementado neste trabalho. Quando a atividade relativa do nodo em questão pertence a margem selecionada é chamada a função *setFixed* do nodo, para selecioná-lo. Quando terminada a busca, todos os nodos que foram selecionados estão prontos para serem manipulados (Figura 32).

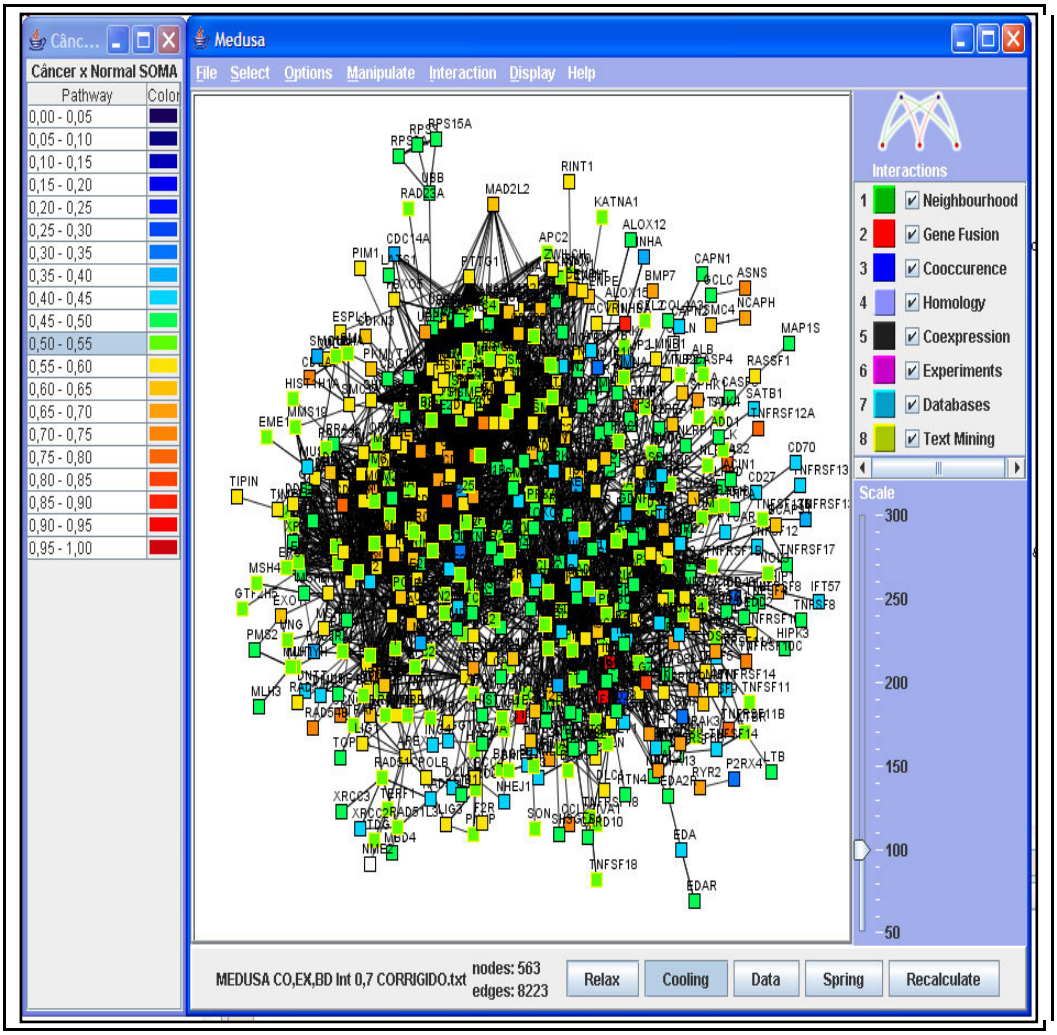


Figura 32 – Proteínas de atividade relativa entre 0,50 e 0,55 selecionadas.

Logo, é possível utilizar a função *Delete Nodes* do menu *Manipulate* para apagar os nodos selecionados. Tornando-se uma funcionalidade muito útil nos casos em que se deseja fazer uma análise do grafo retirando as proteínas menos expressas, por exemplo.

