MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Doutorado em Programa de Pós-Graduação Biologia Computacional e Sistemas

METODOLOGIA DE DESENVOLVIMENTO DE *DATA MARTS* PARA APOIO À DECISÃO BASEADO NO USO DE ONTOLOGIAS E ESTUDO DE CASO PARA A PRIORIZAÇÃO DE ALVOS DE FÁRMACOS EM *TRIPANOSSOMATÍDEOS*

MARLON AMARO COELHO TEIXEIRA

Rio de Janeiro
Março de 2018

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

MARLON AMARO COELHO TEIXEIRA

Metodologia de desenvolvimento de *data marts* para apoio à decisão baseado no uso de ontologias e estudo de caso para a priorização de alvos de fármacos em *tripanossomatídeos*.

Tese apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Doutor em Biologia Computacional e Sistemas

Orientador (es): Prof. Dr. Floriano Paes Silva-Júnior
Profa. Dra. Maria Cláudia Reis Cavalcanti

RIO DE JANEIRO
Março de 2018

Teixeira, Marlon Amaro Coelho .

Metodologia de desenvolvimento de data marts para apoio à decisão baseado no uso de ontologias e estudo de caso para a priorização de alvos de fármacos em tripanossomatídeos / Marlon Amaro Coelho Teixeira. - Rio de janeiro, 2018.
110 f.; il.

Tese (Doutorado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2018.

Orientador: Floriano Paes Silva-Júnior.
Co-orientadora: Maria Cláudia Reis Cavalcanti.

Bibliografia: f. 99-108

1. Sistemas de apoio à decisão. 2. Priorização de alvos de fármacos. 3. Anotação semântica. I. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca de Manguinhos/ICICT com os dados fornecidos pelo(a) autor(a).

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: MARLON AMARO COELHO TEIXEIRA

METODOLOGIA DE DESENVOLVIMENTO DE DATA MARTS PARA APOIO À DECISÃO BASEADO NO USO DE ONTOLOGIAS E ESTUDO DE CASO PARA A PRIORIZAÇÃO DE ALVOS DE FÁRMACOS EM *TRIPANOSSOMATÍDEOS*

ORIENTADOR (ES): Prof. Dr. Floriano Paes Silva-Júnior
Profa. Dra. Maria Cláudia Reis Cavalcanti

Aprovada em: 23/03/2018

EXAMINADORES:

Prof. Dr. Maria Luiza Machado Campos **- Presidente** (DCC/UFRJ)
Prof. Dr. Sérgio Manoel Serra da Cruz (UFRRJ)
Prof. Dr. Ana Carolina Ramos Guimarães (IOC/FioCruz)
Prof. Dr. Fabricio Alves Barbosa (IOC/FioCruz)
Prof. Dr. Kelli de Faria Cordeiro (CASNAV)

Rio de Janeiro, 23 de março de 2018.

AGRADECIMENTOS

À minha mãe Ana e meu irmão Bráulio por todo amor e apoio. À minha esposa Valeska, por todo amor, companheirismo e compreensão. A toda sua família pelo carinho e apoio. Aos meus tios e primos por toda torcida e preocupações.

Aos meus orientadores Drª. Maria Cláudia Cavalcanti e Dr. Floriano Paes Silva Júnior por toda dedicação que tiveram comigo e por estarem sempre presentes com suas orientações.

Aos pesquisadores Dr. Sérgio Manoel Serra da Cruz , Drª Maria Luiza Machado Campos e Dr. Ana Carolina Ramos Guimarães por aceitarem tão prontamente a fazer parte de minha banca. Aos Dr. Fabricio Alves Barbosa e Drª Kelli de Faria Cordeiro por serem suplentes.

À secretária Rose da Pós-Graduação de Biologia Computacional e Sistemas por toda atenção. A toda equipe do LaBECFar que sempre me fizeram sentir parte da equipe.

Aos participantes do convênio IFAC/FioCruz pela amizade e apoio nos momentos difíceis do convênio. Aos professores do IFAC, por toda ajuda e disponibilidade de trocar seus horários de aula, permitindo que eu pudesse viajar.

A todos os amigos do LaRCom/Unicamp que mesmo distantes estão sempre na torcida uns pelos outros. Em especial aos queridos amigos Bruno Zarpelão e Rodrigo Miani por todo incentivo.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

METODOLOGIA DE DESENVOLVIMENTO DE DATA MARTS PARA APOIO À DECISÃO BASEADO NO USO DE ONTOLOGIAS E ESTUDO DE CASO PARA A PRIORIZAÇÃO DE ALVOS DE FÁRMACOS EM TRIPANOSSOMATÍDEOS

RESUMO

TESE DE DOUTORADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Marlon Amaro Coelho Teixeira

A anotação de texto semântico permite a associação de conceitos presentes em uma ontologia a expressões textuais (termos), que são legíveis por agentes de software. No cenário científico, isso é particularmente útil, porque muitas descobertas científicas estão "escondidas" dentro de artigos acadêmicos. A área Biomédica possui mais de 300 ontologias, a maioria composta de mais de 500 conceitos. Estas ontologias podem ser usadas para anotar e/ou indexar artigos científicos. No entanto, no contexto de uma pesquisa científica, uma simples consulta baseada em palavras-chave usando a interface de uma biblioteca de textos digitais, pode retornar mais de mil hits. A análise de um conjunto tão grande de textos anotados com ontologias não é uma tarefa fácil. Neste sentido, este trabalho apresenta um método chamado TOETL, para construir uma visão analítica sobre esses textos. A ideia é fornecer uma maneira sistemática de processar um grande conjunto de artigos científicos e apoiar o pesquisador em uma melhor tomada de decisão em relação aos seus interesses de pesquisa específicos. Para ilustrar a aplicação do método, um cenário científico foi escolhido com foco na pesquisa de essencialidade de gene. Este é um conceito muito importante na busca de genes com potencial para novos alvos de fármacos. Um corpus de artigos foi selecionado e semanticamente anotado com três ontologias diferentes. Este trabalho apresenta como os dados de anotação foram extraídos, organizados e agregados em um esquema dimensional de um data mart chamado TaP DM, aplicando a metodologia proposta. O estudo de caso teve como foco os seguintes protozoários: *Entamoeba histolytica, Leishmania major, Plasmodium falciparum, Trypanosoma brucei* e *Trypanosoma cruzi*. O TaP DM permite aos pesquisadores visualizarem, de forma multidimensional, os diversos conceitos envolvidos nos artigos que abordam a essencialidade de genes. Ao final, são realizadas consultas no TaP DM mostrando algumas estratégias de pesquisa sobre esses dados e discutindo como eles podem ajudar o cientista a priorizar alvos de fármacos em protozoários parasitas.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

DATA MARTS DEVELOPMENT METHODOLOGY TO SUPPORT THE DECISION BASED ON THE USE OF ONTOLOGIES AND CASE STUDY FOR PRIORIZATION OF DRUG TARGETS IN TRIPANOSOMATIDES

ABSTRACT

PHD THESIS IN BIOLOGIA COMPUTACIONAL E SISTEMAS

Marlon Amaro Coelho Teixeira

Semantic text annotation enables the association of semantic information ontology concepts to text expressions (terms), which are readable by software agents. In the scientific scenario, this is particularly useful because a lot of scientific discoveries are \hidden" within academic articles. The Biomedical area has more than 300 ontologies, most of them composed of over 500 concepts. These ontologies can be used to annotate scientific papers and thus, facilitate data extraction. However, in the context of a scientific research, a simple keyword-based query using the interface of a digital scientific texts library can return more than a thousand hits. The analysis of such a large set of texts annotated with such numerous and large ontologies, is not an easy task. Here it is described a method called TOETL, to build an analytical view over such texts. To illustrate the method application, a scientific scenario was chosen with focus on the research of gene essentiality. The later is a key concept to be considered when searching for genes showing potential as anti-infective drug targets. A corpus of selected papers was semantically annotated using three distinct ontologies. This work presents how the annotation data was extracted, organized and aggregated into a dimensional schema of a demo Data Mart. Thus, the idea is to provide a systematic way to process a large set of scientific articles and support the researcher in better decision making with respect to his/her specific research interests. We also present a case study on the design and load of a data mart with focus on gene essentiality for the following five protozoa: *Entamoeba histolytica*, *Leishmania major*, *Plasmodium falciparum*, *Trypanosoma brucei* and *Trypanosoma cruzi*. In addition to the TOETL methodology, this work presents as a case study the TaP DM, which was conceived from the application of the proposed methodology. This data mart allows researchers to view, in a multidimensional way, the various concepts involved in articles that discuss the essentiality of genes.

ÍNDICE

ÍNDICE DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

API	*Application Programming Interface*
BI	*Business Intelligence*
DBMS	*Database management system*
DM	*Data mart*
DS	*Data staging*
DW	*Data Warehouse*
EMBL-EBI	*European Bioinformatics Institute*
ETC	Extração, Transformação e Carga
ETL	*Extract, Transform, Load*
IOC/Fiocruz	Instituto Oswaldo Cruz/ Fundação Oswaldo Cruz
MDX	Multi-dimensional expressions
MSF	*Médecins Sans Frontières*
NCBO	*National Center for Biomedical Ontology*
OBO	*The Open Biological and Biomedical Ontologies*
OLAP	*On-line Analytical Processing*
OLTP	*Online Transaction Processing*
OMS	Organização Mundial de Saúde
OWL/XML	*Web Ontology Language/ eXtensible Markup Language*
PIR	*Protein Information Resource*
RDF/XML	*Resource Description Framework/ eXtensible Markup Language*
RNA	*Ribonucleic acid*
SAD	Sistemas de apoio à decisão
SBC	Sistemas baseados em conhecimento
SIB	*Swiss Institute of Bioinformatics*
SKOS	*Simple Knowledge Organization System*
SQL	*Structured Query Language*
TaP DM	*Prioritization Target Data Mart*
TDR	*Tropical Disease Research*
TI	Tecnologia da informação
TOETL	*Text and Ontology* ETL
UMLS	*Unified Medical Language System*
UniProt	*Universal Protein Resource*

SGBDs	Sistemas de gerenciamento de banco de dados

1 INTRODUÇÃO

Atualmente, novas descobertas na área biomédica vêm surgindo devido ao uso de novas técnicas e equipamentos que são utilizados em experimentos de alto rendimento. Um volume crescente de dados estruturados resultantes destes experimentos são constantemente disponibilizados. De acordo com (1), atualmente existem aproximadamente 1.685 bases de dados online de biologia molecular, tornando-se um desafio substancial para as pesquisas e análise de dados, pois os pesquisadores não utilizam apenas uma base de dados, muitas questões biológicas só podem ser respondidas combinando dados de múltiplas fontes (2).

O grande volume de dados dispersos dificulta a condução de pesquisas no campo biomédico. A coleta de informações precisas e confiáveis de diferentes fontes de dados, de forma eficiente, tornou-se uma atividade complexa e muitas vezes não viável sem o desenvolvimento e reutilização de um conjunto de ferramentas para auxiliar neste processo. Como comprovação desta tendência, nota-se a adoção cada vez maior de iniciativas como *Big data,* onde um enorme conjunto de dados impossibilita o processado por sistemas tradicionais, exigindo o desenvolvimento de novas tecnologias para extrair conhecimento e orientar os processos de tomada de decisão (3,4).

Neste cenário, podemos destacar pesquisas que objetivam encontrar novos alvos de fármacos. Onde as complexidades das pesquisas influenciam diretamente no tempo e no custo para produção de um novo fármaco, impactando no preço final do medicamento (5). Considerando doenças tropicais negligenciadas, com ocorrência predominante nos trópicos, onde as populações mais pobres são mais vulneráveis (6), reduzir o preço dos medicamentos, tornando-os mais acessíveis para estas comunidades é de vital importância (5).

Desenvolver ferramentas computacionais, capazes de integrar e correlacionar diferentes fontes de dados, com o objetivo de descobrir novos alvos de fármacos é uma maneira viável de diminuir o tempo e os custos no desenvolvimento de novos medicamentos, possibilitado que populações mais pobres tenham acesso mais rápido e barato aos tratamentos (2).

Outra importante fonte de informação para pesquisas na área biomédica são os repositórios textuais. Eles armazenam informações relevantes, e a partir delas, pesquisadores podem tomar decisões importantes direcionando suas pesquisas. Um

dos mais importantes repositórios digitais é o PubMed (7), armazenando aproximadamente 26 milhões de textos científicos. Através da seleção e leitura destes artigos, o cientista pode estabelecer quais são os tópicos e qual o escopo de sua pesquisa. Mas mesmo esses repositórios contendo relatórios sobre a maioria das descobertas científicas, realizar uma revisão da literatura geralmente leva muito tempo.

As bibliotecas digitais classificam e indexam grandes conjuntos de artigos científicos de acordo com esses tópicos, facilitando ao cientista encontrar os documentos de seu interesse. No entanto, uma simples consulta baseada em palavras-chave, usando a interface de uma biblioteca digital, pode retornar uma lista de milhares de artigos. Isto pode dificultar a classificação dos artigos de acordo com a sua relevância, podendo o usuário descartar documentos muito importantes devido ao grande número de indexação(8).

Outro ponto envolvendo buscas baseadas em palavras chaves é a sua fragilidade na criação das consultas. Por meio de operadores AND e OR, elas podem correlacionar diferentes termos. Mas essas correlações são vulneráveis, porque só são possíveis com termos explicitamente presentes nos artigos. Se uma variação do termo de interesse está presente em um artigo, a ferramenta não será capaz de identificá-lo(8).

Neste cenário surge a anotação semântica, que vem se tornando muito útil para a comunidade de pesquisa biomédica (9–11). Isso ocorre porque os cientistas biomédicos precisam classificar artigos de acordo com seu interesse de pesquisa. Por exemplo, se um cientista estiver interessado no uso da técnica de *shotgun* para identificar peptídeos, será que a indexação simples desses dois termos traria este texto (e talvez outros) para o topo da lista de resultados? Possivelmente. No entanto, é normal que um cientista possa estar interessado em muitas combinações de técnicas e proteínas distintas. Além disso, outros aspectos também podem ser de interesse, como vias metabólicas, organismos, etc. Portanto, para responder a uma simples pergunta o problema pode torna-se multidimensional. A anotação de texto permite a identificação da ocorrência de tais combinações multidimensionais e, portanto, torna possível classificar esses artigos de acordo com o interesse do cientista.

O uso de um sistema de visão analítica combinado com anotação textual permite correlacionar informações sobre diferentes conceitos. Os dados experimentais sobre alvos de fármacos de organismos pouco estudados, por

exemplo, protozoários, podem ser correlacionados com dados relevantes de outros organismos bem estudados (modelo), podendo direcionar os experimentos dos pesquisadores, tornando as buscas menos onerosas e obtendo resultados relevantes em menos tempo (2).

Diante do desafio de correlacionar informações de organismos na busca de novos alvos de fármacos, emerge o conceito de essencialidade genética (12). Os genes são considerados essenciais para um organismo quando sua repressão, silenciamento ou bloqueio resulta na morte do organismo (12). Por este motivo, a descoberta de genes essenciais é de fundamental importância para o desenvolvimento de novos fármacos. Além do mais, a presença de um gene essencial em um organismo, aumenta as chances deste gene ortólogo, permanecer essencial em outros organismos (13).

O objetivo principal deste trabalho é apresentar uma metodologia de projeto de data marts para análise de dados não estruturados por meio do uso de ontologias, assim, fornecendo uma maneira sistemática para processar um grande conjunto de artigos científicos para apoiar o pesquisador em uma melhor tomada de decisão. Este trabalho utiliza as abordagens de apoio à tomada de decisão, amplamente utilizadas na área de negócios (14), para apoiar pesquisas cientificas.

O objetivo secundário deste trabalho é propor uma estratégia para priorizar novos alvos de fármacos, que são identificados por meio de consultas no *data mart* de essencialidade resultante da aplicação da metodologia TOETL. Os dados utilizados para alimentar o *data mart* foram extraídos de artigos científicos através de anotações de texto baseado em ontologia. Os processos de extração, transformação e de carga no *data mart* são relatados detalhadamente.

A metodologia TOETL proposta neste trabalho é uma extensão do trabalho (2), onde foi discutida e detalhada a etapa de Anotação e Extração. Este trabalho está organizado em 5 capítulos. Este primeiro apresenta uma visão geral dos conceitos centrais. O segundo capítulo apresenta os objetivos geral e específicos. No terceiro capítulo é apresentada a proposta da metodologia TOETL, juntamente com sua aplicação, originando o TaP *Data Mart.* Um proposta de estratégia para priorização de novos alvos também é apresentada, finalizando com a integração do TaP DM com uma interface OLAP Pentaho. No quarto capítulo algumas consultas envolvendo características especificas do *data mart* e das ontologias são apresentadas, demonstrando a capacidade da ferramenta de auxiliar os

pesquisadores na busca por resultados mais refinados. Por fim, o capítulo cinco trás as conclusões e trabalhos futuros..

1.1 Web Semântica e Anotação Semântica

O conceito de Web Semântica não se distingue do conceito da Web (15), na verdade elas se complementam. A Web Semântica se apresenta como uma extensão da Web, onde o conteúdo do texto possui associações com seus significados de forma bem definida, possibilitando que agentes de software realizem atividades complexas para os usuários (15).

Anotação semântica é um dos principais esforços para se alcançar a Web Semântica (16), provendo mecanismos para trazer significado aos documentos, isto é, manipular e associar metadados com o conteúdo. A anotação semântica faz uso de linguagem estruturada para manter anotações em um vocabulário uniforme e enriquecer a descrição dos conceitos, através da especificação das relações dos termos de um domínio de conhecimento (17).

O grande volume de informação textual torna a busca por qualquer informação longa e custosa. Em se tratando da área científica, uma simples classificação de artigos pode demandar muito tempo dos envolvidos, pois a quantidade de textos cresce diariamente. Neste cenário, a anotação semântica por meio de metadados, é capaz de associar a informação com seu significado semântico, permitindo novas formas de recuperação de informação. Esta anotação é feita de forma não ambígua, bem definida e de fácil compreensão do domínio de conhecimento pelos especialistas da área, permitindo assim, uma forma de acesso mais eficiente à informação. Além do mais, a utilização de anotação semântica em fontes textuais permite a extração e estruturação das informações em banco de dados estruturados, tornando possível o cruzamento de dados com outras fontes disponíveis (9,18,19).

1.1.1 Ontologias

A conceptualização é uma definição abstrata e simplificada do ambiente, que por algum motivo se deseja descrever. Deste modo, para se representar formalmente um domínio de conhecimento, uma conceptualização deve se apoiar em elementos essenciais, como objetos, conceitos, agentes e seus relacionamentos, que estão envolvidos em uma determinada área de interesse. Todo sistema

baseado em conhecimento está vinculado a algum tipo de conceptualização, seja ela implícita ou explícita (20,21).

Uma ontologia é uma representação explícita de uma conceptualização. O termo ontologia é usado com diferentes significados nas mais diversas áreas do conhecimento, desde a filosofia a computação. Desta forma, pode-se defini-la como um conjunto de termos de representação, onde os elementos presentes em um determinado universo de discurso são descritos por classes, relações, funções e por outros elementos. Tratando mais especificamente da computação, ontologia é definida como um meio de formular a estrutura de um sistema, ou seja, especificar elementos relevantes, e suas relações, de um determinado domínio do conhecimento (9,21).

Existem diversas representações disponíveis para descrever ontologias como: OWL/XML (*Web Ontology Language/ eXtensible Markup Language*) (22), RDF/XML (*Resource Description Framework/ eXtensible Markup Language*) (23), SKOS (24), UMLS (25) entre outras. Cada formato possui características únicas e permite construir ontologias de maneira sistemática, definindo seus elementos e suas relações de forma clara e não ambígua, tornando essas ferramentas muito úteis e eficientes, principalmente quando manipuladas por sistemas computacionais (10).

1.2 Sistemas de apoio à decisão

O desenvolvimento de tecnologias computacionais com o objetivo de auxiliar na resolução de problemas e na tomada de decisão são conhecidos como sistemas de apoio à decisão (SAD). Os estudos que substanciaram esta área surgiram nas décadas de 50 e 60, sendo na década de 70, onde as técnicas foram significantemente desenvolvidas e projetadas para gerenciar bancos de dados mais complexos capazes de acessar informações dentro e fora de uma corporação (26).

A definição do conceito de sistemas de apoio à decisão é ampla, podendo ser definidos de forma mais genérica como sistemas computacionais interativos que auxiliam os usuários nas atividades de tomada de decisão. Outra nomenclatura utilizada para definir estes sistemas são os sistemas baseados em conhecimento (SBC), que fazem uso de uma descrição formal de um domínio de conhecimento, para aplicar mecanismos de raciocínio automatizado (27).

Os sistemas de apoio à decisão estão sendo aplicados em diferentes áreas de conhecimento como medicina, engenharia e negócios. Estes sistemas são

especialmente úteis e necessários em situações em que a quantidade de informação disponível é muito grande, inviabilizando a análise dos dados, de forma eficiente e confiável, por uma pessoa visando uma possível tomada de decisão. Os SADs permitem a integração de diversas fontes de informação, possibilitando acesso e análise de maior quantidade de dados, ampliando a inclusão de informações relevantes, aumentando o poder da ferramenta e da visão do usuário (27).

Os SADs podem também implementar diversas técnicas e métodos econômicos, estatísticos e heurísticos para que o usuário possa comparar de forma rápida diversos cenários em que o problema em análise está inserido, permitindo uma visão profunda do processo de negócio. Estes sistemas não são apenas um meio de apresentação de diversas possibilidades que um tomador de decisão pode escolher, ele é capaz também de analisar dados de forma mais sofisticada e construir alternativas de cenários. Isto possibilita ao usuário tomar decisões para problemas atuais, além de prenunciar futuros problemas e oportunidades, permitindo que as ações do gestor sejam menos reativas e mais proativas as condições externas e internas (27).

Apesar de toda essa complexidade e flexibilidade, os sistemas de apoio à decisão proveem uma interface amigável, de fácil manipulação, com construção de consultas interativas além de relatórios gráficos, sempre focando em como as informações levantadas podem auxiliar o usuário a tomar melhores decisões (26).

1.2.1 Data Warehouse

Sistemas transacionais são projetados para controlar as atividades relacionadas às regras de negócio da empresa e são projetados para executar tarefas rapidamente e repetidamente. Eles lidam com as atividades do dia a dia da corporação, como cadastro de produtos, cadastro de clientes e alteração de dados. Estes sistemas são responsáveis por gerar dados sobre os processos de negócio de uma empresa, o que atualmente vêm sendo considerado um dos ativos mais importantes de uma corporação. Outra característica intrínseca destes sistemas é a realização de operações de atualização dos dados, com o objetivo de refletir o estado atual da corporação. Com a disseminação destes sistemas, cada vez mais as empresas estão acumulando dados gerados por meio destes sistemas (16,28,29).

Por outro lado, sistemas analíticos ou de apoio à decisão são desenvolvidos para avaliar o desempenho das atividades da corporação. Eles são capazes de gerar informações de alto nível gerencial, através da análise dos dados gerados

pelos sistemas transacionais. Em um sistema de vendas, por exemplo, padrões de compras dos clientes são obtidos, possibilitando que novos produtos sejam desenvolvidos e oferecidos (16,29).

Tipicamente, esses sistemas trabalham com grande quantidade de informação, pois nunca realizam operações de exclusão ou alteração de dados, apenas realizam operações de inserção e leitura. Isto porque estes sistemas não têm o propósito de refletir o estado atual da empresa, mas os diversos estados através do tempo, possibilitando a comparação e análise de seu desempenho (16).

Neste contexto, surge o conceito de data warehouse , que é uma abordagem capaz de integrar diversas fontes de dados não voláteis, muitas vezes vindas de fontes heterogêneas e orientadas sobre um assunto ao longo do tempo em um banco de dados chamado Data Warehouse (DW). Um DW reúne dados operacionais gerados pelos sistemas transacionais, transformando-os em dados analíticos para que se possa extrair informações com objetivo gerencial, auxiliando a tomada decisão. Eles são projetados para lidar com grande quantidade de dados analíticos, capazes de gerar relatórios relacionando diversas informações em uma interface amigável e compreensível para o gestor (16,28). A Tabela 1.1 mostra as diferenças mais latentes entre os bancos transacionais e os *data warehouses*.

Tabela 1.1: Banco transacional X *Data warehouse*

	Banco transacional	***Data Warehouses***
Foco	Nível opercional	Nível estratégico
Operação	Inclusão, Alteração e Exclusão	Inclusão e consultas
Interação com usuário	Fixa, pré-definida	*ad-hoc*
Usuários	Operadores	Gerência
Interface	OLTP (*On-line Transaction Processing*)	OLAP (*On-line Analytical Processing*)
Projeto do banco	Normalizado	Analítico
Dados	Atualizado	Histórico

1.2.2 Data Mart

Data marts são coleções de dados sobre um determinado assunto, organizados com o objetivo de ajudar a tomada de decisão de um processo de negócio. Estes são partes temáticas de um *data warehouse* completo (Figura 1.1), onde o termo *data mart* se refere a um *data warehouse* de menor capacidade e limitado a um escopo de domínio de conhecimento restrito (16,28).

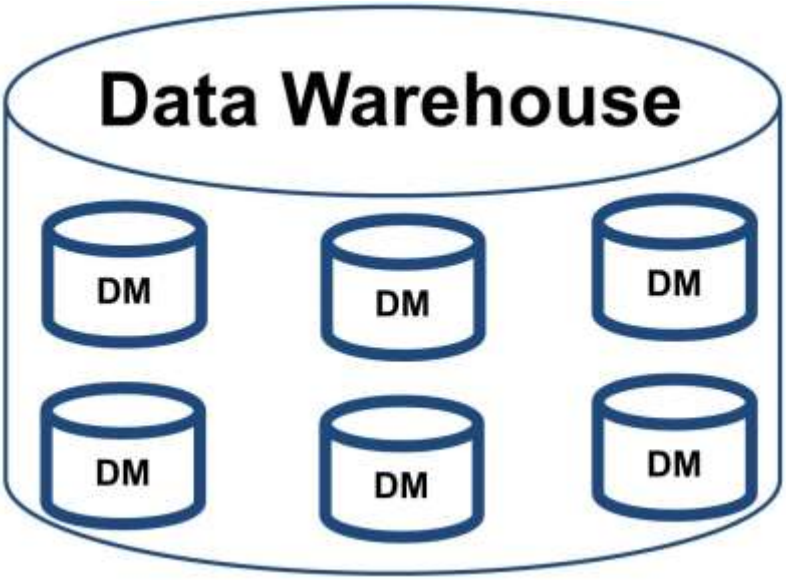


Figura 1.1: Arquitetura do d*ata warehouse.* Figura adaptada (28)**.**

Data marts tratam de áreas de medição (por exemplo, compras de um estabelecimento) e são circundados por entidades descritivas como produtos, clientes, local e tempo. Eles contêm todas as informações nos níveis mais básicos, permitindo por meio de operações, visualizar os dados em todos os níveis (16).

Eles podem ser centralizados ou descentralizados, o que significa que um *data warehouse* coorporativo pode estar fisicamente centralizado em uma única máquina, sendo os *data marts* que o compõem, construídos e integrados ao mesmo tempo, ou desenvolvidos separadamente ao longo do tempo, de acordo com as necessidades da empresa. O último caso ocorre com mais frequência, pois estratégias incrementais e adaptáveis na construção de armazéns de dados refletem mais a realidade das corporações, pois elas estão em constante mudança e expansão, ou seja, novas fontes de dados e novas perspectivas são necessárias ao passar do tempo (16).

1.2.3 Ferramentas OLAP

Em sistemas de apoio à decisão, muitas consultas podem ser criadas utilizando as funcionalidades básicas da linguagem SQL. Em situações mais complexas, a linguagem SQL apresenta limitações, pois nesses casos é necessário a criação de rotinas para obter as respostas, ou exige do usuário um profundo conhecimento (29).

Por estes motivos, *data warehouses* utilizam ferramentas OLAP (*on-line analytical processing*), que possibilitam ao usuário realizar operações como agregação, detalhamento dos níveis das hierarquias, seleção, projeção e

reorientação da visão ao longo de múltiplas dimensões (30,31). Essas ferramentas permitem uma análise mais interativa dos dados, pois possuem diversas extensões da linguagem SQL para dar suporte a estas atividades (29).

Devido ao fato dos bancos de dados analíticos utilizarem ferramentas OLAP para o acesso aos dados, os bancos de apoio à decisão e os transacionais são mantidos separadamente nas corporações, pois bancos transacionais utilizam ferramentas OLTP (*Online Transaction Processing*) para acesso aos seus dados (30).

Ferramentas OLTP automatizam tarefas de processamento de dados como entrada de pedidos e transações bancárias, que são operações do dia a dia das corporações. Essas tarefas são repetidas diversas vezes, consistindo em operações curtas, atômicas e isoladas. Em contrapartida, os bancos analíticos são projetados para o apoio à tomada de decisão. Dados históricos, sumarizados e consolidados são mais importantes que dados pontuais. Isto afeta diretamente o tamanho do banco, pois estes dados representam os estados de uma base operacional ao longo do tempo, tendo consequentemente, o seu tamanho muito maior (30).

Outro motivo de implantar separadamente bancos transacionais e analíticos são os tipos de consultas. As consultas submetidas aos bancos analíticos são mais complexas, pois manipulam uma grande quantidade de registros de diversas tabelas, sendo comum a utilização de percentis, junções e agregações, tornando o desempenho e o tempo de resposta das consultas requisitos fundamentais. Se estas consultas fossem submetidas a um banco transacional, comprometeria o seu desempenho (29,30).

As ferramentas OLAP permitem ao usuário visualizar os dados em qualquer nível de granularidade. Caso ele deseje ver os dados em um nível mais detalhado ele pode realizar uma operação de *drill-down*, ou se desejar acessar um nível mais genérico ele utiliza a operação *roll-up*, possibilitando assim atender as necessidades de diferentes usuários (29).

Umas das ferramentas mais importantes que disponibiliza uma interface OLAP é o Pentaho. O Pentaho possui um conjunto completo de funcionalidades capazes de executarem os processos de extração, transformação, carga (ETC), análise preditiva e várias maneiras de apresentação de resultados. Ele facilita a integração com diversas fontes de dados, além de disponibilizar uma interface web capaz de criar, visualizar e aplicar permissões para geração de relatórios (32).

A interface OLAP disponibilizada pelo Pentaho é chamada Mondrian, onde é possível projetar e construir cubos para realizar a análise dos dados por diferentes perspectivas. A interface é muito intuitiva e permite cruzar informações de diferentes dimensões de maneira rápida e descomplicada. A sua interface web possibilita o acesso ao banco de maneira direta e sem depender de softwares específicos (32), como mostra a Figura 1.2.

Figura 1.2 - Interface Mondrian.

O Mondrian permite gerar de forma simples relatórios e gráficos dos resultados para melhor visualização dos dados. A criação do cubo OLAP é feita por meio da linguagem MDX (*multi-dimensional expressions*). Esta linguagem é um padrão para realizar consultas multidimensionais e permite, de forma poderosa, expressar consultas analíticas (32,33).

1.2.4 Cubo de Dados Multidimensional

Uma maneira de implementar um DW é utilizando os chamados cubos de dados, também conhecidos como hipercubos. Estes cubos são compostos por células e dimensões. As células são responsáveis por armazenar valores numéricos conhecidos como medidas e as dimensões possuem a função de classificá-las (34), como apresentado na Figura 1.3.

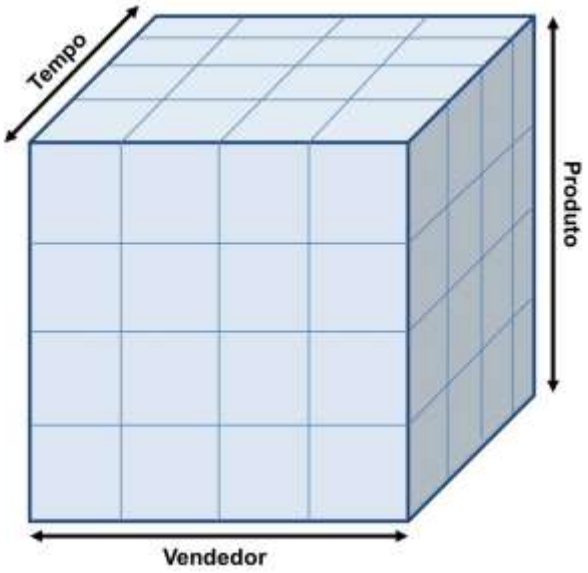


Figura 1.3: Cubo multidimensional.

As medidas, por serem valores numéricos, aceitam operações aritméticas e de cálculos estatísticos, permitindo à ferramenta, de acordo com as necessidades dos usuários, realizar inúmeras análises sobre os dados, aumentando sua flexibilidade e riqueza de detalhe nas suas respostas. Elas representam o nível mais básico da informação, pois permite que de acordo com as dimensões, seja possível calcular outras medidas derivadas (35).

O cubo pode possuir inúmeras dimensões podendo ser compostas por hierarquias e o número de níveis hierárquicos depende da necessidade dos usuários. Se os usuários necessitarem de informações mais detalhadas, a dimensão deve abranger níveis mais baixos. A representação das hierarquias nas dimensões não é obrigatória, mas aumenta o grau de refinamento das respostas e a flexibilidade da aplicação. Por exemplo, considerando um DW de vendas, se for considerado que a dimensão local de venda possui os níveis país, região, estado e cidade, o usuário será capaz de saber quais estados ou quais regiões mais venderam um determinado produto (35).

Existem diversas operações que podem ser realizadas nos cubos de dados para atender às mais variadas necessidades das análises dos usuários. As mais comuns são *slice*, *dice*, *drill-down*, *roll-up* e *pivot* (35).

- A operação de *slice* é utilizada quando o usuário necessita analisar um subconjunto de dimensões. Ela reproduz um efeito de fatiar o cubo, ou seja, para uma ou mais dimensões são fixados valores e então se consegue analisar o comportamento dos dados das outras dimensões.

- O operador *dice* é utilizado quando a resposta de uma operação *slice* retorna muitas tuplas e o usuário necessita filtra-las para se concentrar em poucas linhas. Esse operador define valores fixos para as dimensões de interesse, reduzindo o número de respostas e possibilitando um estudo mais detalhado.
- As operações *drill-down* e *roll-up* permitem aos usuários navegar entre os diferentes níveis da representação hierárquica das dimensões. O *drill-down* é a operação de navegar de um nível mais geral para um nível mais específico ou mais detalhado. Já o *roll-up* é a operação inversa, é a navegação do nível mais específico para o nível mais generalista, realizando uma agregação.
- O operador de *pivot* é capaz de reorganizar a estrutura do cubo de dados, pois dependendo da situação, o usuário pode precisar inverter as dimensões apresentadas em linhas e colunas, possibilitando maior compreensão da situação analisada.

1.2.5 Modelagem relacional para dados multidimensionais

O modelo de dados relacional é o modelo mais popular implementado nos sistemas de gerenciamento de banco de dados (SGBDs) comerciais. Ele representa um banco de dados como um conjunto de tabelas, onde cada uma delas pode ser armazenada separadamente em um arquivo. Eles oferecerem flexibilidade e desenvolvimento rápido de consultas, principalmente quando os requisitos são modificados (36). Existem diversas soluções e suas variações para se implementar um banco de dados analítico em um banco relacional, o esquema estrela é uma das técnicas mais conhecidas e utilizadas para este propósito (37).

O esquema estrela (Figura 1.4) é a representação multidimensional de um cubo de dados em um banco relacional. Ela possui este nome por causa da disposição espacial dos seus 2 principais componentes, a tabela de fatos que ocupa a parte central do diagrama e é envolvida pelas tabelas de dimensão ao seu redor, possuindo um relacionamento 1 para N, tornando mais simples a construção das consultas e permitindo a aplicação de técnicas para melhorar o desempenho (35).

