O Processamento de uma Ontologia sobre a Integração de Dados de Vias de Interação Molecular Envolvidas em Câncer

Heleno Carmo Borges Cabral[1]
Giovani R. Librelotto[2]
Éder M. Simão[2]
Marialva Sinigaglia[3]
Mauro A. A. Castro[3]
José C. M. Mombach[2]

Resumo: O estudo sobre as interações das redes moleculares ligadas ao câncer torna necessária a centralização de dados biológicos, pois as informações estão espalhadas por diversos sistemas públicos de armazenamento. Visando a tal unificação, propõe-se a ontologia Ontocancro[4], para modelar os dados relevantes ao estudo do pesquisador através de uma interface web, que proporciona uma pesquisa direta e consistente, garantindo unicidade das informações.

Palavras-chave: Bioinformática. Câncer. Ontologias.

Abstract: *The study on the interactions of molecular pathways linked to cancer requires the centralization of biological data, because the information is spread over several public systems of storage. Ontocancro ontology aims to provide research data relevant to their study through a Web interface that provides a direct search and consistent, ensuring uniformity of information.*

Keywords: *Bioinformatics. Cancer. Ontologies.*

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 [1]. O tamanho e a complexidade dos dados biológicos coletados durante os últimos anos incluem informações que requerem uma abordagem integradora [13]. Isso impõe aos cientistas da computação e biólogos a procura por métodos inovadores para tratar esses dados, para que se aumente a compreensão dos processos biológicos que operam dentro da célula. Contudo, essa tarefa de integração é difícil, pois os dados biológicos estão disseminados em diversos bancos de dados. Esses bancos possuem diferentes sistemas de gerenciamento, formatos e formas de representar os dados armazenados. A maioria está acessível por arquivos de texto ou por interfaces web que permitem alguns tipos de consultas predefinidas. Os dois maiores problemas envolvidos aqui são a dificuldade de processar os dados quando se está trabalhando com formatos heterogêneos e com inconsistências, devido à ausência de um vocabulário unificado.

[1] Curso de Análise e Desenvolvimento de Sistemas, IFFarroupilha, Campus Alegrete - RS377/KM 27 - Alegrete (RS) - Brasil
{hcabral@iffca.edu.br}
[2] UFSM - Universidade Federal de Santa Maria - Av. Roraima, 1000 - Santa Maria (RS) - Brasil
{librelotto@inf.ufsm.br, eder.simao@terra.com.br, jcmombach@gmail.com}
[3] UFRGS - Universidade Federal do Rio Grande do Sul - Av. Bento Gonçalves, 9500 - Porto Alegre (RS) - Brasil
{megsinigaglia@yahoo.com.br, mauro.a.castro@gmail.com}

[4] Este trabalho foi parcialmente apoiado pelo CNPq (projeto 478432/2008-9).

doi: 10.5335/rbca.2011.009

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 [6]. Enquanto vocabulários controlados somente restringem as palavras a serem utilizadas em um determinado domínio, as ontologias estendem essa característica simples dos vocabulários controlados e permitem uma especificação formal de termos e seus relacionamentos. Isso torna possível compartilhar e reutilizar o conhecimento. Elas suportam a interoperabilidade entre os sistemas e também permitem inferências sobre o conhecimento representado [5].

O mapeamento dos genes de um organismo traz respostas a diversas questões que há anos foram formuladas por cientistas. Essas questões podem ser desde a curiosidade sobre do que os organismos são formados, até a descoberta das causas de uma doença congênita. As pesquisas na área da bioinformática resultaram no 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 artigo 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. Para atingir este objetivo, o artigo está estruturado da seguinte forma: a seção 2 apresenta os conceitos básicos sobre Bioinformática; a seção 3 detalha a ontologia em questão; os resultados obtidos com a criação da Ontocancro se encontram na seção 4, enquanto que os trabalhos relacionados são expostos na seção 5; por fim, a conclusão do artigo se dá na seção 6.