5.3.5. Gerar Arquivo com as Atividades Relativas

Tendo em vista a necessidade de analisar as atividades relativas obtidas na amostra carregada no Medusa, criou-se uma função que gera um arquivo texto com a relação de todas as proteínas envolvidas no grafo com suas respectivas atividades relativas.

Esse arquivo só é possível ser gerado depois de carregado o arquivo contendo os dados com os sinais de expressão e controle. Para isto, o item "Salvar Expressão Genes" presente no menu Ontocancro só é habilitado depois que os dados de atividade relativa já foram gerados.

Depois de gerado, o arquivo pode ser manipulado em editores de texto ou planilhas eletrônicas. Possibilitando assim ordenar ou filtrar o arquivo da maneira que desejar. Por exemplo (Figura 33), ordenar a lista de proteínas pela atividade relativa, tendo no topo da lista a proteína mais ativa na amostra. O que é muito importante para a análise de uma grande amostra de proteínas.

	A	B	C	D
1	IL8	0.9791642427444458		
2	IL1B	0.9260755181312561		
3	IL6	0.90819251537323		
4	BCL2A1	0.907027542591095		
5	INHBA	0.8886247873306274		
6	GZMB	0.8003050088882446		
7	CDC45L	0.7826240658760071		
8	TNFRSF11	0.7781051397323608		
9	GAS2	0.7751649618148804		
10	BNIP3	0.7662078142166138		
11	CDC25B	0.757648229598999		
12	TNFRSF12	0.7556145191192627		
13	ASNS	0.7333166003227234		
14	RAD54B	0.7236078977584839		
15	PLK1	0.7203256487846375		
16	CDC2	0.7168808579444885		

Figura 33 – Arquivo das atividades relativas ordenado em uma planilha eletrônica.

5.4. TESTES E DIFICULDADES NAS IMPLEMENTAÇÕES

Todos os testes realizados nos itens implementados neste trabalho foram feitos também pelos pesquisadores que utilizarão estas ferramentas. O navegador Ontocancro foi testado nos principais *browsers* do mercado, como: *Mozilla Firefox* 3.5.6, *Internet Explorer*

7.0 e *Opera* 10.10, todos com o *JRE* instalados. Em todos os casos os objetivos foram alcançados com o navegador.

Para os testes das novas funções desenvolvidas no *software* Medusa foi utilizado um arquivo gerado no *STRING* com quinhentas sessenta três proteínas e oito mil duzentas vinte três interações. E os equipamentos utilizados para realizar os testes possuíam as seguintes configurações de hardware:

- Processador *Intel Centrino Duo* 1.66 *GHz*, 1 *GB* de Memória *RAM*, 120 *GB* de espaço em disco;
- Processador *Intel Centrino* 1.6 *GHz,* 512 *MB* de Memória *RAM*, 60 *GB* de espaço em disco;
- Processador *Intel Atom* 1.3 *GHz*, 2*GB* de Memória *RAM,* 250 *GB* de espaço em disco.

Em todos os testes a ferramenta obteve um desempenho favorável.

Algumas dificuldades surgiram durante a implementação deste trabalho. Como foi mencionado, no desenvolvimento do *applet* navegador para o site do Ontocancro, uma das questões era a interface de interação com o usuário. A maneira como as informações estariam disponíveis para o pesquisador era um requisito importante para o sucesso da ferramenta.

Para isto, a ferramenta sofreu algumas modificações desde o inicio do desenvolvimento. Primeiramente foi desenvolvida como uma aplicação *off-line* para uso local. Neste momento era possível apenas abrir e visualizar os dados dos genes nas vias disponíveis no site do projeto Ontocancro, através dos arquivos texto.

Assim, verificou-se que o formado de árvore seria uma ótima escolha. Por ser uma interface semelhante à disponível na maioria dos gerenciadores de arquivo, que estão presentes nos principais SO atuais. Desta maneira o usuário tem certa familiaridade com a utilização da ferramenta, sem a necessidade de pré-apresentações.

No desenvolvimento da função que identifica as vias Ontocancro na ferramenta Medusa, realizou-se um estudo do código-fonte do *software* e suas classes para determinar onde seriam feitas as modificações na ferramenta. Os arquivos do código-fonte do Medusa estão bem organizados em pacotes de desenvolvimento, porém envolvem muitos arquivos. E isto demandou certo tempo para descobrir quais as classes deveriam ser alteradas.