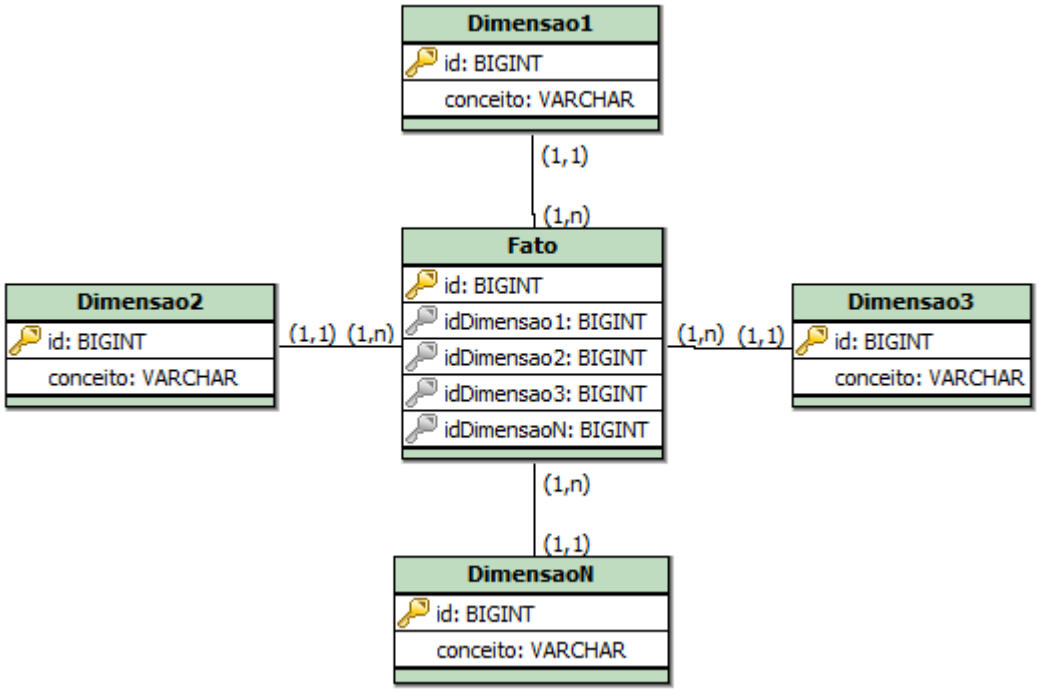


Figura 1.4: Esquema estrela. Adaptadao de (28)**.**

Assim como nos hipercubos, a modelagem relacional de um banco multidimensional também é projetada para possibilitar de maneira eficiente a recuperação de dados que possuam algum tipo de relação, permitindo aos usuários a análise e visualização através de diversas perspectivas, também conhecidas como dimensões (38).

As tabelas de fatos armazenam dados numéricos que representam medidas de eventos ocorridos e estão sempre vinculadas às tabelas de dimensão através de chaves estrangeiras, mantendo a ligação com a tabela de fatos. As tabelas de dimensão, por sua vez, possuem a função de armazenar os dados que descrevem as situações em que os fatos ocorreram (35).

As tabelas de dimensão não devem passar por processos de normalização. Elas normalmente não estão suscetíveis a operações de alteração e exclusão de dados, tornando-se tabelas menores e mais estáveis em comparação a de fatos, pois sua principal função é manter as informações para uma futura recuperação. Desnormalizar tabelas de dimensão de tamanho pequeno não causa grande impacto no desempenho do banco, mas não é incomum ver tabelas de menor tamanho serem normalizadas. Este processo é uma variação do esquema estrela, denominado de esquema floco de neve. Com isso, eliminando as normalizações eliminam-se as operações de união, contribuindo para melhorar o desempenho do banco (35).

1.2.6 Extração, Transformação e Carga (ETC)

Os processos de ETC (Extração, Transformação e Carga) são etapas fundamentais para o sucesso da implantação de um DW. Todos os processos e etapas necessários que envolvem os dados operacionais com o objetivo de transportá-los até o banco analítico fazem parte dos processos do ETC (28).

A primeira etapa é a extração, onde a fonte de dados é lida e compreendida com o objetivo de selecionar quais dados são necessários ou não ao DW. É muito comum nesta etapa, identificar a necessidade de obtenção de dados de mais de uma fonte, além de descartar dados que não fazem parte do escopo do DW (28).

Depois da extração realizada, os dados passam pelo processo de transformação. Eles são submetidos a operações de correção, ajustes ou até mesmo à geração de novos dados, como por exemplo metadados. Por fim, a última etapa é a carga, os dados são armazenados na nova estrutura analítica envolvendo tabelas de dimensões e fatos (28).

1.2.7 Técnicas de modelagem dimensional

Para realizar a modelagem dimensional de um DW, pode se aplicar 2 abordagens básicas: orientada aos dados e orientada pela demanda. A abordagem orientada aos dados necessita de uma análise detalhada e aprofundada dos dados em um processo de reengenharia, com o objetivo de identificar os conceitos dimensionais. Em contraposição, a abordagem orientada à demanda se concentra nos requisitos identificados através das necessidades do usuário para identificar os elementos dimensionais (39).

Na abordagem orientada à demanda, muito antes da equipe de construção do DW se preocupar com os elementos dimensionais, ela precisa compreender as necessidades do negócio. Estes requisitos que devem ser atendidos pelo DW serão descobertos através de entrevistas e reuniões com os especialistas do negócio para o qual será projetado, ou seja, o modelo dimensional deve ser construído em colaboração com os representantes do domínio de conhecimento ou negócio. Existem 4 etapas chaves para se projetar o modelo dimensional (28):

- Escolha do processo de negócio;
- Escolha do grão;
- Identificação das dimensões;
- Identificação dos fatos.

1.2.7.1 Escolhas do Processo de Negócio e do Grão

Os processos de negócio são as atividades operacionais que a corporação realiza. As métricas de desempenho, que são geradas através de eventos ou extraídas dos processos de negócio, são traduzidas em fatos e estes são armazenados na tabela de fatos. Muitas destas tabelas são projetadas para atender apenas um processo de negócio, por este motivo, a escolha do processo de negócio é uma etapa crítica na modelagem multidimensional (28).

A definição do grão é a etapa central no projeto multidimensional, pois ele define o que cada fato da tabela de fatos representa no processo de negócio. Esta definição deve acontecer antes da identificação das tabelas de dimensão e de fatos, pois elas devem ser coerentes com a definição do grão (28).

O grão define o maior nível de detalhe das informações que o DW fornecerá ao usuário. É fortemente recomendado que se utilize o grão no seu nível mais baixo, ou seja, no nível atômico, pois permitem ao usuário realizar consultas mais amplas. Por exemplo, se o grão é definido como vendas diárias de uma loja, por meio da operação de *roll-up* é possível obter as vendas semanais, mas não é possível por meio da operação de *drill-down* obter as vendas horárias. Definir o grão em níveis mais agregados melhora a performance do banco, mas em contrapartida impossibilita a realização de algumas consultas (28).

1.2.7.2 Identificação das Dimensões

As dimensões representam os conceitos e os atributos descritivos que envolvem os eventos de um processo de negócio. As tabelas de dimensão são as “essências” do *data warehouse*, pois descrevem os rótulos que possibilitam a ferramenta de BI realizar suas análises. Os atributos descritos na tabela de dimensão são utilizados para filtrar e agrupar os fatos, sendo recomendado que um valor presente na tabela de fatos deve ser referenciado a um único valor na tabela de dimensão.

Em resumo, cada tabela de dimensão tem uma única chave primária que é utilizada como chave estrangeira para associar o fato às dimensões que o descreve. Nas tabelas de dimensão, o campo escolhido para ser a chave primária não pode ser a chave candidata natural do banco operacional. Deste modo, chaves artificiais

são criadas, desvinculando a chave do DW da chave do banco operacional, tornando-a independente e mais durável (28).

1.2.7.3 Identificação dos fatos

Fatos são resultados de medições realizadas nos eventos de um processo de negócio e na maioria das vezes são representados por valores numéricos. Cada fato deve respeitar o nível do grão definido e cada linha da tabela de fatos possui uma relação de 1 para 1 com as dimensões com que ela se relaciona. É importante salientar que os valores numéricos que representam um fato são obtidos de eventos operacionais realizados e não de dados estimados (28).

Em algumas circunstâncias, o projetista pode se deparar com casos onde uma linha da tabela de fatos possui referência *null* para uma determinada dimensão. Estes casos não prejudicam o correto comportamento do banco, até mesmo as funções de agregação funcionam sem problemas. Mas o valores *nulls* devem ser evitados. Para isso, nas tabelas de dimensão devem possuir valores padrões para que sejam referenciados, quando a tabela de fatos necessitar representar um valor *null* (28).

Outro recurso necessário a ser empregado em um projeto de *data warehouse* são as *factless fact tables*. As tabelas de fato normalmente representam medições numéricas para um determinado evento de interesse, mas em algumas condições a tabela de fatos não mantém nenhuma medição e sim apenas chaves que fazem referências às dimensões. Esses dados não calculáveis descrevem condições em que um fato ocorreu, sendo muito útil para extrair dados estatísticos. As *factless tables* tornam o banco muito mais flexível, pois elas permitem mapear diversas situações em que os fatos ocorreram. Sem esse recurso seria necessário implantar mais de uma tabela de fatos para permitir a análise de mais de um cenário (28,40).

1.3 Doenças de populações negligenciadas

O desenvolvimento de um novo fármaco envolve um grande investimento de recursos. Estudos recentes (41) revelam que em todo o processo de pesquisa e desenvolvimento de uma droga, os custos chegam a atingir aproximadamente 2,5 bilhões de dólares. A Figura 1.5 mostra como a tendência de aumento destes valores vem ocorrendo ao decorrer das décadas.

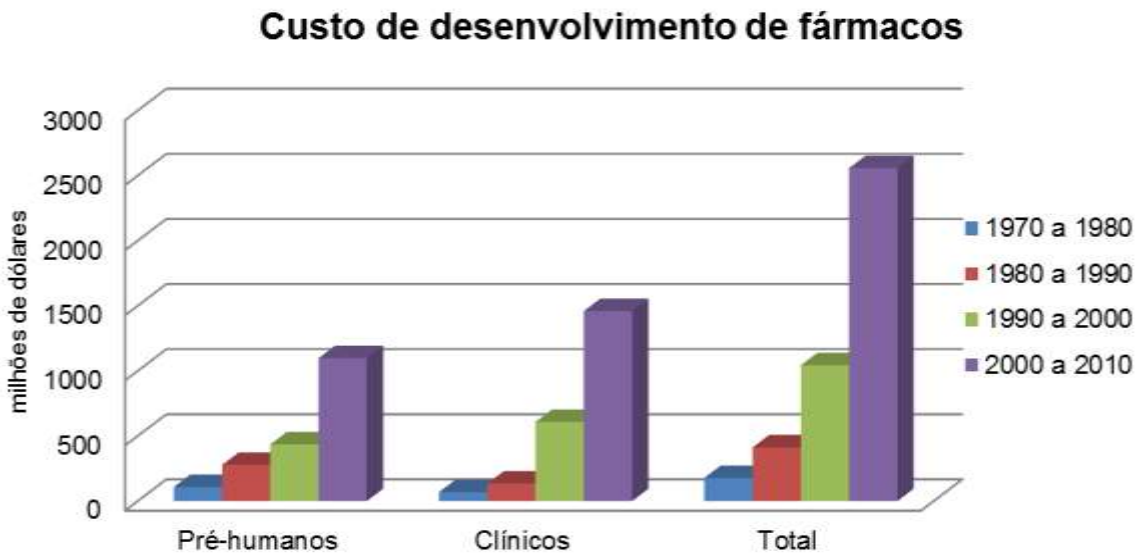


Figura 1.5: Custo de pesquisa e desenvolvimento de fármacos ao longo das últimas décadas, divididos nas fases pré-clínicos e clínicos. Adaptada de (41).

As principais causas para o crescimento dos custos são o aumento da complexidade e do tamanho dos ensaios clínicos (42), influência da inflação no custo dos insumos utilizados no desenvolvimento e na pesquisa, possíveis mudanças nos protocolos para incluir esforços de reunir informação de avaliação das tecnologias de saúde, além de testes em comparadores de fármacos para atender as demandas dos pagadores pelos dados de eficiência comparativa (41).

Este cenário contribui para o surgimento das doenças conhecidas como de populações negligenciadas. Essa denominação faz referência às doenças que despertam pouco interesse das grandes empresas farmacêuticas. Essas instituições não veem nessas doenças potenciais compradores de medicamentos, já que os mais afetados, na grande maioria, são populações pobres e não atenderiam ao lucro esperado por estas corporações (43,44).

O termo “doenças negligenciadas” foi proposto inicialmente na década de 70 pela Fundação Rockefeller. Em 2001, a organização Médicos Sem Fronteiras (MSF) publicou o documento *Fatal Imbalance* (45), onde propõe a divisão em 3 classes de doenças: Globais, Negligenciadas e Mais Negligenciadas. No mesmo ano a Organização Mundial de Saúde OMS também categorizou em 3 classes as doenças no mundo: Tipo I (equivalente às doenças globais dos MSF), Tipo II (Negligenciadas/MSF) e Tipo III (Mais Negligenciadas/MSF). A partir de então esta classificação vem sendo utilizada para se referir a doenças causadas por agentes infecciosos e parasitários endêmicas em populações de baixa renda vivendo,

principalmente em países em desenvolvimento na África, Ásia e nas Américas (43). A incidência das doenças negligenciadas no mundo é apresentada na Figura 1.6.

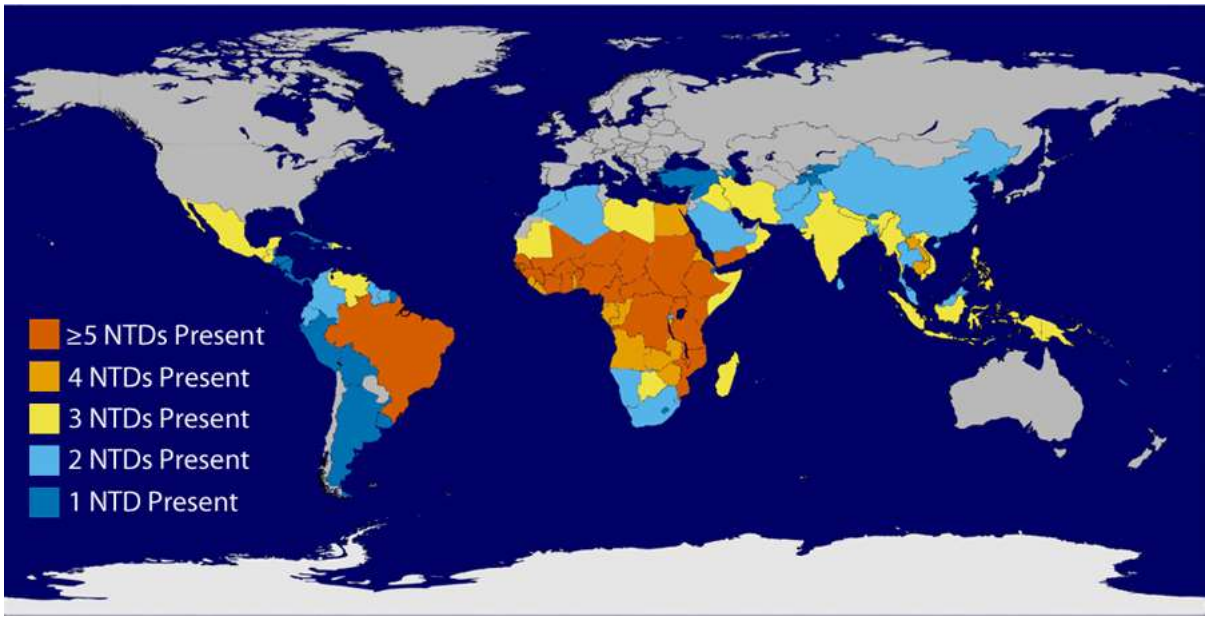


Figura 1.6: Distribuição das doenças negligenciadas[1].

Dentro do conceito de doenças negligenciadas, inicialmente foram inseridas a Doença de Chagas, Doença do Sono, Leishmanioses, Malária, Filarioses, Esquistossomose, Hanseníase, Tuberculose, Dengue, Febre Amarela e HIV/AIDS. Mais recentemente, doenças com menos incidência como Ascaríase, Tricuríase, Necatoríase, Ancilostomíase, Tracoma, Dracunculíase e a Úlcera de Buruli também entraram na lista. Nos últimos anos, doenças como HIV/AIDS, Tuberculose e Malária vem recebendo maior atenção dos governos e financiadores, não podendo ser mais classificadas como negligenciadas (43).

Um levantamento sobre o financiamento mundial de inovação para doenças negligenciadas revelou que menos de 5% do financiamento foi aplicado em doenças extremamente negligenciadas. Portanto, mesmo considerando que mais de 500 milhões de pessoas estejam expostas a enfermidades como Doença do Sono, Leishmaniose Visceral e Doença de Chagas, os investimentos foram baixíssimos (46).

A quantidade de investimento realizado reflete diretamente no número de fármacos desenvolvidos. Entre os anos de 1975 e 2004, apenas 18 fármacos foram produzidos para o combate a doenças tropicais (47), como mostra a Figura 1.7:

[1] Fonte: *Global Health – Division of Parasitic Diseases and Malaria.* *h*ttps://www.cdc.gov/globalhealth/ntd/diseases/ntd-worldmap-static.html.

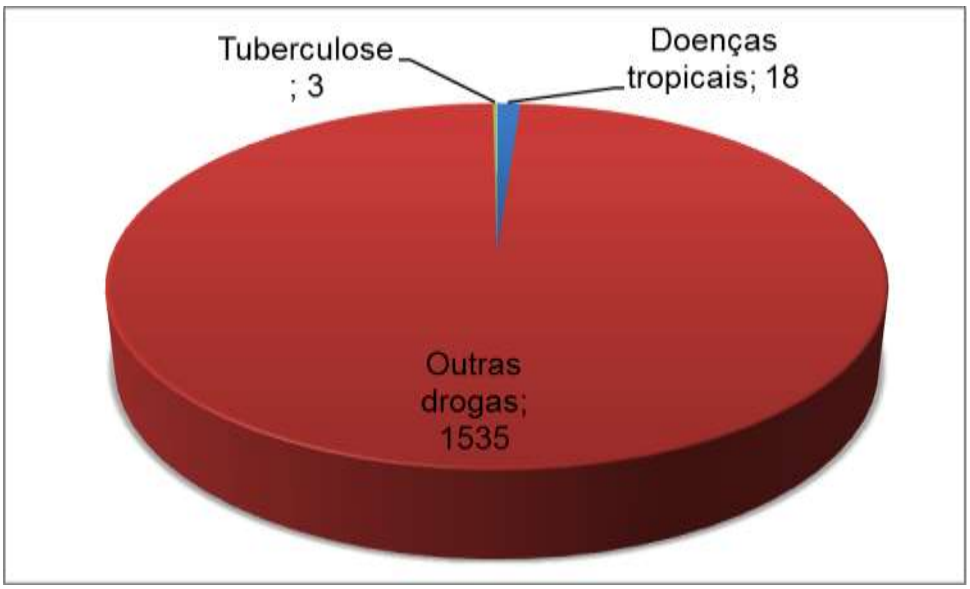


Figura 1.7: Novos medicamentos desenvolvidos entre 1975 e 2004. Adaptada de (47)**.**

Mesmo diante deste cenário desfavorável, os pesquisadores brasileiros sempre se destacaram ao longo dos anos na investigação das doenças causadas por protozoários e helmintos. Na primeira década do século 20, Pirajá da Silva e A. Splendore realizaram importantes estudos sobre os organismos *Schistosoma mansoni* e *Toxoplasma gondii*, respectivamente. Na mesma década, C. Chagas realizou um dos mais importantes trabalhos da ciência brasileira descrevendo o *Trypanosoma cruzi,* e alguns anos após, Gaspar Viana descreveu a *Leishmania braziliensis*.

Todos estes estudos estimularam inúmeros pesquisadores a aprofundar e realizar novas descobertas, e desde então o Brasil vem sendo um dos maiores divulgadores de resultados científicos envolvendo estes agentes, principalmente para o *Trypanosoma cruzi* (43), como mostra a Figura 1.8.

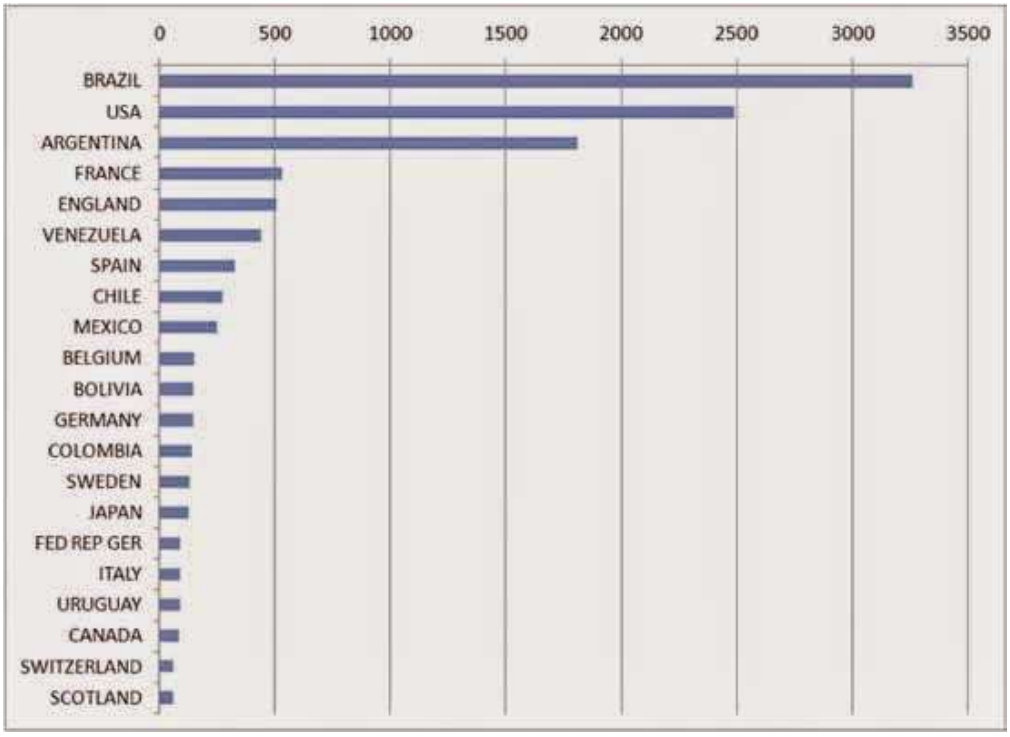


Figura 1.8: Artigos científicos publicados[2].

As ações mais incisivas por parte do governo brasileiro, com relação às doenças negligenciadas, foram feitas somente a partir 2003. Por meio do Ministério da Saúde foi lançado o primeiro edital voltado para atender Tuberculose, seguido pelos editais de Dengue (2004) e Hanseníase (2005) (48). Com o passar dos anos, mais recursos vêm sendo empregados no desenvolvimento de novos fármacos, como mostra a Tabela 1.2.

Tabela 1.2: Recursos investidos pelo governo brasileiro no combate à doenças negligenciadas[3].

Ano	Edital	Recursos
2003	Rede Tuberculose	R$ 1,9 milhões
2004	Dengue	R$ 945 mil
2005	Hanseníase	R$ 2,5 milhões
2006	Doenças negligenciadas	R$ 17 milhões
2008	Doenças negligenciadas	R$ 22 milhões
2009	Rede Malária	R$ 15,4 milhões
2009	Rede Dengue	R$ 22,7 milhões

As informações apresentadas demostram o quanto é urgente a necessidade de desenvolvimento de medicamentos que combatam estas doenças. Outro fato que chama a atenção é a quantidade de artigos científicos publicados, confirmando o empenho da comunidade científica em descobrir novas soluções. Com isso, as

[2] Fonte: Doenças negligenciadas. https://www.abc.org.br/IMG/pdf/doc-199.pdf (43).

[3] Fonte: Doenças negligenciadas: estratégias do Ministério da Saúde.

revistas científicas se tornam um importante meio de publicação de resultados não estruturados e uma ferramenta importante na tomada de decisão dos cientistas na descoberta de novos alvos de fármacos.

1.4 TDR Targets

Outro meio de divulgação de dados de fármacos são os bancos de dados estruturados. O banco de dados TDR[4] (*Tropical Disease Research*) *Targets,* de acesso aberto, é um projeto da Organização Mundial da Saúde e permite aos usuários investigar informações específicas de patógenos em escala genômica, identificando e priorizando a partir de critérios definidos pelos usuários (5,49).

O foco do banco de dados TDR *Targets* é reunir informações de patógenos de doenças tropicais e de outros patógenos filogenéticamente relevantes. Atualmente ele armazena informações sobre 12 organismos: *Mycobacterium leprae, Mycobacterium tuberculosis, Wolbachia endosymbiont of Brugia malayi, Brugia malayi, Caenorhabditis elegans, Schistosoma mansoni, Plasmodium falciparum, Plasmodium vivax, Toxoplasma gondii, Leishmania major, Trypanosoma brucei e T. cruzi* (5,50).

O TDR *Targets* também integra informação de produtos gênicos de bases primárias de informação, reunindo de diversas fontes e estudos publicados informações de ortológos, estruturas 3D, classificação de vias metabólicas e essencialidade. Ele também é alimentado com informações curadas retiradas da literatura sobre validação de alvos. Informações sobre essencialidade de um organismo são reunidas principalmente de dados experimentais (5,49).

Combinar dados genômicos com dados químicos é essencial para o sucesso de novas descobertas de fármacos. A disponibilidade de grande quantidade de compostos é particularmente importante para a pesquisa de doenças tropicais. A importância dessas iniciativas é indiscutível, dada a grande necessidade do desenvolvimento de novos fármacos devido ao rápido aumento dos níveis de resistência aos existentes e o pouco esforço, por motivos comerciais, da indústria farmacêutica no desenvolvimento de fármacos para doenças tropicais (5).

[4] Disponível em http://tdrtargets.org.

1.5 UniProt

O *Universal Protein Resource* (UniProt[5]) é uma colaboração entre as instituições *European Bioinformatics Institute* (EMBL-EBI), SIB *Swiss Institute of Bioinformatics* e o *Protein Information Resource* (PIR), onde aproximadamente 100 pessoas estão envolvidas em diferentes tarefas como validação da base de dados, desenvolvimento de *software* e suporte (51).

Ele é uma fonte central de armazenamento e interconexão de informações sobre proteínas e anotação funcional, contendo dados sobre enzimas específicas, domínios e sítios biologicamente relevantes, localizações subcelulares, características especificas dos tecidos, estrutura, interações, doenças causadas por deficiências ou anormalidades, etc (51).

O banco de dados UniProt é composto de diversos componentes. Uma parte do banco armazena dados curados manualmente e é conhecido como UniProtKB/Swiss-Prot, contendo aproximadamente 550 mil sequências. Todas as outras sequências são armazenadas em um banco não curado conhecido como UniProtKB/TrEMBL, que naturalmente possui um tamanho muito maior que o UniProtKB/Swiss-Prot, com aproximadamente 87 milhões de sequências (52).

Umas das principais atividades do UniProt é realizar a validação de dados sobre proteínas. Esta validação é feita por meio da literatura, provendo alta qualidade nas anotações por proteínas caracterizadas experimentalmente através de diversas famílias e grupos taxionômicos (51).

1.6 Essencialidade de genes

Neste trabalho, definimos que um gene é considerado essencial para um organismo quando a supressão, silenciamento ou bloqueio deste gene veta o crescimento ou implica na morte do organismo. Identificar genes essenciais é um passo muito importante no desenvolvimento de novos fármacos, pois estes genes são potencialmente novos alvos de fármacos (53–55).

Por meio da abordagem de genômica comparativa vem se constatando que genes essenciais possuem alta taxa de conservação durante a evolução dos

[5] Disponível em http://www.uniprot.org/.

organismos, caracterizando-os como genes ortólogos. Genes ortólogos são genes de diferentes espécies, originados de um único gene de um ancestral comum (56).

Muitos trabalhos experimentais vem sendo desenvolvidos com o objetivo de descobrir genes essenciais (57). Estes trabalhos se apoiam em técnicas como nocaute gênico (58) e RNA de interferência (59) para determinar se o gene possui um papel essencial no organismo.

Um dos trabalhos publicados que evidencia a importância da aplicação destas técnicas, concluiu que dentre milhares de genes pertencentes aos organismos *Bacillus subtilis*, *Escherichia coli* e *Mycoplasma genitalium,* apenas 271, 304 e 380 genes, respectivamente, de fato exercem um papel essencial (60–63).

De fato, estas técnicas são fundamentais no estudo de organismos patógenos, mas aplicá-las demanda grande quantidade de esforço, conhecimento e custo por parte dos pesquisadores e de suas instituições. Por estes motivos, existem poucas informações sobre essencialidade de alguns organismos, como é o caso dos protozoários, que são os principais causadores de doenças negligenciadas no mundo (2).

Por outro lado, muitos organismos são muito estudados na busca de genes essenciais, como é o caso dos organismos *Saccharomyces cerevisiae*, *C. elegans* e *Danio rerio* (*Zebrafish*) (64,65). Estes organismos com amplo estudo sobre essencialidade de seus genes são conhecidos como organismos modelo.

Cruzar informações de genes essenciais, por meio de ferramentas computacionais, de organismos modelo com organismos ainda pouco estudados é um jeito mais rápido e menos custoso de descobrir novos potenciais alvos, possibilitando o desenvolvimento de novos fármacos.

1.7 Organismos modelo

A identificação sistemática de genes com funções essenciais foi descrita para vários procariontes e eucariotos. Estes estudos forneceram perspectivas valiosas sobre o conjunto mínimo de genes necessário para funções de células básicas em uma ampla gama de organismos (66).

A compreensão das funções gênicas em alguns organismos menos complexos pode esclarecer o papel de genes em organismos mais complexos, facilitando a compreensão do funcionamento de mecanismos mais sofisticados (67).

É neste cenário que surgem os organismos modelo. Eles são organismos que possuem características próximas à maioria dos organismos e principalmente ao organismo que se deseja estudar. Em geral são organismos que apresentam vantagens na pesquisa experimental em laboratório, possuindo um tamanho pequeno, tempos de geração e ciclo de vida curto, fácil disponibilidade e manutenção (2,67).

Eles são organismos muito estudados, com o genoma quase ou totalmente sequenciado, com abundantes informações sobre regulação gênica, processo evolutivo e doenças. Os organismos modelo tornam acessíveis, por meio da pesquisa experimental, respostas às perguntas sobre biologia que não seriam possíveis em outros organismos (67). Nos itens a seguir, alguns dos principais organismos modelos são descritos:

- *Escherichia coli (E. coli)*

É uma bactéria bacilar encontrada com mais frequência no intestino humano e de alguns animais de sangue quente. Grande parte de suas cepas não causam enfermidades aos humanos, possuindo uma relação de simbiose com seu hospedeiro. Mas algumas variedades podem causar intoxicações. Ela possui a característica de ser produzida de maneira fácil e com baixos recursos financeiros, por estes motivos, vem sendo estudada desde a década de 60 e é um dos organismos procariotos mais estudados (68,69).

Diversos trabalhos experimentais publicados buscam encontrar genes essenciais na bactéria *E. coli*, onde 87% do seu genoma já foram estudados e demonstra que os genes essenciais identificados se mantiveram preservados através do reino das bactérias, especialmente genes ligados à replicação do DNA e síntese de proteínas (70).

- *Saccharomyces cerevisiae*

A levedura *Saccharomyces cerevisiae* é um organismo eucarioto unicelular pertencente ao reino dos Fungos, podendo estar presente na mucosa gastrointestinal, respiratória e urinária. Na década de 70 ganhou destaque como peça importante na expansão de descobertas em organismos procariotos. Ele também foi o primeiro organismo a ter o mapeamento completo do genoma e o único organismo pelo qual uma completa coleção de linhagem nocautes mutantes foi feita (71,72).

Com a evolução dos estudos do genoma do *Saccharomyces cerevisiae*, uma descoberta criou grande impacto na comunidade cientifica. Foram identificados nas leveduras diversos genes conservados através da evolução, viabilizando a ideia de que análises comparativas de organismos modelos poderiam auxiliar a anotação do genoma humano. Em estudos posteriores foi comprovado que aproximadamente 1000 genes de doenças humanas possuíam relação funcional com seus genes (71,73).

Alguns estudos experimentais chegam a avaliar 96,5% do seu genoma, onde é comprovado que 1.105 genes (18,7%) exercem algum tipo de função essencial para o organismo. Foi comprovado também que genes essenciais possuem maior probabilidade de terem ortólogos, onde 82% desses genes essenciais expressam proteínas em outros organismos, contra 67% dos genes não essenciais (58).

- *Arabidopsis thaliana*

É uma planta da família das *Brassicaceae* nativa dos continentes Europeu e Asiático. É um organismo geneticamente muito estudado, sendo em 2000, a primeira planta a ter seu genoma sequenciado. Com apenas 125 milhões de pares de base, o seu genoma pode ser considerado pequeno, se comparado ao de outros organismos (74,75).

Um conjunto de dados sobre essencialidade de genes foi reunido para o organismo *Arabidopsis*, permitindo realizar comparações com outros organismos modelo, facilitando a análise de genes de plantas com funções importantes, mas até então desconhecidas, contribuindo para a compreensão dos processos biológicos essenciais em plantas com flores. Trabalhos experimentais citam que 30% do genoma é composto de genes essenciais (66,76).

- *Caenorhabditis elegans*

É uma espécie de nematódeo com simetria bilateral da família *Rhabditidae.* Medindo cerca de 1 milímetro de comprimento e vivendo em ambientes temperados, ele possui muitos sistemas similares aos dos outros organismos. Tornou-se um importante modelo para o estudo da biologia desde a década de 70, por ser de fácil criação e com ciclo de vida curto (77).

Estudos revelam genes essenciais em diversas funções, desde crescimento até fertilidade em adultos. Estes trabalhos concluem que 15 a 30% (aproximadamente 5700) do genoma exercem algum tipo de função essencial.

Outros trabalhos apontam que este número pode chegar a 8500, devido à evolução das técnicas de identificação de genes essenciais, em que novos genes podem ser inseridos nesta lista ou retirados com o passar do tempo (78).

- *Drosophila melanogaster*

É mais especificamente um inseto díptero da família *Tephritidae* também conhecido como mosca da fruta, um animal invertebrado artrópode. A primeira publicação do seu genoma foi feita em 2000 e desde então seu estudo vem contribuindo para pesquisas em pesquisas genômicas. Criado de forma fácil e a baixo custo, permite por meio do rastreamento gênico identificar quais genes estão envolvidos em um processo biológico. Apesar disso, possui um genoma muito complexo, comparável aos dos mamíferos, e por isso vem auxiliando na compreensão de toda a complexidade genômica (79).

Apesar da *Drosophila melanogaster* possuir aproximadamente 15.000 genes, ou seja, um genoma menor que o *Caenorhabditis elegans*, ela possui mais que o dobro de homólogos em humanos. Neste cenário, 197 dos 287 genes de doenças humanas identificados possuem homólogos na *Drosophila melanogaster* (80).

- *Danio rerio (zebrafish)*

É um vertebrado aquático da família *Cyprinidae*, conhecido no Brasil como peixe paulistinha. Ele tem se tornado um organismo popular para o estudo de funções gênicas em vertebrados. Possui um banco de dados *on-line* próprio, onde são depositadas informações genéticas, embriológicas, genômica e do desenvolvimento, sendo a primeira espécie vertebrada clonada (81).

Ele possui 26.206 genes. Destes, 69% possuem ortólogos em humanos. Reciprocamente, 71% dos genes humanos possuem ortólogos no *Danio rerio*, Dentre o grupo destes ortólogos, 47% possuem uma relação direta de 1 para 1. Estes dados refletem porque os genes do *zebrafish* é muito estudado na compreensão das atividades biológicas dos genes ortólogos em humanos (82).

- *Mus musculus*

É uma espécie de pequeno roedor da família dos murídeos. Com cerca de 8 cm de comprimento, pelagem macia, branca ou cinza-acastanhada é um modelo

de vertebrado mamífero muito estudado, principalmente relacionado a doenças humanas que envolvem metabolismo, controle neurológico e hipertensão.

O seu genoma é aproximadamente 14% menor que o genoma humano. Análises mais criteriosas mostram que 40% do seu genoma pode ser alinhado em nível de nucleotídeos, o que pode acarretar em mais de 1 milhão de elementos conservados. Dentre os genes ortólogos, cerca de 80% possuem uma relação de 1 para 1 com um gene humano e mesmo os 20% restantes estão presentes em outros organismos (83–85).

1.8 Protozoários

Como mencionado no item 1.3, muitas doenças negligenciadas graves que ocorrem nas áreas mais pobres do planeta são causadas por protozoários. Por este motivo, eles são os organismos de interesse deste trabalho. Os protozoários são organismos eucariotos, unicelulares e fazem parte do reino *Protista*. Na natureza, podem ter vida livre ou como parasitas de outros animais. Eles se subdividem nas classes *Sarcodina*, *Flagellata*, *Ciliophora*, *Sporozoa e Mastigophora*, diferenciando-se por suas estruturas locomotoras (pseudópodes, cílios e flagelos). Abaixo seguem descrições mais detalhadas dos cinco protozoários abordados neste trabalho.