2 Bioinformática: conceitos e definições

A definição da estrutura do DNA por Francis Crick e James Watson, em 1953, marcou o início de uma nova etapa de descobertas nas áreas da biologia molecular. Com o tempo, percebeu-se a importância do uso de ferramentas tecnológicas nas pesquisas, que pudessem facilitar e agilizar a manipulação das informações de forma segura [4].

Na década de 90, com o surgimento de computadores com capacidade de armazenar e processar um grande volume de informações oriundas do campo das ciências biológicas, emergiu a Bioinformática, tendo ainda apoio de várias áreas, como a física, a estatística, a química, e a matemática [12]. Um dos objetivos dessa área é estudar meios eficazes de armazenamento, processamento, análise, previsão e modelagem de dados biológicos que definem os seres vivos [11]. A complexidade dessa tarefa implica avanços em áreas como a Ciência da Computação, principalmente, uma vez que o conhecimento mais profundo sobre os princípios universais é de fundamental importância para a descoberta de novas drogas e tratamentos.

2.1 Redes de interações moleculares

As células de um organismo possuem vários agentes moleculares, como proteínas, genes e compostos químicos, que interagem entre si formando uma rede complexa resultante de um longo processo de evolução. Vários processos biológicos podem ser representados como uma rede, um exemplo são as redes metabólicas. Barabási e colaboradores [7] propuseram uma representação gráfica da rede metabólica na qual os nós representam os substratos, que estão ligados uns aos outros através de conexões compostas pelas reações metabólicas. Descobriu-se nessas redes uma estrutura topológica universal entre os seres vivos: a probabilidade de que um determinado composto participe de um certo número de reações (conexões) obedece a uma lei de escala. Outro tipo de rede biológica muito estudado refere-se à rede de interação de proteínas uma vez que esta apresenta uma importância crucial em todos os processos celulares. Dessa forma a informação obtida através dessas interações contribui para um melhor entendimento sobre as doenças, além de proporcionar a base para novos tratamentos.

Atualmente, encontra-se disponível na web um grande volume de informações referentes a muitos processos biológicos que são críticos para a manutenção da estabilidade genômica, incluindo os mecanismos de reparo do DNA, apoptose, ciclo celular, entre outros. Todos esses dados podem ser utilizadas para a construção de redes genéticas com o intuito de extrair conhecimento e tentar compreender o papel dessas vias no processo carcinogênico, o que constitui no interesse biológico deste trabalho.

2.2 Redes de manutenção da estabilidade genômica

Os mecanismos de manutenção da estabilidade genômica são críticos para homeostase celular, uma vez que alterações no funcionamento desses processos podem levar ao surgimento do câncer. As vias do ciclo celular, reparo, apoptose e estabilidade cromossômica desempenham um papel central na manutenção da estabilidade do genoma. A célula possui diferentes mecanismos de reparo para proteger o DNA contra danos, como as quebras de cadeias de DNA ocasionadas pela radiação ultravioleta. Os sistemas de reparo se constituem em redes genéticas especializadas nessa proteção, uma vez que impedem que diferentes tipos de danos sejam fixados no material genético. As células cancerosas apresentam uma série de alerações em seu material genético devido ao mau funcionamento dessas redes de proteção. Sabe-se que os genes de uma das cinco redes de reparo, chamada de Reparo por Excisão de Nucleotídeos (NER), não possui mutações catalogadas casualmente relacionadas a câncer somático, acreditando-se que ela não estaria envolvida no aparecimento de células cancerosas [3].

Castro e colaboradores [2] avaliaram o comportamento das redes de Manutenção da Estabilidade Genômica em amostras de tecidos de câncer e normais disponibilizadas pelo Projeto Genoma do Câncer Humano na Internet. Utilizando a entropia da distribuição de ativação dos genes de todas as redes de reparo, redes energéticas e da rede envolvida em apoptose, os autores verificaram que a rede NER, embora estruturalmente conservada (sem mutações) nos tecidos cancerosos, é a que apresenta a maior alteração funcional em relação às outras redes.