Outra questão no desenvolvimento desta funcionalidade foi o fato de que uma proteína poderia pertencer para uma ou mais vias. Criando assim certa complexidade no procedimento implementado. A solução encontrada foi criar sub-vias com o nome das vias

principais que as formam, e assim relacioná-las com os nodos das proteínas pertencentes a elas.

No momento de identificar as vias surgiu uma dúvida sobre a quantidade de cores que poderiam ser usadas, pois seria extenso o número de sub-vias. Para solucionar este problema, foi determinado cores para as nove (9) vias principais, e as sub-vias são cores diferentes entre elas. Também se criou a possibilidade do usuário trocar as cores determinadas pela função. Tornando assim, a função mais completa.

Já na implementação da função que identifica a atividade relativa das proteínas envolvidas na amostra, era necessário fazer os cálculos da atividade relativa de determinada proteína com base em todas as provas que a mesma existisse no arquivo. Para isso foi necessário relacionar primeiro todas as provas que aquela determinada proteína aparecesse, para então calcular as médias de expressão e controle.

Também houve dúvidas quanto ao modo como representar as atividades relativas no grafo. O modo como o grafo está disposto na ferramenta não possibilitaria uma análise precisa da atividade relativa. Para isto foi utilizado uma tabela com graduação de cores, sendo o azul escuro representando os nodos menos ativos e o vermelho para os mais ativos. E quando o usuário seleciona um grau de atividade relativa no índice a função seleciona todos os nodos inseridos neste grau. Facilitando a identificação.

Por final, uma funcionalidade que não estava prevista no trabalho era gerar um arquivo com a lista das proteínas envolvidas na amostra e suas respectivas atividades relativas. Mas durante o desenvolvimento da função, viu-se a importância para os pesquisadores que a ferramenta oferecesse também este procedimento.

É importante destacar que este capítulo todo foi baseado no trabalho final de graduação do aluno da UFSM, Jonathan Luis da Silveira Alves, com resultados publicados na Jornada Acadêmica Integrada da UFSM [ALVES, 2009]. Portanto, a conclusão do mesmo só foi realizada devido à dedicação do Jonathan em desenvolver este aplicativo.

6. TRABALHOS RELACIONADOS À ONTOCANCRO

Dentre as centenas de bancos de dados biológicos acessíveis via *Web*, a Ontocancro se difere por ser a única abordagem que permite buscas exclusivamente sobre vias metabólicas relacionadas a câncer.

Dentre os bancos de dados biológicos mais conhecidos, pode-se citar o *GenBank*, que é a base de dados mais completa sobre informações de RNA mensageiro, DNA complementar, DNA genômico, EST (sequências curtas de DNA complementar retiradas de células em desenvolvimento e usadas para identificação rápida de genes) e *GSS* (*Genome Survey Sequence* – um conjunto de anotações genéticas hipotéticas com um alto grau de proporção de erros de sequenciamento) [GIBAS et al., 2001].

Apesar de ser o maior, o *GenBank* não é um banco de dados curado, diferentemente do Swiss-Prot. Um banco de dados curado é um banco que tem suas informações validadas por especialistas da área. Isso significa que o *GenBank* pode conter inconsistências em suas informações, ao contrário do Swiss-Prot e da Ontocancro, os quais são curados por profissionais especialistas na área genética

Entretanto, tanto o *GenBank* quanto o Swiss-Prot não possuem uma estruturação de seu conteúdo em forma de vias metabólicas, como apresenta a Ontocancro. Da forma como está estruturada, esta ontologia permite com que se consiga encontrar os genes que estão inseridos dentro de uma mesma via, assim como descobrir o nível de confiança entre estes genes (informação essa oriunda do banco *STRING*).

A Ontocancro também se diferencia de bancos de dados de estruturas de proteínas, como o *PDB*. Este é um repositório central para registro de dados de moléculas de proteína e seus dados tridimensionais. Entretanto, não associa estas proteínas entre si, no que diz respeito às vias metabólicas relativas ao câncer.

Em relação aos bancos de dados de *microarrays*, como o *ArrayExpress*, tanto estes quanto a Ontocancro armazenam dados de *microarrays* coletados por tecnologias genômicas. No caso aqui proposto, os dados dos *microarrays* foram obtidos a partir dos experimentos da plataforma *Affymetrix*.