- *Entamoeba histolytica*

É o causador da doença denominada amebíase e pertence ao filo *Sarcodina*. A amebíase atinge, em sua maioria, humanos e outros primatas. Seu ciclo de vida não é complexo e a sua forma infecciosa é o cisto. As pessoas são contaminadas através do contato com o cisto, seja pela contaminação da água, alimentos ou das mãos. Quando o cisto entra no corpo humano, ele se aloja no intestino delgado, transformando em trofozoítos. Logo após migram para o intestino grosso, onde vivem e se multiplicam (86).

Os sintomas podem não se manifestar, mas os mais comuns são disenteria, vômitos e cólicas intestinais, podendo afetar qualquer pessoa, embora seja mais presente em pessoas que vivem em áreas tropicais, com pouco acesso a boas condições sanitárias. No entanto, estudos recentes de análise da água, já identificaram o protozoário em regiões do globo que até então não havia relatos (87).

Os investimentos no desenvolvimento de novos fármacos para o combate a esta doença é muito importante, pois de acordo com estimativas, 50 milhões de

casos em todo o mundo são detectados, com taxas de mortalidades significativas que variam de 50 mil a 110 mil mortes, sendo a terceira doença causada por protozoários que mais mata no mundo (88,89).

- *Leishmania major*

É uma espécie de protozoário flagelado da família *Trypanosomatidae*. A doença causada por este protozoário denomina-se leishmaniose, manifestando de diferentes formas: cutânea, muco-cutânea, cutânea difusa e visceral (forma mais grave, podendo levar os pacientes a óbito) (90).

O protozoário é transmitido pela picada do vetor flebotomíneos (inseto hematófago) fêmeas. A partir daí a forma infecciosa, conhecida como promastigota, se aloja. Com a infecção, em vertebrados, os promastigotas passam por um processo de fagocitose por macrófagos e se transformam em amastigotas. No estágio de amastigotas, os parasitos se multiplicam por divisão binária, rompendo a célula hospedeira. Os parasitos são liberados no meio intercelular ou na corrente sanguínea, infectando outras células (91,92).

A contaminação do vetor flebotomíneo fêmea se dá por meio da picada e da ingestão do sangue contaminado com o parasita no estágio de amastigotas. No organismo do flebotomíneo fêmea, eles evoluem para o estágio de promastigotas e multiplicam-se no intestino do inseto, migrando posteriormente para a região da probóscide (aparelho picador-sugador dos insetos) (91,92).

A ocorrência da leishmaniose está intimamente ligada às condições precárias de vida (moradia e nutrição) das pessoas. Aproximadamente entre 1,5 a 2 milhões de novos casos ocorrem anualmente e o número de mortes está por volta de 60 mil por ano (92). No Brasil, as espécies predominantes são as *L. chagasi*, *L. brasiliensis*, *L. guyanensis* e *L. amazonenies.* Trabalhos mais recentes identificaram nos estados das regiões Norte e Nordeste as espécies *L. lainsoni*, *L. naiffi*, *L. lindenberg* e *L. shawi* (93,94).

Não foram encontrados ainda ocorrências da espécie *Leishmania major* no Brasil. As regiões mais atingidas estão no norte da África, Oriente Médio, China e Índia, mas seu genoma é o melhor estudado e anotado. O diagnostico da doença se dá pela identificação das formas amastigotas, por meio de isolamento e o tratamento é feito prioritariamente por derivativos do pentavalentes (95).

- *Plasmodium falciparum*

É um protozoário pertencente ao filo *Sporozoa* ou *Apicomplexa*. É o causador da manifestação mais agressiva da malária, uma das principais doenças parasitárias do mundo. A malária está presente em grandes regiões tropicais e subtropicais dos países subdesenvolvidos. Cerca de 40% da população mundial está exposta a esta doença. Dados mostram que em 2010 aproximadamente entre 6 a 7 milhões de pessoas foram infectadas, onde 660 mil vieram a óbito. Destes 660 mil óbitos, 78% são crianças de até 5 anos de idade. (96).

As diferentes espécies de *Plasmodium* possuem muitas características em comum em seu ciclo de vida, onde um inseto vetor e um hospedeiro humano estão envolvidos. No ser humano o ciclo inicia pela picada do mosquito fêmea pertencente ao gênero *Anopheles*. Deste modo os esporozoítos são liberados a partir da glândula salivar e entram na corrente sanguínea, atingindo rapidamente as células do fígado, onde se multiplicam e evoluem para diferentes estágios até atingirem a forma de merozoítos. Após este estágio, as células do fígado se rompem e parte dos merozoítos, atingindo as células hemácias, onde os novos estágios para trofozoíta, esquizonte e novamente para merozoítos acontecem.

Este ciclo se repete de 36 a 72 horas, dependendo da espécie, e após algum tempo, o estágio da evolução se estabiliza na forma de gametócitos. É neste estágio de infecção, por meio da picada, que vetor ingere e transmite a outra pessoa, dando continuidade ao ciclo (97,98).

Após a picada do mosquito, durante 10 a 15 dias, a pessoa infectada apresenta os sintomas de febre, dor de cabeça e vômitos. O diagnostico é feito por meio da identificação dos sintomas ou por meio de análises laboratoriais. Os fármacos mais comuns no combate à doença envolvem o uso de quinina, cloroquina, mefloquina, primaquina, doxiciclina e malarone, mas principalmente na espécie *P. falciparum*, algum tipo de resistência aos medicamentos é manifestada (94,96,98).

- *Trypanosoma brucei*

É uma espécie de protozoário flagelado também da família *Trypanosomatidae*. Ele é o causador da doença conhecida como doença do sono (tripanossomíase africana) (99). Ela possui duas subespécies, a *Trypanosoma brucei gambiense* e a *Trypanosoma brucei rhodesiense*, onde a primeira é responsável por 98% dos casos da doença (96).

O vetor é a mosca *tsé-tsé* que, quando infectada com os tripomastigotas metacíclicos, os transmite através da picada, entrando na corrente sanguínea e atingindo diversos fluidos, como o fluido linfático e o fluido espinhal. Na corrente sanguínea elas passam para o estágio de tripomastigotas e se multiplicam por fissão binária. A doença é transmitida quando outra mosca ingere o sangue infectado (98).

No vetor, o ciclo de vida tem duração de três semanas. Os tripomastigotas sanguíneos ingeridos de um vertebrado contaminado se transformam em tripomastigotas procíclicos no intestino da mosca e se multiplicam. Eles se transformam em epimastigotas e migram para as glândulas salivares. Após isso, evoluem para tripomastigotas metacíclicos e multiplicam-se uma vez por fissão binária (98).

A doença atinge somente países da África, principalmente na região Subsaariana, onde a mosca tsé-tsé está presente. As pessoas mais expostas à doença pertencem a comunidades rurais, dependentes economicamente da agricultura, pesca, caça e criação de animais. Em sua maioria, estas populações vivem em lugares de difícil acesso com recursos de saúde limitados, dificultando o combate e a vigilância dos casos. O número estimado de casos da doença é de 30 mil, mas estimativas chegam a mencionar aproximadamente 70 milhões de pessoas vivendo em regiões com alto risco de infecção, sendo 5 milhões com altíssimo risco (96,99).

Os principais sintomas são dores de cabeça, febre, fraqueza, dor nas articulações e rigidez, sendo o terceiro estágio o mais perigoso, podendo levar a óbito. O tratamento é baseado nos sintomas e nos resultados laboratoriais. Quatro fármacos (pentaminida, suramina, melarsoprol e eflornitina) são utilizados para o combate à doença (100,101).

- *Trypanosoma cruzi*

Juntamente com *L. major* e T*. brucei,* compõe parte da família *Trypanosomatidae*. Ele é o agente causador da doença de Chagas (tripanossomíase americana). A doença passa por duas fases, uma aguda, podendo ou não ser identificada e uma fase crônica (102).

O vetor do *T. cruzi* é o inseto barbeiro (triatomíneo). O ciclo se inicia quando o barbeiro se alimenta do sangue de um hospedeiro e elimina em suas fezes ou urina o parasito no estágio de tripomastigota metacíclica. Com as fezes ou urina contaminada entrando em contato com mucosas ou ferimentos na pele, os parasitas

entram no organismo do hospedeiro infectando suas células. Dentro do organismo, o parasita evolui para o estágio de amastigota e multiplica-se por divisão binária, até a célula encher e um novo estágio ser atingido, chamado de tripomastigota. A célula se rompe, lançando na corrente sanguínea os parasitas e infectando novos tecidos e órgãos (103).

Os barbeiros se contaminam ingerindo sangue contaminado. Os parasitas chegando ao intestino dos barbeiros assumem a forma epimastigota e migram para a parte posterior do sistema excretor, assumindo a forma tripomastigota metacíclica, completando então o ciclo (104).

A América Latina é a região com maior incidência da doença e os casos ocorrem principalmente no Brasil. No entanto, devido a emigração de pessoas, diversos casos já foram registrados em outros países. A proliferação da doença está vinculada as más condições de moradia, pois favorecem a multiplicação do barbeiro. Estimativas apontam que aproximadamente 8 milhões de pessoas estão expostas a infecção no mundo (94,96).

Na primeira fase da doença, a fase aguda, os sintomas são febre, dores de cabeça e musculares, aumento dos gânglios linfáticos, dificuldade em respirar e palidez. Na segunda fase, a fase crônica, os parasitas estão em circulação pela corrente sanguínea. Nesse estágio, até 30% dos infectados desenvolvem alterações cardíacas e 10% alterações digestivas e neurológicas (96).

O diagnóstico da doença é feito por teste imunoenzimático e testes sorológicos. O melhor modo de combate à doença é realizar o controle do barbeiro. Os principais fármacos de tratamento é o benzonidazol e o nifurtimox que podem ser muito eficientes na cura da doença se diagnosticada no início (96,105).

1.9 Justificativa

Por meio do aprimoramento das técnicas de estudos experimentais genômicos, cada vez mais dados sobre organismos vêm sendo disponibilizados. Estes dados contribuem para descobertas de novos alvos de fármacos, que são essenciais no desenvolvimento de fármacos para o combate à doenças.

Estes dados ficam armazenados e disponíveis em bases de dados específicas de cada domínio de conhecimento. Outra forma de disponibilizar dados são os artigos científicos publicados em jornais e revistas especializadas, disponíveis em repositórios ou bases textuais destinados ao seu armazenamento. O

grande desafio é a grande quantidade de documentos armazenados, tornando impossível para a capacidade humana processá-los e classificá-los.

Por estes motivos, desenvolver ferramentas computacionais capazes de integrar e correlacionar diferentes fontes de dados e domínios de conhecimento é essencial para o sucesso das pesquisas. Assim é possível correlacionar informações de diferentes áreas de conhecimento, aumentando a capacidade dos pesquisadores de tomarem boas decisões em menor espaço de tempo.

Este trabalho propõe uma metodologia de construção de *Data marts*, apresentando uma abordagem de integração de dados orientada por assunto, onde os dados de interesse do usuário são carregados em um modelo dimensional e agrupados de forma a facilitar a análise dos mesmos, concedendo um acesso de alto desempenho.

As técnicas de construção de *data marts* a partir de dados estruturados já são bem difundidas, mas técnicas de construção de *data marts* para dados não estruturados e de fontes heterogêneas, como por exemplo, dados extraídos de textos utilizando ontologias, ainda não foram muito exploradas. Alguns trabalhos científicos vem abordando esta nova temática, mas muito ainda precisa ser discutido para que novas técnicas possam surgir.

2 OBJETIVOS

2.1 Objetivo Geral

Apresentar uma metodologia de projeto de *data marts*, denominada TOETL (*Text and Ontology* ETL), para análise de dados não estruturados (textuais) por meio do uso de ontologias, fornecendo uma ferramenta para processar um grande conjunto de artigos científicos para apoiar o pesquisador em uma melhor tomada de decisão, no direcionamento de sua pesquisa. Através da TOETL, é também um objetivo deste trabalho, priorizar o estudo de possíveis proteínas na busca por novos alvos de fármacos.

2.2 Objetivos Específicos

- Disponibilizar um *data mart* de essencialidade (TaP DM) para 5 protozoários: *Entamoeba histolytica*, *Leishmania major*, *Plasmodium falciparum*, *Trypanosoma brucei* e *Trypanosoma cruzi*;
- Disponibilizar uma ferramenta OLAP para que os pesquisadores acessem o DM;
- Identificar proteínas como possíveis alvos para fármacos;
- Definir as etapas para projeto do DM como uma estratégia genérica, que possa ser seguida para atender outros domínios de descoberta científica;

3 MATERIAIS E MÉTODOS

Neste capítulo é apresentada uma contextualização da metodologia proposta (item 3.1), comparando as suas principais diferenças e as vantagens de se utilizar uma abordagem que combina visão dimensional com analise semântica. No item 3.2 é apresentada a metodologia TOETL (*Text and Ontology ETL*), detalhando suas etapas e passos. O item 3.2 serve de embasamento para o item 3.3, onde a metodologia é aplicada gerando como resultado o TaP DM (*Target Prioritization Data Mart*) e uma metodologia de priorização de proteínas alvo, projetada sobre as funcionalidades e flexibilidade que a TOETL provê. Por fim, no item 3.4, uma interface OLAP é integrada ao TaP DM para tornar a interface mais amigável ao usuário, facilitando a realização de consultas.

3.1 Contextualização

Data marts são ferramentas muito exploradas no meio corporativo como forma de obterem vantagens competitivas, pois por meio das visões analíticas, possibilitam os gestores descobrirem tendências de comportamento. As metodologias de construção de DM são muito difundidas e aceitas pela comunidade científica. Como mencionado na Introdução, a utilização destes sistemas de visão analítica juntamente com anotação textual, permite correlacionar informações sobre diferentes conceitos, possibilitando encontrar ligações ocultas entre elementos de interesse.

Por estes motivos, desenvolver ferramentas capazes de correlacionar conceitos e obter informações mais refinadas, com o objetivo de auxiliar os pesquisadores na tomada de decisão de suas pesquisas, cruzando dados de maneira rápida e confiável é muito importante. Estes quesitos são atingidos combinando as abordagens de anotação semântica com *data mart*, permitindo explorar os diversos cenários do problema de diferentes perspectivas, com toda a riqueza de informações contidas nas ontologias, ao longo do tempo.

O desenvolvimento deste trabalho se baseia nas abordagens de *data marts* e análise semântica utilizando ontologias. *Data mart* foi escolhido porque são bancos orientados a assunto com o objetivo de auxiliar na tomada de decisão. Diferentemente dos bancos de dados relacionais convencionais que servem apenas

para armazenar dados operacionais. O uso de ontologias foi motivado porque elas descrevem explicitamente os conceitos e os seus relacionamentos. Características que não estão presentes, por exemplo, na classificação facetada, onde não é possível expressar relações entre facetas (106,107). Outro critério de escolha de ontologias é a disponibilização em diversos formatos, por parte de grupos, de ontologias ricas em informação da área biomédica com alto grau de maturidade, facilitando muito a implementação de ferramentas e a confiabilidade de informações.

Este trabalho objetiva desenvolver uma metodologia de projeto de *data marts* para análise de dados não estruturados denominada TOETL, onde elementos de interesse do pesquisador podem ser quantificados e correlacionados com o uso de ontologias. Uma maneira de obter estas informações é medir a ocorrência dos termos pertencentes aos elementos de interesse em um conjunto de artigos. Os aspectos para esta medição também devem ser definidos, mas isso depende muito da amostra de textos que se tem em mãos.

Também é proposto nesse trabalho uma estratégia de priorização de novos alvos, onde é utilizado o *data mart* resultante da aplicação da metodologia TOETL para refinar os dados armazenados nos artigos. Essa estratégia é flexível e permite que o usuário estabelece seus próprios parâmetros de busca.

Diferentemente dos métodos tradicionais de projeto de DM, neste trabalho o cenário não envolve uma corporação e seus processos, mas um foco de pesquisa, onde se busca determinar o teor do que se publica em torno de um foco de pesquisa, onde é orientada tanto pela demanda dos usuários quanto pelos dados, pois as questões que devem ser respondidas quanto as ontologias interferem diretamente no projeto do *data mart* (Figura 3.1).

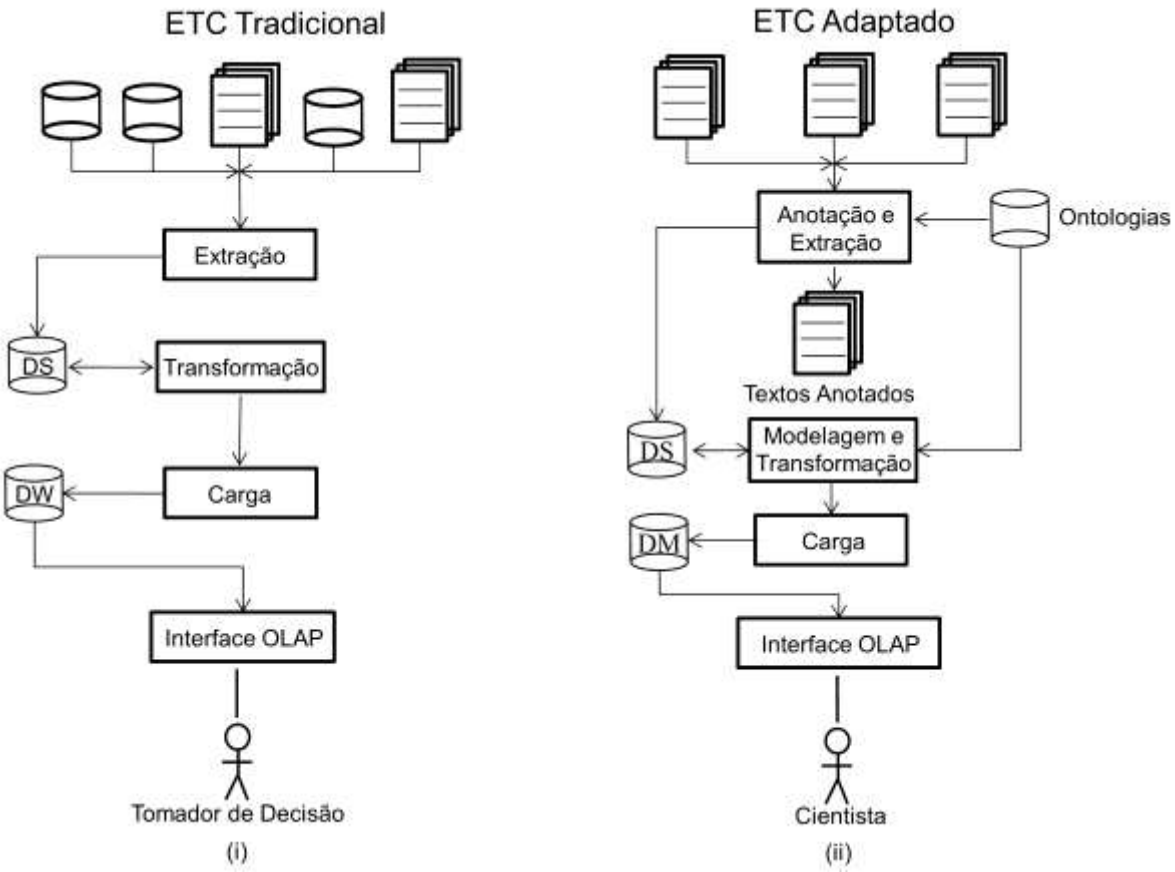


Figura 3.1 : ETC Tradicional *versus* ETC Adaptado (TOETL).

3.2 Metodologia TOETL

A metodologia TOETL é dividida em 3 etapas principais (Anotação e Extração, Modelagem e Transformação e por fim Carga), sendo cada uma delas subdivididas em passos, como mostra a Figura 3.2.

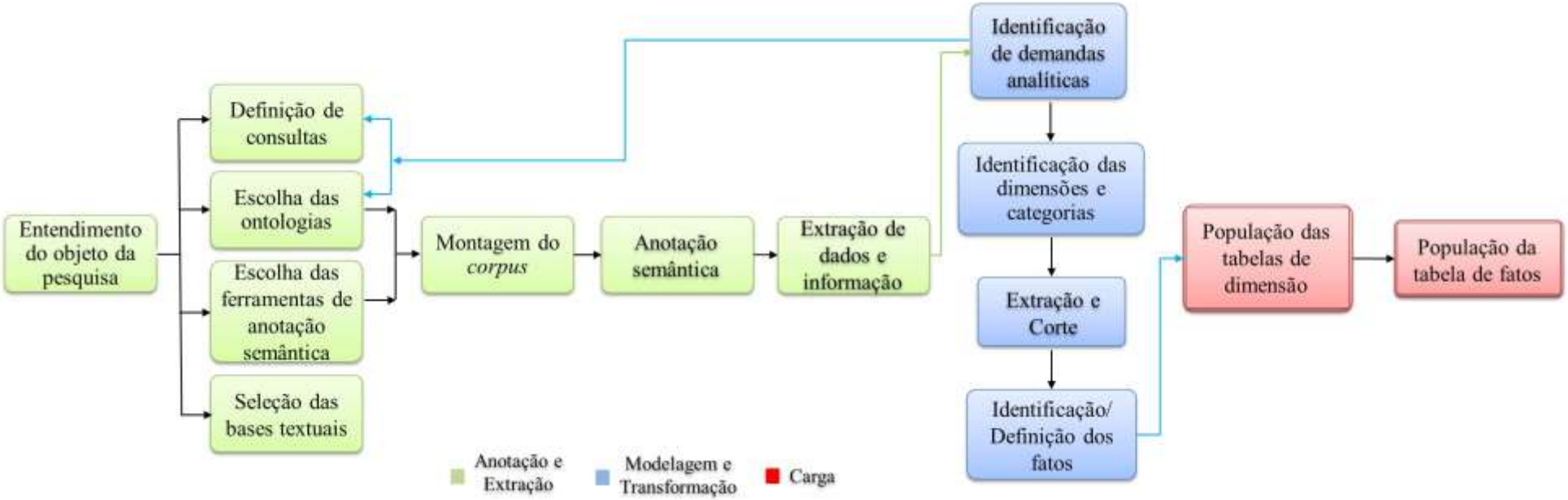


Figura 3.2: Visão geral da metodologia proposta (TOETL).

A etapa de Anotação e Extração envolve a compreensão dos aspectos que compõem os objetivos da pesquisa dos cientistas, construção das consultas que serão submetidas à biblioteca digital resultando no *corpus* de artigos, Escolha da ferramentas de anotação semântica, Escolha das ontologias e finalmente, anotação dos artigos, sendo seus dados extraídos e armazenados em um banco de dados. Esta etapa foi descrita no trabalho (2) e precisou ser reproduzida com o objetivo de atualizar e aumentar o *corpus* de artigos. A TOETL contribui acrescentando as etapas de Modelagem e Transformação (definição das dimensões e fatos) e Carga (população do *data mart*) que são discutidas e apresentadas detalhadamente nas seções 3.2.2 e 3.2.3

3.2.1 Anotação e Extração

Como mencionado na introdução, a etapa de Anotação e Extração foi apresentada e discutida no trabalho (2). Neste item é apresentado uma visão geral desta etapa (Figura 3.3) para melhor compreensão deste estudo.

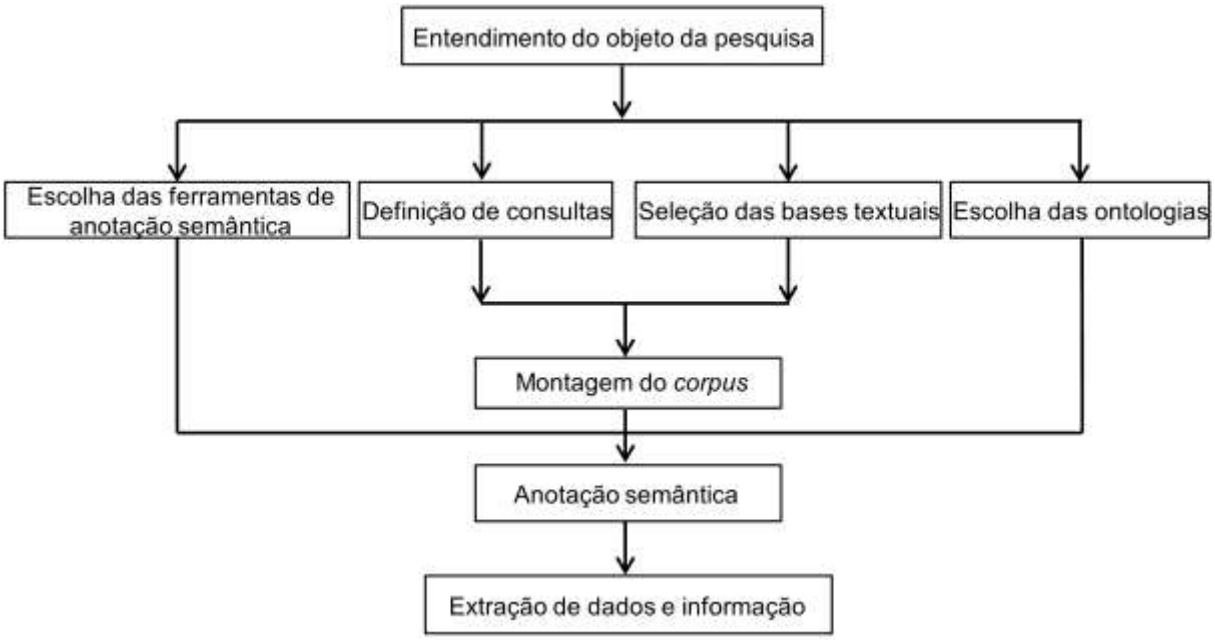


Figura 3.3: Passos da etapa de Anotação e Extração. Adaptada de (2).

O estágio *entendimento do objeto da pesquisa* envolve o estudo da literatura clássica da área de atuação do DM, bem como entrevistas com os usuários (pesquisadores). Um conjunto de termos, expressões e seus sinônimos são levantados a fim de identificar na literatura uma quantidade significativa de textos científicos (*corpus*). Este conjunto deve abranger todos os domínios do tema da investigação.

Uma vez definidos os termos, eles são utilizados para prosseguir para outros dois passos: a *definição das consultas* e a *seleção das bases textuais*. O passo de *definição das consultas* usa os termos e expressões para compor consultas de pesquisa de palavras-chave em uma biblioteca digital. As consultas genéricas e específicas devem ser construídas, a fim de não perder itens importantes da literatura na composição do *corpus*.

Portanto, a ideia é organizar termos de acordo com domínios específicos e construir expressões lógicas com o operador *AND*. Primeiro em pares de termos, cada um de um domínio diferente, isso trará um grande número de respostas, mas garante uma boa abrangência. Em seguida, consultas mais específicas são formadas, combinando mais termos usando o operador *AND*. Além disso, para cada operando nessas expressões, um parêntese com alternativas ou sinônimos deve ser formado, usando o operador *OR*. Por exemplo, para um tema de pesquisa que envolve 3 domínios, o conjunto de termos $S = \{t_{1,1},t_{1,2},t_{2,1},t_{2,2},t_{2,3},t_{2,4},t_{3,1},t_{3,2}\}$, as seguintes consultas podem ser formadas:

- $t_{1,1}$ *AND* ($t_{2,1}$ *OR* $t_{2,2}$)
- $t_{1,2}$ *AND* ($t_{2,1}$ *OR* $t_{2,2}$)
- $t_{1,1}$ *AND* ($t_{2,3}$ *OR* $t_{2,4}$)
- $t_{1,2}$ *AND* ($t_{2,3}$ *OR* $t_{2,4}$)
- $t_{1,1}$ *AND* ($t_{2,1}$ *OR* $t_{2,2}$) *AND* ($t_{3,1}$ *OR* $t_{3,2}$) ...

No passo de *seleção das bases textuais*, a ideia é escolher os repositórios digitais (um ou mais) que possuam trabalhos que envolvam os domínios de interesse. É importante que a maioria dos textos esteja totalmente disponível (textos completos, não apenas os resumos), pois alguns artigos estão disponíveis somente sob pagamento.

Uma vez que as expressões de consultas estejam formadas e as bibliotecas selecionadas, o *corpus* pode ser construído. Neste passo, a *construção do corpus*, as consultas são submetidas à biblioteca digital e os textos salvos. Depois, todos eles são passam por um processo de limpeza, onde figuras, referências e *stop words* são retirados.

Paralelamente a estes passos iniciais, outros dois podem ser realizados: *escolha das ontologias* e *escolha das ferramentas de anotação semântica*. Conforme mencionado anteriormente, o uso de ontologias é recomendado para manter anotações baseadas em um vocabulário uniforme. No entanto, não é uma

tarefa fácil escolhê-las. Em primeiro lugar, mais de uma ontologia pode ser necessária, uma vez que o tema da pesquisa geralmente envolve vários domínios. Os repositórios de ontologias como o OBO *Foundry* e o NCBO BioPortal citados anteriormente, podem ser usados para facilitar a pesquisa.

Nesses repositórios, podem-se encontrar muitas ontologias que cobrem o mesmo domínio, mas sua cobertura pode ser diferente, ou seja, algumas podem ter uma cobertura profunda e estreita, enquanto outras podem ter uma cobertura grande e superficial. Algumas anotações de amostragem anteriores podem ser usadas para facilitar a escolha. Isso é feito tomando um conjunto pequeno e diversificado de textos do *corpus*, e submetê-los a anotações iniciais com um primeiro conjunto de ontologias selecionadas. As ontologias redundantes devem ser eliminadas, assim como aquelas que apresentam anotações irrelevantes ou apenas um pequeno número de ontologias relevantes.

Devido ao grande número de textos e termos, utilizar anotações manuais se torna inviável, por isso, é necessário escolher uma ferramenta de anotação automática. Além disso, dois outros recursos são necessários para a ferramenta de anotação: (i) deve usar qualquer ontologia arbitrária para anotação, ou seja, o usuário deve ser livre para escolher a ontologia de sua preferência; (ii) o texto anotado resultante deve estar pronto para processamento posterior. Algumas das ferramentas de anotação disponíveis não abordam ambos os requisitos, como relatado em (108,109).

Uma funcionalidade desejada, mas não estritamente necessária para ferramentas de anotação, é anotar usando inferência hierárquica. Algumas ferramentas de anotação fornecem essa funcionalidade (109). Depois que um termo é anotado com um conceito de ontologia, uma anotação baseada em inferência ocorre quando esse conceito pertence a um ramo de conceitos. Neste caso, pode também ser anotado com todos ou alguns dos seus pais. Esta característica garante que um texto científico poderia ser associado não apenas a conceitos específicos, mas também às suas generalizações.

Uma vez definidos ontologias, *corpus* e ferramenta de anotação, a próximo passo é a *anotação semântica*. Ele pode ser muito demorado, podendo levar dias, pois depende do tamanho do *corpus* e das ontologias. Alguns trabalhos na literatura já propõem soluções que podem reduzir consideravelmente este tempo de processamento (109,110), utilizando modularização de ontologias.

Finalmente, após a anotação, a *extração dos dados e informação* é o passo que prepara o banco de dados da área de armazenamento (*Data Store* - DS) para a fase de construção do *data mart*. Basicamente ele consiste em um processamento automático dos textos anotados, onde cada anotação em um texto irá gerar uma tupla de banco de dados com (i) a expressão de texto anotada, (ii) a classe de ontologia (id e rótulo) usada para anotá-la e (iii) o id de texto (identificação do artigo).

No processo de anotação as ontologias são usadas separadamente. Consequentemente, um conjunto de artigos anotados com a mesma quantidade de arquivos do *corpus* é gerado para cada ontologia. Todos os termos anotados são extraídos e armazenados na área de armazenamento.

3.2.2 Modelagem e Transformação

É importante entender que, no contexto de uma pesquisa científica, embora a fonte textual permaneça a mesma, o interesse e o conteúdo de um cientista durante uma pesquisa é dinâmico, ou seja, o foco ou sua pesquisa pode mudar com o decorrer do tempo. Os aspectos úteis para a análise científica são definidos somente após a extração dos dados de anotação.

Portanto, diferentemente do estágio de Transformação presente nos métodos tradicionais de construção de *data marts*, onde as dimensões são pré-definidas e projetadas, no cenário de uma pesquisa científica, algumas dimensões só serão definidas nesta etapa. Em outras palavras, o projeto do *data mart* ocorre também durante o etapa de Modelagem e Transformação do processo de ETL, isto significa que os processos de ETL e de projeto do DM ocorrem em paralelo.

Tipicamente nesta etapa de Modelagem e Transformação, nos cenários das pesquisas científicas quando artigos são utilizados como fonte de dados, as dimensões tempo e artigo estarão sempre presentes. O que altera de um cenário para outro são as dimensões ontológicas, pois estas seguem o escopo do problema, os interesses dos pesquisadores e as informações contidas nas ontologias. Mesmo dependentes destas situações, algumas informações dificilmente não estarão presentes nas dimensões ontológicas, como mostra a Figura 3.4:

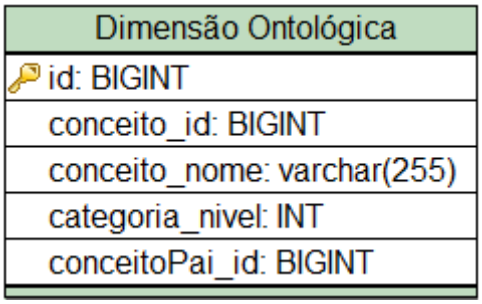


Figura 3.4: Modelagem geral das dimensões baseadas em ontologias.

Nesta etapa, os passos de *identificação das demandas analíticas*, *identificação das dimensões e categorias* e *identificação/definição do fato* estão diretamente ligados a modelagem do *data mart*, já o passo *extração e corte* está ligado ao processo de transformação, pois os dados contidos nas ontologias também servem para alimentar o *data mart*. Ao final desta etapa, o banco a modelagem é finalizada, como mostra de forma genérica a Figura 3.5.

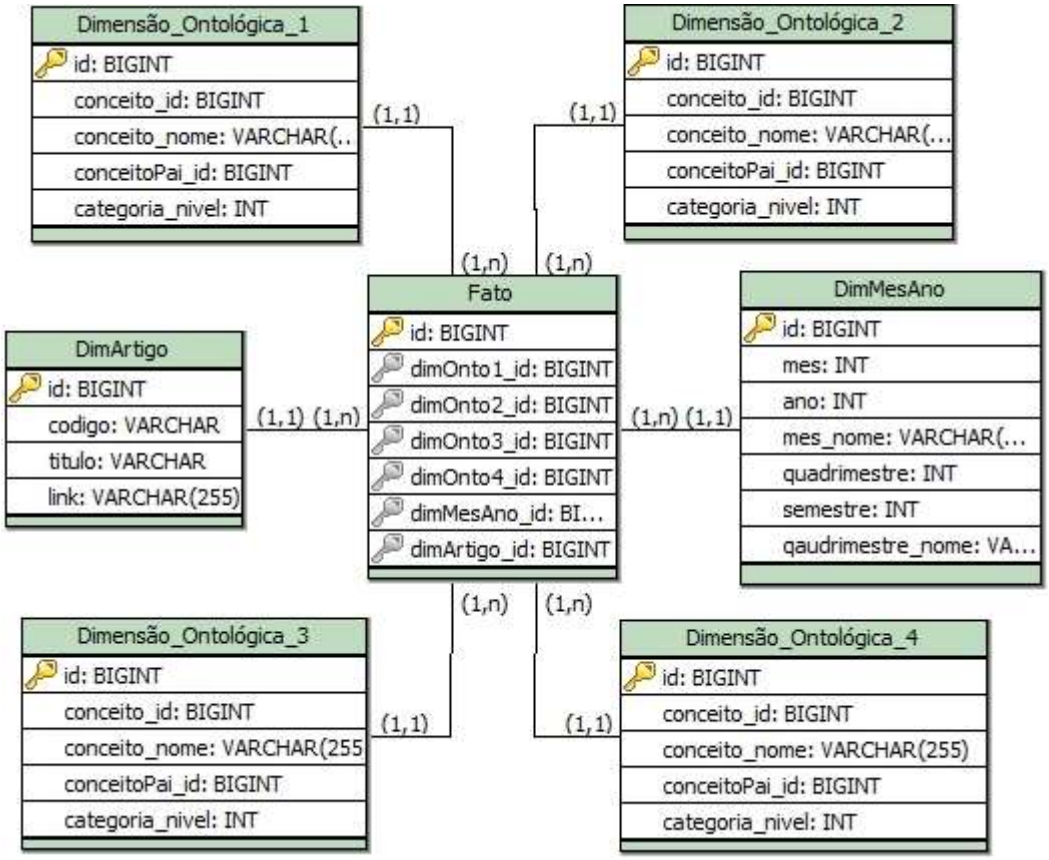


Figura 3.5: Modelagem genérica do *data mart*.