Através da construção da rede de interação entre esses genes e da projeção dos dados de expressão sobre a mesma, os autores propuseram que o mau funcionamento da rede NER em câncer era ocasionado pela disfunção da rede de apoptose que se comunica com esta via, principalmente através do gene TP53. Este gene possui um papel chave em vários processos, como ciclo celular, apoptose e reparo, e encontra-se frequentemente mutado em vários tumores. O grafo de interações entre a rede dos genes de reparo e da rede de apoptose pode ser visto na Figura 1.

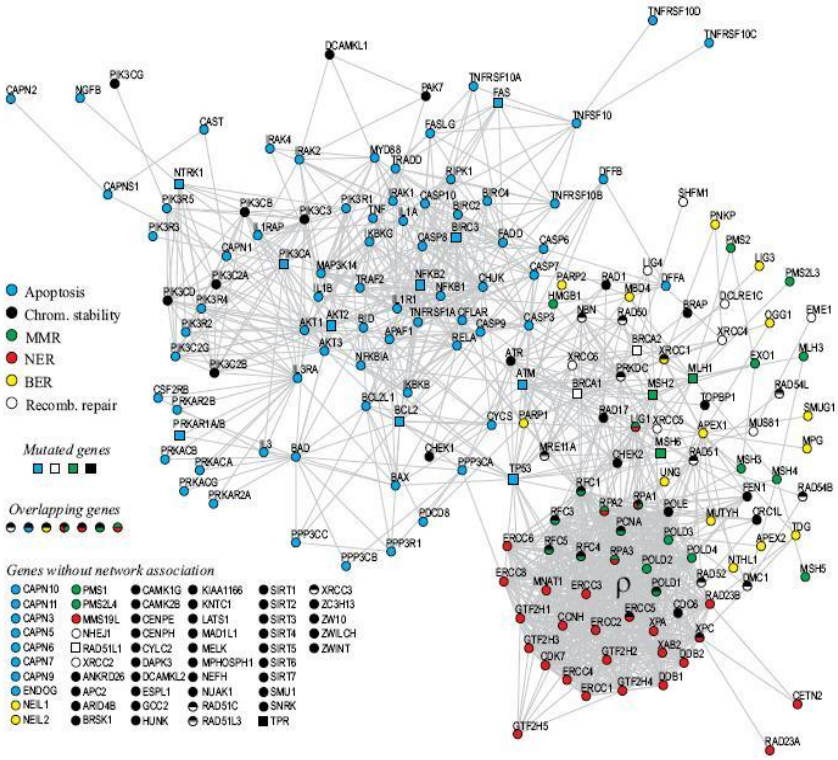

Figura 1. Redes genéticas envolvidas na apoptose, reparo de DNA e estabilidade cromossômica. Os genes de apoptose estão representados em azul, os da rede de Reparo por Excisão de Nucleotídeos (NER) em vermelho. Esta figura é uma adaptação autorizada do original publicado em [2].

2.3 Armazenamento de dados biológicos

A enorme quantidade de informações extraídas dos estudos realizados pela bioinformática e áreas relacionadas fez com que diversos campos de estudos fossem criados para analisar e dar significado a todo este conhecimento. Um desses campos refere-se ao armazenamento de dados resultantes das pesquisas em torno do DNA, das proteínas e de seus produtos.