7. CONCLUSÃO

O câncer é uma das doenças mais preocupantes hoje em dia, com isso o investimento na busca de uma cura demanda muitas pesquisas nesta área. E a bioinformática agrega ferramentas que, cada dia mais, se torna imprescindível na contribuição destas pesquisas que se beneficiam com suas aplicações.

Por outro lado, a integração de dados biológicos é uma tarefa complexa, pois exige que o pesquisador busque as informações importantes ao seu trabalho, em diversos locais. Visto que ainda não há um padrão que caracterize de forma exata uma informação biológica, ocorrem transtornos quando se tenta unificar estes dados.

A ontologia Ontocancro foi desenvolvida para contribuir na integração das informações de interatoma e transcriptoma de câncer, dispersas em diversos bancos de dados biológicos espalhados pela Web, propondo uma abordagem integradora para o estudo das redes genéticas mais importantes do processo carcinogênico, como o ciclo celular, o reparo do DNA, a apoptose, entre outras.

Outra importante aplicação na área de bioinformática é o *software* Medusa, que foi desenvolvido para visualizar e manipular os grafos gerados a partir da base de dados do *STRING.* Apesar de ser uma ferramenta muito utilizada nesta área, muitas vezes os usuários precisavam trabalhar com outros aplicativos para concluir procedimentos de análise no grafo que o próprio Medusa gerava. Sendo esta uma ferramenta de código aberto, foi um fator que contribuiu para que se implementasse na própria aplicação novas funções que os pesquisadores necessitavam para que a ferramenta se torna-se mais completa.

Funções estas que agilizaram o processo de pesquisa e análise das amostras de proteínas. A primeira função implementada no Medusa, foi a que identifica as vias Ontocancro no grafo. Esta função vem a diminuir o número de aplicativos que eram necessários para o pesquisador obter o resultado final para a análise da amostra. Pois necessitava da utilização de planilhas eletrônicas e ferramentas de desenho para filtrar os dados e então redesenhar o grafo com as vias identificadas. Logo, a nova função implementada neste trabalho automatizou o procedimento que levaria dias ao pesquisador.

A função realmente cumpriu com seu objetivo e faz ainda mais, pois fornece suporte para que o pesquisador possa manipular as vias no grafo, alterando a cor, o formato e inclusive deletando os nodos que representam as proteínas das vias.

Também foi possível agregar ao Medusa uma função que pode expressar no grafo a atividade relativa das proteínas envolvidas na amostra. Este tipo de análise, que também necessitava o uso de outras ferramentas, demandava tempo ao pesquisador tendo que abrir a mesma amostra em vários aplicativos para obter dados da atividade relativa das proteínas. Com a implementação feita neste trabalho da função que calcula a atividade relativa destas proteínas, o pesquisador pode rapidamente carregar o arquivo das provas com sinais de expressão e sinais de controle, e em alguns segundos obter de modo visual o grau de atividade relativa das proteínas envolvidas no grafo.

O desenvolvimento desta função permitiu que se fosse além do que foi proposto neste trabalho, pois permite visualizar a atividade relativa das proteínas representadas pelos nodos no grafo, e também permite que o usuário manipule e/ou apague os nodos que estão envolvidos em determinados graus de atividade relativa, caso seja necessário. Além disso, a função disponibiliza que seja gerado um arquivo contendo a lista de proteínas envolvidas na amostra com suas respectivas atividades relativas, vindo a contribuir ainda mais para os profissionais que estudam esta área, podendo com este arquivo trabalhar em conjunto com outros aplicativos, como planilhas eletrônicas para ordenar seus dados e/ou filtrá-los.

Nas alterações implantadas no Medusa, tomou-se o cuidado para que não alterasse o desempenho da ferramenta. Nos testes feitos no aplicativo, após as alterações, é possível verificar que se obteve resultado satisfatório inclusive em processadores como o *Intel Atom*, desenvolvido para rodar aplicações que não necessitem um alto poder de processamento.