Caso ocorram novas extrações e o foco da pesquisa permaneça o mesmo, o banco será atualizado. Mas se houver mudança no foco da pesquisa ou novos elementos chamarem a atenção do pesquisador, um novo banco deve ser projetado, impactando mudanças diretas nas dimensões ontológicas. A etapa de Modelagem e Transformação consiste em quatro passos sequenciais, descritos a seguir.

3.2.2.1 Identificação de demandas analíticas

Este passo tem o objetivo de identificar as questões analíticas que devem ser abordadas. Neste ponto, entrevistas com os usuários do DM devem ser realizadas. O seu propósito é identificar as informações relevantes a serem obtidas por meio do processo de anotação. Geralmente, para responder a estas questões, é necessário correlacionar conceitos de diferentes domínios de conhecimento, por exemplo: "*Quais os efeitos colaterais mais citados com o tratamento da doença de Chagas*?". Claramente neste caso, a resposta depende da correlação de três aspectos (conceitos) distintos: artigos, efeitos colaterais e doenças.

Depois de levantar tais questões, cada uma é analisada e os termos citados são classificados. As ontologias podem auxiliar neste processo, realizando a anotação dos termos, as ontologias agrupam os termos em domínios de conhecimento. Por exemplo, na questão "*Quantos artigos citam doença de chagas, febre e dor corporal*?", o termo doença de chagas pode ser classificado como uma doença, e os termos febre e dor corporal podem ser classificados como sintomas ou efeitos colaterais, isto vai depender do foco da pesquisa.

3.2.2.2 Identificação das dimensões e categorias

O seu objetivo principal é usar conceitos genéricos para representar um conjunto de conceitos específicos identificados no passo anterior. Por exemplo, a doença de Chagas pode ser incluída num conjunto de doenças chamado doenças infecciosas. Já os conceitos como *Archaea* e *Eukaryota* podem ser representados por um único conceito genérico denominado organismos celulares. Portanto, é possível definir um conjunto reduzido de dimensões utilizando a estrutura hierárquica das ontologias.

Normalmente, as ontologias das áreas biomédicas são muito grandes, incluindo conceitos de muitos domínios e subdomínios de conhecimento. Por exemplo, a ontologia *National Cancer Institute* (NCI) *Thesaurus* (111) descreve conceitos de organismos, fármacos, produtos químicos, genes, atividades, processos biológicos, etc. O projetista deve identificar e selecionar as hierarquias de interesse presentes nas ontologias. Utilizando as perguntas analíticas como referência, os conceitos identificados que precisam ser combinados devem ser mantidos em dimensões distintas. Por exemplo, se a questão analítica precisa identificar combinações de organismos e produtos químicos, essas hierarquias

devem corresponder a dimensões distintas, mesmo se estiverem presentes na mesma ontologia.

De acordo com o projeto tradicional de DM, uma vez que as dimensões são definidas, é hora de caracterizá-las (28). Isto significa que é necessário identificar atributos que poderiam descrever cada instância de dimensão, incluindo categorias de dimensão para atender às demandas de agregação. As ontologias também ajudam neste passo, já que as hierarquias de seus conceitos já incorporam uma categorização, permitindo que as dimensões baseadas em ontologias possuam caraterizações semelhantes.

Os atributos básicos que uma tabela de dimensão (Figura 3.4) deve possuir além de uma chave substituta, usada como a chave primária (atributo id), são quatro atributos: conceito_id (identifica o conceito da ontologia, normalmente utiliza o identificador sugerido pela ontologia), conceito_nome (descrição do conceito), categoria_nivel (descreve o nível hierárquico do conceito, por exemplo, espécie, gênero, família, etc.) e o conceitoPai_id (representa uma relação recursiva, usada para permitir que cada tupla aponte para a tupla que representa um conceito mais genérico). É por meio desta informação hierárquica que o DM é capaz de associar a ocorrência de termos mais genéricos, mesmo quando estes não estão explicitamente presentes no artigo.

3.2.2.3 Extração e Corte

Uma vez definidas as dimensões de interesse presentes na ontologia, o próximo passo envolve realizar a eliminação dos conceitos que não fazem parte do escopo do *data mart*, a fim de preparar a população das tabelas de dimensão na etapa de Carga. Isso pode ser feito manualmente ou através do uso de ferramentas de segmentação de ontologia como o *Locality Module Extractor* (112) e o SEGMENTATION (113).

Como as ontologias podem ser muito abrangentes, onde diversos conceitos são descritos, este passo tem o objetivo de retirar da ontologia os conceitos que não fazem parte do interesse da pesquisa. Os cortes devem estar alinhados com as perguntas analíticas, por exemplo, se consideramos a pergunta: "*Quais processos patológicos são mais citados com algum químico*?", neste caso, a correlação entre processos biológicos e químicos é necessária para responder a esta pergunta. Se a ontologia que descreve um dos conceitos identificados também descrever outros conceitos não abordados na consulta analítica, então eles devem ser removidos,

como mostra a Figura 3.6, onde a ontologia NCI *Thesaurus* descreve o conceito de processos patológicos e diversos outros.

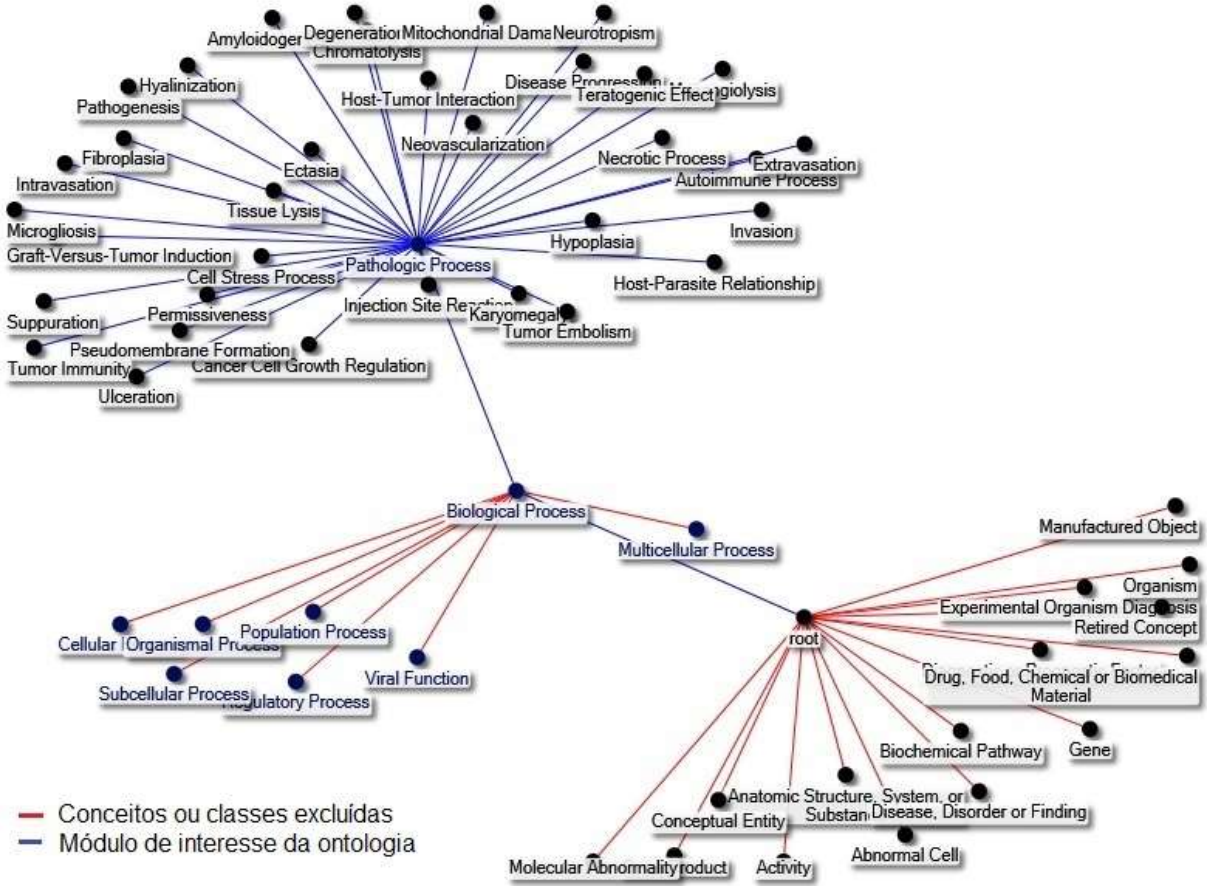


Figura 3.6: Árvore da ontologia NCBI *thesaurus*.

Uma importante dimensão de qualquer DM é a dimensão tempo, pois é ela que permite realizar análises temporais dos dados. No caso de textos científicos onde a informação temporal é a data de publicação (dia/mês/ano), o grão de tempo que faz sentido realizar análises é meses, e a partir das operações de agregação, informações sobre semestres, anos, décadas, etc., poderiam ser extraídas. Então, os atributos que caracterizam informações sobre mês/ano seriam suficientes para a análise temporal dos dados e podem ser populados previamente, seguindo as necessidades do intervalo de tempo dos projetistas.

3.2.2.4 Identificação / Definição do Fato

As demandas analíticas geralmente envolvem encontrar o número de ocorrências de termos em um *corpus* de artigos. Exemplos de tais questões são: (i) "*Quantos artigos mencionaram um determinado termo*?", (ii) "*Com que frequência foi citado um termo específico*?", (iii) "*Quantas vezes dois (ou mais) termos são*

mencionados conjuntamente no mesmo artigo?", (iv) "*Quantas ocorrências de um termo (ou dois ou mais termos) estão dentro de cada artigo*?". Todas essas consultas visam observar o número de ocorrências de termos, seja em todo o corpus, seja dentro de um artigo.

No primeiro caso (i), uma presença única/múltipla de um termo em um artigo é registrada. Portanto, o fato a ser observado é a contagem para cada combinação de valores de *d1,. . . , dn* e mês/ano, onde *di* corresponde a cada hierarquia de ontologia selecionada (dimensão), definida no passo anterior.

Para o segundo caso, para questões como (iv), é importante contar o número de ocorrências dentro do artigo. Isto seria útil para calcular a relevância do termo em relação ao todo corpus, como a métrica TF-IDF (114). Neste caso, o próprio artigo emerge como uma dimensão. Então, o fato observaria a contagem para cada combinação de valores de *d1,. . . , dn*, mês/ano e artigo onde *di* corresponde a cada hierarquia de ontologia selecionada (dimensão), definida no passo de *identificação das dimensões e categorias*.

Uma terceira alternativa para a identificação do fato é ter uma situação sem fatos (*factless*). Suponha que é necessário representar simplesmente a ocorrência ou não de um termo, o que significa que esta é a única informação que importa. Neste caso, nenhuma contagem do número de presenças é necessária, mas ainda assim, o artigo deve ser mantido como uma dimensão.

A dimensão do artigo pode ser implementada com atributos que caracterizam a publicação, como seu título, seu código de identificação atribuído em seu repositório original, o *link* para seu arquivo pdf, etc. Não é necessário manter dados sobre a data de publicação dos artigos nesta dimensão, uma vez que já está representada pela tabela de dimensão Mês/Ano.

3.2.3 Carga

Esta etapa consiste em dois passos principais: a população das tabelas de dimensão e a população da tabela de fatos. A população de cada tabela de dimensão pode ser automatizada usando como entrada as ontologias correspondentes (ou módulo das ontologias) em um formato estruturado, como os formatos XML RDF/OWL.

A população da dimensão Artigo é baseada no conjunto dos textos anotados. Seus ids, títulos e links são carregados na tabela, estas informações facilitam o acesso ao conteúdo do artigo, que é solicitado com bastante frequência.

Finalmente, a população da tabela de Fatos é feita da seguinte forma. Para cada artigo armazenado na tabela de dimensão Artigo, uma lista é criada com todos os termos distintos anotados. Esta lista contém termos de todas as dimensões presentes no artigo. Em seguida, esta lista é separada em sublistas de termos para cada dimensão. Cada tupla da tabela de Fatos trás a combinação de todos os elementos presentes nestas listas. Por exemplo, considerando três dimensões e se um artigo tiver dois termos de cada dimensão (ou seja, as listas de cada dimensão têm dois elementos), 8 tuplas serão necessárias para representar os fatos. A Figura 3.7 apresenta de maneira geral como é a identificação dos termos das dimensões em um artigo e como os fatos são gerados.

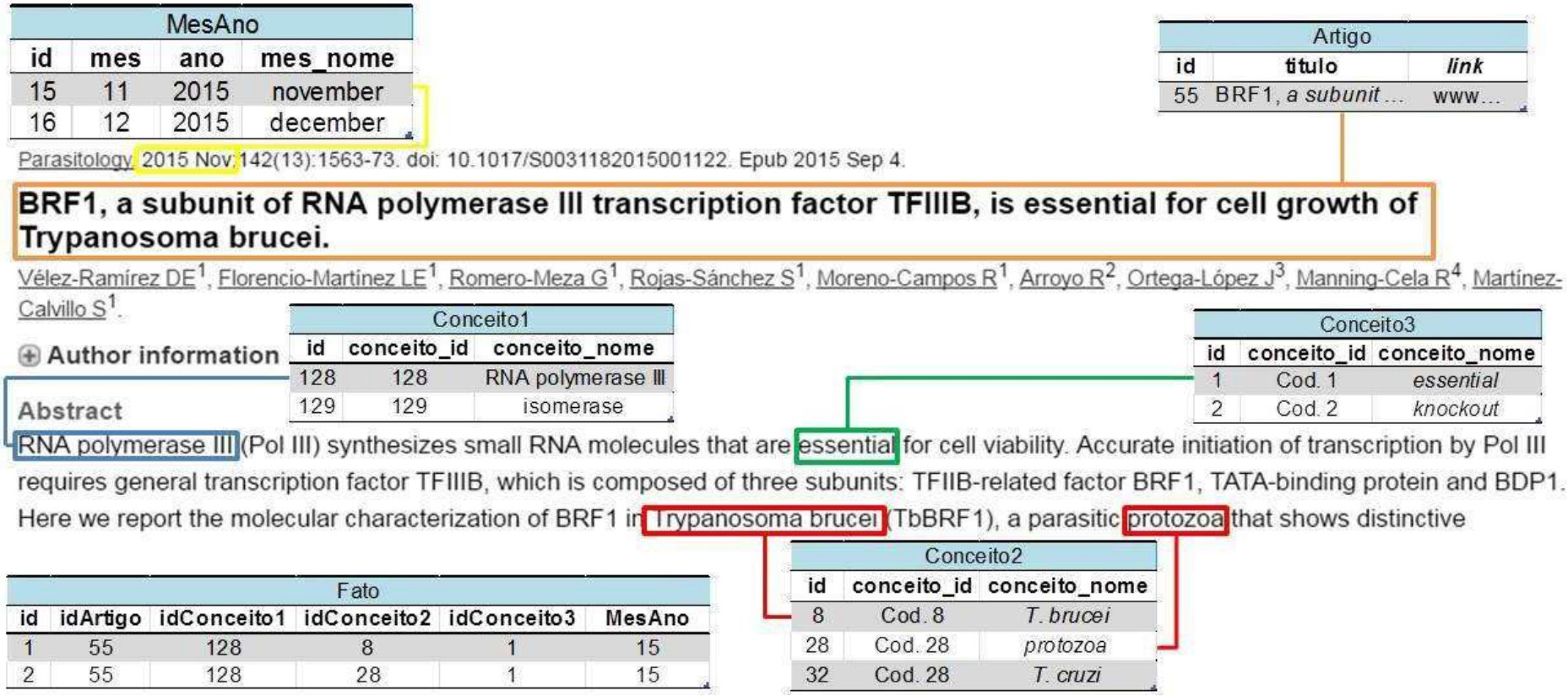


Figura 3.7: Visão geral do processo de identificação das dimensões em um artigo científico.

Podem ocorrer casos em que não há anotação de termos de uma dimensão específica. Para tratar essa exceção, uma tupla especial rotulada NA (não anotada) é inserida em todas as dimensões, quando um artigo não possui um termo da dimensão, na tupla dos fatos referente a esta dimensão recebe a referência do termo NA.

O item 3.4 apresenta a aplicação da metodologia em um caso de estudo sobre priorização de novos alvos de fármacos. Todo o processo é apresentado, desde a obtenção dos artigos para a extração dos dados até a realização das consultas no *data mart*.

3.3 Método de priorização de novos alvos de fármacos para *tripanossomatídeos*

Como mencionado nas seções 1.3 e 1.6, a priorização de alvos de fármacos é um tema muito relevante, pois pode auxiliar pesquisadores a encontrarem novos alvos de fármacos, tornando as pesquisas mais rápidas e baratas. Neste contexto, a busca por proteínas essenciais em organismos patógenos é fundamental, pois proteínas essenciais exercem funções imprescindíveis no ciclo de vida desses organismos e, portanto, são candidatas a possíveis alvos.

Juntamente com a importância de pesquisas por novos alvos de fármacos, emergem as doenças negligenciadas, por chamarem pouca atenção das indústrias farmacêuticas, ainda há pouca pesquisa e informação. Por estes motivos, este esta secção tem foco nos *tripanossomatídeos,* por serem os principais causadores de doenças negligenciadas nos países subdesenvolvidos.

Este item se dedica a aplicar a metodologia TOETL apresentada na seção 3.2. Por meio desta iniciativa, foi possível projetar, desenvolver e disponibilizar o TaP DM (*Target Prioritization Data Mart*), além de gerar metodologias, que por meio de consultas ao TaP DM, priorizam novos alvos de fármacos, auxiliando os pesquisadores na sua tomada de decisão.

No final desta seção, descrevemos alguns exemplos de consultas construídas e submetidas ao TaP DM. Para facilitar a manipulação do banco, uma interface OLAP é integrada à ferramenta, permitindo que os usuários realizem consultas de maneira mais intuitiva.

3.3.1 Anotação e Extração

Como mencionado no início desta seção, a etapa de *anotação e extração* foi apresentada e discutida no trabalho (2). Como a metodologia TOETL é uma extensão deste trabalho, neste item é apresentada uma visão geral desta etapa, pois algumas informações são utilizadas nas etapas posteriores e facilitam a compreensão deste trabalho.

Para construir um corpus de artigos que servirão como fonte de dados para o *data mart*, os termos envolvendo organismos de interesse e técnicas de essencialidade foram identificados. Este passo, denominado *entendimento do objeto da pesquisa*, foi realizado através de reuniões com usuários (pesquisadores) e resultou na seguinte lista de termos:

- Essencialidade de gene: *gene, protein, essential, essentiality, reverse genetic, knockout, knockdown, RNA interference, rnai, lethal phenotype, survival, null mutants*.
- Organismos (protozoários alvos e modelos): *Entamoeba histolytica, Leishmania major, Plasmodium falciparum, Trypanosoma brucei, Trypanosoma cruzi, Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Danio rerio, zebrafish, Mus musculus, Saccharomyces cerevisiae, Escherichia coli*.

A partir da combinação desses termos identificados e os operadores *AND/OR*, consultas foram construídas (passo de *definição de consultas*), como por exemplo: *essentiality AND ("trypanosoma cruzi" OR "trypanosoma brucei" OR "leishmania major" OR "entamoeba histolytica" OR "plasmodium falciparum")*.

O *PubMed Central* (PMC) foi escolhido como base textual, pois concentra milhares de citações e resumos sobre temas nas áreas de saúde e biomédicina (passo de *seleção das bases textuais*). As consultas acima foram submetidas ao *Pubmed* no segundo semestre de 2015, resultando em 1383 artigos completos (passo *montagem do corpus*). É importante salientar que as consultas retornaram muito mais que os 1383 textos obtidos, mas como alguns artigos são de acesso restrito, não foi possível adicioná-los ao *corpus*.

Para realizar a anotação semântica, foi necessário selecionar o conjunto de ontologias que descreviam os domínios de conhecimento presentes nos artigos (passo de *escolha das ontologias*). Os critérios de seleção são apresentados no trabalho (2) e 3 ontologias foram selecionadas:

- *Molecule Role* (115): representa um vocabulário controlado de proteínas, químicos e suas famílias;
- NCBI *Organismal Classification* (NCBI *Taxon*) (116): representa a classificação taxonômica de organismos vivos;
- NCI *Thesaurus* (NCIt) (111): contém um vocabulário para representar cuidados médicos, investigação básica e translacional, informação para o público e atividades administrativas. Inclui termos relacionados às técnicas de essencialidade gênica.

Com os conjuntos de artigos e ontologias definidos, o passo de *anotação semântica* pode começar. Mas antes, foi necessário submeter os artigos a um processo de limpeza, onde figuras, referências e *stop words* são removidos. Em seguida, usando a ferramenta de anotação Autômeta (117), os artigos foram anotados intrusivamente, ou seja, os termos citados no texto que pertencem à ontologia foram marcados por uma tag RDFa (*Resource Description Framework in Attributes*).

Após o passo de *anotação semântica*, segue-se para o passo de *extração de dados e informação*, que foi realizada através da API RDFa e armazenada em uma tabela do banco de dados. Todos os termos anotados são armazenados no banco de dados, se um termo é anotado mais de uma vez em um artigo, ele será inserido mais de uma vez no banco. Essa duplicidade de informação não compromete as respostas do DM, pois este armazenamento inicial é um processo temporário que possibilita a população do DM. Essas repetições serão eliminadas no momento da população da tabela de fatos.

Assim, cada termo anotado é registrado em uma tabela (anotação) contendo as seguintes colunas: id (identificador de tupla), id do artigo (identificador do artigo), rótulo (termo anotado), id da classe (identificador anotado da classe), conforme mostrado na Tabela 3.1.

Tabela 3.1: Tabela anotação onde os dados extraídos dos artigos são armazenados

Anotação				
id	**id_artigo**	**termo**	**id_classe**	**nome_classe**
1	5	Protein	IMR_0000001	MoleculeRole
2	2	small GTPase	IMR_0000914	GTP-binding protein
3	38	transferase	IMR_0000207	enzyme

A próxima seção descreve o estágio de construção do *data mart*, que inclui a identificação dos principais elementos dimensionais (dimensões e tabelas de fatos) que irão compor a modelagem do *schema* estrela.

3.3.2 Modelagem e Transformação

Conforme mencionado no item 3.2.2.1, o primeiro passo na concepção e construção do *data mart* é a identificação de demandas analíticas. Neste passo, as questões que o DM deve ser capaz de responder são levantadas. Através de entrevistas, pesquisadores do Laboratório de Peptídeos e Bioquímica do IOC/Fiocruz mencionaram questões como: "*Qual a técnica mais citada com algum organismo específico*?", "*Quais organismos foram mais citados com alguma proteína específica*?", "*Qual organismo é citado com pelo menos duas técnicas de essencialidade genética*?" e "*Qual químico foi mais frequentemente citado com algumas técnicas de essencialidade genética*?".

Utilizando estas questões como referência, o próximo passo (Identificação de Dimensões e Categorias) identificou 6 conceitos (proteína, organismo, técnicas de essencialidade genética, químico, artigo e tempo) candidatos para exercerem o papel de dimensões no TaP DM. Observe que a maioria deles é mencionada nas consultas que foram levantadas.

A Figura 3.8 representa uma consulta feita no cubo multidimensional, onde ela busca o número de artigos que citam técnicas, químicos e proteínas conjuntamente. Note que esta consulta apresenta os resultados ao longo do ano, então o usuário pode visualizar o cenário do problema da perspectiva do tempo. Outro fato a se destacar é a ausência da dimensão organismos, pois cubo permite adicionar e remover dimensões de interesse do usuário.

1989		Proteínas		
1988		Proteínas		
1987		Proteínas		
Técnicas	**Químicos**	P1	P2	P3
T1	Q1	20	10	60
	Q2	40	15	21
	Q3	10	1	55
T2	Q1	8	52	2
	Q5	21	14	8
	Q7	35	24	42

Figura 3.8: Representação da resposta de um cubo multidimensional.

Para cada um destes conceitos foi criada uma tabela de dimensão correspondente: DimProteina, DimQuimico, DimOrganismo, DimTecnica, DimArtigo e DimMesAno. As tabelas DimProteina, DimQuimico, DimOrganismo e DimTecnica foram projetadas para representar os termos presentes nas ontologias e, consequentemente, elas refletem as relações hierárquicas destes termos, como mostrado anteriormente no item 3.2.2.

No terceiro passo (*extração e corte*), uma análise de cada ontologia foi feita para identificar se cortes seriam necessários. Esta não é uma tarefa fácil, já que algumas ontologias são muito grandes. É importante observar que as tabelas de dimensão devem ser preenchidas somente com termos que correspondam ao domínio de representação da dimensão. Um corte realizado de maneira errada pode levar a resultados analíticos equivocados.

Este cenário foi observado na ontologia *Molecule Role*. Esta ontologia foi editada para gerar dois segmentos (módulos). Um módulo incluiu termos relacionados a químicos e o outro incluiu termos de proteínas.

A separação da ontologia em módulos para que possam popular dimensões diferentes, além de evitar que a ferramenta realize interpretações erradas dos resultados, tem o objetivo de permitir que os conceitos presentes nos módulos sejam correlacionados. Desta forma, perguntas analíticas que envolvem descobrir a relação (co-ocorrência) de proteínas e químicos podem ser atendidas.

A ontologia NCI *Thesaurus* descreve diversos domínios de conhecimento, incluindo as técnicas de essencialidade. Mas analisando os termos presentes na ontologia, concluiu-se que as informações sobre as hierarquias dos elementos não seriam úteis para os usuários. Realizar operações de *roll-up* e *drill-down* nesta dimensão não acrescentaria informações úteis à análise dos dados.

Além disso, a ontologia NCI *Thesaurus* é muito ampla e descreve vários conceitos de diferentes áreas do conhecimento. As classes que descrevem as técnicas de essencialidade representam uma parte muito pequena da ontologia e, essas classes estão dispersas em vários ramos da ontologia.

Considerando essas características específicas da ontologia, um módulo da ontologia foi selecionado. Este módulo contém todas as classes de interesse, neste caso de estudo, as classes que descrevem os termos levantados sobre essencialidade (item 3.3.1). A construção do módulo é apresentada de forma geral na Figura 3.9.

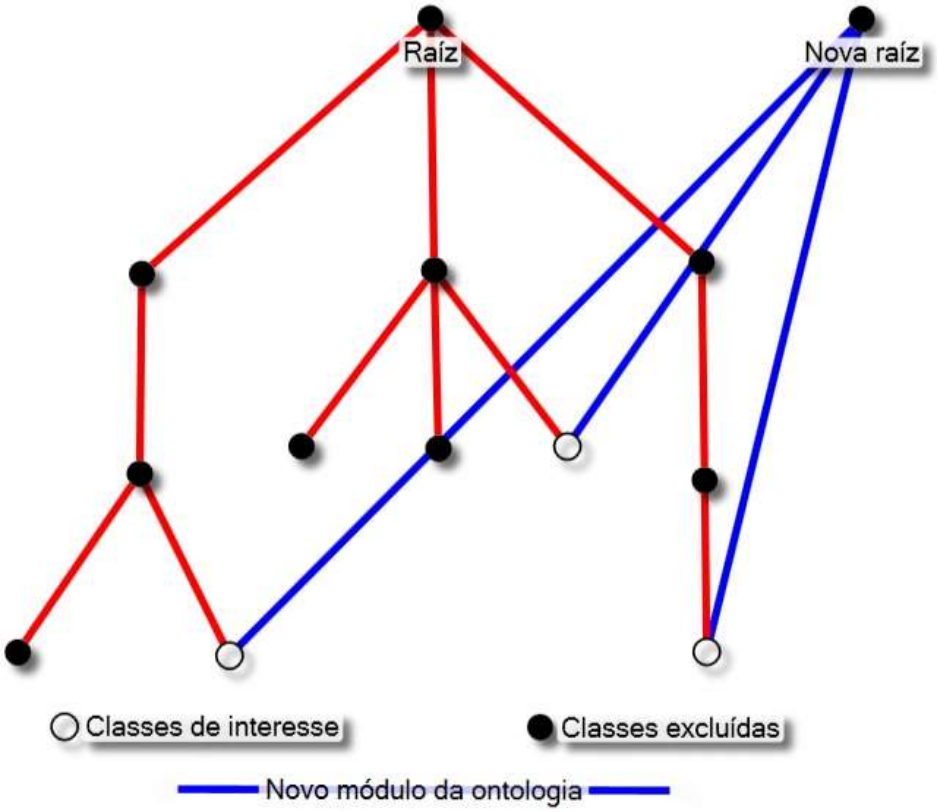


Figura 3.9: Criação do módulo de interesse de uma ontologia.

Com relação à ontologia NCBI *Taxon* não foi necessário realizar cortes, pois a ontologia descreve apenas organismos e suas classificações taxionômicas, não possuindo outros domínios de conhecimento.

O próximo passo é a Identificação/Definição de Fatos. Neste caso de estudo, o fato não inclui a observação do número de ocorrência dos termos dentro do artigo, o que deve ser registrado é apenas se ele está presente ou não. Para atender a estes requisitos uma *factless table* foi utilizada.

Por consequência, usando uma *factless table*, as ocorrências dos termos de dimensões são feitas pelo registro de suas referências. Portanto, cada tupla da tabela de fatos armazena os identificadores dos termos de cada dimensão presentes em um artigo. Deste modo, cada tupla da tabela de fatos consiste em 7 campos: o campo chave e os 6 campos que se ligam as 6 tabelas de dimensão. Considerando um caso hipotético no qual um artigo é anotado com um termo de cada dimensão, uma tupla da tabela de fatos seria suficiente para representar a presença das tabelas de dimensões neste artigo, como representa a Figura 3.10.

2000		Artigo 325						Artigo 340			
1993		Artigo 60						Artigo 65			
1992		Artigo 1						Artigo 2			
		Organismos/Proteinas						Organismos/Proteinas			
		O1			O2			O1		O3	
Técnicas	Químicos	P1	P2	P3	P1	P2	P3	P2	P5	P2	P5
T1	Q1	1	1	1	1	1	1	1	1	1	1
	Q2	1	1	1	1	1	1	1	1	1	1
	Q3	1	1	1	1	1	1	1	1	1	1
	Q4	1	1	1	1	1	1	1	1	1	1
T2	Q1	1	1	1	1	1	1	1	1	1	1
	Q5	1	1	1	1	1	1	1	1	1	1
	Q7	1	1	1	1	1	1	1	1	1	1

Figura 3.10: Identificação dos fatos.

Mas na maioria dos casos, em um único artigo ocorre a presença de mais de um termo de uma dimensão. Assim, todos esses elementos são combinados com os elementos citados de outras dimensões, conforme mencionado no item 3.2.2.4. Casos em que não há anotação de termos de uma dimensão específica, a referência 99999999 foi adotada, apontando para o termo NA (não anotado) registrado na dimensão.

Neste estudo de caso, um artigo é registrado na tabela de fatos se ao menos 3 dimensões forem citadas, excluindo as dimensões DimMesAno e DimArtigo, já que são inerentes de cada artigo. Esta foi uma escolha de implementação do TaP DM, uma vez que para os pesquisadores, o registro de artigos onde apenas uma ou duas dimensões ocorrem não é significativo para suas análises.

É importante enfatizar a definição do grão no estudo de caso. O grão representa, em um artigo, a presença de pelo menos 3 dimensões em um determinado mês do ano. As informações do mês são tomadas a partir da data de publicação do artigo. A unidade de tempo mês foi estabelecida como a menor unidade, pois não há necessidade de realizar análises para espaços temporais menores (hora, dia, etc.). A partir da unidade de mês e por meio das operações de *drill-up*, é possível apresentar informações mais generalistas, como por exemplo análises bimensais, trimestrais, anuais e etc.

Alguns detalhes na implementação das dimensões foram omitidos nesta seção, pois eles estão envolvidos com aspectos específicos das ontologias e da ferramenta Pentaho. Por estes motivos, informações mais específicas são discutidas e apresentadas no Apêndice B - . Ao final desta etapa, o TaP DM está projetado, como mostra a Figura 3.11, iniciando assim a terceira e última etapa de Carga da metodologia.

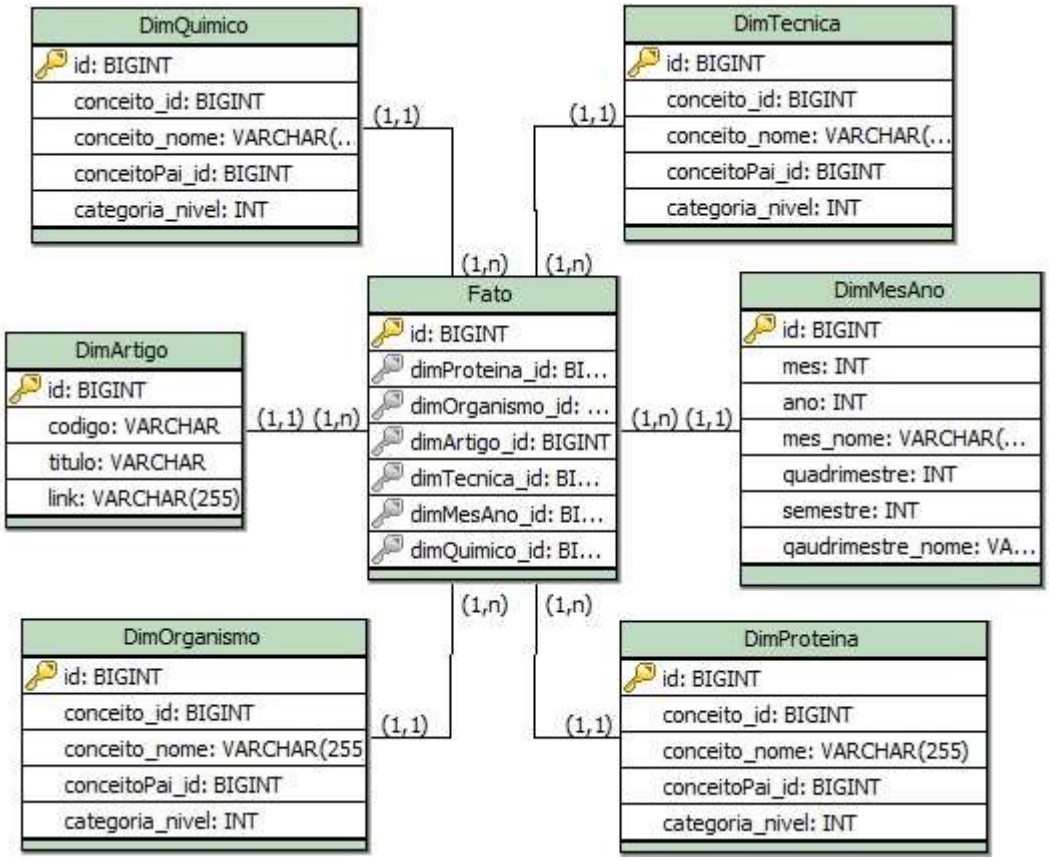


Figura 3.11: Projeto final do TaP DM.

3.3.3 Carga

Após realizados os cortes nas ontologias, elas foram usadas para popular as tabelas de dimensão. A tabela DimOrganismo foi populada utilizando toda a ontologia de NCBI *Taxon*, fornecida pelo *National Center for Biotechnology Information* (NCBI), que incorpora uma completa representação hierárquica dos organismos.

Da mesma forma, as tabelas DimProteina e DimQuimico também foram projetadas para armazenar informações extraídas de uma ontologia específica, a *Molecule Role*. Entretanto, cada tabela armazenou um módulo da ontologia, obtidos através da extração dos ramos *Protein* e *Chemical*.

A tabela de DimTecnica foi populada utilizando um módulo da ontologia NCI *Thesaurus.* Como a NCI *Thesaurus* é uma ontologia que descreve diversos conceitos e os termos referentes a técnicas de essencialidade se encontram dispersos em diversos ramos, o módulo foi criado a partir da identificação dos termos na ontologia e então extraídos, como mostrou a Figura 3.9. Neste caso, os termos não foram categorizados, porque sua hierarquia é superficial e neste caso de estudo, não possuem utilidade do ponto de vista analítico, como foi mencionado no item 3.3.2.

Diferentemente, a população das tabelas de dimensão DimArtigo e DimMesAno não foi baseada em ontologias. Em vez disso, foi realizada por meio da extração de dados do conjunto de artigos selecionados e anotados. A tabela DimArtigo foi preenchida com dados (id, título, *link*, etc.) extraídos dos artigos Pubmed, permitindo que fossem recuperados na sua forma original. A DimMesAno foi preenchida com os meses e anos do intervalo de tempo que cobriu todo o conjunto de artigos selecionados.