2.3.1 Banco de dados biológicos

O uso de banco de dados na bioinformática é essencial para o armazenamento e gerenciamento da crescente quantidade de informações extraídas de pesquisas realizadas na área. Sua importância é percebida na dificuldade em analisar de forma construtiva esses dados, de modo que se possam retirar valiosos conhecimentos de sequências de caracteres. Uma das pesquisas pioneiras de análise e armazenamento de dados genéticos foi o Projeto Genoma Humano (PGH), que teve seu início em 1990 e término em 2003, contando com o apoio de mais de 5.000 cientistas em países como Estados Unidos (organizador), Japão, Alemanha, Reino Unido, entre outros. A iniciativa deste projeto foi da NIH (National Institutes of Health), juntamente com o Departamento de Energia Norte-Americana, dirigido por James Watson, um dos responsáveis pela definição do DNA em dupla hélice. Assim como o PGH, outros resultados de projetos encontram-se armazenados em bancos de dados públicos, podendo ser acessados por interessados que buscam unir seus conhecimentos com os de outros pesquisadores, evitando análises de sequências de DNA ou proteínas que já foram estudadas e publicadas anteriormente.

2.3.2 Problemas encontrados no gerenciamento de bases de dados biológicos

Devido à falta de um padrão de alguns procedimentos, referentes à atualização dos bancos de dados biológicos, foram criados diversos meios de armazenar as informações, o que causa hoje uma enorme dificuldade em integrar diferentes bancos com o mesmo fim. Por exemplo, a nomenclatura definida para cada gene varia dependendo do banco em que se está pesquisando. Outro problema refere-se à atualização constante dos dados, uma vez que os dados biológicos crescem exponencialmente e estão constantemente sendo depositados nos diferentes bancos. Visando solucionar alguns dos problemas levantados até este ponto, este trabalho propõe a ontologia Ontocancro, a qual integra dados de vias de interação molecular relacionadas ao câncer. Esta ontologia é descrita em detalhes na próxima seção.

3 Criação de uma ontologia para integração de dados de interatoma e transcriptoma de câncer: a 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 desses 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 gerenciamento, como os bancos de dados públicos, que 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 termos que caracterizam alguma informação, o que provoca transtornos quando há uma tentativa de unificar tais dados.

3.1 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 dessas 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. Nesse contexto, um dos principais objetivos da ontologia é mapear o maior número possível de genes envolvidos nas vias de Manutenção de Estabilidade Genômica, disponibilizando ao usuário uma base de dados

mais completa e manualmente curada. Ela 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 2.

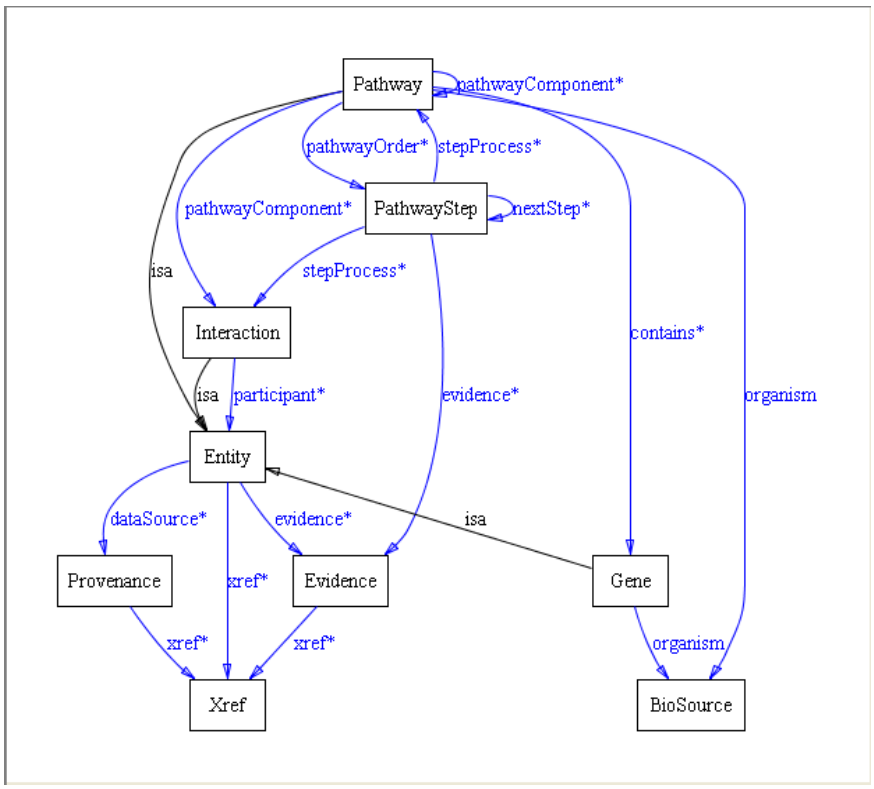

Figura 2. A ontologia Ontocancro

Conforme apresentado na Figura 2, os dois principais elementos da ontologia Ontocancro são os pathways e os genes que compõem 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. Essas 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 se referem 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.

3.2 Arquitetura do sistema para o processamento da ontologia Ontocancro

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, 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. Os 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/
- 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 [8], 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 [10].

O ambiente valida a ontologia gerada 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 ideia 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 [9].

Após a escolha da modelagem e do sistema gerenciador, partiu-se para a integração dos dados. Esta fase foi dividida em três etapas:

- *Aquisição dos dados*: o acesso aos dados dos bancos públicos é feita 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 científica.

Ao final dessas 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, é possível construir uma rede de manutenção de estabilidade genômica e analisar dados públicos de microarranjos de DNA (disponíveis no GEO "Gene Expressiom Omnibus") sobre essa rede. Este tipo de análise permite um maior entendimento de como estas vias se comportam em diferentes tipos e estágios de câncer. A arquitetura do sistema que processa a ontologia e constrói a interface da Ontocancro está representada na Figura 3.

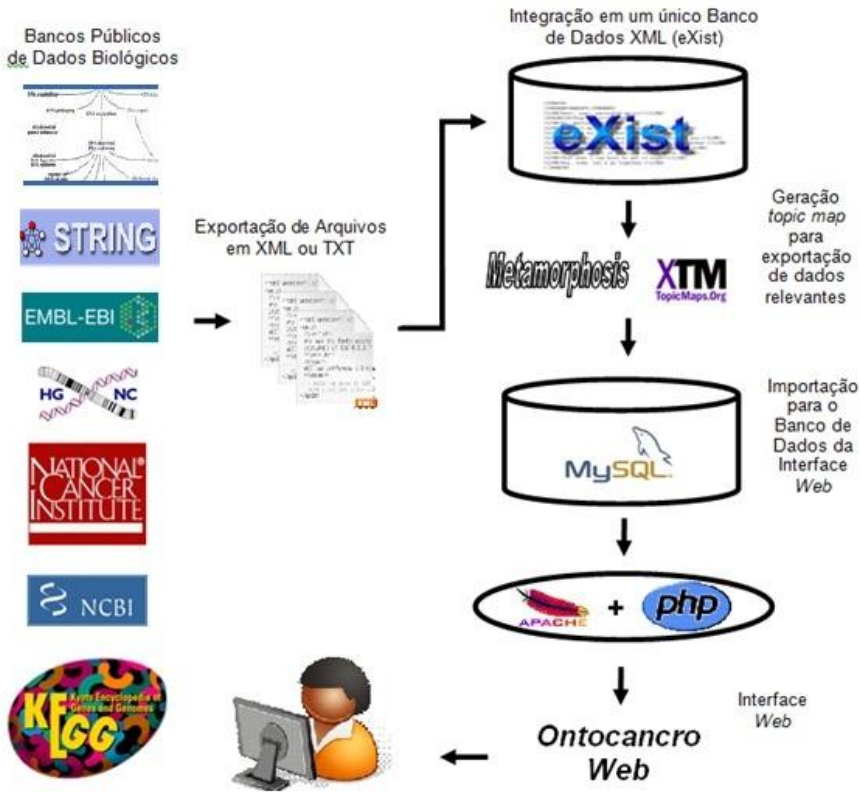

Figura 3. Arquitetura do sistema da Ontocancro

3.3 Integração dos dados dos bancos de dados biológicos

A Ontocancro é composta de informações de bancos de dados biológicos. Como se percebe na Figura 3, 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 referentes a cada banco. Os arquivos de texto são posteriormente tratados para que seus dados estejam em formato XML e, dessa 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.

Nesse banco, os arquivos XML obtidos de cada fonte são mantidos 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 130 arquivos XML. Dentro deste montante destacam-se 32 arquivos oriundos da Biocarta e 65 arquivos obtidos do Gene Ontology; estes são os bancos biológicos que mais contribuem com vias de interação molecular 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 desses 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 de interação molecular, encontrados nos diversos bancos, são utilizados principalmente dados como:

- o código EntrezGene de cada gene, gerenciado pelo NCBI;
- o símbolo e o nome oficial do gene aprovado pelo HGNC;
- o identificador do gene 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 se criar 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. Dessa 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.

3.4 Uma interface para o acesso à 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 *on line* 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:

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

Na tela de importação têm-se os dados da via de interação molecular, 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.

4 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 1.428 genes distribuídos em 130 vias. Todos esses dados foram extraídos dos principais bancos de dados públicos de genes: NCINature, BioCarta, KEGG, Reactome, Prosite, GO e STRING, além dos demais citados anteriormente. Devido à falta de consenso na definição dos conjuntos de genes das vias de estabilidade genômica nas diferentes bases de dados pesquisadas, o projeto disponibiliza aos usuários as vias Ontocancro, que estão relacionadas da seguinte forma:

- Apoptose – 491 genes
- Reparo por Excisão de Base (BER "*Base excision repair*") – 44 genes
- Ciclo Celular (CC) – 286 genes
- Estabilidade Cromossômica (CS "*Chromosome Stability*") – 76 genes
- Apoptose expandida – 955 genes
- Recombinação homologa (HR "*Homologous Recombination*") – 34 genes
- Reparo por mau pareamento de bases (MMR "*Mismatch Repair*") – 28 genes
- Junção de pontas não homólogas (NHEJ "*Non-homologous end-joining*") – 14 genes
- Reparo por Excisão de Nucleotídeos (NER "*Nucleotide Excision Repair*") – 51 genes

É um fato bem estabelecido que a disfunção nas vias de manutenção da estabilidade genômica pode levar ao desenvolvimento do câncer. A via do ciclo celular controla a proliferação celular normal, garantindo que a replicação cromossômica apropriada e a segregação sejam atendidas. A estabilidade de um genoma depende não só de um mecanismo de replicação acurado do DNA, mas também de mecanismos que reparem eficientemente os danos que estão sendo gerados no material genético. Existem vários sistemas de reparo que podem atuar na célula a fim de impedir que um dano genético seja fixado, dentre os quais se encontram o BER, NER, HR, MMR e o NHEJ.

O mecanismo de reparo que é recrutado está relacionado com o tipo de lesão e com a fase do ciclo celular em que a célula se encontra. No caso das vias de reparo falharem na remoção dos danos a célula será encaminhada para a apoptose, também conhecida como morte celular programada, que se constitui em um mecanismo de autodestruição das células danificadas. É importante salientar que a imortalidade é um ingrediente essencial para o processo carcinogênico uma vez que as células necessitam burlar este mecanismo de morte celular para se tornarem cancerosas. Outra via importante é a via de estabilidade cromossômica, que é responsável pela estabilidade dos cromossomos.

O acesso à base de dados da Ontocancro é feito 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. Isso 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. Essas informações estão descritas também no arquivo texto da via, que pode ser visualizado e mais facilmente descrito quando aberto em uma planilha eletrônica.

5 Trabalhos relacionados

Dentre as centenas de bancos de dados biológicos acessíveis via Web, a Ontocancro destaca-se por disponibilizar dados manualmente curados sobre vias envolvidas na manutenção da estabilidade genômica. 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) [4].

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 de genética. Entretanto, tanto o GenBank quanto o Swiss-Prot não possuem uma estrutura de seu conteúdo em forma de vias de interações moleculares, como apresenta a Ontocancro.

Da forma como está estruturada, a Ontocancro permite buscar os genes que estão inseridos dentro de uma mesma via, a fonte de onde vieram esses genes, bem como a rede de interação da via com o nível de confiança de interação entre esses genes (informação essa oriunda do banco String). Todas essas informações permitem ao usuário, inserir ou descartar algum gene da via de acordo com o seu critério de curagem. Além disso, todos os genes listados na Ontocancro apresentam o identificador da plataforma da Affymetrix, facilitando, assim, a utilização dos dados de expressão disponibilizados em bancos de dados de microarranjos.

6 Conclusão

O câncer é uma das doenças mais preocupantes da humanidade. O investimento na busca de alternativas de tratamento demanda muita pesquisa na área e a bioinformática agrega ferramentas que cada dia mais se tornam imprescindíveis nesta pesquisa. 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 único que caracterize de forma exata a informação biológica, pois ocorrem problemas quando se tenta unificar estes dados.

A ontologia Ontocancro foi desenvolvida para contribuir na integração das informações de interatoma e transcriptoma de câncer relacionadas às vias de estabilidade genômica dispersas em diversos bancos de dados espalhados pela Web. Ela propõe uma abordagem integradora para o estudo das redes genéticas diretamente envovidas no processo carcinogênico, como o ciclo celular, o reparo do DNA, a apoptose e outras. Atualmente, a Ontocancro tornou-se uma fonte de pesquisa para os profissionais que estudam o processo carcinogênico.

Atualmente, existem alguns projetos que estão prevendo suas expansões. Um deles planeja a construção de Web Services que permitam a atualização automática da ontologia a partir dos bancos de dados biológicos que lhe servem de fonte, sem a necessidade de uma interferência humana. Outro projeto tem como objetivo a construção de um visualizador gráfico das vias, permitindo a visualização de todas as suas informações (como os relacionamentos entre os genes) em uma mesma tela do computador.

Referências

[1] BARABÁSI, A. L.; OLTVAI, Z. Network biology: understanding the cell's functional organization. s.l.: *Nature Reviews Genetics*, 2004. p. 101-113.

[2] CASTRO, M. A. A. et al. Impaired expression of NER gene network in sporadic solid tumors. s.l.: *Nucleic Acids Research*, 2007. p. 1859-1867.

[3] FUTREAL, P. A. et al. A census of human cancer genes. s.l.: *Nature Reviews Cancer*, 2004. p. 177-183.

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

[5] GOBLE, C. et al. *TAMBS:* Transparent Access to Multiple Bioinformatics Information Sources. s.l.: IBM Syst, p. 532-552.

[6] GRUBER, T.R. Toward principles for the design of ontologies used for knowledge sharing. In: GUARINO N.; POLI R. (Ed.). *Formal ontology in conceptual analysis and knowledge representation.* Dordrecht, Netherlands: Kluwer Academic. 1995.

[7] JEONG, H., et al. The large-scale organization of metabolic networks. s.l.: *Nature,* 407, p. 651-654, 2000.

[8] LIBRELOTTO, G.R.; RAMALHO, J.C.; HENRIQUES, P.R. A *Topic Maps Based Environment to Handle Heterogeneous Information Resources*. s.l.: Lecture Notes in Computer Science, Springer-Verlag GmbH, p. 14-25, 2006.

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

[10] PARK, J.; HUNTING, S. *XML Topic Maps:* Creating and Using Topic Maps for the Web. Addison-Wesley, 2003.

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

[12] SETUBAL, J. C. *ComCiência.* A origem e o sentido da bioinformática. Disponível em: <http://www.comciencia.br/reportagens/bioinformatica/bio10.shtml>. Acesso em: 10 ago. 2003.

[13] 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.