Ressalta-se então que todos os objetivos propostos foram alcançados, como a criação da ontologia para a representação do conhecimento de redes moleculares e expressão, agregadas de forma a permitir um fácil acesso ao pesquisador, pois agora, o mesmo tem ao seu dispor todas as informações que antes estavam espalhadas por diversos bancos biológicos, bem como ferramentas gráficas para análise gênica das vias metabólicas da informação que deseja, agregados a uma base local, agora integrada, antes disposta em diversos bancos de dados semeados ao longo da rede.

BIBLIOGRAFIA

ALVES, J. L. S.; LIBRELOTTO, Giovani R. **Integrando novas funções ao software Medusa de análise gênica em Vias Ontocancro no combate ao Câncer**. In: XXIV Jornada Acadêmica Integrada da UFSM. Santa Maria/RS. 2009.

APWEILER, R.; BAIROCH, A.; FERRO S.; NATALE, D. A.; YEH, L.S. ***UniProt: the Universal Protein Knowledgebase.*** Nucleic Acids Res. 32: D115-D119, 2004.

BARABÁSI, A. L. and OLTVAI, Z. **Network biology: understanding the cell's functional organization.** s.l.: Nature Reviews Genetics, 2004. pp. 101-113, 2004.

BAUER-MEHREN, A.; FURLON, L. I.; SANZ, F. **Pathway databases and tools for their exploitation: benefits, current limitations and challenges.** Revista Mol Syst Biol. 2009; 5: 290.

BEDELL, M. A.; CLEVELAND, L. S.; O'SULLIVAN, T. N.N.; COPELAND, G. and JENKINS, N. A. **Deletion and Interallelic Complementation Analysis of Steel Mutant Mice.** Blood, November 15, 2003; 102(10): 3548 - 3555.

BIEZUNSKY, M.; BRYAN, M.; NEWCOMB S. **ISO/IEC 13250 - Topic Maps.** ISO/IEC JTC 1/SC34. 1999. Disponível em: *http://www.y12.doe.gov/sgml/sc34/document/0129. pdf*. Acesso em: dezembro de 2007.

CASTRO, M. A. A. **Impaired expression of NER gene network in sporadic solid tumors**. s.l.: Nucleic Acids Research, 2007. pp. 1859-1867, 2007.

FILHO, F. C.; PROSDOCIMI, F.; CERQUEIRA, G. C.; BINNECK, E.; SILVA, A. F.; REIS, A. N.; JUNQUEIRA, A. C. M., SANTOS, A. C. F.; JÚNIOR, A. N.; WUST, C. I.; KESSEDJIAN, J. L.; PETRETSKI, J. H., CAMARGO, L. P., FERREIRA, R. G. M.; LIMA, R. P.; PEREIRA, R. M.; JARDIM, S.; SAMPAIO, V. S., FOLGUERASFLATSCHART, A.

V. **Bioinformática: Manual do Usuário**, 2002. Disponível em: *http://www.biotecnologia.com.br/revista/bio29/bioinf.pdf*. Acesso em: 11 de outubro de 2009.

FISHER, R. A. **On the interpretation of X^2 from contingency tables and the calculation of p. Journal** of the Royal Statistical Society, v. 85, p. 87–94, 1922.

FUTREAL**,** P. A.; COIN, L.; MARSHALL, M.; DOWN, T.; HUBBARD, T.; WOOSTER, R.; RAHMAN, N.; STRATTON, MR. **A census of human cancer genes**. s.l.: Nature Reviews Cancer, 2004. pp. 177-183, 2004.

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

GOBLE, C.; Stevens, R., Ng, G. e Bechhofer, S. **TAMBS: Transparent Access to Multiple Bioinformatics Information Sources**. s.l.: IBM Syst, pp. 532-552, 2001.

GRUBER, T.R. **Toward principles for the design of ontologies used for knowledge sharing**. In: Formal ontology in conceptual analysis and knowledge representation (Guarino N and Poli R, eds.). Kluwer Academic, Dordrecht, Netherlands, 1993.

HOOPER, S. D.; BORK P**. Medusa: a simple tool for interaction graph analysis.** Bioinformatics 2005.

JENSEN L. J.; KUHN M.; STARK M.; CHAFFRON S.; CREEVEY C.; MULLER J.; DOERKS T.; JULIEN P.; ROTH A.; SIMONOVIC M.; BORK P.; von MERING C. **STRING 8--a global view on proteins and their functional interactions in 630 organisms,** 2008.