Finalmente, a população da tabela de fatos contou com as dimensões e o DS (Base de Dados de Anotações) já preenchidos. Conforme descrito anteriormente, para cada id de artigo, ele obtém o conjunto de tuplas relacionadas na tabela de anotação, identifica as tuplas correspondentes nas tabelas de dimensão e combina suas referências para formar uma tupla na tabela de fatos.

A Tabela 3.2 mostra um trecho da tabela de fatos gerada. Observe que a tupla 4 aponta para a tupla NA da tabela DimTecnica (valor 9999999). Isso significa que o artigo 22 não menciona termos referentes as técnicas de essencialidade.

Tabela 3.2: Tabela de fatos carregada.

Fato						
id	**idDimArtigo**	**idDimQuimico**	**idDimProteina**	**idDimTecnica**	**idDimOrganismo**	**idDimMesAno**
1	20	5	10	3	6	5
2	21	2	15	3	7	8
3	21	2	15	4	7	8
4	22	8	12	99999999	5	12

No final desse processo, o *data mart* está projetado, construído e carregado. O sistema de gerenciamento de banco de dados MySQL (DBMS) foi usado para hospedar o TaP DM, permitindo aos usuários submeterem consultas.

A seguir, no item 3.3.4, são apresentadas algumas consultas que foram submetidas ao TaP DM com o objetivo de demonstrar como a ferramenta é flexível e capaz de atender à perguntas mais complexas, onde mais de um conceito são correlacionados.

3.3.4 Consultas submetidas ao Data Mart

Com o TaP DM projetado e carregado, este está pronto para o passo de Submissão das consultas. Diversas consultas foram construídas em SQL com o objetivo de testar a ferramenta. Outro objetivo é certificar que as perguntas de interesse possam ser respondidas. Abaixo estão alguns exemplos de consultas realizadas e também os resultados obtidos.

a) Quais proteínas foram citadas com o organismo *T. cruzi* em 2015 que nunca tinham sido citadas anteriormente?

Na Figura 3.12 é apresentado o *script* SQL construído para atender à pergunta. É utilizado o identificador (5691) do organismo *T. brucei* na consulta com o objetivo de melhorar o tempo de resposta do banco, evitando assim mais um *join* com a tabela de dimensão organismos. A resposta obtida está representada na Tabela 3.3.

```
select c1.proteina from (
    (select distinct(dp.nome) proteina, dp.idProteina id from fatos f
        inner join dimMonthYear dt inner join dimProteina dp
        on f.idDimProteina = dp.idProteina and f.idDimMonthYear = dt.id
        and f.idDimOrganismo = 5691 and dt.ano = 2015
    )c1 left join

    (select distinct(dp.nome) proteina, dp.idProteina id from fatos f
        inner join dimMonthYear dt inner join dimProteina dp
        on f.idDimProteina = dp.idProteina and f.idDimMonthYear = dt.id
        and f.idDimOrganismo = 5691 and dt.ano < 2015
    )c2 on c1.id = c2.id
) where c2.proteina is null order by c1.proteina;
```

Figura 3.12: Consulta SQL submetida ao TaP DM. Esta consulta busca todas as proteínas que foram citadas em 2015 com *T. brucei* que nunca tinham sido anteriormente.

Tabela 3.3: Resposta do TaP DM. Proteínas citadas em 2015 com *T. brucei* que nunca tinham sido citadas anteriormente.

Resposta
Aspartate aminotransferase
DECA_DROME
ENTP5_HUMAN
GAST_RAT
KCNC1_RAT
KTHY_MOUSE
KV401_HUMAN
MDHM_HUMAN
NDKA_MOUSE
PUR8_HUMAN
PURA2_HUMAN
RAB14_HUMAN
rhomboid
RHOM_DROME
ribonucleoside-diphosphate reductase
RPC5_HUMAN
RPIA_HUMAN
Sar

b) Quais químicos foram citados em comum com os organismos *Leishmania major*, *Schizosaccharomyces pombe* e *Trypanosoma brucei*?

Na Figura 3.13 é apresentado o script SQL construído para atender à pergunta. São utilizados os identificadores (5691 do *T. brucei, 4896 do Schizosaccharomyces pombe e 5664 do Leishmania major*) dos organismos na consulta com o mesmo objetivo mencionado no item a). A resposta obtida está representada na Tabela 3.4.

```
select c1.idq as id, c1.nome from (
    (select distinct(f.idDimQuimico) idq, dq.nome nome from fatos f inner join
        dimQuimico dq on f.idDimQuimico = dq.id
        and f.idDimOrganismo = 5664) c1
        inner join
    (select distinct(f.idDimQuimico) idq, dq.nome nome from fatos f inner join
        dimQuimico dq on f.idDimQuimico = dq.id
        and f.idDimOrganismo = 4896) c2
        inner join
    (select distinct(f.idDimQuimico) idq, dq.nome nome from fatos f inner join
        dimQuimico dq on f.idDimQuimico = dq.id
        and f.idDimOrganismo = 5691) c3
    on c1.idq = c2.idq and c3.idq = c1.idq
);
```

Figura 3.13: Consulta SQL submetida ao TaP DM. Esta consulta busca os químicos citados em comum com os organismos *Leishmania major, Schizosaccharomyces pombe e Trypanosoma brucei*.

Tabela 3.4: Resposta do TaP DM. Químicos citados em comum com os organismos *Leishmania major, Schizosaccharomyces pombe e Trypanosoma brucei.*

Resposta			
amino acid	phosphatidylinositol	adenosine	sphingolipid
nucleotide	phospholipid	Uracil	sphingosine
peptide	lipid	carbohydrate	NH3
CO2	diacylglycerol	cAMP	Lanosterol
glutathione	Formaldehyde	AMP	L-Proline
ATP	dCTP	Acetic acid	rapamycin
CTP	chemical	Uridine	H+
UTP	gas	Acceptor	dTTP
GTP	ion	Folic acid	Thymidine
Creatine	ADP	CoA	guanosine
Glutathione	Guanine	GMP	NADH
glutamic acid	GDP	Pyrophosphate	NADPH
Urea	ceramide	IMP	Ubiquinone
nucleoside	H2O	phosphatidylserine	Iron
Triphosphate	inositol phospholipid	vitamin	H2O2
glycine	Adenine	Orotic acid	Squalene

Como apresentado nos exemplos acima, o TaP DM permite construir consultas para questões mais abrangentes. Porém, a construção das consultas se apresentou relativamente complexa, exigindo do usuário um conhecimento mais

profundo da linguagem SQL. A complexidade aumenta principalmente para questões que envolvem manipulação dos níveis hierárquicos.

Outro ponto importante são as interfaces de manipulação do banco. Elas não são projetadas para que a construção e resultados das consultas sejam feitas de forma intuitiva e amigável, impossibilitando a utilização da ferramenta por pessoas sem treinamento específico. Por estes motivos surgiu a necessidade de implantar uma ferramenta OLAP, que é apresentada no item 3.4.

Estas consultas mostram a capacidade da ferramenta de buscar respostas para as mais diversas e específicas perguntas que os pesquisadores possam ter. Com isso, no próximo item (3.3.5), é apresentado uma estratégia para ajudar o pesquisador a priorizar novos alvos de fármacos, mostrando que o TaP DM permite criar consultas ainda mais complexas e flexíveis.

3.3.5 Estratégia de priorização de novos alvos de fármacos

Neste item é apresentada uma estratégia de busca de novos alvos de fármacos para um determinado organismo. Desta forma, é possível explorar parte das possibilidades de consulta do *data mart*, permitindo aos usuários correlacionar elementos na busca por alguma questão de seu interesse.

É importante salientar que esta metodologia busca priorizar novos alvos para concentrar esforços. Como o TaP DM é alimentado por informações contidas em artigos e ontologias, os resultados obtidos precisam ser validados por outros processos. A análise de artigos científicos por parte do pesquisador é apenas um dos processos na busca por novos alvos. Esta metodologia busca enriquecer os resultados obtidos dos artigos, encontrando relações e informações importantes que podem passar desapercebidos na leitura, ainda mais se for considerado a grande quantidade de artigos disponíveis. Analisar uma fonte tão importante de dados, que trás informações de fronteira de conhecimento dos pesquisadores, pode contribuir muito para as futuras tomadas de decisões.

Diferentemente das perguntas específicas apresentadas no item 3.3.4, este item apresenta um exemplo de pergunta mais ampla como “*Quais proteínas possuem mais chances de serem novos alvos para um determinado o organismo*?”. O TaP DM permite ao usuário estabelecer e aplicar uma estratégia de busca, conseguindo filtrar grande quantidade de dados e descobrir relações implícitas entre os elementos, o que não é possível por meio de outras ferramentas.

A pergunta de interesse apresentada neste exemplo é: "*Quais são os melhores novos alvos para um organismo de interesse*?". Para responder a esta questão, algumas definições são feitas:

- P: conjunto proteína (representado pela tabela DimProteina);
- O: conjunto organismo (representado pela tabela DimOrganismo);
- F: conjunto fato (representado pela tabela Fatos).

Se a pergunta objetiva encontrar novos possíveis alvos, então as proteínas candidatas nunca foram citadas com o organismo de interesse anteriormente. A Figura 3.14 apresenta uma visão geral da abordagem e ilustra as relações entre as proteínas com maior probabilidade de ter um papel importante no estudo de sua essencialidade (EP conjunto) e as proteínas nunca mencionadas com ele (NTP conjunto), observe que estes são conjuntos disjuntos. É na busca por conjuntos disjuntos que esta estratégia se apoia para encontrar relações ocultas entre elementos de interesse. No Apêndice C - é apresentada em maiores detalhes as consultas criadas em álgebra relacional.

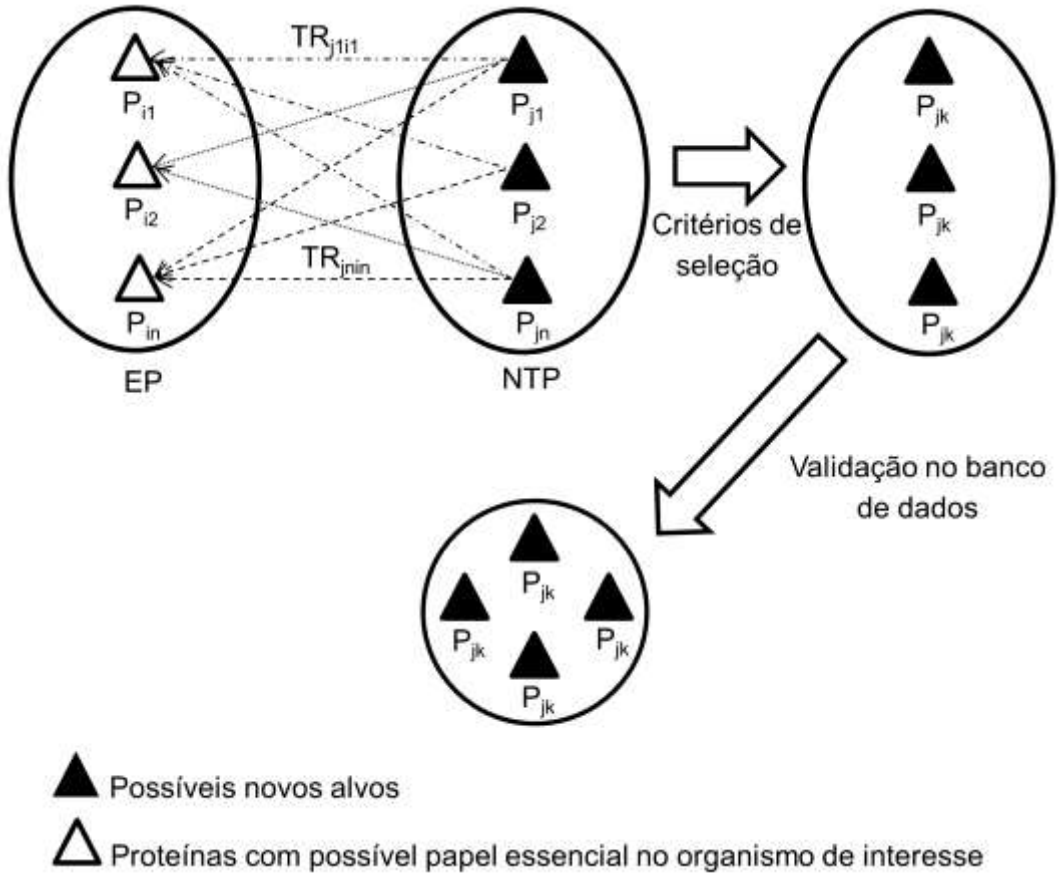


Figura 3.14: Visão geral da estratégia de priorização de novos alvos de fármacos.

Depois de definir os conjuntos EP e NTP, é necessário calcular uma taxa de relação entre os seus elementos. Como mencionado anteriormente, o objetivo é correlacionar elementos do conjunto NTP com elementos do conjunto EP, correlacionando proteínas EP com probabilidade de serem essenciais para um organismo e que são citadas com proteínas do conjunto NTP.

Isto pode ser feito calculando para cada par de proteínas (P_{in}, P_{jn}) que pertencem aos conjuntos {EP x NTP}, uma taxa de co-ocorrência no *corpus* de artigos. Na consulta (8), $cNTP_{jn}$ é o número de artigos em que P_{jn} foi citado e $cNTP_{jn}EP_{in}$ é o número de artigos que P_{jn} foi citado com P_{in}. Quanto maior a co-ocorrência entre P_{jn} e P_{in}, maior é a taxa TR_{jnin}. Se TR_{jnin} for igual a 100.00, significa que P_{jn} sempre ocorre com P_{in}, tornando possível P_{jn} exercer ou participar de um processo essencial para o organismo de interesse.

$$TR_{jnin} = cNTP_{jn} * 100 / cNTP_{jn}EP_{in} \quad (8)$$

Para avaliar o comportamento desta estratégia aplicada ao TaP DM, o organismo *T. brucei* foi selecionado como o organismo de interesse. As consultas (3) e (4) foram construídas e executadas, sendo o conjunto NTP constituído de 1159 proteínas e o conjunto EP de 23 proteínas.

Os parâmetros nTec e nAno foram ajustados para os valores 3 e 7, respectivamente. Assim, para que uma proteína pertença ao conjunto NTP, além de nunca ter sido mencionada com o organismo de interesse, em todos os artigos onde ela é citada, deve ser citado também pelo menos 3 técnicas de essencialidade. Para pertencer ao conjunto EP, a proteína deve ser citada por pelo menos 7 anos com *T. Brucei* e com as técnicas de essencialidade.

Em seguida, a consulta (6) foi submetida ao TaP DM e a taxa de TR foi calculada para cada par de proteínas pertencentes ao conjunto {EP X NTP}. Foram selecionados pares de proteínas com TR iguais a 100.00, resultando em 302 proteínas (P_{jk}). Estas proteínas podem pertencer a diferentes organismos e é necessário selecionar quais proteínas pertencem ao organismo de interesse (*T. brucei*). Com este propósito em mente, as bases de dados UniProt e Tdr *Targets* foram utilizadas para identificar quais proteínas pertencem a *T. brucei*. Ao final desta validação, 92 proteínas foram selecionadas, sendo as proteínas mais prováveis de serem novos alvos potenciais para *T. brucei*.

As bases de dados UniProt e Tdr *Targets* foram utilizadas neste processo porque alguns termos de proteínas empregados por estas bases de dados e a

ontologia que alimentou o DM são diferentes. Algumas proteínas são validadas por apenas um desses bancos de dados. Por exemplo, o termo da *PI3-kinase* está presente no Tdr *Targets* e no DM, mas o Uniprot utiliza o termo *fosfatidilinositol 3-quinase, putativa*. Se apenas um banco de dados fosse utilizado para validar as proteínas, o número de proteínas não validadas em consequência desse problema, seria maior. Usando os dois bancos de dados para a validação não elimina totalmente esse problema, mas ele é minimizado.

Para validar as 92 proteínas resultantes como possíveis alvos, utilizou-se o banco de dados Tdr *Targets*. Das 92 proteínas 35 foram positivamente confirmadas como alvos. Mas o mesmo problema mencionado anteriormente, sobre os diferentes termos adotados pelas bases de dados, também acontece neste caso. Portanto, as 57 proteínas que foram apontadas pelo TaP DM como possíveis alvos, mas não foram confirmadas pelo Tdr *Targets*, ainda podem revelar-se alvos validados. Dentre essas proteínas está a *fosfofrutoquinase*, ela não foi validada porque o Tdr *Targets* utiliza o termo *ATP-dependente fosfofructoquinase*. Outro aspecto a ser explorado é o fato que, dentre essas 57 proteínas existam alvos que não foram cadastrados na no banco Tdr *Targets* ou são alvos ainda desconhecidos de *T. brucei*. A lista das proteínas resultantes desta estratégia está na Tabela 3.5.

Tabela 3.5: Lista das 57 possíveis proteínas alvos para o organismo *T. brucei*.

	Nomes preferenciais
1	KPYM_HUMAN
2	AAPK2_HUMAN
3	AAPK1_HUMAN
4	RAF1_HUMAN
5	KS6A3_HUMAN
6	PAK1_MOUSE
7	ERCC2_MOUSE
8	RAF1_RAT
9	ERCC3_MOUSE
10	RPCY
11	DNA-directed RNA polymerase III subunit 22.9 kDa polypeptide
12	AKT1_RAT
13	PDPK1_DROME
14	KS6B1_HUMAN
15	KS6B1_MOUSE
16	DPOLB_HUMAN
17	GNPI1_HUMAN
18	PTEN_HUMAN
19	phosphofructokinase

20	Q9V3L4_DROME
21	SGK1_CAEEL
22	Glyceraldehyde 3-phosphate dehydrogenase, liver
23	GSK3_CAEEL
24	PGS1_HUMAN
25	CG21_YEAST
26	NMT1_MOUSE
27	NMT1
28	ENOA_HUMAN
29	HDAC3_MOUSE
30	PTPA_HUMAN
31	PTN11_HUMAN
32	TRAF2_MOUSE
33	PSN2_MOUSE
34	Rab7
35	PSN2_HUMAN
36	TOR1_SCHPO
37	PSN2_RAT
38	TOR2_SCHPO
39	TRAF6_HUMAN
40	DHSA_MOUSE
41	METK2_HUMAN
42	PMM1_HUMAN
43	UBE2N_MOUSE
44	ADHX_MOUSE
45	XPO1_MOUSE
46	MK08_MOUSE
47	XPO1_RAT
48	MAAI_HUMAN
49	PLK1_HUMAN
50	HHAT_MOUSE
51	UBE3A_HUMAN
52	NDKA_RAT
53	F16P1_HUMAN
54	HHAT_HUMAN
55	EP300_HUMAN
56	DNLI4_HUMAN
57	MDM2_DANRE

Entre os 1383 artigos presentes no *corpus* distribuídos ao longo de 79 anos, 4187 termos de proteínas, 2116 termos de organismos e 249 termos de químicos foram citados. Possibilitar estabelecer relacionamentos entre estes elementos em um conjunto tão grande de dados e conseguir reduzir, em pouco tempo, os elementos de estudo em algumas poucas dezenas de possibilidades, priorizando

esforços, é de vital importância. Utilizando estas abordagens, o pesquisador consegue ter a visão ampliada sobre o problema, permitindo que estabeleça parâmetros tornando suas buscas ora menos ora mais criteriosas. Esta liberdade e flexibilidade são fundamentais para auxiliar o pesquisador a tomar decisões de forma mais eficiente e rápida.

Para facilitar a compreensão, a Figura 3.15 mostra detalhadamente como a seleção estratégica de alvos foi realizada em nível dos dados armazenados no *data mart.*

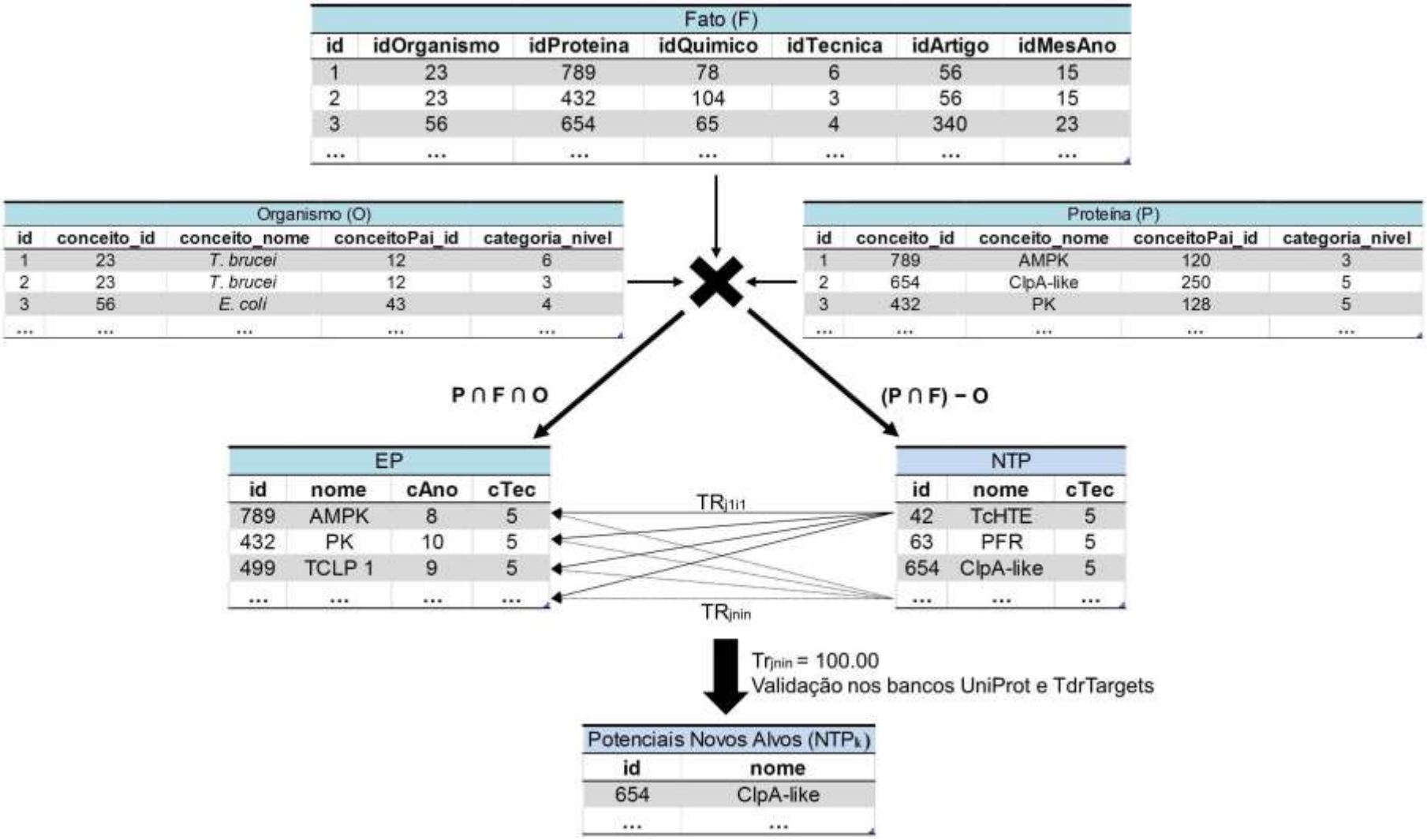


Figura 3.15: Visão geral da estratégia de obtenção de novo alvos considerando as operações e tabelas envolvidas.

3.4 Ferramenta OLAP

Para facilitar a utilização do TaP *DM* e tornar a visualização das informações mais intuitivas, foi implantada a ferramenta OLAP Pentaho. Ela é capaz de prover uma interface *web* amigável, permitindo a construção de consultas de forma interativa e a visualização gráfica dos resultados, possibilitando a manipulação do banco sem a necessidade de aprender outras tecnologias.

O Pentaho foi escolhido por ser amplamente utilizado, possui licença gratuita, código aberto e todas as facilidades de uso mencionadas anteriormente. Isto facilita a implantação da ferramenta pelo desenvolvedor e utilização do *data mart* pelos usuários.

Com esta ferramenta (Figura 3.16), o usuário é capaz de interagir com o banco por meio de uma interface mais amigável e intuitiva. Um exemplo de consulta é apresentado no item 4.1, onde a navegação nas hierarquias é simples, pois a interface utiliza ícones e dinâmicas comuns aos sistemas de uso cotidiano dos usuários, além de realizar consultas com bom desempenho. A interface também permite criar relátorios e visualizar as respostas gráficamente, proporcionando uma melhor análise das informações obtidas.

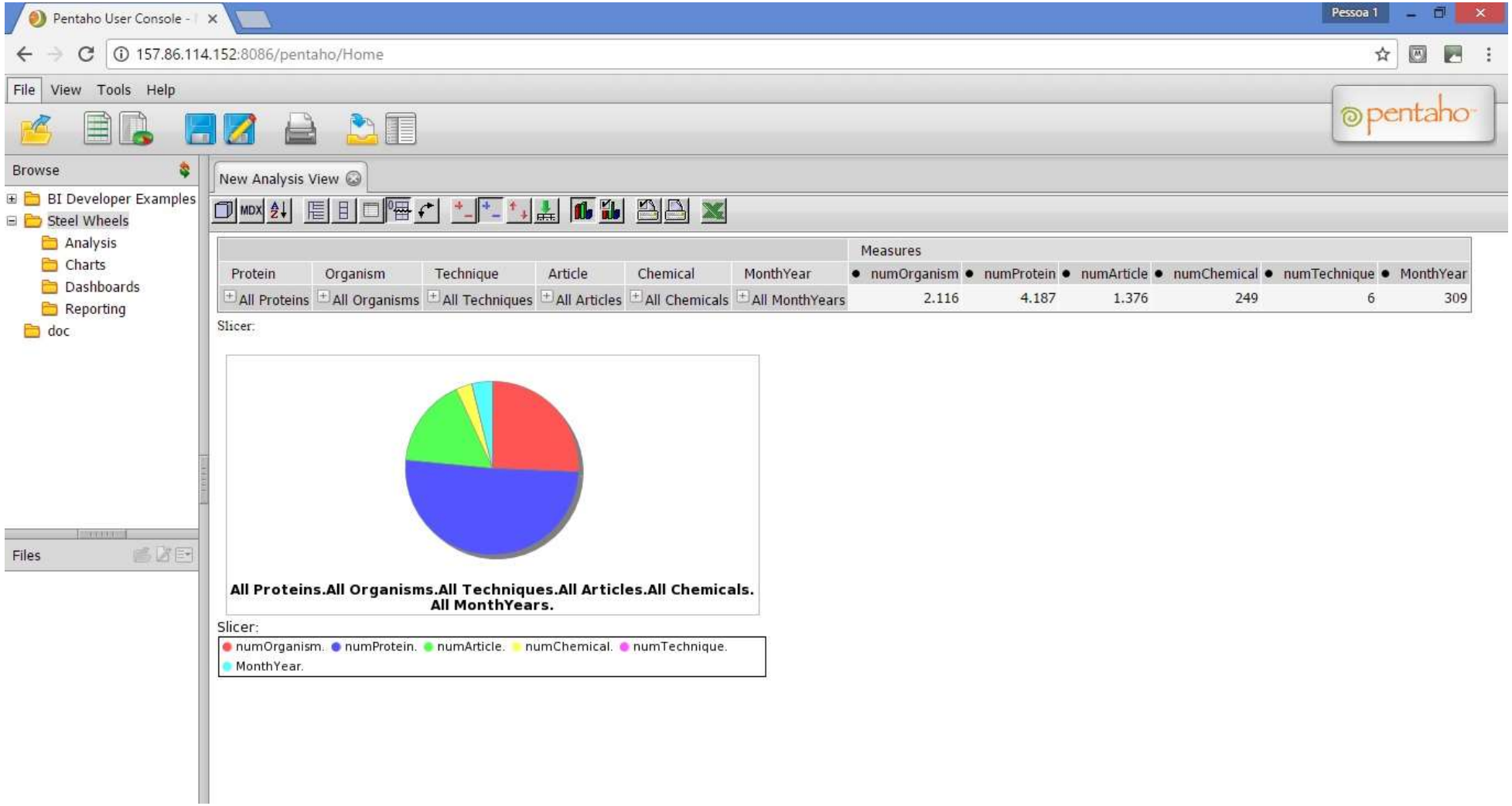


Figura 3.16: Interface do Mondrian acessando o TaP DM.

4 RESULTADOS E DISCUSSÃO

Os principais resultados apresentados neste trabalho são a metodologia de construção de *data marts* baseadas no uso de ontologias para análise de dados não estruturados e a proposta para priorização de novos alvos de fármacos.

Existem poucos estudos que exploram metodologias específicas para desenvolver DMs a partir de dados não estruturados para análise e suporte à decisão (118–120). Alguns deles (39,121–123) propõem abordagens baseadas em ontologia, que afirmam ser úteis para lidar com fontes de dados textuais. Mas diferentemente da abordagem TOETL aqui descrita, nenhum deles detalha como lidar e selecionar grandes quantidades de dados textuais. Como mencionado anteriormente, esse recurso é fundamental para apoiar pesquisas científicas, especialmente no campo biomédico.

Um deles (122) reforça a necessidade de utilizar ontologias no planejamento do DW para superar deficiências importantes. Em (121), os autores usam ontologias para orientar o projeto do DW, analisando cada fonte de dados e os requisitos de análise. Outros estudos usaram ontologias diretamente e automaticamente no design do *data mart*, onde uma ontologia foi usada como entrada e, no final do processo, um diagrama dimensional foi obtido (121,123).

No entanto, estes trabalhos se concentram apenas em uma ontologia. Devido ao fato de que a modelagem de problemas biomédicos precisa usar múltiplas ontologias, de acordo com os diferentes subdomínios envolvidos, essas não são abordagens adequadas. Além disso, em nossa abordagem, nos concentramos em um interesse de análise específico e identificamos os cortes de um conjunto de ontologias selecionadas. Diferentemente, sua ideia é explorar completamente uma única ontologia.

Como mencionado anteriormente, as ontologias de biomedicina são bastante amplas, e explorar toda a ontologia pode levar a múltiplas dimensões. Além disso, é importante enfatizar que essas iniciativas relacionadas discutem ou propõem o uso de ontologias no design do *data mart* e não no processo ETL. Em nossa abordagem, as ontologias são usadas para anotar textos extraídos de fontes de dados selecionadas (ETL primeira etapa) e, em seguida, para orientar o projeto do DM. No contexto de fontes científicas de dados textuais, é necessário identificar o foco da

pesquisa para avançar para o projeto DM. Portanto, nossa abordagem integra os processos de design ETL e DM.

Outro trabalho importante (124) aborda a construção de DW a partir de dados não estruturados, detalhando um processo ETL de 11 etapas. Mas este trabalho propõe uma abordagem baseada em PNL para o ETL e aplica técnicas, como o reconhecimento de frases, a filtragem de palavras e a substituição de sinônimo. Diferentemente, ele não usa anotações com base em ontologia para definir dimensões, nem usa ontologias para enriquecer hierarquias de dimensão. Portanto, de acordo com o nosso conhecimento, não há trabalho semelhante ao presente trabalho.

Finalmente, a ideia de explorar o banco de dados PubMed para identificar alvos de fármacos, para o tratamento de doenças negligenciadas não é nova. Em (125) os autores informam sobre o uso de etapas de mineração de texto para extrair termos relacionados a nomes de proteínas ainda não explorados como alvos para essas doenças. No entanto, eles não se concentram na construção de um DM para análise posterior, nem usam ontologias para classificar ou organizar termos de acordo com seu contexto.

Além disso, encontrar links entre conceitos pode não ser suficiente para tomar uma decisão melhor. Os tomadores de decisão (pesquisadores) precisam analisar os fatos e obter respostas para suas perguntas específicas (126). *Data warehouse* é uma abordagem madura que é altamente explorada pelo mundo corporativo, capaz de responder a questões envolvendo quem, o que, quando, onde entre outras questões (127).

A proposta de uma metodologia para priorização de novos alvos a partir de consultas feitas no TaP DM tem o objetivo de complementar e facilitar o trabalho do pesquisador na análise de artigos científicos, buscando através das ontologias e do *data mart* enriquecer os resultados e suas análises. Realizar a leitura de artigos científicos é apenas uma das etapas na busca por novos alvos. Procedimentos posteriores como simular redes metabólicas na busca genes essenciais e buscar associações a partir da similaridade na sequência de aminoácidos ou na estrutura 3D das proteínas são importantes na validação.

O principal foco da metodologia para priorização de novos alvos é ampliar o nível e as condições de percepção dos dados contidos em um grande número de artigos. Possibilitar ao pesquisador correlacionar e analisar seus objetos de interesse, presentes em uma fonte tão rica de informação diante de uma visão

dimensional, trás outras perspectivas para sua tomada de decisão nas etapas posteriores de sua pesquisa.

Outro resultado obtido do presente trabalho, é o projeto, construção e disponibilização do TaP DM, testando e validando a metodologia proposta. As consultas iniciais submetidas evidenciaram o quanto o TaP DM é flexível e poderoso para análise de dados não estruturados.

Uma característica da metodologia proposta é a necessidade de criar um novo *data mart* para cada cenário de pesquisa a ser explorado. Isto pode impactar muito no tempo e no esforço de construção dos *data mart*. Estratégias para tornar a metodologia mais genérica podem ser adotadas, diminuindo os impactos nas mudanças de foco das pesquisas.

Nos próximos itens são apresentadas funcionalidades do TaP DM com o objetivo de mostrar que ele consegue obter respostas e estabelecer relações nas quais outras ferramentas não conseguem. No item 4.1 é apresentada as diferenças entre o TaP DM e as interfaces baseadas em palavras chaves. No item 4.2 é apresentado um conjunto de consultas que manipulam dados históricos, ampliando a visão do usuário.

4.1 TaP DM x Palavras chaves

Neste item é apresentado um exemplo de consulta que permite comparar o TaP DM com as interfaces baseadas em palavras chaves. Na Figura 4.1 é apresentada a interface do TaP DM, onde a hierarquia do organismo da espécie *Acari* é explorada. A ferramenta indica que este termo foi citado nos artigos PMC2929727 e PMC2836984, com a presença das técnicas de essencialidade *survival* e *RNA interference*.

Ecdysozoa	DimTechnique	DimArticle	6
Nematoda	DimTechnique	DimArticle	2
Panarthropoda	DimTechnique	DimArticle	4
Arthropoda	DimTechnique	DimArticle	4
Chelicerata	DimTechnique	DimArticle	1
	RNA Interference	DimArticle	1
		PMC2836984	1
		PMC2929727	1
	survival	DimArticle	1
		PMC2836984	1
		PMC2929727	1
Arachnida	DimTechnique	DimArticle	1
Acari	DimTechnique	DimArticle	1
	RNA Interference	DimArticle	1
		PMC2836984	1
		PMC2929727	1
	survival	DimArticle	1
		PMC2836984	1
		PMC2929727	1
Mandibulata	DimTechnique	DimArticle	3
cellular organisms	DimTechnique	DimArticle	20

Figura 4.1: Consulta utilizando o Mondrian detalhando a proteína *Ran.*

Por meio das informações hierárquicas contidas na ontologia NCBI *Taxon* e que foi armazenada na dimensão Organismo, o TaP DM consegue estabelecer a relação entre os conceitos *Chelicerata* e *Acari*. Isto permite que o TaP DM determine que o organismo *Acari* é também um *Chelicerata,* possibilitando determinar que nos artigos PMC2929727 e PMC2836984 também existe a referência ao conceito *Chelicerata* (mesmo que não explicitamente).

Utilizando a interface baseada em palavras chaves do PubMed, duas consultas foram realizadas. A primeira consulta (PMC2836984 *AND Acari AND RNA interfecence*) busca no artigo PMC2836984 o termo *Acari* juntamente com o termo *RNA interference* (Figura 4.2).

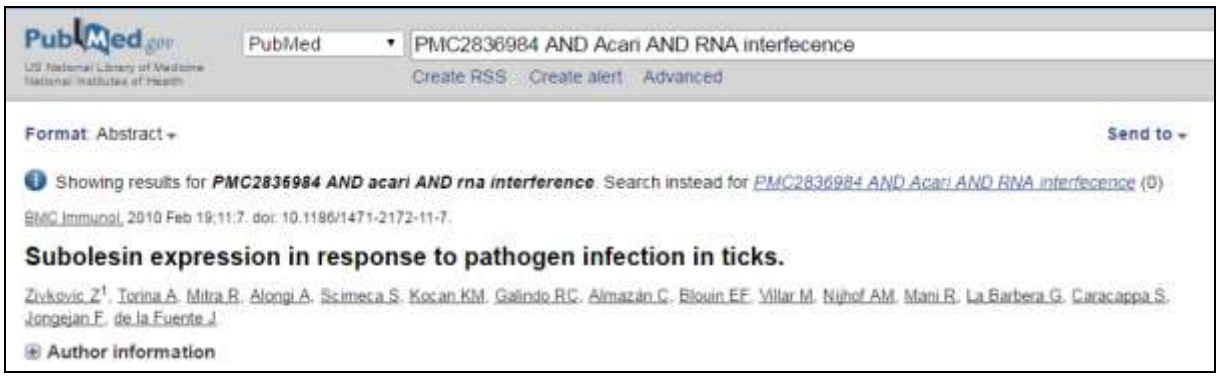


Figura 4.2: Busca realizada no PubMed. Correlação da presença dos termos *Acari* e RNA *interfecence* no artigo PMC2836984.

A Figura 4.2 confirma o que o TaP DM havia mostrado. O termo *Acari* realmente está presente no artigo PMC2836984. A segunda consulta (PMC2836984 *AND Chelicerata AND RNA interfecence*) busca no mesmo artigo o termo *Chelicerata* juntamente com o termo *RNA interfecence* (Figura 4.3).

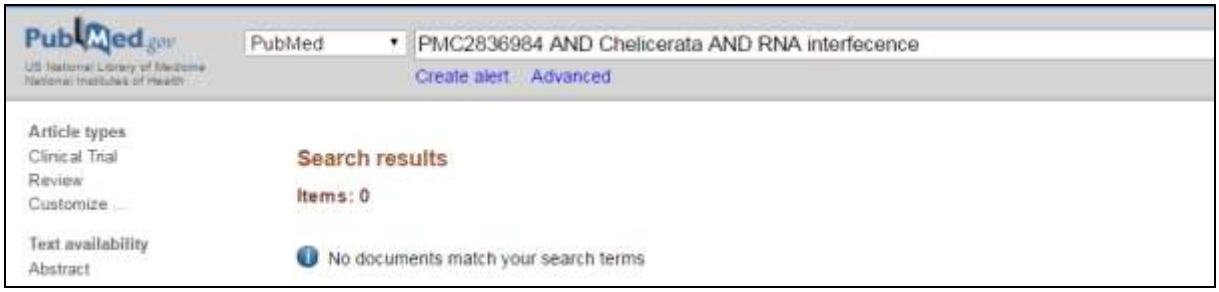


Figura 4.3: Busca realizada no PubMed. Correlação da presença dos termos *Chelicerata* e RNA *interfecence* no artigo PMC2836984.

A Figura 4.3 mostra que o termo *Chelicerata* não está presente no artigo PMC2836984, o que é comprovado abrindo o artigo completo e realizando a busca pelo termo (Figura 4.4).

Figura 4.4: Busca pelo termo *Chelicerata* no artigo completo PMC2836984.

Estas consultas comprovam que os resultados obtidos pelas interfaces baseadas em palavras chaves são muito dependentes dos termos utilizados na construção da busca. Isto porque elas verificam se os termos utilizados na sua construção estão explicitamente presentes nos artigos. Nos exemplos apresentados, esta fragilidade é confirmada. Se o pesquisador utilizar o termo *Chelicerata* ao invés do *Acari,* ele não teria acesso ao artigo PMC2836984. Se este artigo contiver informações importantes, isso poderia comprometer a tomada de decisão na sua pesquisa.

Em contra partida, esta fragilidade é superada pelo TaP DM. Por meio da riqueza de informação contida nas ontologias e que é incorporada pelas dimensões, o TaP DM consegue estabelecer relações entre os termos *Chelicerata e Acari,* deste modo, ele encontra conceitos que não estão explícitos no texto. As buscas baseadas no uso de ontologia são mais flexíveis e poderosas, pois objetivam encontrar conceitos e não simplesmente termos. Por estes motivos, o TaP DM aponta a presença do conceito *Chelicerata* no artigo PMC2836984 (Figura 4.1), pois *Acari* é um *Chelicerata*.

4.2 Análise histórica de dados

Com o objetivo de explorar as características de armazenamento de dados históricos, foi criada uma estratégia para visualizar o comportamento dos dados ao longo do tempo. O exemplo criado tem o objetivo de comparar proteínas utilizando como parâmetro os organismos citados com elas. Neste exemplo foi utilizado organismo, mas pode-se construir a mesma estratégia utilizando quaisquer elementos do TaP DM.

A estratégia é identificar os artigos que citam as proteínas de interesse e dividir estes artigos em faixas de tempo, onde o intervalo de tempo entre as faixas é de no máximo 36 meses.

Para realizar uma análise mais detalhada, toda vez que uma faixa possuir mais de 30 organismos, uma nova faixa é criada, independentemente de ter ou não atingido os 36 meses de intervalo. Com as faixas definidas para cada proteína, os organismos contidos em cada faixa são identificados e comparados com todas as faixas das outras proteínas.

Para exemplificar a construção das faixas, é apresentada a Tabela 4.1, onde uma determinada proteína é citada em 10 artigos entre os anos 2000 e 2015, e o número de organismos em cada artigo é mostrado.

Tabela 4.1: Tabela de exemplo sobre citação de organismos junto com uma determinada proteína.

Ano	Artigo1	Artigo2	Artigo3	Artigo4	Artigo5	Artigo6	Artigo7	Artigo8	Artigo9	Artigo10
2000	20									
2001										
2002		15								
2003										
2004			10							
2005				5						
2006					60					
2007										
2008						12				
2009										
2010										
2011							31			
2012								6		
2013									10	
2014										
2015										2

Executando o algoritmo para o exemplo acima, as faixas ficam distribuídas com apresentado na Tabela 4.2:

Tabela 4.2: Definição das faixas e quantidade de organismos em cada faixa.

Ano	faixa0	faixa1	faixa2	faixa3	faixa4	faixa5
2000- 2002	35					
2004-2006		75				
2008-2010			12			
2011				31		
2012-2014					16	
2015						2

Depois de identificados os organismos de cada faixa, eles são comparados criando uma taxa de relação. Essa taxa é calculada por meio da comparação dos organismos citados em comum de cada proteína. Comparando 2 proteínas P1 e P2, a taxa de relação é calculada entre $R_{P1->P2}$ e $R_{P2->P1}$ entre todas as faixas. Deste modo, tem-se uma visão histórica dos organismos estudados com as proteínas de interesse.

Esta estratégia foi adotada para realizar 3 comparações entre proteínas (i) FGF8_HUMAN x WNT1_HUMAN, (ii) NMT1 x Wee1 e (iii) *Glyceraldehyde* 3-*phosphate dehydrogenase, liver* x BAX_HUMAN. A seguir, os resultados são apresentados graficamente, juntamente com as discussões.

i. WNT1_HUMAN x FGF8_HUMAN

A Figura 4.5 mostra que a relação entre as proteínas WNT1_HUMAN e FGF8_HUMAN ($R_{(WNT1\_HUMAN \rightarrow FGF8\_HUMAN)}$) é mais próxima (possui mais organismos em comum) se comparada com a relação entre FGF8_HUMAN e WNT1_HUMAN ($R_{(FGF8\_HUMAN \rightarrow WNT1\_HUMAN)}$).

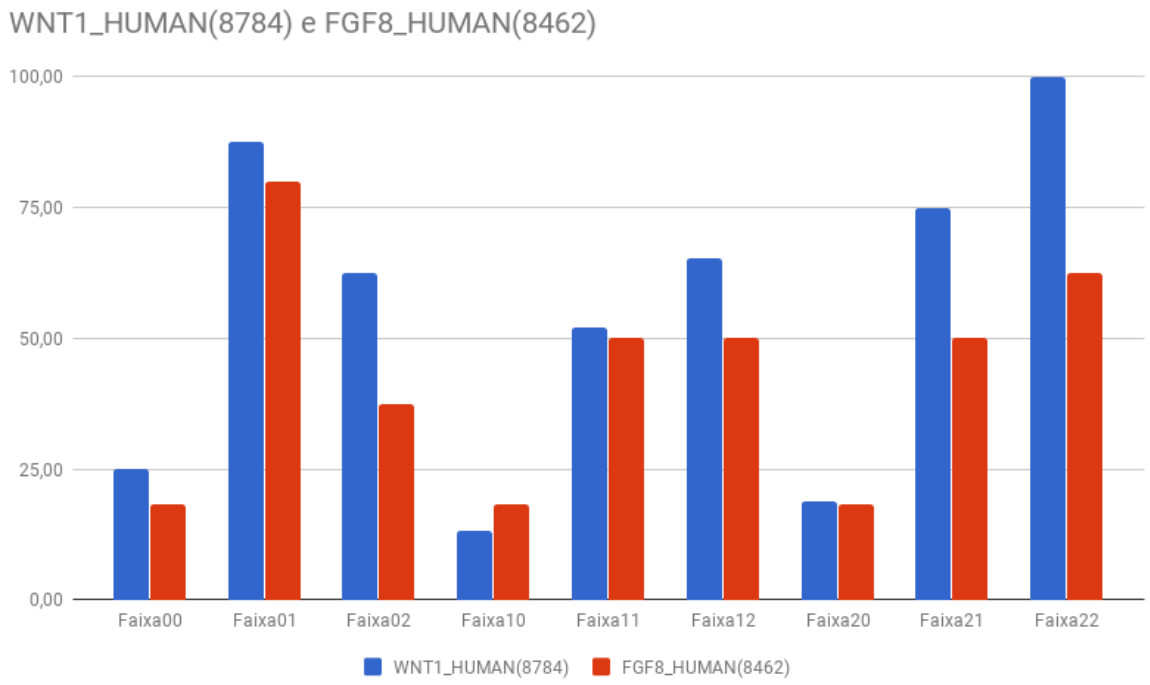


Figura 4.5: Comparação do histórico de citação de organismos entre as proteínas FGF8_HUMAN e WNT1_HUMAN.

Analisando os primeiros artigos publicados (Faixa00), nota-se poucos organismos em comum. Os valores de $R_{(WNT1\_HUMAN \rightarrow FGF8\_HUMAN)}$ = 25% e $R_{(FGF8\_HUMAN \rightarrow WNT1\_HUMAN)}$ = 18% demonstram isso. Avaliando as relações na Faixa01, os organismos em comum atingem mais de 80%.

Os valores das relações entre as faixas subsequentes variam de 13,04% a 65,22%, mostrando pouca intersecção de organismos. Analisando os artigos mais recentes (Faixa22), o valor de $R_{(WNT1\_HUMAN \rightarrow FGF8\_HUMAN)}$ atinge 100%. Uma explicação para este fato seria que toda vez que a proteína WNT1_HUMAN foi citada em um artigo, também foi citada a proteína FGF8_HUMAN.

A comprovação desta hipótese é apresentada na Tabela 4.3, onde por meio de consultas realizadas no TaP DM, foi capaz de identificar em quais artigos as 2 proteínas foram citadas conjuntamente ou não. Como esperado, nos artigos mais

recentes, a proteína WNT1_HUMAN é sempre citada juntamente com a proteína FGF8_HUMAN.

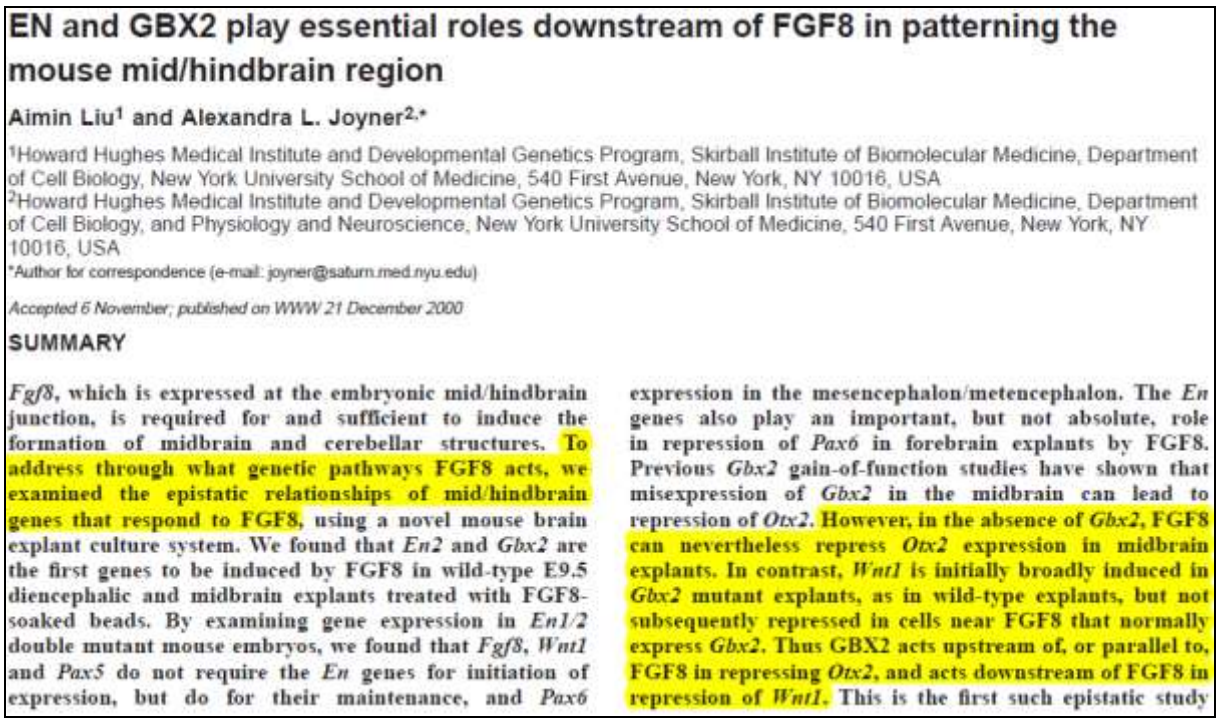
Tabela 4.3: Artigos durante os anos onde são citadas as proteínas FGF8_HUMAN(8784) e WNT1_HUMAN(8462).

		Anos							
		1991	**2005**	**2007**	**2008**	**2010**	**2011**	**2012**	**2013**
Artigos	PMC359739	8784							
	PMC1361118		8462/8784						
	PMC1964797			8462					
	PMC2323570				8784				
	PMC4092012					8784			
	PMC3206183						8462		
	PMC3519931						8462		
	PMC3229493						8462		
	PMC3338518						8462		
	PMC3920936							8462	
	PMC4132851								8462
	PMC3824680								8462/8784
	PMC3698544								8462/8784

A Figura 4.5 mostra também que a recíproca não é verdadeira, ou seja, o valor de $R_{(FGF8\_HUMAN \rightarrow WNT1\_HUMAN)}$ é de 65,50%, comprovando que FGF8_HUMAN foi citada em artigos onde não necessariamente a proteína WNT1_HUMAN esteve presente.

O valor de $R_{(WNT1\_HUMAN \rightarrow FGF8\_HUMAN)}$ igual a 100% poderia representar que a proteína FGF8_HUMAN possui um papel importante nas funções biológicas exercidas pela proteína WNT1_HUMAN, como por exemplo, estarem envolvidas em uma mesma via metabólica, onde uma proteína está diretamente relacionada a função biológica da outra.

A estreita relação biológica entre as proteínas WNT1_HUMAN e FGF8_HUMAN é comprovada por meio do artigo PMID 11124114 (Figura 4.6), que não está presente no *corpus* de artigos.



EN and GBX2 play essential roles downstream of FGF8 in patterning the mouse mid/hindbrain region

Aimin Liu[1] and Alexandra L. Joyner[2,*]

[1]Howard Hughes Medical Institute and Developmental Genetics Program, Skirball Institute of Biomolecular Medicine, Department of Cell Biology, New York University School of Medicine, 540 First Avenue, New York, NY 10016, USA
[2]Howard Hughes Medical Institute and Developmental Genetics Program, Skirball Institute of Biomolecular Medicine, Department of Cell Biology, and Physiology and Neuroscience, New York University School of Medicine, 540 First Avenue, New York, NY 10016, USA
*Author for correspondence (e-mail: joyner@saturn.med.nyu.edu)

Accepted 6 November; published on WWW 21 December 2000

SUMMARY

Fgf8, which is expressed at the embryonic mid/hindbrain junction, is required for and sufficient to induce the formation of midbrain and cerebellar structures. To address through what genetic pathways FGF8 acts, we examined the epistatic relationships of mid/hindbrain genes that respond to FGF8, using a novel mouse brain explant culture system. We found that *En2* and *Gbx2* are the first genes to be induced by FGF8 in wild-type E9.5 diencephalic and midbrain explants treated with FGF8-soaked beads. By examining gene expression in *En1/2* double mutant mouse embryos, we found that *Fgf8*, *Wnt1* and *Pax5* do not require the *En* genes for initiation of expression, but do for their maintenance, and *Pax6* expression in the mesencephalon/metencephalon. The *En* genes also play an important, but not absolute, role in repression of *Pax6* in forebrain explants by FGF8. Previous *Gbx2* gain-of-function studies have shown that misexpression of *Gbx2* in the midbrain can lead to repression of *Otx2*. However, in the absence of *Gbx2*, FGF8 can nevertheless repress *Otx2* expression in midbrain explants. In contrast, *Wnt1* is initially broadly induced in *Gbx2* mutant explants, as in wild-type explants, but not subsequently repressed in cells near FGF8 that normally express *Gbx2*. Thus GBX2 acts upstream of, or parallel to, FGF8 in repressing *Otx2*, and acts downstream of FGF8 in repression of *Wnt1*. This is the first such epistatic study

Figura 4.6: Trecho do artigo PMID 11124114 onde é comprovado a relação biológica entre as proteínas FGF8 e Wnt1.

A descoberta de relações entre proteínas é muito importante no estudo da essencialidade, pois se WNT1_HUMAN é essencial para algum organismo, a proteína FGF8_HUMAN pode também exercer um papel importante nesta função biológica.

ii. NMT1 x Wee1

A Figura 4.7 mostra que nos primeiros artigos (Faixa00) as proteínas possuíam alta intersecção de organismos, mas com o passar dos anos, esta interseção tendenciou a diminuir. Outra informação que a figura mostra é o fato da proteína NMT1 ter sido mais estudada, pois ela possui muito mais faixas que a Wee1, consequentemente foi muito mais citada com diferentes organismos.

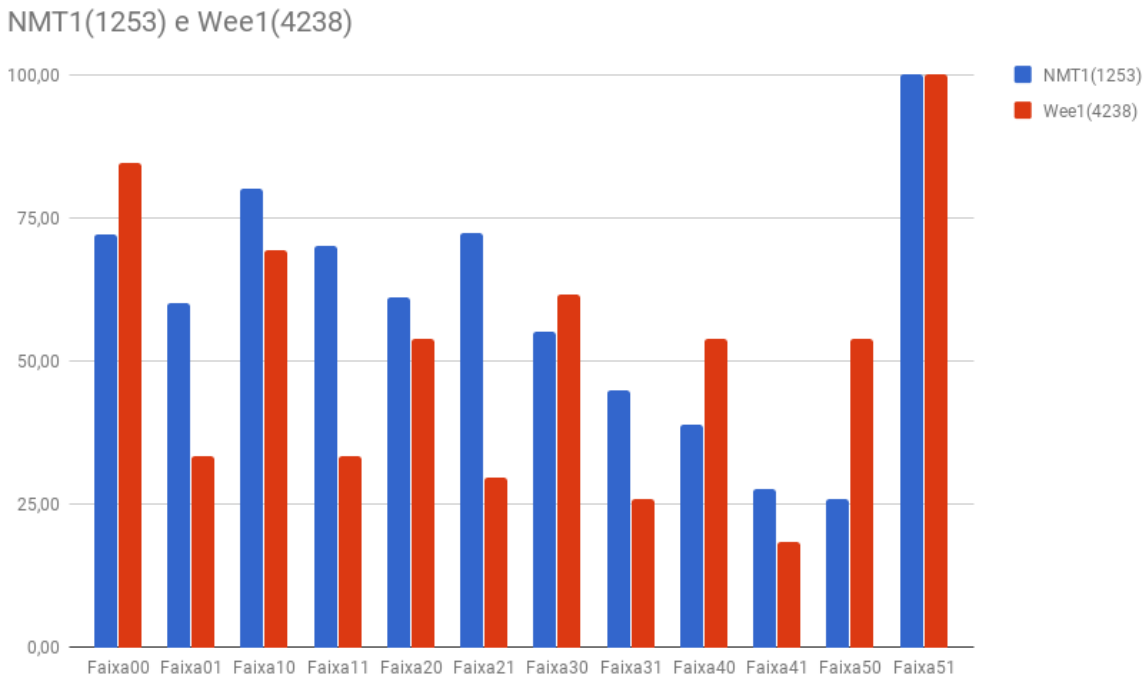


Figura 4.7: Comparação do histórico de citação de organismos entre as proteínas NMT1 e Wee1.

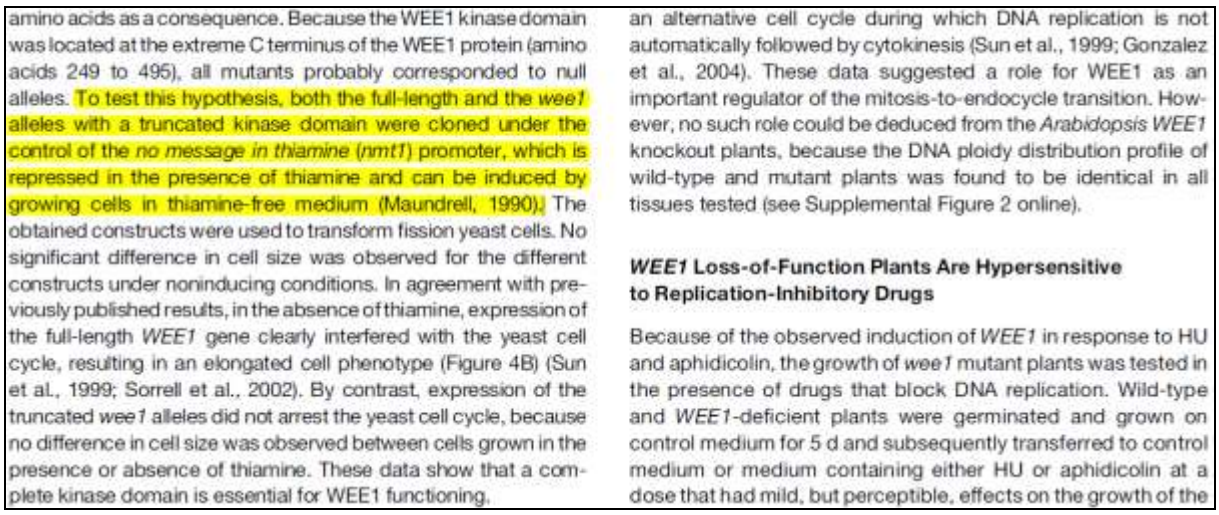
Os artigos mais recentes (Faixa51) mostram que elas citaram exatamente os mesmos organismos ($R_{(NMT1 \rightarrow Wee1)}$ = 100% e $R_{(Wee1 \rightarrow NMT1)}$ = 100%), tendo a possibilidade de estarem presentes sempre nos mesmos artigos.

O fato da proteína Wee1 ser menos estudada que a proteína NMT1 é comprovado por meio da Tabela 4.4, onde uma grande diferença do número de artigos que citam as 2 proteínas é apresentado.

Tabela 4.4: Artigos durante os anos onde são citadas as proteínas NMT1(1253) e Wee1(4238).

		Anos											
		1997	1998	1999	2000	2001	2002	2003	2004	2006	2007	2008	2013
Artigos	PMC1460106	1253											
	PMC2139860	1253											
	PMC317080		4238										
	PMC25529			1253									
	PMC310237				1253								
	PMC1461765					1253							
	PMC102257						1253						
	PMC181578							1253					
	PMC1061621								1253				
	PMC1694798									1253			
	PMC1900001										1253		
	PMC2643609											1253	
	PMC3818516												1253/4238

Este resultado pode mostrar que existe uma relação entre as 2 proteínas que não era conhecida anteriormente e que foi descoberta em artigos mais recentes. Esta relação é comprovada pelo artigo PMC1820959 (Figura 4.8), que não está presente no *corpus* de artigos.



amino acids as a consequence. Because the WEE1 kinase domain was located at the extreme C terminus of the WEE1 protein (amino acids 249 to 495), all mutants probably corresponded to null alleles. To test this hypothesis, both the full-length and the *wee1* alleles with a truncated kinase domain were cloned under the control of the *no message in thiamine* (*nmt1*) promoter, which is repressed in the presence of thiamine and can be induced by growing cells in thiamine-free medium (Maundrell, 1990). The obtained constructs were used to transform fission yeast cells. No significant difference in cell size was observed for the different constructs under noninducing conditions. In agreement with previously published results, in the absence of thiamine, expression of the full-length *WEE1* gene clearly interfered with the yeast cell cycle, resulting in an elongated cell phenotype (Figure 4B) (Sun et al., 1999; Sorrell et al., 2002). By contrast, expression of the truncated *wee1* alleles did not arrest the yeast cell cycle, because no difference in cell size was observed between cells grown in the presence or absence of thiamine. These data show that a complete kinase domain is essential for WEE1 functioning.

an alternative cell cycle during which DNA replication is not automatically followed by cytokinesis (Sun et al., 1999; Gonzalez et al., 2004). These data suggested a role for WEE1 as an important regulator of the mitosis-to-endocycle transition. However, no such role could be deduced from the *Arabidopsis WEE1* knockout plants, because the DNA ploidy distribution profile of wild-type and mutant plants was found to be identical in all tissues tested (see Supplemental Figure 2 online).

***WEE1* Loss-of-Function Plants Are Hypersensitive to Replication-Inhibitory Drugs**

Because of the observed induction of *WEE1* in response to HU and aphidicolin, the growth of *wee1* mutant plants was tested in the presence of drugs that block DNA replication. Wild-type and *WEE1*-deficient plants were germinated and grown on control medium for 5 d and subsequently transferred to control medium or medium containing either HU or aphidicolin at a dose that had mild, but perceptible, effects on the growth of the

Figura 4.8: Trecho do artigo PMC1820959 onde é comprovado a relação entre as proteínas NMT1 e Wee1.

Mesmo tendo poucos artigos citando as 2 proteínas conjuntamente, a Figura 4.7 mostra que nas primeiras faixas, as 2 proteínas possuem alta taxa de organismos citados em comum, indicando que poderiam estar presente nos mesmos organismos. Este resultado é importante, pois consegue estabelecer indícios de ligações entre elementos que não estão explicitamente citados nos artigos, já que até o ano de 2012, não havia a presença das 2 proteínas no mesmo artigo.

Semelhante ao exemplo i, estabelecer relações biológicas entre proteínas pode auxiliar a compreensão de mecanismos de essencialidade em organismos pouco estudados. Como a proteína NMT1 é muito mais estudada que a proteína Wee1, descobrir que Wee1 exerce um papel relevante nas funções biológicas de NMT1 abre caminho para estudar a proteína Wee1 nos mesmos organismos onde NMT1 possui papel essencial.

iii. *Glyceraldehyde* 3-*phosphate dehydrogenase, liver* x BAX_HUMAN

O mesmo estudo foi aplicado às proteínas *Glyceraldehyde* 3-*phosphate dehydrogenase, liver* e BAX_HUMAN. A Figura 4.9 mostra que os organismos citados com as proteínas não possuem grande interseção, chegando a 0% quando se compara as suas últimas faixas.

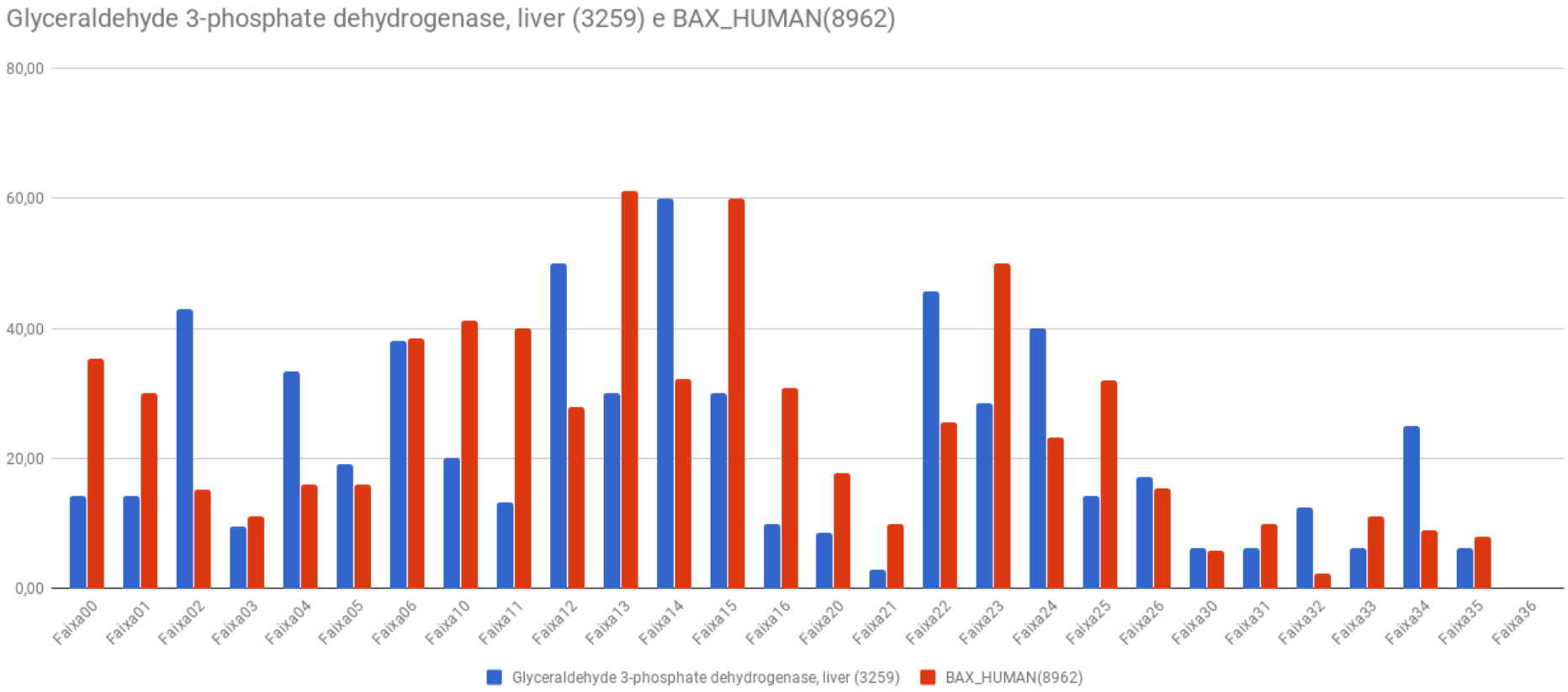


Figura 4.9: Comparação do histórico de citação de organismos entre as proteínas *Glyceraldehyde* 3-*phosphate dehydrogenase, liver* x BAX_HUMAN

A Tabela 4.5 comprova que em 31 artigos entre os anos de 2000 e 2014 as proteínas nunca foram citadas conjuntamente e durante todo esse tempo, as taxas de organismos comuns citados se mantiveram muito baixas, indicando uma baixa probabilidade das 2 proteínas possuírem algum tipo de relação biológica.

Tabela 4.5: Artigos durante os anos onde são citadas as proteínas *Glyceraldehyde* 3-*phosphate dehydrogenase, liver* (3259) e BAX_HUMAN (8962)

		Anos										
		2000	**2002**	**2005**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
Artigos	PMC2169366	8962										
	PMC150003		8962									
	PMC133809		8962									
	PMC1667098			3259								
	PMC1838530				8962							
	PMC2080898				8962							
	PMC2577361					8962						
	PMC2561060					8962						
	PMC2493216					8962						
	PMC3098719					8962						
	PMC2700382						8962					
	PMC2633045						8962					
	PMC2962645							8962				
	PMC2951503							8962				
	PMC2944066							8962				
	PMC3028639								8962			
	PMC3190110								8962			
	PMC3181418								8962			
	PMC3153934								8962			
	PMC3149646								3259			
	PMC3147419								3259			
	PMC3858082									8962		
	PMC3267717									8962		
	PMC3465800									3259		
	PMC3438091									3259		
	PMC3402961									3259		
	PMC3920936										8962	
	PMC3535848										3259	
	PMC4249318											8962
	PMC4210727											8962
	PMC3914405											8962

Não foram encontrado artigos estabelecendo vínculos entre as 2 proteínas. É importante salientar que os resultados apresentados pelo TaP DM não são capazes de determinar uma relação biológica entre as proteínas. A ferramenta se baseia apenas na presença de conceitos nos artigos para estabelecer relações, mas podem ser importantes indícios, principalmente na análise de grande quantidade de artigos.

5 CONCLUSÕES

Pesquisas na literatura revelaram muitos trabalhos que buscam, através de dados não estruturados, extrair informações e correlações ocultas úteis para os usuários. No entanto, nenhuma dessas abordagens foi capaz de promover soluções flexíveis ao ponto de analisar elementos de interesse em diferentes cenários e perspectivas ao longo do tempo.

Este trabalho propôs o desenvolvimento de um método de construção de *data marts* baseada em ontologias denominada TOETL (*Text and Ontology ETL*). Esta metodologia permite analisar dados não estruturados de forma analítica, possibilitando cruzar dados de diferentes conceitos (domínios de conhecimento) de interesse do pesquisador. A TOETL foi apresentada de forma sistemática e dividida em etapas, possibilitando a sua replicação e aplicação em qualquer domínio de conhecimento.

Com o objetivo de analisar os resultados da metodologia proposta, a mesma foi aplicada no cenário de essencialidade de genes, com foco na priorização de novos alvos de fármacos. Esta iniciativa teve como resultado a geração de um *data mart* de essencialidade, denominado TaP DM (*Target Prioritization Data Mart*), sendo populado com dados extraídos de artigos científicos e de ontologias que descrevem os conceitos de essencialidade gênica.

Este trabalho também contribuiu propondo um método de priorização de alvos de fármacos para protozoários. Esta metodologia foi baseada nas características do TaP DM, resultante da aplicação da metodologia TOETL, onde a riqueza de informação contida nas ontologias, juntamente com a visão dimensional do *data mart,* promoveram consultas flexíveis de diferentes perspectivas ao longo do tempo.

Desse modo, os usuários (pesquisadores) obtiveram uma visão analítica dos dados, quantificando a presença dos conceitos de interesse nos artigos e não somente a citação explicita de termos. A busca por conceitos utilizando ontologias potencializou os resultados encontrados nos artigos científicos, pois permitiu localizar conceitos que não estão explicitamente presentes, possibilitando aos usuários descobrir relações ocultas entre diversos elementos.

O *data mart* resultante da aplicação da metodologia permitiu aos usuários definir e aplicar estratégias de busca. Essas estratégias permitiram ao usuário cruzar informações de qualquer conceito presente no *data mart*, assim, possibilitando

estudar o comportamento de muitos elementos e de suas relações ao longo do tempo, auxiliando o pesquisador a priorizar e focar suas atenções em um conjunto menor de objetos de interesse.

Portanto, a metodologia proposta mostrou-se útil e seus resultados coerentes, permitindo aos usuários definirem estratégias de busca por novos alvos de fármacos. Ficou claro também, neste trabalho que a metodologia pode ser aplicada nas mais diversas áreas de conhecimento, onde a análise de grande quantidade de dados não estruturados se faz necessária.

5.1 Principais contribuições

- Concepção de um projeto genérico de um *data mart* baseado em ontologias;
- Implantação do TaP DM com foco em 5 proteozoários: *Entamoeba histolytica*, *Leishmania major*, *Plasmodium falciparum*, *Trypanosoma brucei* e *Trypanosoma cruzi*.
- Integração de uma ferramenta OLAP ao TaP DM para facilitar o acesso e a realização de consultas;
- Definição de uma metodologia para priorização de novos alvos de fármacos;
- Priorização de 57 proteínas para o organismo *T. brucei*;
- Sistematização da metodologia TOETL como uma estratégia genérica, permitindo sua aplicação em outros domínios da pesquisa científica;

5.2 Trabalhos futuros

Na intenção de continuidade, para trabalhos futuros, em curto prazo, artigos de outros repositórios serão armazenados no *data mart*. As etapas e passos serão refinados e otimizados, assim como, estratégias alternativas de priorização de novos alvos serão criadas e sistematizadas.

Na criação do TaP DM, a dimensão tempo foi simplificada pois a aplicação não exigia um maior grau de detalhamento. Isto pode ser um aspecto de melhoria para os próximos trabalhos.

Já trabalhos futuros, em longo prazo, discutirão em maiores detalhes o corte das ontologias, buscando um alinhamento com pesquisas que discutam sua modularização, assim como, aplicar a metodologia TOETL em outros campos de

pesquisa cientifica, permitindo a integração de *data marts* de diferentes áreas de conhecimento, objetivando a construção de um *data warehouse*.

6 REFERÊNCIAS BIBLIOGRÁFICAS

1. Rigden DJ, Fernández-Suárez XM, Galperin MY. The 2016 database issue of nucleic acids research and an updated molecular biology database collection. Nucleic Acids Res. 2016;44(D1):D1–6.

2. Belloze KT. Priorização de alvos para fármacos no combate a doenças tropicais negligenciadas causadas por protozoários. 2013.

3. Provost F, Fawcett T. Data Science and Its Relationship to Big Data and Data-Driven Decision Making Data Science and its relationship to Big Data and data-driven decision making. Data Sci Big Data. 2017;(March 2013).

4. Marx V. Biology: The big challenges of big data. N ature. 2013;498:255–60.

5. Magariños MP, Carmona SJ, Crowther GJ, Ralph SA, Roos DS, Shanmugam D, et al. TDR targets: A chemogenomics resource for neglected diseases. Nucleic Acids Res. 2012;40(D1):1118–27.

6. Feasey N, Wansbrough-Jones M, Mabey DCW, Solomon AW. Neglected tropical diseases. Br Med Bull [Internet]. 2010;93(1):179–200. Available from: http://bmb.oxfordjournals.org/cgi/doi/10.1093/bmb/ldp046

7. PubMed.gov: US National Library of Medicine National Institutes of Health Search databaseSearch term Search [Internet]. 2016. Available from: https://www.ncbi.nlm.nih.gov/pubmed

8. Soldatos TG, Donoghue SIO, Satagopam VP, Barbosa-silva A. Caipirini : using gene sets to rank literature. BioData Min [Internet]. BioMed Central Ltd; 2012;5(1):1. Available from: http://www.biodatamining.org/content/5/1/1

9. Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. Int J Hum Comput Stud [Internet]. 1995;43(5–6):907–28. Available from: http://www.sciencedirect.com/science/article/pii/S1071581985710816

10. Gruber TR. A translation approach to portable ontology specifications. Knowl Acquis [Internet]. 1993;5(2):199–220. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.7493

11. Obitko M, Snášel V, Smid J. Ontology design with formal concept analysis. Proc Int Work Concept Lattices their Appl (CLA 2004) [Internet]. 2004;111–9. Available from: http://ceur-ws.org/Vol-110/paper12.pdf

12. Elia MAD, Pereira MP, Brown ED. Are essential genes really essential ? 2009;433–8.

13. Bergmiller T, Ackermann M, Silander OK. Patterns of evolutionary conservation of essential genes correlate with their compensability. PLoS Genet. 2012;8(6).

14. Bose R, Sugurnaran V. Application of Intelligent Agent Technology for Managerial Data Analysis and Mining. 1999;30(1):77–94.

15. Durgin JK, Sherif JS. The semantic web: a catalyst for future e-business. Kybernetes. 2008;37(1–2):49–65.

16. Chaudhuri S, Dayal U, Hamel L, Hall T, Negash S, Sauter VL, et al. Building the Data Warehouse, Fourth Edition [Internet]. The Encyclopedia of Data Warehousing and Mining. 2005. 543 p. Available from: http://190112.8m.com/Bibliografia.pdf\nhttp://doi.wiley.com/10.1002/9780470634431\nhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.9494&rep=rep1&type=pdf

17. Belloze KT, Monteiro DISB, Lima TF, Silva FP, Cavalcanti MC. An evaluation of annotation tools for biomedical texts. CEUR Workshop Proc. 2012;938:108–19.

18. Gomes PC e C, Moura AM de C, Cavalcanti MC. A multi-ontology approach to annotate scientific documents based on a modularization technique. J Biomed Inform. 2015;58:208–19.

19. Jonquet C, Lependu P, Falconer S, Coulet A, Noy NF, Musen MA, et al. NCBO Resource Index: Ontology-based search and mining of biomedical resources. J Web Semant. 2011;9(3):316–24.

20. Genesereth MR, Nilsson NJ. Logical Foundations of Artificial Intelligence [Internet]. The Journal of Symbolic Logic. 1987. 405 p. Available from: http://www.loc.gov/catdir/description/els032/87005461.html

21. Guarino N, Oberle D, Staab S. What Is an Ontology? Handb Ontol [Internet]. 2009;1–17. Available from: http://link.springer.com/10.1007/978-3-540-92673-3

22. McGuinness D, Harmelen F Van. OWL web ontology language overview [Internet]. W3C recommendation. 2004. p. 1–22. Available from: http://www.academia.edu/download/30759881/5.3-B1.pdf

23. World Wide Web Consortium (W3C). RDF [Internet]. 2014 [cited 2017 Sep 5]. p. 3–5. Available from: https://www.w3.org/2001/sw/wiki/RDF

24. Bechhofer S, Miles A. Using OWL and SKOS [Internet]. W3C Recommendation 18 August 2009. 2008. p. 1–40. Available from: http://www.w3.org/TR/2009/REC-skos-reference-20090818/#concepts\npapers2://publication/uuid/67B635A0-CD7E-48A0-8730-EC86DB6CC3A1

25. Library USN. UMLS Quick Start Guide [Internet]. 2016. p. 1–3. Available from: https://www.nlm.nih.gov/research/umls/quickstart.html

26. Shim JP, Warkentin M, Courtney JF, Power DJ, Sharda R, Carlsson C. Past, present, and future of decision support technology. Decis Support Syst. 2002;33(2):111–26.

27. Tripathi KP. Decision Support System Is a Tool for Making Better Decisions in the Organization. Indian J Comput Sci Eng [Internet]. 2011;2(1):112–7. Available from: http://www.ijcse.com/docs/IJCSE11-02-01-054.pdf

28. Kimball R, Ross M. The Data Warehouse Toolkit, 3rd Edition Dimentional

Modeling. 3rd ed. Indianapolis: Jonh Wiley & Sons, Inc.; 2013. 147 p.

29. Silberschatz A, Korth HF, Sudarshan S. Database System Concepts - 6th. ed. Database. 2011. 1376 p.

30. Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology. ACM SIGMOD Rec. 1997;26(1):65–74.

31. Sathe G, Sarawagi S. Intelligent Rollups in Multidimensional OLAP Data. Vldb [Internet]. 2001;531–40. Available from: http://citeseer.ist.psu.edu/453749.html

32. Patil MR, Thia F. Pentaho for Big Data Analytics. 2013. 118 p.

33. Casters M, Bouman R, Dongen J van. Pentaho Kettle Solutions: Building Open Source ETL Solution with Pentaho Data Integration. Wiley Publishing, Inc; 2010. 722 p.

34. Sarawagi S, Agrawal R, Megiddo N, Univ Politecn Valencia GVAV, Edbt Fdn ETHZOSSI. Discovery-driven exploration of OLAP data cubes. 6th Int Conf Extending Database Technol (EDBT 98) [Internet]. 1998;168–82. Available from: <Go to ISI>://000078840500012

35. Mannino M V. Projeto, Desenvolvimento de Aplicações e Administração de Banco de Dados [Internet]. 3rd ed. AMGH Editora, editor. 2008. 717 p. Available from: http://www.americanbanker.com/issues/179_124/which-city-is-the-next-big-fintech-hub-new-york-stakes-its-claim-1068345-1.html

36. Elmasri R, Navathe SB. Sistemas de Banco de Dados. 4th ed. Pearson Education; 2004.

37. Moody D, Kortink MA. From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design. Proc Int Work Des Manag Data Warehouses. 2000;2000:5–16.

38. Kenan, Corporation S. An Introduction To Multidimensional Database Technology. Data Warehous. 1993;29.

39. Romero O, Abelló A. Automating multidimensional design from ontologies. Proc ACM tenth Int Work Data Warehous Ol - Dol ’07 [Internet]. 2007;1–8. Available from: http://portal.acm.org/citation.cfm?doid=1317331.1317333

40. Group Kimball. Kimball Dimensional Modeling Techniques. 2013;1–24.

41. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ [Internet]. Elsevier B.V.; 2016;47:20–33. Available from: http://dx.doi.org/10.1016/j.jhealeco.2016.01.012

42. Getz K a, Wenger J, Campo R a, Seguine ES, Kaitin KI. Assessing the impact of protocol design changes on clinical trial performance. Am J Ther [Internet]. 2008;15(5):450–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18806521

43. Souza W de. DOENÇAS NEGLIGENCIADAS [Internet]. Academia Brasileira

de Ciências. 2015. 1551-1564 p. Available from: https://www.abc.org.br/IMG/pdf/doc-199.pdf

44. Ehrenberg JP, Ault SK. the determinants of health in Latin America and the Caribbean. 2005;(April 2014).

45. Commission of Macroeconomics and Health. Macroeconomics and health: investing in health for economic development. Rev Panam Salud Pública. 2002;12(December):143–4.

46. Moran M, Guzman J, Ropars AL, McDonald A, Jameson N, Omune B, et al. Neglected disease research and development: How much are we really spending? PLoS Med. 2009;6(2):0137–46.

47. Chirac P, Torreele E. Global framework on essential health R&D. Lancet. 2006;367(9522):1560–1.

48. De S, Insumos T. Doenças negligenciadas: estratégias do Ministério da Saúde. Rev Saude Publica. 2010;44(1):200–2.

49. Agüero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, Campbell RK, et al. Genomic-scale prioritization of drug targets: the TDR Targets database. Nat Rev Drug Discov [Internet]. 2008;7(11):900–7. Available from: http://www.nature.com/doifinder/10.1038/nrd2684

50. The TDR Drug Targets Network. The TDR Targets Database [Internet]. A chemogenomics resource for neglected tropical diseases. 2015 [cited 2017 Aug 6]. p. 1–5. Available from: http://www.tdrtargets.org

51. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res [Internet]. 2004;32(Database issue):D115-9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308865&tool=pmcentrez&rendertype=abstract

52. UniProt Consortium. UniProt [Internet]. 2017 [cited 2017 Dec 6]. Available from: http://www.uniprot.org/

53. Riddle DLB. What are Essential Genes ? Cold Spring Harb. 1997;3–4.

54. Jordan IK, Rogozin IB, Wolf YI, Koonin E V. Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. Genome Res. 2002;962–8.

55. Dickerson JE, Zhu A, Robertson DL, Hentges KE. Defining the role of essential genes in human disease. PLoS One. 2011;6(11).

56. Koonin E V. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol [Internet]. 2003;1(2):127–36. Available from: http://www.nature.com/doifinder/10.1038/nrmicro751

57. Gustafson AM, Snitkin ES, Parker SCJ, DeLisi C, Kasif S. Towards the identification of essential genes using targeted genome sequencing and

comparative analysis. BMC Genomics [Internet]. 2006;7:265. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1624830&tool=pmcentrez&rendertype=abstract

58. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature [Internet]. 2002;418(6896):387–91. Available from: http://www.nature.com/doifinder/10.1038/nature00935

59. Cullen LM, Arndt GM. Genome-wide screening for gene function using RNAi in mammalian cells. Immunol Cell Biol. 2005;83(3):217–23.

60. Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, et al. Essential Bacillus subtilis genes. Proc Natl Acad Sci [Internet]. 2003;100(8):4678–83. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0730515100

61. Zhang R, Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. 2009;37(SUPPL. 1):455–8.

62. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, et al. Essential genes of a minimal bacterium. Proc Natl Acad Sci [Internet]. 2006;103(2):425–30. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0510013103

63. Pender M, Wotherspoon L, Sa'Don NM, Orense R. Macro Element for Pile Head Cyclic Lateral Loading. Geotech Geol Earthq Eng. 2012;16:129–45.

64. Biology C, Medical O, City O. Large-Scale Screening for Targeted Knockouts in the *Caenorhabditis elegans* Genome. G3: Genes|Genomes|Genetics [Internet]. 2012;2(11):1415–25. Available from: http://g3journal.org/lookup/doi/10.1534/g3.112.003830

65. Hegemann JH, Güldener U, Köhler GJ. Gene disruption in the budding yeast Saccharomyces cerevisiae. Methods Mol Biol. 2006;313:129–44.

66. Meinke D, Muralla R, Sweeney C, Dickerman A. Identifying essential genes in Arabidopsis thaliana. Trends Plant Sci. 2008;13(9):483–91.

67. Rine J. A future of the model organism model. Mol Biol Cell [Internet]. 2014;25(5):549–53. Available from: http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E12-10-0768

68. Gritsenko VA, Bukharin O V. The ecological and medical aspects of the symbiosis between Escherichia coli and man. J Microbiol Epidemiol an Immunobiol. 2000;(3):92–9.

69. Vogt RL, Dippold L. Escherichia coli O157:H7 outbreak associated with consumption of ground beef, June-July 2002. Public Health Rep. 2005;120(2):174–8.

70. Gerdes S, Scholle MD, Campbell JW, Balázsi G, Ravasz E, Daugherty MD, et al. Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol [Internet]. 2003;185(19):5673–84.

Available from: http://jb.asm.org/cgi/content/full/185/19/5673?view=long&pmid=13129938

71. Giaever G, Nislow C. The yeast deletion collection: A decade of functional genomics. Genetics. 2014;197(2):451–65.

72. Walker LJ, Aldhous MC, Drummond HE, Smith BRK, Nimmo ER, Arnott IDR, et al. Anti-Saccharomyces cerevisiae antibodies (ASCA) in Crohn's disease are associated with disease severity but not NOD2/CARD15 mutations. Clin Exp Immunol. 2004;135(3):490–6.

73. Heinicke S, Livstone MS, Lu C, Oughtred R, Kang F, Angiuoli S V., et al. The Princeton Protein Orthology Database (P-POD): A comparative genomics analysis tool for biologists. PLoS One. 2007;2(8).

74. Stinchcombe JR, Caicedo AL, Hopkins R, Mays C, Boyd EW, Purugganan MD, et al. Vernalization sensitivity in Arabidopsis thaliana (Brassicaceae): The effects of latitude and FLC variation. Am J Bot. 2005;92(10):1701–7.

75. Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, et al. Evolution of genome size in Brassicaceae. Ann Bot. 2005;95(1):229–35.

76. Lloyd J, Meinke D. A Comprehensive Dataset of Genes with a Loss-of-Function Mutant Phenotype in Arabidopsis. Plant Physiol [Internet]. 2012;158(3):1115–29. Available from: http://www.plantphysiol.org/cgi/doi/10.1104/pp.111.192393

77. Brenner S. The genetics of Caenorhabditis elegans. Genetics. 1974;77(1):71–94.

78. Kempheus K. Essential genes. WormBook [Internet]. 2005;1–7. Available from: http://www.wormbook.org/chapters/www_essentialgenes/essentialgenes.html

79. Ashburner M, Bergman CM. Drosophila melanogaster : A case study of a model genomic sequence and its consequences. Cold Spring Harb Perspect Biol. 2005;15:1661–7.

80. St Johnston D. The Art and Design of Genetic Screens: Drosophila Melanogaster. Nat Rev Genet [Internet]. 2002;3(3):176–88. Available from: http://www.nature.com/doifinder/10.1038/nrg751

81. Howe DG, Bradford YM, Eagle A, Fashena D, Frazer K, Kalita P, et al. The Zebrafish Model Organism Database: New support for human disease models, mutation details, gene expression phenotypes and searching. Nucleic Acids Res. 2017;45(D1):D758–68.

82. Howe K, Clark M, Torroja C, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature [Internet]. 2013;496(7446):498–503. Available from: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature12111.html?utm_source=feedly

83. Gu JL, Gu JL, Receive R, Receive R, Research G, Research G, et al. The mouse genome. Genome Res. 2005;1729–40.

84. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature [Internet]. 2002;420(6915):520–62. Available from: http://www.nature.com/doifinder/10.1038/nature01262\n<Go to ISI>://WOS:000179611600053\nhttp://www.nature.com/nature/journal/v420/n6915/pdf/nature01262.pdf

85. Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, Rubin E, et al. Strategies and Tools for Whole-Genome Alignments. 2003;73–80.

86. Ximénez C, Morán P, Rojas L, Valadez A, Gómez A, Ramiro M, et al. Novelties on amoebiasis: a neglected tropical disease. J Glob Infect Dis [Internet]. 2011;3(2):166–74. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125031&tool=pmcentrez&rendertype=abstract

87. Berglund B, Dienus O, Sokolova E, Berglind E, Matussek A, Pettersson T, et al. Occurrence and removal efficiency of parasitic protozoa in Swedish wastewater treatment plants. Sci Total Environ [Internet]. Elsevier B.V.; 2017;598:821–7. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0048969717308434

88. Petri WA, Haque R, Lyerly D, Vines RR. Estimating the impact of amebiasis on health. Parasitol Today. 2000;16(8):320–1.

89. Haque R. Human intestinal parasites. J Heal Popul Nutr. 2007;25(4):387–91.

90. Alvar J, Vélez ID, Bern C, Herrero M, Desjeux P, Cano J, et al. Leishmaniasis worldwide and global estimates of its incidence. PLoS One. 2012;7(5).

91. Leite NR. Estudos moleculares de duas triptofanil tRNA sintetases do parasita Leishmania major e de uma cisteíno protease da bactéria Xylella fastidiosa . Universidade de Sào Paulo; 2007.

92. World Health Organization. Control of the leishmaniases. World Health Organization technical report series. 2010.

93. Gontijo CMF, Melo MN. Leishmaniose visceral no Brasil: quadro atual, desafios e perspectivas. Rev Bras Epidemiol [Internet]. 2004;7(3):338–49. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-790X2004000300011&lng=pt&nrm=iso&tlng=en

94. Ministério da Saúde. Guia de Vigilância Epidemiológica [Internet]. 7th ed. Série A. Normas e Manuais Técnicos. 2009. 819 p. Available from: http://bvsms.saude.gov.br/bvs/publicacoes/guia_vigilancia_epidemiologica_7ed.pdf

95. Monzote L. Current treatment of leishmaniasis: a review. Open Antimicrob Agents J [Internet]. 2009;1(August):9–19. Available from: http://ftp.benthamscience.com/open/toantimj/articles/V001/9TOANTIMJ.pdf

96. World Health Organization. Chagas disease (American trypanosomiasis) [Internet]. 2017. p. 11–4. Available from: http://www.who.int/mediacentre/factsheets/fs340/en/

97. Kasprzyk a, Keefe D, Smedley D, London D, Spooner W, Melsopp C, et al. Ingenuity Pathway Analysis Tool\rEnsMart: a generic system for fast and flexible access to biological data. Genome Res [Internet]. 2004;14:160–9. Available from: http://www.ingenuity.com/

98. Center for Disease Control and Prevention. DPDx - Laboratory Identification of Parasitic Diseases of Public Health Concern [Internet]. 2017. p. 7–9. Available from: https://www.cdc.gov/dpdx/

99. Simarro PP, Cecchi G, Franco JR, Paone M, Fèvre EM, Diarra A, et al. Risk for human African trypanosomiasis, Central Africa, 2000-2009. Emerg Infect Dis [Internet]. 2011;17(12):2322–4. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3311175&tool=pmcentrez&rendertype=abstract

100. Alsford S, Kelly JM, Baker N, Horn D. Genetic dissection of drug resistance in trypanosomes. Parasitology [Internet]. 2013;140(12):1478–91. Available from: http://www.journals.cambridge.org/abstract_S003118201300022X

101. World Health Organization. World Health Statistics 2010. World Health [Internet]. 2010;1:177. Available from: http://www.who.int/gho/publications/world_health_statistics/EN_WHS10_Full.pdf

102. Coura JR. Chagas disease: what is known and what is needed--a background article. Mem Inst Oswaldo Cruz. 2007;102 Suppl(August):113–22.

103. Gürtler RE, Ceballos LA, Ordóñez-Krasnowski P, Lanati LA, Stariolo R, Kitron U. Strong Host-Feeding Preferences of the Vector Triatoma infestans Modified by Vector Density: Implications for the Epidemiology of Chagas Disease. PLoS Negl Trop Dis [Internet]. 2009;3(5):e447. Available from: http://dx.plos.org/10.1371/journal.pntd.0000447

104. Neto LL. Caracterização de inibidores de complemento liberados pelas formas metacíclicas de Tripanosoma cruzi e sua função na evasão da imunidade inata. Instituto Oswaldo Cruz; 2013.

105. Villarreal D, Nirde P, Tibayrenc M. Differential Gene Expression in Benznidazole-Resistant. Society. 2005;49(7):2701–9.

106. Choi Y. Making Faceted Classification More Acceptable on the Web : A Comparison of Faceted Classification and Ontologies Problem with Faceted Classification RDF vs . Faceted Classification. In: Proceedings of the Association for Information Science and Technology. 2008. p. 1–6.

107. KWASNIK BH. The Role of Classification in Knowledge Representation and Discovery ’. Libr Trends. 1999;48(1):22–47.

108. Belloze KT, Monteiro DISB, Lima TF, Silva-jr FP, Cavalcanti MC. Analyzing Tools for Biomedical Text Annotation with Multiple Ontologies. :2.

109. Da Silva MAA, Cavalcanti MC, Belloze KT, Silva-Junior F. Agile semantic annotation of scientific texts at the biomedical scenario. Proc - 2014 IEEE 10th Int Conf eScience, eScience 2014. 2014;1:100–7.

110. Silva MAA da, Cavalcanti MC. Combining Ontology Modules for Scientific Text Annotation. J Inf Data [Internet]. 2015;27(1):288–311. Available from: https://seer.ufmg.br/index.php/jidm/article/view/724

111. Sioutos N, Coronado S de, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007;40(1):30–43.

112. Grau BC, Horrocks I, Kazakov Y, Sattler U. Modular reuse of ontologies: Theory and practice. J Artif Intell Res. 2008;31:273–318.

113. Seidenberg J. Web Ontology Segmentation: Extraction, Transformation, Evaluation. In: Modular Ontologies: Concepts, Theories and Techniques for Knowledge Modularization [Internet]. 2009. p. 211–43. Available from: http://www.springer.com/computer/database+management+&+information+retrieval/book/978-3-642-01906-7

114. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing and Management. 1988. p. 513–23.

115. Yamamoto S, Asanuma T, Takagi T, Fukuda KI. The Molecule Role Ontology: An ontology for annotation of signal transduction pathway molecules in the scientific literature. Comp Funct Genomics. 2004;5(6–7):528–36.

116. Federhen S. The NCBI Taxonomy. Nucleic Acids Res [Internet]. 2012;40(D1):D136--D143. Available from: http://www.ncbi.nlm.nih.gov/Taxonomy/

117. Fontes CA. Explorando Inferência em um Sistema de Anotação Semântica. Instituto Militar de Engenharia; 2011.

118. João Luiz Moreira, Kelli de Faria Cordeiro MLMC. JointOLAP – Sistema de Informação para Exploração Conjunta de Dados Estruturados e Textuais : Um estudo de caso no ... 2013;(June).

119. Wongthongtham P, Salih BA. Ontology and Trust based Data Warehouse in New Generation of Business Intelligence. (Idc).

120. Abdullah MF, Ahmad K. Business Intelligence Model for Unstructured Data Management. 2015;473–7.

121. M.Thenmozhi KV. A Tool for Data Warehouse Multidimensional Schema Design using Ontology. IJCSI Int J Comput Sci Issues [Internet]. 2013;10(2):8. Available from: www.IJCSI.org

122. Pardillo J, Mazon JN. Using Ontologies for the Design of Data Warehouses. Int J Database Manag Syst. 2011;3(2):73–87.

123. Gulic M. Transformation of OWL ontology sources into data warehouse. 2013;1143–8.

124. Inmon W.H. KK. Building the Unstructured Data Warehouse: 688 Architecture, Analysis, and Design. Technics Publications; 2011. 3050 p.

125. Guimarães M, Caffarena E, Cruz OG, Silva F. Identifying Drug Repositioning Targets Using Text Mining. In: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (IC3K 2014). 2014. p. 348–53.

126. Srinivasa K, Prasad N. Text Analytics to Data Warehousing. Int J Comput Sci Eng. 2010;2(6):2201–7.

127. Gao L, Chang E, Han S. Powerful Tool to Expand Business Intelligence : Text Mining. Int J Comput Inf Eng. 2007;1(8):2666–71.

128. Senseman B. Parent - Child Hierarchies with Abnormal Genealogy [Internet]. Inquidia Consulting. 2014 [cited 2017 Feb 6]. p. 1–5. Available from: http://inquidia.com/news-and-info/parent-child-hierarchies-abnormal-genealogy

129. Hyde J. How to Design a Mondrian Schema [Internet]. Mondrian Documentation. 2017 [cited 2017 Jun 2]. p. 1–41. Available from: http://mondrian.pentaho.com/documentation/schema.php#Parent_child_hierarchies

130. Chodnicki S. Adventures with Open Source BI Musings of a BI-Developer [Internet]. Analyzing Hierarchical Data Using Bridge Tables. 2010 [cited 2017 Jun 2]. p. 2000–3. Available from: http://geekswithblogs.net/darrengosbell/Default.aspx

APÊNDICE A - DETALHAMENTO DA FERRAMENTA TAP OLAP.

Este documento tem o objetivo de mostrar com maior nível de detalhe a manipulação da ferramenta TaP OLAP e explicar algumas interpretações dos dados obtidos, bem como exemplos de consultas.

1. Acesso à ferramenta

O link do TaP DM está disponível em *http://157.86.114.152:8086/pentaho*. Depois de clicar no *link* "*Pentaho User Console Login*", selecione o usuário "Joe (admin)", a senha será preenchida automaticamente (se isso não acontecer, a senha é "*password*"). Ao clicar em "*Login*" novamente, ele irá redirecioná-lo para outra página, clique no *link* "*New Analysis View*" e as opções de *Schema* e *Cube* serão preenchidas com o valor "TaPDM", basta clicar em OK e, em seguida, o TaP DM pode ser manipulado, como mostra a Figura 6.1.

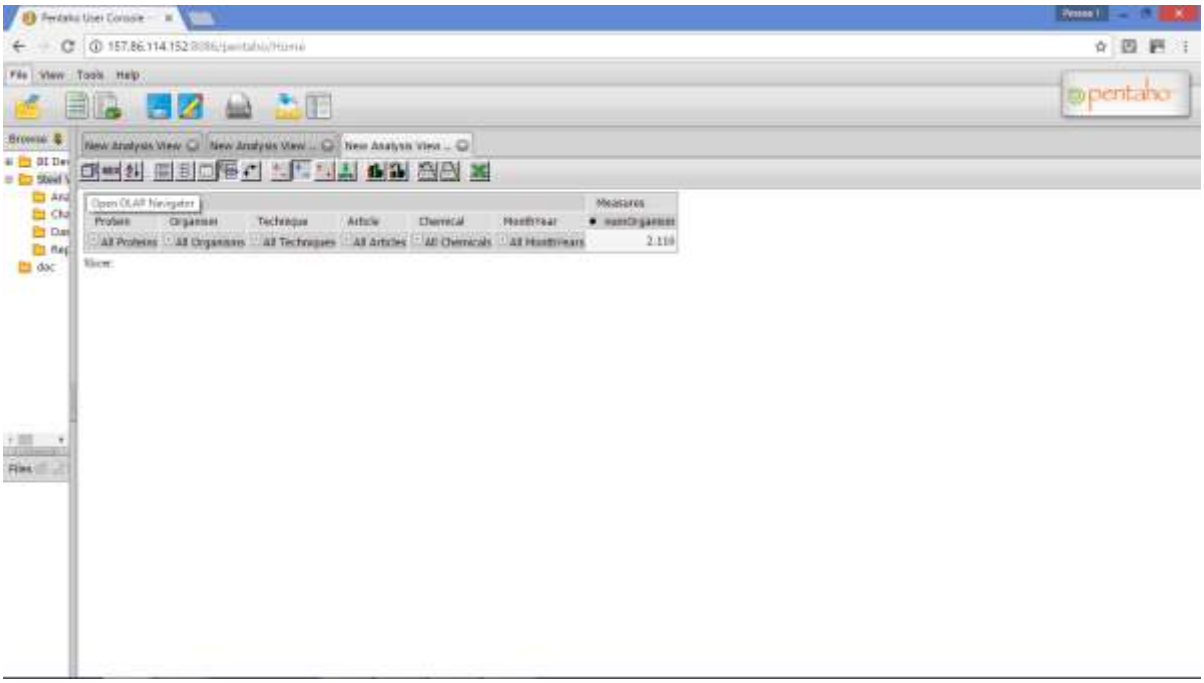


Figura 6.1: Tela inicial do TaP DM.

2. Métricas e dimensões

A Figura 6.1 é a tela inicial da ferramenta. Ao clicar em  (*OPEN OLAP NAVIGATOR*), pode-se selecionar as dimensões e métricas de acordo com suas necessidades. Na Figura 6.2, todas as métricas foram selecionadas. Ela mostra o

total de cada dimensão (conceito) encontrada no *corpus* de artigos. Podemos ver que 2.116 organismos, 4.187 proteínas, 1.376 artigos, 249 produtos químicos, 6 técnicas e 309 meses estão carregados no TaP DM.

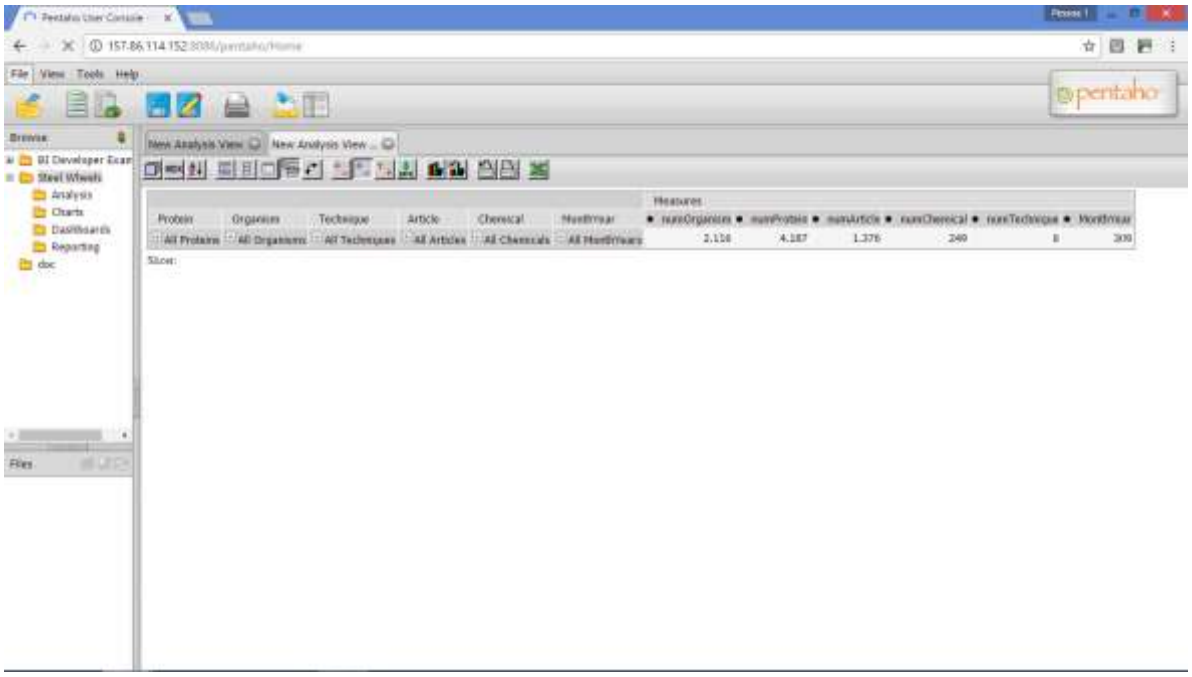


Figura 6.2: Tela com a inserção de todas as métricas.

Se as hierarquias de "*All proteins*" e "*proteins*" forem expandidas (clicando em ⊞), aparece a seguinte tela (Figura 6.3):

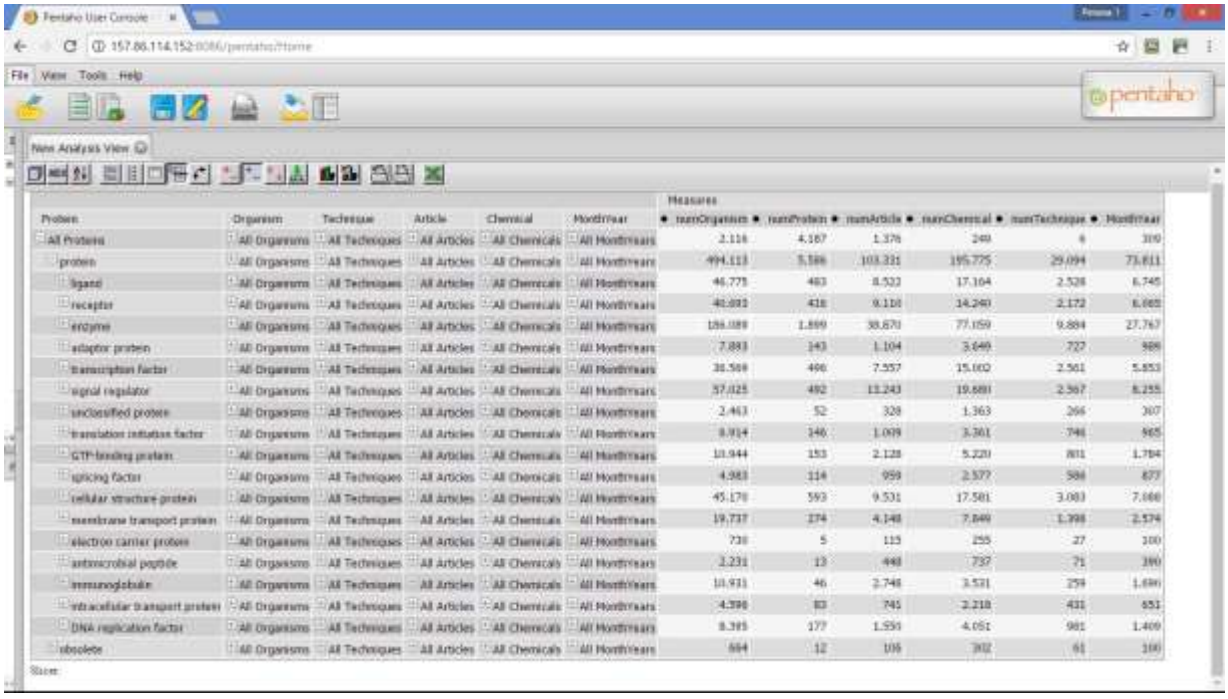


Figura 6.3: Tela mostrando a hierarquia de proteínas.

Analisando os resultados, vemos que, no conjunto de artigos, o conceito de proteína:

- Ocorreu 494.113 vezes junto com algum organismo;
- Ocorreu 5,586 vezes em conjunto com alguma proteína;
- Ocorreu 103.331 vezes no conjunto de artigos;
- Ocorreu 195,775 vezes em conjunto com algum produto químico;
- Ocorreu 29.094 vezes em conjunto com alguma técnica.

Essas estatísticas de citação pertencem ao conceito de proteína. Se desejar visualizar as estatísticas para outras dimensões, basta alterar a ordem das dimensões no botão  (*OPEN OLAP NAVIGATOR*). A próxima Figura 6.4 mostra as estatísticas de citações para a dimensão químico.

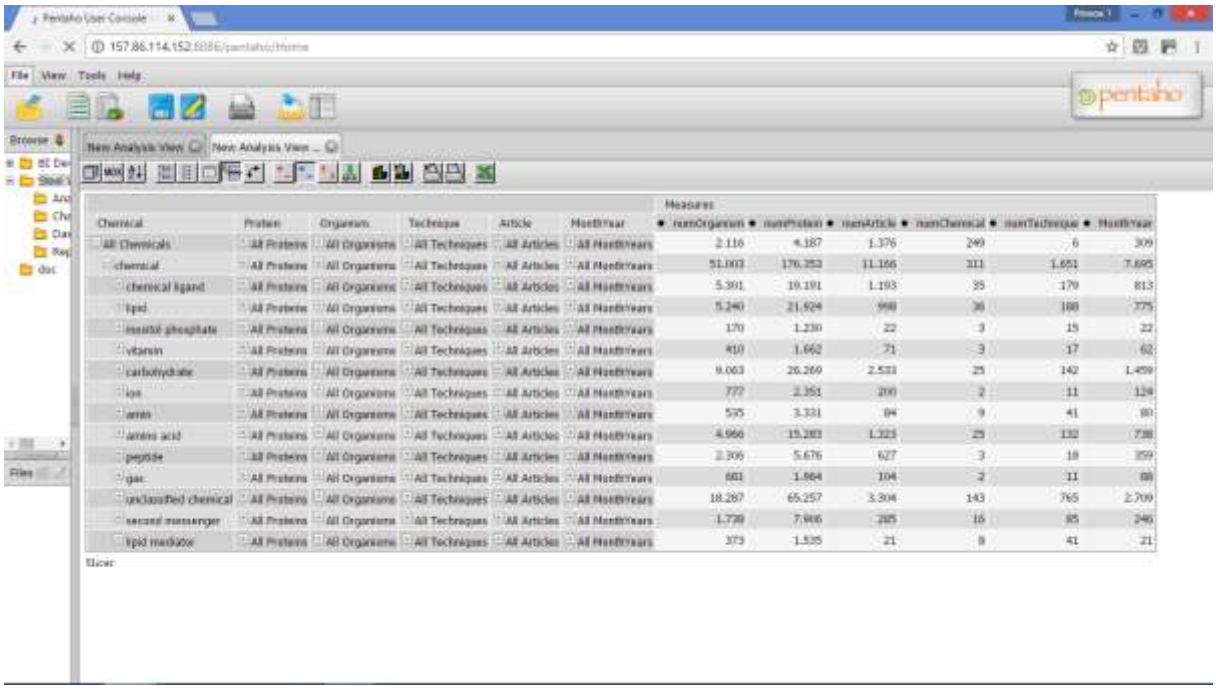


Figura 6.4: Estatísticas para a dimensão químicos.

Analisando as Figura 6.3 e Figura 6.4, observa-se que o conceito de proteína e químico foram citados, respectivamente, 494.113 e 51.003 vezes em conjunto com a dimensão do organismo

Esses valores são a soma do número de vezes que os termos pertencentes as ontologias de proteína (Figura 6.3) e químico (Figura 6.4) apareceram nos artigos com algum termo da ontologia do organismo.

Por exemplo, se um artigo citou os termos *protein*, *enzyme* e *T. cruzi*, o TaP DM registra as combinações de *protein*/*T. cruzi* e *enzyme*/*T. cruzi*. Esta é a razão

pela qual a contagem de termos de proteínas e químicos são tão superiores à quantidade de artigos.

Essa habilidade permite que o DM registre a presença de diferentes subconjuntos de conceitos de uma ontologia dentro do mesmo artigo. Olhando para a Figura 6.3, vemos que o número de conceitos de *proteins* que aparecem com organismos é de 494.113 e o número de conceitos *enzyme* citados com organismos é 186.089, o que é totalmente consistente, porque as enzimas são um subconjunto de proteínas.

3. Análise dos dados

Para melhor visualizar este cenário, analisamos a família de proteínas GADD45 com mais detalhes, clicando , temos a seguinte tela (Figura 6.5):

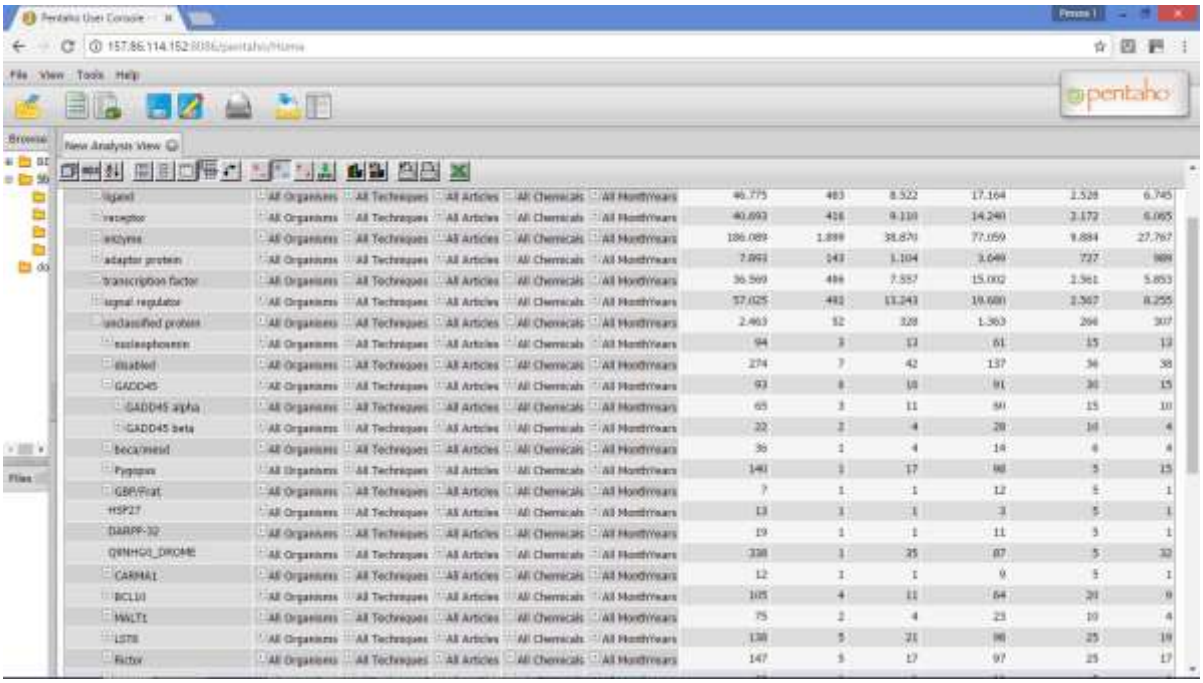


Figura 6.5: Análise da hierarquia da proteína GADD45.

A família GADD45 é composta de 2 subfamílias: GADD45 *alfa* e GADD45 *beta*. A família GADD45 *beta* é composta pelas proteínas GA45B_HUMAN e GA45B_MOUSE, sendo citadas 11 vezes cada, somando estas ocorrências resulta nas 22 citações do conceito GADD45 *beta*.

A medida de citação do termo mais genérico é a soma da citação dos seus termos mais específicos. Então, neste caso, concluímos que o termo GADD45 *beta* não aparece explicitamente nos artigos, porque se tivesse ocorrido, o número de

citação do termo GADD45 *beta* seria maior do que a soma da citação de termos mais específicos.

Analisando o conceito GADD45, ele foi citado 93 vezes com organismos. Isto é devido a conceitos GADD45 *beta* e GADD45 *alpha* terem sido citados 22 e 65 vezes, respectivamente. Isso significa que o termo GADD45 aparece explicitamente com algum organismo 6 vezes nos artigos, já que 93 - 87 = 6.

É importante enfatizar que o TaP DM não registra repetições do mesmo termo, portanto, se um termo é citado várias vezes ou apenas uma vez em um artigo, a ferramenta registra sua presença apenas uma vez. As ontologias que possuem hierarquias e se os termos pertencentes a diferentes níveis são mencionados em um artigo, o DM registrará a presença dos termos mais específicos e genéricos, que é o motivo pelo qual, no mesmo artigo, o TaP DM pode registrar as várias presenças de um conceito.

Isso é comprovado pela Figura 6.6, onde a ordem das dimensões foi alterada para mostrar a hierarquia das proteínas presentes em um artigo.

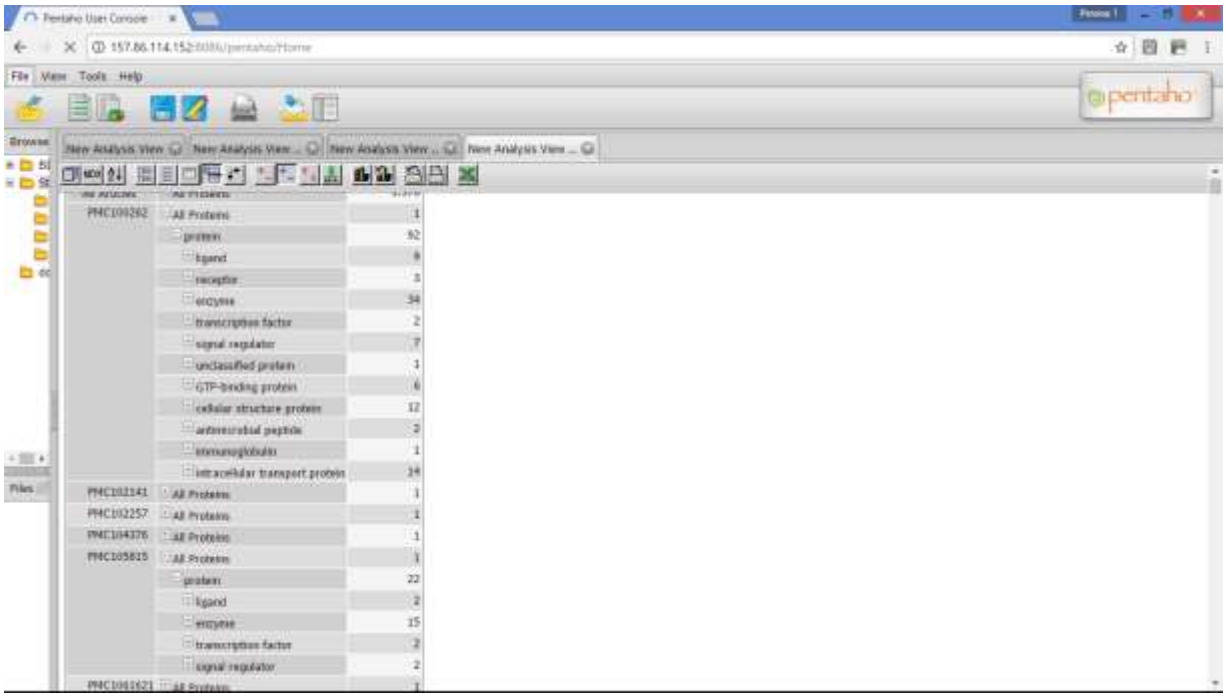


Figura 6.6: Analisando a hierarquia de proteínas em um artigo.

Analisando o artigo PMC100262, o conceito de proteína está presente 92 vezes. Isto é resultado da soma da presença de todos os níveis hierárquicos mais específicos (*ligand*, *receptor*, *enzyme*, ...), 9 + 3 + 34 + 2 + 7 + 1 + 6 + 12 + 2 + 1 + 14 + 1 (*protein*) = 91. Se o termo *protein* não estivesse explicitamente no texto, o

resultado seria 91. Esse tipo de contabilidade de conceitos e subconjuntos permite ao usuário comparar a presença de domínios de conhecimento nos artigos.

Para implementação desta contagem é muito importante devido ao fato de que a presença de um termo específico e de um termo genérico indica maior confiabilidade em relação ao domínio de interesse.

Por exemplo, o termo '*STEP*' possui alta taxa de citação no *corpus* de artigos. Este termo representa o nome de uma proteína presente na ontologia *Molecule Role*, mas este termo também é muito utilizado nos artigos experimentais para indicar as etapas de um experimento.

O TaP DM foi projetado para armazenar dados de proteínas, e considerar o termo '*step*' como proteína quando na verdade ele é utilizado com outro significado, acarreta um erro na resposta da ferramenta.

Considerando o registro na tabela de fatos dos termos genéricos e específicos em um mesmo artigo, a ferramenta é capaz de indicar a presença do termo '*STEP*' e também das suas classes mais genéricas, propiciando um indicativo de que o termo específico citado está relacionado ao domínio de interesse, no caso as proteínas, e não simplesmente um termo de outro domínio, por este motivo foi decidido esta implementação.

Caso deseja-se que o DM apresente o número exato do número de artigos que o termo aparece, é necessário realizar um simples ajustes nos dados. Quando houver em um mesmo artigo a citação de um termo específico e a citação de um termo mais genérico, então a citação deste nível mais genérico não deve ser considerado na tabela de fatos, sendo registrado apenas o nível mais específico. Mesmo com este ajuste, o DM será capaz de relacionar os termos mais específicos aos termos menos específicos, a dimensão armazena toda a hierarquia da ontologia, mantendo assim o registro de todas as relações entre os termos.

Por outro lado, realizar esta alteração empobrece os resultados da ferramenta, pois ela perde o poder de registrar os diversos sub conceitos presentes em um mesmo artigo.

4. Exemplo de consulta

Considerando a questão: Quais proteínas da família da *electron carrier protein* não foram citadas com técnicas de essencialidade?

Ao clicar em *electron carrier protein*, a subfamília *cytochrome c* aparece. Ao clicar no *cytochrome c*, surgem as proteínas CYC_MOUSE, CYC_RAT e CYC_HUMAN. Expandindo a hierarquia na coluna *techniques*, clicando em ⊞, a ferramenta lista as técnicas citadas, como mostra a Figura 6.7:

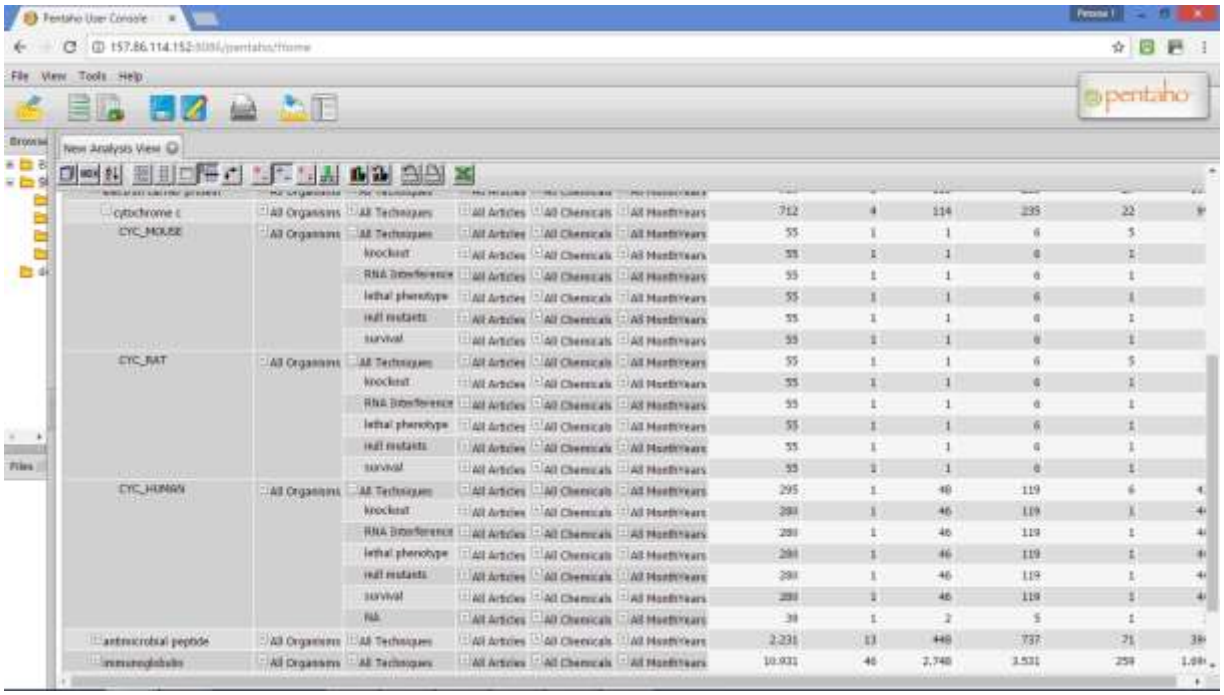


Figura 6.7: Análise da família *electron carrier protein.*

A proteína CYC_HUMAN possui uma técnica NA, indicando que é mencionada em alguns artigos onde não há citações de técnicas de essencialidade. CYC_MOUSE e CYC_RAT não possuem este item, indicando que, quando citado, há sempre a citação de alguma técnica de essencialidade. Para descobrir em que artigos isso acontece, basta clicar na coluna do artigo, conforme mostrado na Figura 6.8.

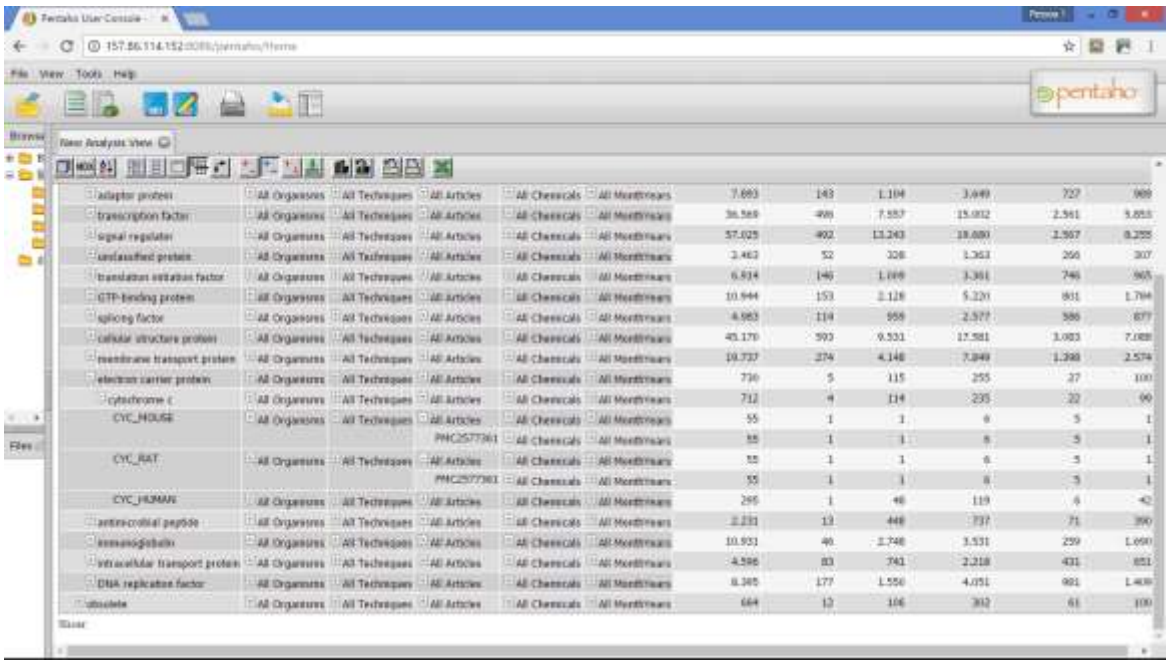


Figura 6.8: Artigos onde ocorrem a citação de proteínas CYC_MOUSE e CYC_RAT.

APÊNDICE B - IMPLEMENTAÇÃO DO CUBO OLAP NO MONDRIAN

Este trabalho realiza a implantação do cubo OLAP em um banco de dados relacional. Mas está implantação em alguns casos merece um pouco mais de atenção, principalmente quando se trata da implementação de hierarquias. Neste apêndice serão apresentados algumas soluções utilizadas para criar e popular o banco de dados, pois a ferramenta Mondrian, assim como outras ferramentas, possui algumas particularidades e limitações.

É muito comum os dados armazenados no banco possuírem algum tipo de relacionamento. Mas quando este relacionamento representa algum tipo de hierarquia, é necessário implementar soluções menos convencionais, apesar de ser extremamente utilizadas no dia a dia (128,129).

As hierarquias mais comuns são as hierarquias onde o elemento possui apenas um antecessor, ou seja, ele tem um único pai (130), como mostra a Figura 6.9.

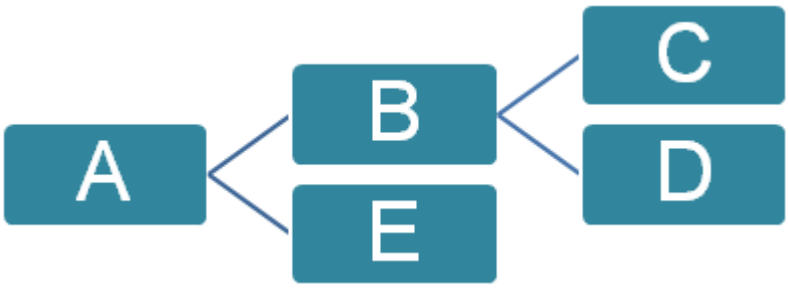


Figura 6.9: Hierarquia convencional.

Considerando que os elementos A, B, C e D são empregados de uma corporação e possuem uma relação hierárquica de supervisão de outros empregados, a implementação da tabela no banco de dados seguiria a tabela 10.1:

Tabela 6.1: Exemplo de população de uma hierarquia

id	supervisor_id	nome	salário
1	0	A	80000
2	1	B	30000
3	2	C	25000
4	2	D	23000

No caso de um projeto de *data mart* seguindo o *schema* estrela, a tabela de fatos recebe como chave estrangeira a chave primaria da tabela de dimensão.

Esta abordagem possui algumas desvantagens que merecem ser abordadas. Primeiro, o desempenho não é muito bom se uma hierarquia contenha mais de cem membros, se considerarmos as ontologias da área biomédica, este valor é superado com facilidade (129). Em segundo lugar, como o Mondrian implementa o agregador de contagem distinta ao gerar SQL, você não pode definir uma medida de contagem distinta em qualquer cubo que contenha uma hierarquia pai-filho (129).

Para superar estes problemas, é possível implementar no Mondrian as *closure tables*. Uma *closure table* é uma tabela SQL que contém os registros das relações empregado/supervisor (pai/filho), independentemente da profundidade. Ela também implementa um campo distância, ela não é essencial para estabelecer as relações, mas facilita na sua construção (129). A Tabela 6.2 mostra como ficaria o exemplo anterior implementado com uma *closure table*.

Tabela 6.2: Exemplo de população de uma *closure table.*

supervisor_id	empregado_id	distância
1	1	0
2	2	0
1	2	1
3	3	0
2	3	1
1	3	2
4	4	0
2	4	1
1	4	2

Este caso exemplo de herança foi utilizado para implementar a dimensão organismos e a modelo relacional entre as tabelas é apresentado na Figura 6.10:

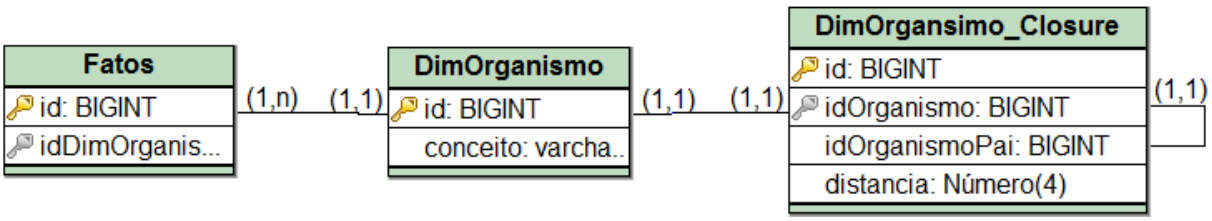


Figura 6.10: Modelagem dimensão organismos.

A implementação do cubo OLAP para este caso foi feita como mostra a Figura 6.11:

```
<Dimension type="StandardDimension" visible="true" foreignKey="idDimOrganismo"
    highCardinality="false" name="Organismo">
      <Hierarchy name="Organismo" visible="true" hasAll="true" primaryKey="id">
        <Table name="DimOrganismo"></Table>
        <Level name="New Level 0" visible="false" table="DimOrganismo" column="id"
            nameColumn="nome" parentColumn="idOrganismoPai" nullParentValue="0"
            type="Numeric" uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
              <Closure parentColumn="idOrganismoPai" childColumn="idOrganismo">
                <Table name="DimiOrganismo_closure"></Table>
              </Closure>
        </Level>
      </Hierarchy>
</Dimension>
```

Figura 6.11: Trecho da implementação da dimensão organismos no cubo OLAP.

Em alguns casos, a relação de herança entre os elemento pode ser mais complexa, como é o caso das heranças múltiplas. Herança múltipla é quando um elemento possui mais de um elemento pai, como mostra a Figura 6.12.

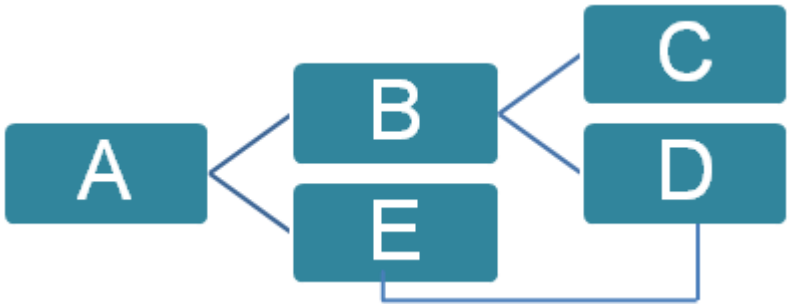


Figura 6.12: Exemplo de herança múltipla.

Nestes casos é preciso fazer algumas adaptações para que a ferramenta OLAP reconheça a herança múltipla. Estas modificações podem ser feitas no código da aplicação, mas como algumas ferramentas possuem código fechado e, mesmo para ferramentas de código aberto, realizar modificações no código fonte não são tarefas triviais.

Por outro lado, existem adaptações na modelagem relacional das tabelas de dimensão e fatos que podem atender a esta situação. A dimensão proteína foi projetada para permitir heranças múltiplas entre seus elementos, o diagrama relacional é apresentado na Figura 6.13:

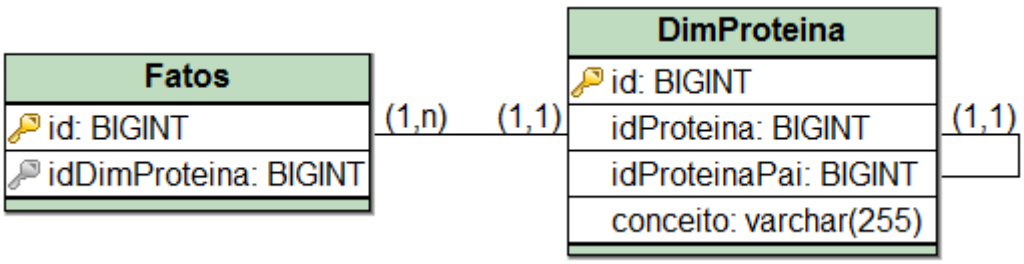


Figura 6.13: Modelagem da dimensão proteína.

O campo idDimProteina da tabela de fatos não possui vínculo com a o campo id da tabela DimProteina, como é natural de se pensar. Na verdade, na verdade o campo idDimProteina da tabela de fatos possuí vínculo com o campo idProteina da tabela DimProteina. A **Erro! Fonte de referência não encontrada.** mostra como o cubo OLAP implementou a tabela de dimensão proteína.

```
<Dimension type="StandardDimension" visible="false" foreignKey="idDimProteina"
    highCardinality="false" name="Proteina">
      <Hierarchy name="Proteina" visible="false" hasAll="true" primaryKey="idProteina">
        <Table name="dimproteina"></Table>
        <Level name="nivel1" visible="true" table="dimproteina" column="idProteina"
            nameColumn="nome" parentColumn="idProteinaPai" nullParentValue="0"
            type="Numeric" uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
</Dimension>
```

Figura 6.14: Trecho de implantação da dimensão proteína no cubo OLAP.

APÊNDICE C - ÁLGEBRA RELACIONAL

As consultas (1) e (2) (apresentadas em álgebra relacional) foram criadas e submetidas ao *data mart*:

$$PC \leftarrow \pi_{concept\_id,concept\_label}((\sigma_{concept\_label='particular\_organism'}(O)) * F * P) \quad (1)$$

$$NPC \leftarrow \pi_{concept\_id,concept\_label}(P - PC) \quad (2)$$

De acordo com a premissa principal deste estudo de caso, para que estas novas proteínas tenham uma maior probabilidade de serem novos alvos para o organismo de interesse, devem ser citadas com as técnicas de essencialidade, assim, as chances destas proteínas exercerem um papel de essencialidade aumentam.

Para flexibilizar esse requisito, o parâmetro nTec foi usado na consulta e pode ser alterado de acordo com as necessidades do usuário. Ele determina a quantidade de técnicas distintas de essencialidade que devem ser citadas com a proteína para que possa ser considerada um possível novo alvo. Deste modo, aumentando o valor do parâmetro nTec, a probabilidade de que a proteína tenha um papel importante no estudo de essencialidade também aumenta.

As consultas (3) e (4) são utilizadas para extrair, a partir do conjunto NPC, o conjunto das proteínas que são citadas com as técnicas de essencialidade, determinando o conjunto NTP (conjunto de proteínas alvo).

$$Q_0 \leftarrow {}_{concept\_id,concept\_label}\xi_{countdistinct(idDimTechnique)}(NPC * F) \quad (3)$$

$$NTP \leftarrow \pi_{id,label}(\sigma_{tottech \geq nTec}(\rho Q_{0\ (id,label,tottech)}(Q_0)) \quad (4)$$

Uma vez definido o conjunto NTP, como devemos correlacionar suas proteínas ao organismo de interesse, já que estas proteínas nunca foram mencionadas com o organismo de interesse? A estratégia adotada foi correlacionar proteínas do conjunto NTP com as proteínas do conjunto PC (proteínas com probabilidade de exercerem um papel importante no estudo de essencialidade gênica para o organismo de interesse). Em outras palavras, se existirem proteínas do conjunto NTP que são muito frequentemente citadas com proteínas de PC,

possivelmente as proteínas de NTP podem ter um papel essencial para o organismo de interesse.

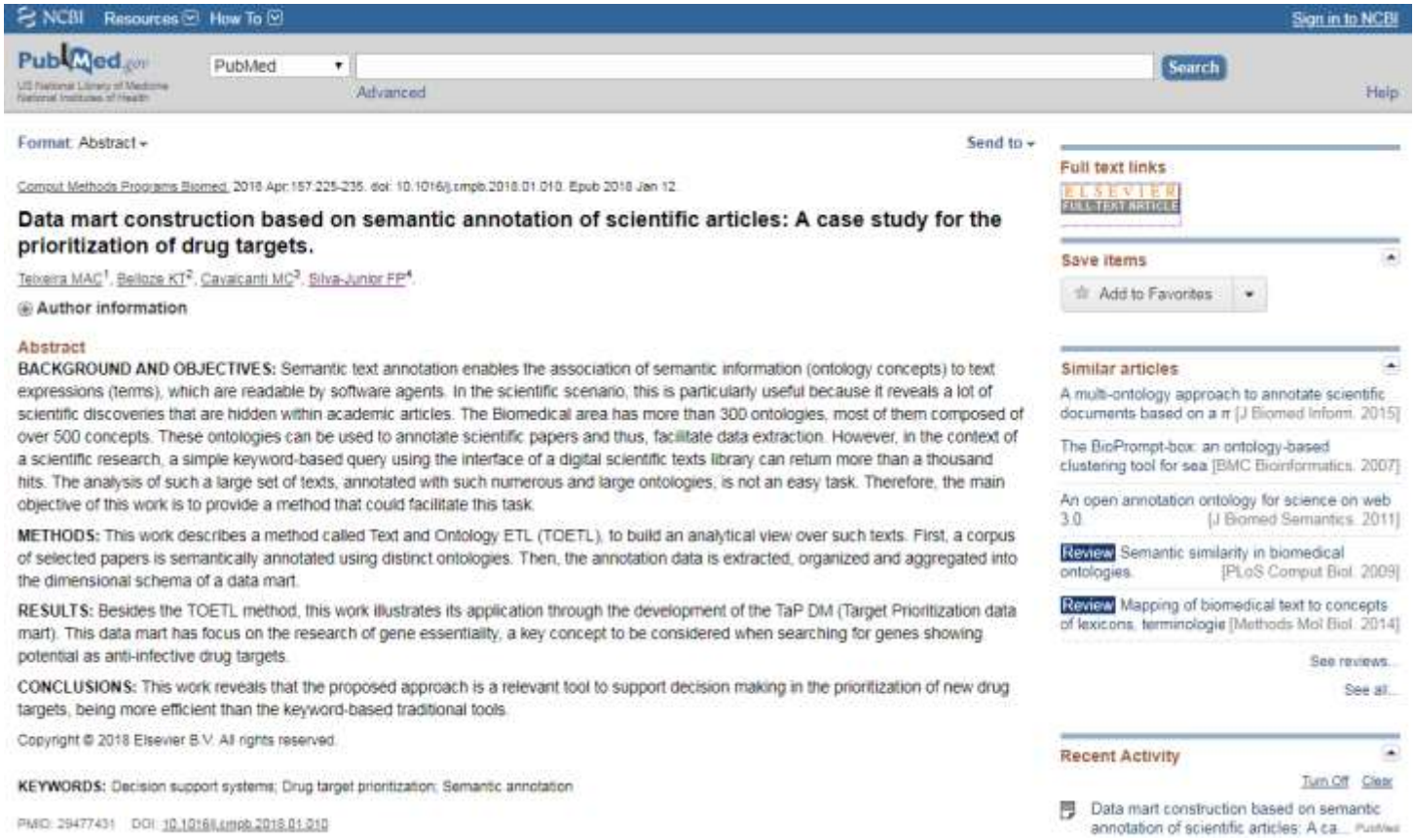
Para isso assumimos que as proteínas têm um papel importante na essencialidade de um organismo, quando forem citadas com esse organismo e as técnicas de essencialidade ao longo de um determinado tempo. Nesse contexto, o parâmetro nAno foi criado para determinar o número de anos que a proteína deve ser citada com as técnicas de essencialidade para que possa ser considerada uma proteína que desempenha um papel importante no organismo. Semelhante ao parâmetro nTec, aumentando o valor do parâmetro nAno, a probabilidade de que a proteína tenha um papel importante no estudo de essencialidade também aumenta. As consultas (5-7) recuperam o conjunto das proteínas relacionadas à essencialidade (EP).

$$Q_1 \leftarrow {}_{\text{concept\_id,concept\_label}}\,\xi_{\text{countdistinct(idMonthYear),countdistinct(idDimTechnique)}}\,(P * F) \tag{5}$$

$$Q_2 \leftarrow \sigma_{\text{cYear} \geq \text{nYear}}\,(\rho_{Q_1\,(\text{id,label,cYear,cTec})}\,(Q_1)) \tag{6}$$

$$EP \leftarrow \pi_{\text{id,label}}\,(\sigma_{\text{cTec} \geq \text{nTec}}\,(Q_2)) \tag{7}$$

APÊNDICE D - Publicação científica relacionada à tese



NCBI Resources How To Sign in to NCBI

PubMed.gov US National Library of Medicine National Institutes of Health

PubMed Search Advanced Help

Format: Abstract Send to

Comput Methods Programs Biomed. 2018 Apr;157:225-235. doi: 10.1016/j.cmpb.2018.01.010. Epub 2018 Jan 12.

Data mart construction based on semantic annotation of scientific articles: A case study for the prioritization of drug targets.

Teixeira MAC[1], Belloze KT[2], Cavalcanti MC[3], Silva-Junior FP[4].

Author information

Abstract

BACKGROUND AND OBJECTIVES: Semantic text annotation enables the association of semantic information (ontology concepts) to text expressions (terms), which are readable by software agents. In the scientific scenario, this is particularly useful because it reveals a lot of scientific discoveries that are hidden within academic articles. The Biomedical area has more than 300 ontologies, most of them composed of over 500 concepts. These ontologies can be used to annotate scientific papers and thus, facilitate data extraction. However, in the context of a scientific research, a simple keyword-based query using the interface of a digital scientific texts library can return more than a thousand hits. The analysis of such a large set of texts, annotated with such numerous and large ontologies, is not an easy task. Therefore, the main objective of this work is to provide a method that could facilitate this task.

METHODS: This work describes a method called Text and Ontology ETL (TOETL), to build an analytical view over such texts. First, a corpus of selected papers is semantically annotated using distinct ontologies. Then, the annotation data is extracted, organized and aggregated into the dimensional schema of a data mart.

RESULTS: Besides the TOETL method, this work illustrates its application through the development of the TaP DM (Target Prioritization data mart). This data mart has focus on the research of gene essentiality, a key concept to be considered when searching for genes showing potential as anti-infective drug targets.

CONCLUSIONS: This work reveals that the proposed approach is a relevant tool to support decision making in the prioritization of new drug targets, being more efficient than the keyword-based traditional tools.

Copyright © 2018 Elsevier B.V. All rights reserved.

KEYWORDS: Decision support systems; Drug target prioritization; Semantic annotation

PMID: 29477431 DOI: 10.1016/j.cmpb.2018.01.010

Full text links

ELSEVIER FULL-TEXT ARTICLE

Save items

Add to Favorites

Similar articles

A multi-ontology approach to annotate scientific documents based on a π [J Biomed Inform. 2015]

The BioPrompt-box: an ontology-based clustering tool for sea [BMC Bioinformatics. 2007]

An open annotation ontology for science on web 3.0 [J Biomed Semantics. 2011]

Review Semantic similarity in biomedical ontologies. [PLoS Comput Biol. 2009]

Review Mapping of biomedical text to concepts of lexicons, terminologie [Methods Mol Biol. 2014]

See reviews...

See all...

Recent Activity

Turn Off Clear

Data mart construction based on semantic annotation of scientific articles: A ca... PubMed