JEONG, H.; BARABÁSI, ALBERT-LÁSZLÓ; OLTVAI, ZOLTÁN N.; TOMBOR, B.; ALBERT, R. **The large-scale organization of metabolic networks**. s.l.: Nature, 2000.

KANEHISA, M.; GOTO, S.; KAWASHIMA, S.; OKUNO, Y.; HATTORI, M. ***The KEGG Resource for Deciphering the Genome.*** Nucleic Acids Research, 32(Database issue): D277-80, 2004.

LESK, A. M. **Introdução à Bioinformática**, 2. ed. – Porto Alegre, Artmed, 2008.

LIBRELOTTO, G.R.; RAMALHO, J.C. and HENRIQUES, P.R**. - A Topic Maps Based Environment to Handle Heterogeneous Information Resources.** Lecture Notes in Computer Science, Springer-Verlag GmbH, v. 3873, p. 14-25, 2006.

MAGATÃO, M. G. S.; JÚNIOR, E. F. S. **Educação para a Ciência: Uma proposta de intervenção diferenciada no ensino de biologia.** 2008.

MEIER, W. **eXist: An Open Source Native XML Database**. Lecture Notes in Computer Science, vol. 2593/2009, pp 169-183, 2003.

MURRAY, D.; DORAN, P.; MACMATHUNA, P.; MOSS, A. C. **In silico gene expression analysis–an overview.** Mol Cancer, General Clinical Research Unit, UCD School of Medicine and Medical Sciences, Mater Misericordiae University Hospital, Dublin 7, Ireland. dmurray@mater.ie, v. 6, p. 50, 2007.

NHGRI. **The Human Genome Project Completion: Frequently Asked Questions**, 2009. Disponível em *http://www.genome.gov/11006943*. Acesso em 20 de nov. 2009.

NLM. **The DNA Structure**, 2009. Disponível em *http://ghr.nlm.nih.gov/handbook/ illustrations/dnastructure*. Acesso em 22 de nov. 2009.

PARK, J. E.; HUNTING, S. **XML Topic Maps: Creating and Using Topic Maps for the Web**. s.l. : Addison-Wesley, 2003. 0-201-74960-2.

PINHEIRO, D. G. **Desenvolvimento de uma Plataforma Integrativa para Depuração e Análise de Dados de Expressão Gênica.** Tese (Doutorado – Programa de Pós-Graduação em Genética. Área de Concentração: Genética) – Faculdade de Medicina de Ribeirão Preto, 2009.

PINAGÉ, K. **Gene Ontology.** Trabalho do curso de mestrado da Universidade Federal do Amazonas, 2005. Disponível em http://www.dcc.ufam.edu.br/~ontologias/exercicios/ ex1_kellen.pdf. Acesso em 04 de jan. 2010**.**

ROCHA, M. **Bioinformática: passado, presente e futuro!** Bragança, Portugal : s.n., 2006.

SCHULER, G. D. **Pieces of the puzzle: expressed sequence tags and the catalog of human genes.** J Mol Med, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA., v. 75, n. 10, p. 694–698, Oct 1997.

SETUBAL, J. C. **A origem e o sentido da bioinformática**, Revista eletrônica ComCiência.br, 2003. Disponível em *http://www.comciencia.br/reportagens/ bioinformatica/bio10.shtml*. Acesso em 08 de out. 2009**.**

SPLENDORE, A. **Para que existem as regras de nomenclatura genética?** Rev. bras. hematol. hemoter. 2005; 27(2): 148-152.

KASAHARA, K.; KINOSHITA, K.; TAKAGI, T. **Ligand-binding site prediction of proteins based on known fragment–fragment interactions.** Bioinformatics, v. 26, n. 12, p. 1493-1499, Jun 2010.

TÔRRES, J. A. **Protein Classification Tool: Uma ferramenta para anotação de proteínas utilizando bases secundárias.** Dissertação (Mestrado – Departamento Ciência da Computação - UFMG. Área de Concentração: Informática) – Universidade Federal de Minas Gerais, 2006.

UETZ, P.; IDEKER, T.; SCHWIKOWSKI, B. **Visualization and integration of protein-protein interactions. Protein-protein interactions - a molecular cloning manual.** NY, USA : Cold Spring Harbor Laboratory, 2002.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo