MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Mestrado em Programa de Pós-Graduação Biologia Computacional e Sistemas

REPOSICIONAMENTO DE FÁRMACOS PARA MALÁRIA: UM MÉTODO QUE IDENTIFICA ALVOS NO DOMÍNIO DAS DOENÇAS NEGLIGENCIADAS

EDUARDO BARÇANTE

Rio de Janeiro
Março de 2015

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

EDUARDO BARÇANTE

Reposicionamento de fármacos para malária: Um método que identifica alvos no domínio das doenças negligenciadas

Dissertação apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Mestre em Ciências.

Orientador (es): Prof. Dr. Fabricio Alves Barbosa da Silva
Prof. Dr. Ernesto Raúl Caffarena

RIO DE JANEIRO
Março de 2015

Ficha catalográfica elaborada pela
Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

B242 Barçante, Eduardo

Reposicionamento de fármacos para malária: um método que identifica alvos no domínio das doenças negligenciadas / Eduardo Barçante. – Rio de Janeiro, 2015.
xviii,125 f. : il. ; 30 cm.

Dissertação (Mestrado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2015.
Bibliografia: f. 99-112

1. Recuperação de informação. 2. Mineração de textos. 3. Documento digital. 4. Reposicionamento de fármacos. 5. Ontologia. 6. Anotação semântica. 7. Proteínas. I. Título.

CDD 614.532

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: EDUARDO BARÇANTE

REPOSICIONAMENTO DE FÁRMACOS PARA MALÁRIA: UM METODO QUE IDENTIFICA ALVOS NO DOMÍNIO DAS DOENÇAS NEGLIGENCIADAS

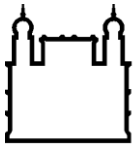
ORIENTADOR (ES): Prof. Dr. Fabrício Alves Barbosa da Silva
Prof. Dr. Ernesto Raúl Caffarena

Aprovada em: 13/03/2015

EXAMINADORES:

Prof. Dr. Marcelo Ribeiro Aves **- Presidente**	(FIOCRUZ/RJ)
Prof. Dr. Roney Coimbra	(FIOCRUZ/MG)
Prof. Dr. Alberto Martin Rivera Dávila	(FIOCRUZ/RJ)
Prof. Dr. Fábio Passetti	(INCA)

Rio de Janeiro, 13 de março de 2015.



Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

Este trabalho é dedicado a Cleusa Maria dos Santos. Uma pessoa simples e uma grande mulher.

AGRADECIMENTOS

À Dra. Maria Cristina Soares Guimarães por me fazer trilhar um novo caminho.

À Dra. Maria Luiza Campos e Dra. Maria Luiza Machado Campos pela luz e inspiração nesse novo caminho.

Ao Dr. Carlos Henrique Marcondes por toda paciência durante a minha jornada na UFF.

Ao Dr. Oswaldo Gonçalves Cruz por me aceitar prontamente e permitir o meu ingresso no curso de Biologia Computacional e Sistemas.

Ao Dr. Carlos Medicis Morel pelas primeiras aulas com *softwares* de mineração de textos.

Ao Dr. Luciano Moreira pela oportunidade em projetos sobre doenças negligenciadas.

À Dra. Maria de Lourdes de Souza Maia pela experiência no Programa Nacional de Imunização (PNI).

Ao meu orientador Dr. Fabrício Barbosa da Silva por me receber como aluno no seu grupo de pesquisa.

Ao meu orientador Dr. Ernesto Caffarena por suas palavras de apoio, atenção e disponibilidade no auxílio dos meus estudos.

À Mara Lucia Alves Leitão Correa, bibliotecária UFRJ por sua ajuda profissional.

Aos amigos do COPPEAD/UFRJ, em especial a Tatiana Kaus Sarkis, uma amiga importante na minha trajetória acadêmica.

À Regina Esch por todo o carinho e competência profissional, Dr. Adelzon Assis, Marcelo Marciano, Tereza M. M. Costa, aos bibliotecários do ICICT e da ENSP, PROCC e à turma da BCS.

Ao DATASUS Ricardo, Gustavo, José Carlos Jorge e Consuelo Freiria. Salve Jorge!

Ao Dr. Marcio Argollo de Menezes, Dra. Nicole Scherer, Dr. Fabio Passetti, Dra. Renata Schama, Dr. Marcos Catanho, Dr. Alberto Dávila e a todos os outros que colaboraram para a construção desse trabalho.

A todos aqueles que desejam o meu sucesso, a todos aqueles que nada fizeram e a todos aqueles que criaram obstáculos para que este trabalho não fosse realizado. Tudo isso faz parte da vida!

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

À FAPERJ pelo apoio financeiro.

Aos meus amigos visíveis e invisíveis.

Eduardo Barçante.

“Não beba catuaba, Eduardinho!
Colatino tem uma ótima garrafada...”
Dona Irene dirigiu a palavra ao seu neto de seis anos. Enquanto este questionava seu avô, sobre os motivos de uma família tão grande e muitas tias para tomar conta das suas artes.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

BARÇANTE, Eduardo. **REPOSICIONAMENTO DE FÁRMACOS PARA MALÁRIA: UM MÉTODO QUE IDENTIFICA ALVOS NO DOMÍNIO DAS DOENÇAS NEGLIGENCIADAS.** Manguinhos, 2015.

RESUMO

DISSERTAÇÃO DE MESTRADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Eduardo Barçante

O atual cenário da Biologia Computacional conta com o *know-how* de diversas áreas tecnológicas voltadas para informação, computação e, especialmente, para construção e uso de bancos de dados na Internet como MEDLINE, PubMed, PDB. Na medida em que essas bases de dados possibilitam o aumento no registro de produção, estoque e circulação de dados genéticos, também viabilizam, em anos recentes, ambientes para acessar, integrar e produzir o novo conhecimento. O reuso desses dados, isto é, a transformação e o emprego desses dados em um processo diferente do qual os dados foram originalmente concebidos, torna-se um desafio em pesquisas na área biológica. Os problemas surgem pela falta de estrutura textual ou marcação para processamento por computadores. Neste trabalho, foi desenvolvido um método que identifica nomes de proteínas que servirão de substrato às práticas laboratoriais de reposicionamento de medicamentos. Dentre as fontes citadas, foram empregados, inicialmente, documentos textuais digitais do PubMed, a principal fonte de informação das ciências da saúde da Biblioteca Nacional de Medicina (*National Library of Medicine*). Esta base foi explorada com métodos e recursos da mineração de textos que são amplamente empregados para extrair termos relacionados aos nomes de proteínas no domínio das doenças negligenciadas. Os resultados obtidos após o processamento de 8444 artigos relacionados ao termo *malaria* do PubMed revelaram 254 fármacos candidatos ao reposicionamento. Dentre estes, três fármacos (Amitriptyline, Trimethoprim e Methrotrexate) foram comprovados, por meio de uma busca não exaustiva na literatura biomédica, que são utilizados em experimentos para o tratamento de malária. Por outro lado foi possível sugerir um conjunto de proteínas ou moléculas que poderão servir de insumos na fase inicial da cadeia de produção de medicamentos que é o *screening* de moléculas.

Palavras-chave: Recuperação de informação; Mineração de textos; Documento digital; Reposicionamento de fármacos; Ontologia; Anotação Semântica; Proteínas.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

DRUG REPOSITIONING TO MALARIA DISEASE: A METHOD TO A METHOD TO IDENTIFYING TARGETS IN DOMAIN NEGLECTED DISEASES

ABSTRACT

MASTER DISSERTATION IN COMPUTATIONAL BIOLOGY AND SYSTEMS

EDUARDO BARÇANTE

The current scenario of computational biology relies on the know-how of many technological areas, with focus on information, computing, and, particularly on the construction and use of existing Internet databases such as MEDLINE, PubMed and PDB. In recent years, these databases provide an environment to access, integrate and produce new knowledge by storing ever increasing volumes of genetic or protein data The transformation and management of these data in a different way from the one that were originally thought can be a challenge for research in biology. The problems appear by the lack of textual structure or appropriate markup tags. In this study, It aimed to develop a method that identifies proteins names that will serve as a substrate to laboratory practices for drug repositioning. Among the sources cited, It was initially explored digital text documents from PubMed, the main source of information about Health Sciences of the National Library Medicine. This base was explored with text mining methods and resources that are widely used to extract terms related to proteins names in domain of neglected diseases. The results obtained after the processing of 8444 articles related to term malaria in PubMed revealed 254 drugs candidates for repositioning. Among these, three drugs (Amitriptyline, Trimethoprim and Methrotrexate) were confirmed by a non-exhaustive search in the biomedical literature, that are used in experiments to treat malaria. On the other hand we can suggest a set of proteins or molecules that can serve as inputs in screening of molecules, the early stage of the drug production chain.

Keywords: Information retrieval; Text mining; Search interface; Digital document; Drugs repositioning; Ontology; Semantic annotation, Proteins.

ÍNDICE

ÍNDICE DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

ASCII – American Standard Code for Information Interchange

API – Application Programming Interface (Interface de programação de aplicativos)

CDC – Center for Disease Control and Prevention

CID-10 – Classificação Internacional de Doenças versão 10

FIOCRUZ – Fundação Oswaldo Cruz

GO – Gene Ontology

MEDLINE - Medical Literature Analysis and Retrieval System Online

MeSH - Medical Subject Headings

MSF – Médicos sem Fronteiras

NCBI – National Center for Biotechnology Information

NCBI Taxon – NCBI Taxonomy

NCBO – The National Center for Biomedical Ontology

NCIt – National Cancer Institute Thesaurus

NIH – National Institutes of Health

NLM – U.S. National Library of Medicine

NTD – Neglected Tropical Disease (Doenças tropicais negligenciadas)

OBO – The Open Biological and Biomedical Ontologies

OMS – Organização Mundial de Saúde

ONU – Organização das Nações Unidas

OWL – Ontology Web Language

PDB – Proteins Data Bank

PROCC – Programa de Computação Científica

SQL – Structured Query Language

UMLS – Unified Medical Language System

UniProt – Universal Protein Resource

UniProtKB – UniProt Knowledgebase

URL – Localizador-Padrão de Recursos (Uniform Resource Locator)

W3C – World Wide Web Consortium

WHO – World Health Organization

WWW – World Wide Web

1 INTRODUÇÃO

O progresso alcançado nas pesquisas genômicas é visto atualmente como grande orientador de uma nova dinâmica socioeconômica que, junto a outras medidas políticas adotadas por alguns países, propõe saúde como um requisito para o desenvolvimento sustentável tanto no ponto de vista econômico como social. Os desafios propostos por meio da declaração da Organização das Nações Unidas (ONU) (ONU, 2000) e a demanda por nova informação, decorrente naturalmente da evolução tecnológica, obrigam as nações a redimensionar seus sistemas de pesquisa em saúde cuja orientação procure garantir que os resultados obtidos em pesquisas sejam incorporados às ações de saúde e, consequentemente, levar à sustentabilidade do progresso social.

Nesse contexto, é dever buscar cada vez mais, mecanismos e estratégias com intuito de viabilizar avanços na área de pesquisa biotecnológica e promover os meios necessários para que estes sejam reutilizados na base de conhecimento corrente. Segundo Markus (2001), isto é a reutilização do conhecimento, a base necessária que é composta pela matéria-prima dos processos inovadores que trarão as melhorias na saúde e em todos os sistemas que dela dependem.

Contudo, inúmeros são os desafios que acompanham esse conjunto de metas pré-estabelecidas, principalmente na Biologia, onde bases de dados possuem diferentes estruturas, diferentes representações de dados e normas (SCHIMITT, 2011) que conflitam com diferentes culturas, modelos interdisciplinares, limites tecnológicos e respostas urgentes aos problemas específicos vividos pelos povos de cada país.

Tais desafios nascem em contextos diferentes, pertinentes a diversos setores produtivos. Nessa relação de forças, o ambiente científico nada difere em alguns pontos quando se observa o crescimento da produção de dados biológicos, na quantidade da produção intelectual representada pela literatura científica amplamente divulgada sob a forma de teses, artigos, dissertações e também nas bases de dados genômicas e proteicas.

A motivação para o presente trabalho foi explorar este conjunto com material recuperado em ambientes e bases de dados diversificados e temática voltada para

as doenças negligenciadas[1] como a malária (WHO, 2013) e a doença de chagas (DNDi, 2010), que são pontos fortes e pactuados nos objetivos do milênio para 2015 (ONU, 2000).

As doenças negligenciadas compõem um grupo de doenças presentes em climas tropicais (DNDi, 2010). Embora, até meados da década de 1970, estivessem associadas aos locais de extrema pobreza, atualmente são vistas como uma barreira no desenvolvimento dos países da mesma faixa equatorial repelindo investimentos e, consequentemente, levando à redução dos lucros dos investidores de capital (INCT, 2014).

São denominadas negligenciadas porque os investimentos destinados à pesquisa, principalmente na indústria farmacêutica, não traduzem um ciclo sustentável voltado para a prevenção e produção de novos medicamentos, uma vez que essas indústrias visam apenas o lucro (DNDi, 2010; BRASIL, 2010).

Contudo, há esforços na busca de soluções para o problema que atualmente afeta a vida de cerca de um bilhão de pessoas que permanecem concentradas em áreas rurais, áreas remotas, favelas e em grande parte populações também reconhecidas pelo baixo poder aquisitivo.

A Organização Mundial da Saúde (OMS) reconhece e trata o tema doenças negligenciadas (DNDi, 2010) como um problema de saúde pública. Após uma fase de aprendizado e reconhecimento do problema, redigiu um relatório que caracteriza o agrupamento de doenças e um novo quadro conceitual que permite atualmente a setores de saúde pouco eficazes ou com escassez de recursos atingir pessoas mais pobres.

O relatório também estimula os Estados, parceiros de interesses em comum, a contribuir com recursos científicos, tecnológicos e financeiros com o objetivo de controlar ou erradicar algumas dessas doenças.

A OMS reconheceu dezessete doenças tropicais negligenciadas que são encontradas em 149 países. Para cem países essas doenças são endêmicas, isto é, específicas de determinada região. Desses, 100 países são endêmicos para duas ou mais dessas doenças negligenciadas e 30 países para seis ou mais (WHO, 2010).

Em 2001 foram sugeridas outras duas classificações, a saber: divisão das doenças em Globais, Negligenciadas e Mais Negligenciadas, classificações sugeridas pela organização não governamental Médicos Sem Fronteiras (MSF)

[1] Doenças que são subfinanciadas e possuem baixo reconhecimento, mas são as principais ameaças em países pouco desenvolvidos.

(SMITH et al., 2001); e outra divisão que introduz uma classificação similar para essas mesmas doenças com o Tipo I (equivalente às doenças globais dos MSF), Tipo II (Negligenciadas/MSF) e Tipo III (Mais Negligenciadas/MSF), sugerida pela OMS no Relatório da Comissão sobre Macroeconomia e Saúde de 2001 (*Commission on macroeconomics and health. Macroeconomics and health: investing in health for economic development.*) (WHO, 2001). Nesse sentido, o presente utilizou o termo doenças negligenciadas no decorrer da pesquisa.

O Ministério da Saúde Brasileiro, em parceria com a FIOCRUZ, priorizou, para medidas de prevenção e políticas de saúde pública, sete doenças negligenciadas, que são: dengue, doença de chagas, leishmanioses, malária, esquistossomose, tuberculose e hanseníase (BRASIL, 2010; INCT, 2014).

Inúmeras são as contribuições científicas feitas pelo Brasil no âmbito das doenças negligenciadas (SOUZA, 2010). Ao destacar a gravidade do problema e caracterizar bem o quadro endêmico, Alecrim, em sua pesquisa denominada "Tratamento de crianças com malária pelo *Plasmodium falciparum* com derivados da artemisinina". Alecrim (2003) comprova que: "Malária é a principal endemia da Amazônia brasileira. No Estado do Amazonas, no período compreendido entre janeiro de 1996 a dezembro de 1998, foram diagnosticados 279.914 casos, sendo que o *Plasmodium falciparum* contribuiu com 68.021 dos diagnósticos".

A identificação de nomes de proteínas nesses artigos e as relações desses nomes com outras bases de dados biológicas podem colaborar para o reposicionamento de medicamentos. Este está definido na base de descritores em Ciências da Saúde como prática intencional e metódica de encontrar novas aplicações para drogas já existentes (DECS, 2014).

Sir James Black (1998) antevia que a base para se chegar a um novo fármaco era começar por um antigo: *The most fruitful basis for the discovery of a new drug is to start with an old drug* (Laureado, em 1988, Prêmio Nobel de Fisiologia e Medicina). Para Wermuth (2003), partir de moléculas estrutural e terapeuticamente diversas, que já tiveram a sua biodisponibilidade e toxicidade testadas, é o caminho mais vantajoso para se chegar a um novo fármaco. Ambos os autores defendem que a base do método é a segurança de um produto que já se encontra disponível no mercado proporcionando considerável redução de tempo e custos na fase-pesquisa.

> Embora os fármacos possam, teoricamente, ligar-se a quase qualquer tipo de alvo tridimensional, a maioria dos fármacos produz seus efeitos desejados (terapêuticos) através de uma interação

> seletiva com moléculas-alvo, que desempenham importantes papéis na função fisiológica e fisiopatológica. Em muitos casos, a seletividade da ligação do fármaco a determinados receptores também estabelece os efeitos indesejáveis (adversos) de um fármaco. Em geral, os fármacos são moléculas que interagem com componentes moleculares específicos de um organismo, produzindo alterações bioquímicas e fisiológicas dentro desse organismo. Os receptores de fármacos são macromoléculas que, através de sua ligação a determinado fármaco, medeiam essas alterações bioquímicas e fisiológicas. (GOLAN et al., 2009).

Diante desse contexto, emergem os seguintes questionamentos: seria possível fazer a caracterização das proteínas segundo a sua função e, por outro lado, fazer a caracterização dos ligantes segundo as propriedades terapêuticas, e depois cruzar os bancos para propor reposicionamento de medicamentos[2]? Quais metodologias existem hoje para identificar nomes de proteínas e estruturas de proteínas em bases de documentos textuais digitais[3]? Repositórios com resumos e artigos completos da Literatura Biomédica, como o PubMed[4], por exemplo, podem ser utilizados como fonte de informação para descoberta de alvos de reposicionamento de fármacos?

Acredita-se que as respostas para essas questões e a elaboração de um método baseado no reuso (MARKUS, 2001) de documentos textuais digitais sejam um potencial promissor no sentido de facilitar o reposicionamento de medicamentos.

Para atingir o objetivo, construiu-se um método baseado em algoritmos computacionais que permitem identificar, na literatura científica, proteínas similares e seus respectivos fármacos com vista ao reposicionamento de medicamentos para o tratamento de doenças negligenciadas.

1.1 O problema em estudo e a perspectiva de análise

O atual cenário da Biologia Computacional conta com a colaboração e *know-how* de diversas áreas tecnológicas voltadas para informação, estatística, matemática, computação e especialmente áreas voltadas para a construção e uso

[2]Prática intencional e metódica de encontrar novas aplicações para drogas já existentes. Descritor Inglês: drug repositioning. Fonte: DeCS.

[3] No presente trabalho, reconheceremos os dados digitais armazenados em bancos de dados como Documentos Textuais Digitais, isto é, todo e qualquer arquivo independente do seu suporte e formato digital que contenha textos.

[4] PubMed compreende mais de 24 milhões de citações de literatura biomédica do MEDLINE, revistas de ciências da vida e livros *online*. As citações podem incluir *links* para conteúdo de texto completo da PubMed Central e *sites* de editores. Fonte: http://www.ncbi.nlm.nih.gov/pubmed/

de bancos de dados na Internet como MEDLINE (PubMed)[5], PDB[6], UniProt[7], SWISS-PROT[8], bases que detêm alguma especificidade como a informação de genes ortólogos (SBC, 2013) e interfaces para recuperar seus dados (SCHIMITT et al., 2011).

Na medida em que essas bases de dados possibilitam o aumento no registro de produção, estoque e circulação de dados genéticos e proteicos em formato digital elas têm construído, em anos recentes, ambientes para acessar, integrar e produzir o novo conhecimento. O reuso desses dados, isto é, a transformação e o emprego desses dados em um processo diferente do qual os dados foram originalmente concebidos, torna-se um desafio em pesquisas na área biológica. Os problemas surgem pela falta de um padrão de registro, falta de uma estrutura textual ou marcação para processamento por computadores, ausência de palavras-chave adequadas ao processo de busca e recuperação ou pelo grande número de relações possíveis que devem ser construídas para individualizar a informação desejada em mais de uma base de dados etc.

> Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully (BERNERS-LEE et al., 2001).

Dentre as possíveis fontes de dados citadas, utilizou-se inicialmente o PubMed que é a principal fonte de informação das ciências da saúde da *National Library Medicine* (NLM) e o sistema de busca bibliográfica do PubMed desenvolvido pelo *National Center for Biotechnology Information* (NCBI) (PUBMED, 2013). Denominado *Entrez,* esse sistema de busca integra várias bases de dados que mantêm informação sobre Biologia Molecular em seus domínios.

A busca por informação em arquivos no formato texto digital[9], segundo Lancaster (1993), tornou-se uma ideia promissora quando computadores foram utilizados em grande escala para essa atividade. O autor segue afirmando que:

[5] MEDLINE. Principal banco de dados bibliográficos da National Library of Medicine. Disponível em: http://www.nlm.nih.gov/pubs/factsheets/medline.html.
[6] Banco de dados de proteínas. Fonte: http://www.rcsb.org/pdb/home/home.do
[7] Recurso abrangente e de livre acesso a sequências de proteínas e informações funcionais de alta qualidade. Fonte: http://www.uniprot.org/
[8] Bases de Dados de Sequências de Proteínas. Fonte DeCS.

[9] Um arquivo texto digital é um conjunto de linhas formadas por caracteres ASCII que traduzem um conteúdo legível para seres humanos. O arquivo normalmente recebe uma extensão TXT usualmente reconhecida como um arquivo cujo padrão é textual.

> [...] Ao se estudar a história dos sistemas informatizados de recuperação da informação, reconhecem-se duas linhas principais de desenvolvimento. Uma delas tem sua origem nos grandes sistemas, desenvolvidos por certas instituições como a National Library of Medicine (NLM), o Department of Defense (DOD) e a National Aeronautics and Space Administration (NASA), que funcionavam com bases em termos de indexação extraídos de um vocabulário controlado e atribuídos aos documentos por indexadores humanos. A outra linha de desenvolvimento teve seu início no campo do direito, e envolvia a colocação de textos completos (por exemplo, leis) em formato eletrônico e a utilização do computador para fazer buscas de palavras ou combinações de palavras nesses textos. Trabalhos dessa natureza antecederam, na realidade, o desenvolvimento de tesauros e o surgimento dos grandes sistemas baseados na indexação feita por seres humanos [...].

Concomitantemente, há um esforço da comunidade científica no sentido de construir e fazer uso de ontologias, vocabulários controlados, descritores, dentre outros, para que sistemas destinados a extrair informação de documentos textuais localizem a informação desejada para o usuário e compreendam o conteúdo dos textos em linguagem natural.

Contudo, a *Web*[10] com seus artigos científicos e bases de conhecimento, ainda exigem dos seres humanos fases de avaliação, classificação e seleção da informação destacando principalmente a sua citação.

Apesar de serem disponibilizados em formato digital, esses artigos são concebidos para leitura humana. Portanto, há um longo processo que deve ser acrescido às exigências quando se refere à busca do conhecimento de interesse e o processamento por computadores desses artigos.

Esta é a *Web* Sintática, assim denominada por Moura (2002) e Breitman (2005), onde os computadores desempenham a função de apresentar a informação, uma vez que cabe aos humanos o processo de interpretação.

> There is so much information available that we simply no longer know what we know, and finding what we want is hard – too hard. The knowledge we seek is often fragmentary and disconnected, spread thinly across thousands of databases and millions of articles in thousands of journals. The intellectual energy required to search this array of data-archives, and the time and money this wastes, has led several researchers to challenge the methods by which we traditionally commit newly acquired facts and knowledge to the scientific record. (ATTWOOD et al., 2009, p. 317).

[10] Web é um conceito empregado no âmbito tecnológico para nomear uma rede de informática, e de um modo geral, a Internet.

Por outro lado, a *Web* Semântica (W3C, 2014) permite que as pessoas construam mecanismos que, além de armazenarem os dados, também viabilizem a escrita de regras para interação entre esses dados com outras bases de dados. A *Web* Semântica apresenta-se em oposição a *Web* Sintática e cria o ambiente propício para que computadores e homens possam processar, relacionar e interpretar conteúdos de diversas fontes.

> Uma das aplicações mais importantes para a tecnologia da *Web* Semântica é a ciência e a busca por dados. O aprimoramento dos meios de comunicação, o baixo custo de computadores e o desenvolvimento de *softwares* favorecem cada vez mais espaços onde a *Web* Semântica pode consolidar esses dados que estão distribuídos em diversos ambientes computacionais. (BARÇANTE, 2011).

No âmbito da ciência da computação, o estudo sobre ontologias ganha força no segmento da inteligência artificial. A sua aplicação em recursos computacionais, mais especificamente na busca de métodos para representar o conhecimento sob a forma de programas de computador, torna-se um atrativo quando pesquisadores explicitam o papel destas na organização e representação do conhecimento. Gruber (1993) define ontologias como a especificação de um conceito e acrescenta:

> In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application. In the context of database systems, ontology can be viewed as a level of abstraction of data models, analogous to hierarchical and relational models, but intended for modeling knowledge about individuals, their attributes, and their relationships to other individuals. (GRUBER, 2008).

Na visão da *Web* Semântica, as ontologias são centrais, pensadas como uma camada explícita (BERNERS-LEE et al., 2001), e atuam como fornecedores dos meios pelos quais os termos usados são compreendidos no resto do contexto (O'HARA et al., 2009). "*Uma ontologia é, assim, um conjunto de conceitos padronizados, termos e definições aceitas por uma comunidade particular*."

(CAMPOS et al., 2009). E na busca por tornar interoperável todo esse conhecimento registrado, as instituições científicas têm dedicado recursos humanos e financeiros para criar e manter ontologias em larga escala para esse e outros fins (BARÇANTE, 2011).

O Centro Nacional para Ontologia Biomédica - *The National Center for Biomedical Ontology* (NCBO)[11] tem como objetivo apoiar pesquisadores biomédicos em seu trabalho de busca por conhecimentos, por meio de ferramentas *online* e um portal *Web* que lhes permite acessar, analisar e integrar recursos ontológicos diferentes e em todos os aspectos da investigação biomédica e clínica. O foco principal do trabalho desenvolvido pelo NCBO envolve o uso de ontologias biomédicas para auxiliar na gestão e análise de dados provenientes de experimentos complexos.

Dentre outras ontologias, são exemplos na área biomédica: *Gene Ontology*[12] que é o resultado da colaboração para abordar a necessidade de descrições consistentes de produtos de genes em bancos de dados, um vocabulário controlado desenvolvido pela *Open Biomedical Ontologies* (OBO)[13] (OBO, 2005); *Protein Ontology*[14] (PR), que oferece uma representação ontológica das entidades relacionadas com proteínas e as define explicitamente, mostrando as relações entre elas; BioPortal NCBO disponibiliza 405 ontologias, métodos de relacionamento entre ontologias e serviços no formato URL[15] (MUSEN et al., 2012).

A finalidade é identificar quais documentos textuais na sua forma completa são empregados no método para encontrarmos proteínas similares na sua estrutura ou função. A ontologia *Protein Ontology* (PR) disponibilizada pelo portal NCBO é utilizada em outros projetos, (Figura 1) e tem papel fundamental no presente trabalho porque pertence ao domínio biomédico, portanto, inerente ao tema da pesquisa em questão. A ontologia fornece, na etapa oportuna, os termos candidatos

[11] http://www.bioontology.org
[12] *Gene Ontology* – iniciativa importante no campo da bioinformática com o objetivo de uniformizar a representação do gene, seu produto e atributos por meio das suas espécies e o registro em bancos de dados.
[13] OBO Foundry é uma experiência de colaboração que envolve os desenvolvedores de ontologias com uma abordagem científica. Visa estabelecer um conjunto de princípios para o desenvolvimento e posterior criação de um conjunto de ontologias de referências ortogonais e interoperáveis no domínio da área biomédica.
[14] Fornece uma classificação formal, baseada em lógica de classes de proteínas específicas, incluindo representações estruturadas.
[15] *URL* representa um *Uniform Resource Locator*, um ponteiro para um "recurso" na *World Wide Web*. Um recurso pode ser algo tão simples como um arquivo ou uma pasta, ou pode ser uma referência a um objeto mais complexo, como uma consulta a um banco de dados ou a um motor de busca.[31]

a nomes de proteínas ou evidências que permitam ao *pipeline*[16] reconhecer esses nomes.

Projeto	Descrição	Responsável	Instituição
Banco de Dados Genômico - Células Beta	Pesquisas de células beta e ferramentas para explorar banco de dados com temática relacionada a genômica. Fornece informação detalhada sobre: genes, transcrições, interações entre genes, regiões genômicas e beta de células relacionadas a estudos sobre genômica funcional.	Consórcio de Biologia Celular Beta	Universidade da Pensilvânia
Ontologia Sanguínea	A Ontologia saguínea é um conjunto de ontologias criada para reunir e representar os dados sobre o sangue de acordo com princípios ontológicos bem fundamentados.	HTLV Grupo de Pesquisa interdisciplinar	UFMG
ChEMBL-RDF	Versão RDF do banco de dados 200-BY-AS ChEMBL RDF do grupo de John Overington.		Universidade de Maastricht
Reposcionamento de Fármacos	Reposicionamento de Fármacos para doenças negligenciadas	Eduardo Barçante	FIOCRUZ
Estrutura e Informação sobre neurociências	Estruturas e Informação sobre neurociência (NIF; http://nif.nih.gov) é um inventário dinâmico baseado na Web de recursos de neurociência, dados e ferramentas acessíveis por meio de qualquer computador conectado à Internet.	Maryann Martone Jeffrey Grethe Amarnath Gupta Gordon Shepherd Giorgio Ascoli Paul Sternberg David Van Essen	
PACHTS	É um projeto sobre medicamentos inovadores (IMI), uma parceria inédita entre a Comunidade Europeia e a Federação Europeia das Associações e Indústrias Farmacêuticas (EFPIA)		

Figura 1 - Projetos usando *Proteins Ontology (PR)*.
Fonte: http://bioportal.bioontology.org/ontologies/PR

Segundo Feldman et al. (1998), extrair informação é uma das técnicas mais importantes usadas na mineração de textos e a define como o processo de análise de textos não estruturados ou em linguagem natural com objetivo de recuperar informação e conhecimento que dificilmente se obteria numa leitura feita por humanos. Portanto, a mineração de textos caracteriza-se como um processo de análise de textos que identifica e extrai informação de texto em linguagem natural e o transforma em índices significativos destinados à diversas finalidades tais como a predição, análise de clusters, etc. (WITTEN et al., 2004).

A mineração de textos nos domínios da área biomédica tem se mostrado muito útil e com diversas aplicações. Inúmeros trabalhos são desenvolvidos incluindo conceitos como extração de termos (URAMOTO, 2004), regras de associação para descobertas (CREIGHTON, 1993; HRISTOVSKI, 2001) e extração de relacionamentos entre vários conceitos (ADAMIC, 2002; ARZUCAN, 2008; PALAKAL, 2002; SRINIVASAN, 2004; WEEBER, 2003; WREN, 2004). Além disso, outras técnicas de processamento de linguagem natural foram aplicadas na literatura biomédica, por exemplo, na extração de informação, na extração do nome do gene e interações proteicas, no reconhecimento de nomes de entidades, *families protein*

[16] Conjunto de instruções que determinam um fluxo de dados.

(ANDRADE, 1998) e na desambiguação de nomes de genes e proteínas (COIMBRA et al. 2010; GINTER et al., 2004) dentre várias outras (AL-MUBAID, 2005).

Sendo assim, formado o *corpus* textual pertinente ao domínio das doenças negligenciadas, é possível dispor das técnicas mencionadas, além de fazer uso das técnicas estatísticas e grafos disponíveis no ambiente R[17]. Acredita-se que o uso de ontologias e anotações semânticas contribuem nas etapas de visualização, identificação e classificação de compostos proteicos e suas funções para o reposicionamento de medicamentos[18].

1.2 Definições em Biologia

Nesta seção, são descritos sucintamente alguns conceitos básicos no campo da Biologia. São definições operacionais importantes relacionadas às propriedades estruturais e funcionais das proteínas e que são relevantes para o entendimento do presente trabalho.

1.2.1 Proteínas

Proteínas são compostos orgânicos nitrogenados caracterizados por sequências de aminoácidos (HUNTER, 1993), os quais se encontram unidos covalentemente por ligações peptídicas. As proteínas são sintetizadas a partir da leitura do RNAm pelos ribossomos presentes no espaço citosólico[19] (ALBERTS, 1997; LEAL, 2012).

As proteínas possuem funções tão variadas quanto as suas estruturas as quais podem ser caracterizadas em diferentes níveis de complexidade: a estrutura primária das proteínas é conferida às cadeias polipeptídicas, isto é, a cadeia principal da proteína formada pela ligação de aminoácidos e a sequência na qual eles são apresentados; a estrutura secundária é dada pelo arranjo de ligações de hidrogênio entre diferentes regiões da sequência formando estruturas hélice alfa, folha beta ou *coils*; A estrutura terciária, é responsável pela estrutura tridimensional da molécula, ocorre quando as proteínas se dobram sobre si mesmas, a exemplo do

[17] R é um ambiente de software livre para computação estatística e gráficos.
[18] Prática intencional e metódica de encontrar novas aplicações para drogas já existentes. Fonte DeCS.
[19] Conteúdo do compartimento principal do citoplasma, excluindo organelas ligadas à membrana, como o retículo endoplamatico e a mitocôndria. Originalmente definido como a fração celular restante após a remoção de componentes do citoesqueleto, da mebrana e outras organelas, por centrifugação em baixa rotação.

que ocorre quando há ligações covalentes dissulfeto, determinando o empacotamento das estruturas secundárias em um ou múltiplos domínios; por último, a estrutura quaternária ocorre em caso de proteínas poliméricas, a exemplo de hemoglobina e das imunoglobulinas (ALBERTS, 1997; SCITABLE, 2014).

Proteínas são compostos orgânicos formadas pelo encadeamento de aminoácidos e realizam a maior parte das funções da célula assumindo diferentes estruturas e funções em um único organismo.

Uma classe especial de proteínas, as denominadas enzimas, promovem reações químicas específicas, acelerando a sua cinética. Sem elas, as reações ocorreriam lentamente ou seriam inviáveis do ponto de vista termodinâmico (HUNTER, 1993; ICSP, 2015).

As proteínas exercem uma miríade de funções, dentre as quais destacam-se: o suporte estrutural, a exemplo dos colágenos; a capacidade de reconhecimento célula-célula e de antígenos, a exemplo das imunoglobulinas e dos receptores celulares; contração (miosina), que participam do metabolismo celular (citocromos), além de atuarem como sensores elétricos, químicos e de propagação de estímulos (CREIGHTON, 1993).

Buscar nomes de proteínas em documentos textuais digitais, fazer a caracterização segundo a sua função e relacionar o seu nome a outras bases de dados são os objetivos do presente trabalho. Acredita-se que seja o caminho para relacionar proteínas, ligantes e fármacos associados ou não aos nomes das doenças negligenciadas.

1.2.2 Ligantes

A capacidade de uma célula responder a sinais é importante para desenvolvimento de suas atividades no organismo. O processo de sinalização se dá pelas proteínas enquanto os ligantes disparam ou inibem processos de sinalização.

Em alguns casos, ligantes são moléculas lipossolúveis que, portanto, atravessam as membranas. Já outros tipos de ligantes, como os polipeptídios e moléculas polares, não são capazes de atravessar diretamente a camada lipídica. Então, esses outros ligantes precisam interagir com proteínas presentes nas biomembranas[20]. Tais proteínas reconhecem ligantes muito especificamente, portanto, são denominados receptores (ALBERTS, 1997).

[20] *Os compartimentos das células e organelas são delimitados pelas biomembranas. Formadas por uma bicamada de fosfolídio, cuja expessura varia entre 6 a 10 nm, as biomembranas apresentam*

Segundo Alberts (1997), uma interação efetiva entre uma molécula de proteína e um ligante necessita que muitas ligações fracas se concretizem ao mesmo tempo. As moléculas ligantes que podem ligar-se fortemente às proteínas são as únicas que encaixam perfeitamente na superfície. A área de ligação é denominada sítio de ligação, sendo descrita como uma cavidade formada por um conjunto de aminoácidos específicos sobre a superfície da proteína (COX, 2002).

1.2.3 Fármacos

De acordo com a Portaria nº 3.916/MS/GM, de 30 de outubro de 1998 do Conselho Federal de Farmácia, fármaco é a substância química que é o princípio ativo do medicamento (BRASIL, 1998).

Se o remédio é qualquer substância ou meio usado para curar uma moléstia, convém substituir o termo que designa 'remédio' por 'medicamento' quando se deseja falar especificamente de uma formulação ou produto farmacêutico que contém um ou vários princípios ativos denominados fármacos. (IVF, 2006; BRASIL, 2010, p.14).

A ligação entre fármacos e ligantes é uma questão mais complexa. Primeiro, é importante saber quais moléculas se ligam (interagem) porque a força de atração entre elas é maior do que a repulsão (ALBERTS, 1997).

A indústria farmacêutica e de imunobiológicos busca novos ligantes ou incrementos aos já existentes (FERREIRA, 2011). Refere-se aqui à ideia de que qualquer coisa que interagir com proteínas será denominada ligante, incluindo, outras proteínas. A ideia da associação é ter a finalidade de inibir ou produzir algum efeito, por exemplo, uma proteína que induz a morte celular é associada a um ligante que anulará o seu efeito (ALBERTS, 1997).

1.2.4 Medicamento

O medicamento é uma fórmula farmacêutica que contém o fármaco ou conjunto de fármacos. O medicamento por definição é o produto farmacêutico concebido com a finalidade de curar, amenizar ou para fins de diagnóstico (Brasil, 1973).

---

também colesterol, proteínas e carboidratos. As biomembranas possibilitaram uma grande especialização nos organismos ao longo da evolução da vida por oferecer a formação de compartimentos semipermeáveis, flexíveis [54].

1.3 Aspectos técnicos em Computação

Nesta seção, são descritos importantes aspectos técnicos em informática que foram empregados na construção do *pipeline* e orientam a compreensão da metodologia empregada no presente estudo. O conteúdo contempla: as ferramentas, bibliotecas, ambiente computacional e outras características e técnicas de programação.

1.3.1 A linguagem de programação e ambiente

Atualmente, em computação estatística, a linguagem R (R FOUDATION, 2002) é considerada uma linguagem e um ambiente de computação voltado para estatística e para gráficos. O ambiente R representa um conjunto de *softwares* livres[21] com a finalidade de: manipular e armazenar dados de forma eficaz, calcular e apresentar graficamente resultados; possuir um conjunto de operadores que permitem manipulação de vetores e cálculos com matrizes; conter uma coleção de ferramentas coesas e integradas para análise de dados.

1.3.2 O PostgreSQL

O PostgreSQL é um gerenciador de banco de dados *open source*[22]. Durante um período de mais de 15 anos de desenvolvimento ativo, o PostgreSQL ganhou uma arquitetura comprovadamente forte, com boa reputação e confiabilidade, exatidão e integridade de dados. Além disso, suporte para os tipos de dados do SQL: 2008 e contempla os tipos: inteiro, numérico, lógico, caracter, palavras, data, e intervalos com configurações opcionais de precisão. O PostgreSQL também suporta o armazenamento de grandes objetos binários, incluindo imagens, sons ou vídeo. Possui interfaces de programação nativas para C / C ++, Java, Net, Perl, Python, Ruby, Tcl, ODBC, entre outros, e documentação excepcional (POSTGRESQL, 2015).

No presente trabalho, dentre outros recursos oferecidos pelo banco de dados, o PostgreSQL tem a finalidade de fornecer a estrutura das tabelas, armazenar os

[21] O projeto R para computação estatística é um software livre sob os termos da GNU *General Public License* e o seu código fonte é compilado e executado nas plataformas UNIX ou sistemas similares.
[22] *Software* de código aberto utilizado livremente. Ao seguir determinadas regras, qualquer pessoa pode alterar e compartilhar livremente o software (modificado ou não).

resultados obtidos durante a execução do *pipeline* e viabilizar as consultas em *SQL Language*.[23]

1.3.3 Extensible Markup Language (XML)

Extensible Markup Language (XML) é um formato de texto simples, flexível e derivado do SGML (ISO 8879),[24] originalmente concebido para enfrentar os desafios da publicação eletrônica em grande escala. Sendo o formato XML reconhecido como um padrão de linguagem de marcação, desempenha atualmente um papel importante na troca de dados na *Web* e em outros ambientes (W3C, 2014).

1.3.4 Programas Analisadores *(Parsers)*

Permitir o processamento e compreensão de documentos textuais digitais em larga escala por máquinas é um grande desafio. Isso ocorre uma vez que ler e interpretar esses documentos textos é uma habilidade humana. Porém, as máquinas demandam para esta ação um texto estruturado ou, no mínimo, que o documento textual traga algum sentido ou a temática, explícita, que se deseja alcançar.

Nesse sentido, segundo Maia (2001), os programas analisadores são programas de computador com capacidade de processar textos automaticamente e identificar sua estrutura sintática − surgem como analisadores para uma gramática que aceita sentenças como entrada e cria para essas sentenças a sua árvore gramatical.

Contudo, no segmento da Ciência da Computação, um analisador, geralmente compõe um compilador ou pré-compilador que recebe, como parâmetros de entrada, instruções advindas de outros programas, instruções sequenciais, comandos interativos, *tags* marcadores ou outra interface definida. O analisador atua como um divisor ao transpor dados de um ambiente para outro ou, por exemplo, subdividir um texto em uma estrutura de substantivos (objetos), verbos (métodos ou ações) e seus atributos que serão controlados por outros recursos de programação (TECHTARGET, 2014).

Ambos os conceitos apresentados aqui são úteis porque analisadores funcionam como programas de interface entre bases de dados e páginas no formato *html* ou com toda sua complexidade oculta nas ontologias construídas e empregadas nesse trabalho.

[23] SQL Language - *SQL – Structured Query Language* ou Linguagem de Consulta Estruturada.
[24] Standard Generalized Markup Language (SGML)

Os programas analisadores tornam-se fundamentais e necessários, dadas as idiossincrasias existentes na construção de inúmeras bases de dados, recursos e aplicações disponíveis atualmente na *Web*.

1.3.5 Mineração de textos

O presente trabalho explora bases de dados de documentos textuais digitais buscando desenvolver um método que identifique nomes de proteínas que servirão de substrato às práticas laboratoriais de reposicionamento de medicamentos. A mineração de textos científicos é empregada para extrair termos que são relacionados a proteínas no domínio das doenças negligenciadas em suas respectivas bases de dados.

Hotho et al. (2005) definem como se dá o processo de descoberta de conhecimento em bases de textos por meio da mineração de texto. Segundo os autores, a mineração de textos consiste em usar técnicas para recuperar informação, extrair informação, bem como processar linguagem natural com os algoritmos e métodos de descoberta de conhecimento, mineração de dados e aprendizado de máquina.

De acordo com Zweigenbaum (2007), extrair informação é uma forma de análise de linguagem natural e está se tornando uma tecnologia central para diminuir a distância entre um texto não estruturado e conhecimento formal expresso em ontologias. Os métodos de processamento de linguagem natural fornecem a fundamentação para as investigações no campo da mineração de textos biomédicos.

A mineração de texto abrange um vasto campo de abordagens teóricas e métodos cujo objeto em comum de entrada de informação é o texto. Isso permite várias definições, que vão desde uma extensão de mineração de dados clássica de textos às formulações sofisticadas aplicadas às grandes coleções *online* para descobrir novos fatos e tendências sobre o próprio mundo (HEARST, 1999).

A mineração de textos aparece em campo interdisciplinar com atividades que buscam explorar agrupamentos de documentos, ontologias e anotação semântica, documentos completos, resumo e fragmentos de resultados. Na Biologia Computacional, o esforço visa relacionar termos da área biomédica, palavras-chave, códigos de bases proteicas que possam auxiliar a recuperação de informação, principalmente nomes de proteínas, e associar bases de dados biomédicas (BARBOSA-SILVA, 2011).

O ambiente de programação R é contemplado com um grande número de bibliotecas construídas para dar suporte a demandas específicas como acesso aos bancos de dados, manipulação de *strings* e outros. Contudo, o conjunto de programas denominado *TM package* (FEINERER, 2014; FEINERER, HORNIK, MEYER, 2008) oferece uma estrutura de mineração de textos que permite dentre outras atividades manipular o *corpus* textual, pré-processar, gerenciar metadado[25] e criar matrizes de termos em documento (FEINERER, 2014).

1.3.6 Ontologia

Ontologia é uma definição de características de um conceito e busca representar um domínio de conhecimento procurando adequar um conjunto de termos. A adequação desses termos ocorre por meio de axiomas cujas regras pré-estabelecidas permitem uma nova organização do conhecimento que se deseja compartilhar.

E, nesse contexto de partilhar o conhecimento, o uso do termo ontologia significa uma especificação de uma conceituação, isto é, uma ontologia é uma descrição (como uma especificação formal de um programa) dos conceitos e das relações que possam existir para um agente ou uma comunidade de agentes (GRUBER, 1993). Esta definição é coerente com o uso de ontologias como *set-of-concept-definitions* (GRUBER, 2008).

A especificação formal decorre da necessidade de ser processável por máquinas, premissa definida de acordo com Bernes Lee et al. (2001), isto é, a formalidade para usar vocabulários, ou seja, solicitar consultas e fazer afirmações de maneira consistente, porém não completa. Torna-se necessário que ontologias sejam especificadas em linguagens para descrever o significado do objeto que se deseja representar. Portanto, as linguagens de definição de ontologias são mais expressivas e possuem alto nível de semântica, o que permite empregá-las em soluções de problemas para integração de bases de dados e interoperabilidade de sistemas distintos (LIMA, 2007; JARDIM, 2013).

Se uma ontologia é uma das tecnologias para implementação da Web Semântica (LIMA, 2005) o W3C *Web Ontology Language* (OWL) é a linguagem da

[25] O metadado é um conjunto de dados sobre outros dados, informação adicional que permite a identificação de ou crítica sobre um conteúdo normalmente textual. O metadado tem o papel na anotação semântica de viabilizar recursos para prover dados (Lima, 2007) [71]. É onde se registram conceitos descritos numa ontologia; palavras previamente definidas (descritores, vocabulários controlados, tesauros); conceitos no domínio de determinada língua (Pisanelli et al., 1998) [73]; (Oliveira et al., 2003) [74].

Web Semântica projetada para representar o conhecimento rico e complexo sobre as coisas, grupos de coisas, e as relações entre as coisas. OWL é uma linguagem baseada em lógica computacional de tal forma que o conhecimento expresso em OWL pode ser explorado por programas de computador. Documentos OWL reconhecidos como ontologias podem ser publicados na Web e podem fazer referências ou serem direcionados a outras ontologias OWL. OWL é parte do conjunto que representa a Web Semântica (Figura 3), a pilha de tecnologias Web do W3C, que inclui RDF, RDFS e outros (W3C, 2012).

Há para a área biomédica um esforço significativo na construção e manutenção de ontologias. Iniciativas como o (*OBO Foundry*) - *The Open Biological and Biomedical Ontologies* ou simplesmente OBO*.* OBO é uma comunidade aberta e, ao aderir à iniciativa, os autores de uma ontologia deverão comprometer-se a sua manutenção, tendo em conta o progresso científico e a divisão do trabalho com os outros membros para assegurar a melhoria desses princípios ao longo do tempo. O *OBO Foundry* é uma tentativa de aplicar o método científico para a tarefa de desenvolvimento de ontologias, e o método científico repousa sobre críticas constantes e na premissa de que nenhum recurso venha a existir de uma forma em que não pode ser melhorado (OBO, 2014).

OBO nada mais é que uma linguagem utilizada para ontologias nos segmentos biológicos e biomédicos. O Portal OBO recomenda seis ontologias que seguiram os princípios (OBO, 2006) adotados pela comunidade, tais quais: *Chemical entities of biological interest,* Chebi; GO, *Gene Ontology*; *Phenotypic quality*, PATO; PRO, *Protein Ontology*; XAO, *Xenopus Anatomy*; ZFA, *Zebrafish anatomy and development*. A ontologia Protein Ontology (PRO) está disponível no *BioPortal Annotator* e é utilizada nas fases de processamento do *pipeline*.

A combinação dos serviços, ontologias, sites, repositórios e outros citados anteriormente, aliadas aos serviços de anotação semântica disponibilizados na *WEB,* viabilizam novos recursos para processamento de documentos textuais digitais em repositórios como o PubMed. *OBO Foundry* e o *BioPortal Annotator* do NCBO são exemplos dessa combinação de serviços e recursos tecnológicos. O NCBO não só oferece serviços e suporte no segmento de ontologias como mantém em seu portfólio ontologias que são disponibilizadas pelo OBO Foundry, buscando integrá-las a outras ontologias, tais como: *Protein Ontology (PR)*, *Gene Ontology (GO)* e outras.

A Figura 2, a seguir, mostra a integração de bases de dados e ontologias da área biomédica centrada no *Protein Ontology (PR)*. São formadas coleções de ontologias e estas são componentes que se mantêm estruturados. Esta é uma forma lógica de estruturar e relacionar ontologias que procura dar significado a um vocabulário formal e muito especifico baseado no modelo em camadas proposto por Berners-Lee. A Figura 3, a seguir, mostra o modelo em camadas que é aceito na comunidade científica como uma forma de representação da arquitetura da Web Semântica.

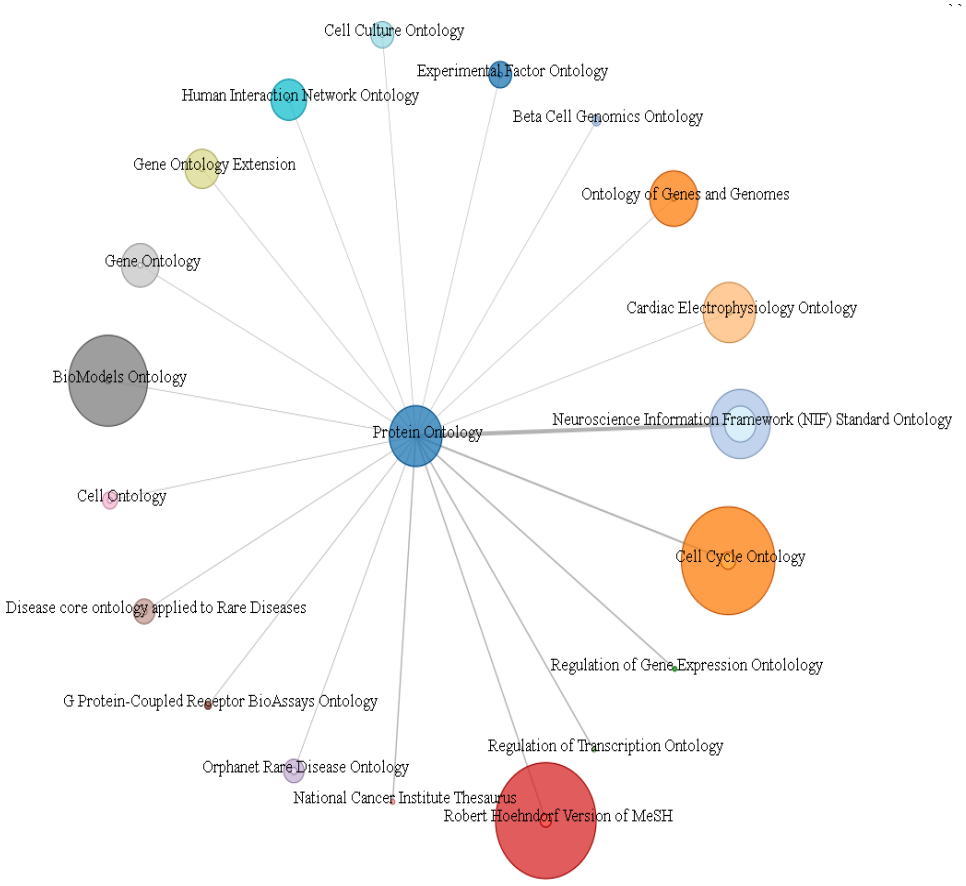


Figura 2 - Mapeamento de classes entre diferentes ontologias.
Fonte: **http://bioportal.bioontology.org/mappings - Ontologia: Proteins Ontology (PR)**

O desenvolvimento de ontologias baseado no modelo em camadas (ver 1.3.7 Web Semântica) precisa de uma linguagem que permita a leitura de dados por homens e o processamento por máquinas. Assim, o desafio é manipular os dados e representar o conhecimento em ambiente global, isto é, a exportação e uso na *WEB.*

1.3.7 Web Semântica

Segundo Berners-Lee et al. (2001), a *Web* Semântica é uma extensão da *Web*, e nesta é dada à informação um significado, e é desta associação que têm

origem a cooperação e o desenvolvimento do trabalho com interação entre pessoas e computadores. O autor afirmava que o nível de complexidade e estrutura de um sistema advém do número de relações entre os componentes. Por esse motivo, a arquitetura e como foi projetada a *Web* Semântica tiveram como fim a estruturação dos conteúdos registrados em documentos textuais e a Web Semântica é o mecanismo facilitador para o reuso desses dados (BELLOZE, 2013).

Por esse motivo, a arquitetura da Web Semântica tem por objetivo a estrutura de conteúdos cujo crescimento é exponencial, e cada vez mais complexo, dado as diferentes relações existentes (BERNERS-LEE, 2001). O nível sintático serve de base para outros níveis estruturados em XML (W3C, 2014), os quais aumentam o potencial de processamento semântico, formando conteúdos estruturados.

A Figura 2 mostra as estruturas inter-relacionadas de tecnologias da Web Semântica e, conforme descrito, é possível perceber que há uma semântica computacional que está ligada à crescente estruturação de conteúdos na Web.

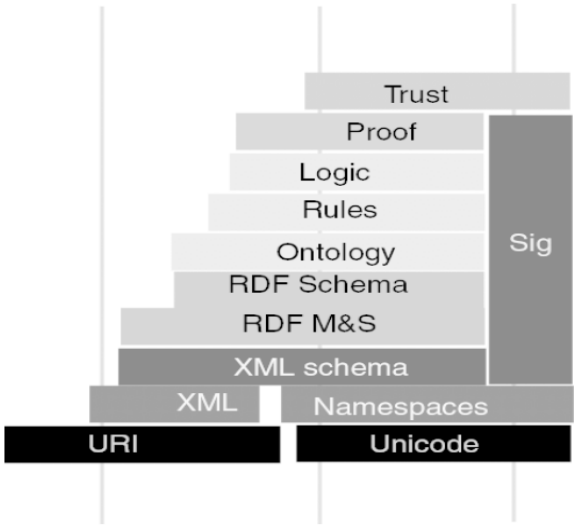


Figura 3 - Estruturas inter-relacionadas de tecnologias da WEB Semântica.
Fonte: www.w3.org/2001/ - W3C.

As primeiras definições para a estruturação de conteúdos no escopo proposto à *Web* Semântica (Figura 3) foram assim descritas:

- *UNICODE* - O padrão Unicode é um sistema de codificação de caracteres projetado para suportar o intercâmbio mundial de dados, processamento e visualização dos textos escritos das diversas linguagens e disciplinas técnicas do mundo moderno. Além disso, ele

suporta textos clássicos e históricos de muitas línguas escritas (UNICODE. 2014);

- *Uniform Resource Identifier URI* - são cadeias de caracteres usadas para identificar ou denominar um recurso disponível especialmente, porém não exclusivamente, na Web. O propósito é permitir a interação com representações do recurso na Web. Isto é viável por meio de protocolos específicos. URIs são identificados em grupos definindo uma sintaxe específica e protocolos associados. A URI tem duas funções principais e pode ser utilizada, como um localizador, como um nome ou ambos, seguem definidas: URN - Uniform Resource Name, são utilizadas para nomear algo, mesmo que seja um objeto abstrato que não está disponível na Web; URL - Uniform Resource Locator, funciona como um endereço, isto é, um método para encontrar um objeto (W3C, 2014);

  - UNICODE e URI formam a base das tecnologias da Web Semântica.

- *Namespaces* XML – fornece o método simples para qualificar um elemento. Atribui prefixos aos nomes utilizados em documentos *Extensible Markup Language* e *namespaces* com objetivo de evitar conflito entre nomes “idênticos” de elementos (BRAGANHOLO, 2001; W3C, 1989);

- XML Extensible Markup Language (XML) é um formato de texto simples, muito flexível derivado do SGML (ISO 8879). Originalmente concebido para enfrentar os desafios da publicação eletrônica em grande escala, XML também está desempenhando um papel cada vez mais importante na troca de uma ampla variedade de dados na Web e em outros lugares. XML deve ser entendida como uma representação sintática para outras linguagens, sintaxe superficial para estruturar documentos e não traz informação semântica, isto é, não dá sentido ao conteúdo (W3C, 2014);

- *Namespaces* e a linguagem de marcação XML estão situados sobre a camada URI. *Namespaces* de um XML é definido como uma coleção de nomes usados em documentos XML que é identificada por URI. (W3C, 1998);

- XML *Schema* (linguagem para restrições à estrutura de documentos XML e definições aos tipos de dados). É um padrão de alta complexidade para especificar estruturas, tipos e validar conteúdos dos elementos do XML. Os esquemas fornecem um meio para definir com mais detalhes a estrutura, conteúdo e semântica de documentos XML. XML Schema 1.0 foi aprovado como uma recomendação no W3C em 2 de Maio de 2001 e uma segunda edição publicada em 28 de Outubro de 2004 (W3C 2014);

- RDF é um modelo padrão para o intercâmbio de dados na Web (RDF, 2004). RDF tem características que facilitam a fusão de dados, mesmo se os esquemas subjacentes diferem. Especificamente suportam a evolução ao longo do tempo, isto é, as mudanças de esquemas ocorrem sem a necessidade de que todos os usuários de dados também mudem seus esquemas e seus dados. RDF amplia a estrutura de ligação da Web para usar URIs com a finalidade de nomear as relações entre as coisas, bem como as duas extremidades da ligação, isto é uma tripla, onde cada tripla tem sujeito, verbo e objeto. Usando esse modelo simples, é permitido que os dados estruturados e semiestruturados se misturem, sejam expostos, e compartilhados entre aplicações diferentes. (LASSILA, 1999; RDF, 2004; W3C, 2014);

- RDF *Schema* é uma linguagem de propósito geral e define não somente as propriedades dos recursos, mas também os conjuntos de recursos que estão sendo descritos. Com isso, representam-se objetos e classes de objetos. Descrevem metadados sobre recursos como documentos XML, por exemplo. RDF Schema é um sistema de classes extensível e genérico que pode ser utilizado como base para esquemas de um domínio específico. Esses esquemas podem ser compartilhados e estendidos por meio de refinamento de subclasses.

Além disso, definições de metadados podem ser reutilizadas por meio do compartilhamento de esquemas. O vocabulário de RDF está definido em dois namespaces: rdf e rdfs. Declarações RDF são precedidas de um dos dois prefixos, dependendo de qual das duas especificações definiu a propriedade em questão (BRAGANHOLO, 2001; BRICKLEY, 2000; W3C, 2014);

- Ontologias em linguagem OWL (W3C, 2004) e vocabulário OBO (OBO, 2005), assim definidas: a OWL Web Ontology Language se destina a fornecer uma linguagem que pode ser usada para descrever as classes e as relações entre elas que são inerentes em documentos da Web e suas aplicações (W3C, 2004); *OBO Foundry* é uma experiência de colaboração que envolve os desenvolvedores de ontologias baseadas na ciência que buscam estabelecer um conjunto de princípios para o desenvolvimento de ontologias com o objetivo de criar um conjunto de ontologias ortogonais e interoperáveis na área biomédica (OBO, 2005). A relação entre ontologias OBO, OWL e a camada RDF está na compatibilidade e portabilidade dos recursos oferecidos pelas duas iniciativas para construir e manter ontologias OBO e OWL por meio do recurso RDF;

  - Ontologias são especificações de conceitos e representam um domínio de conhecimento. Adicionam mais vocabulário para descrever propriedades e classes com mais informação (OWL, 2008; W3C, 2004).

1.3.8 Processos de anotação semântica

Anotação semântica é um procedimento que permite ao usuário ou sistema assinalar e vincular expressões, anotações ou comentários a um determinado texto; há métodos ou linguagens, como RDF e RDF *Schema* (BRICKLEY, 2000; W3C, 2004), que fazem uso das tecnologias da *Web* Semântica para anotar textos fornecendo um simples modelo de dados e um padrão de escrita para declarações.

O exemplo a seguir, conforme Fontes (2011, p.26), mostra como é possível representar um bloco simples de informação, como uma “tripla sujeito-propriedade-objeto”, a partir de um simples arquivo de marcação XML, cuja sintaxe é

reconhecida pela linguagem RDF. O RDF é utilizado como a estrutura ou *framework* para descrever os recursos.

```
<rdf: Description rdf:about=”#methotrexate”>
        <pdbText>methotrexate<pdbText />
        <nomeDB>DB00563<nomeDB/>
        <proteinaRelacionada rdf:resource=”#4OCX”  />
<rdf:Description>
<rdf: Description rdf:about=”#40CX”>
        <nomeCompleto>Fab complex with methotrexate<nomeCompleto/>
        <autor>Longenecker<autor>
<rdf:Description>
```

O uso da tripla pode descrever uma declaração da seguinte forma: “*O fármaco methotrexate, está no DrugBank com o nome DB00563, é relacionado a proteína 4OCX cujo nome completo é Fab complex with methotrexate, registrada por Longenecker.*” A figura 4, a seguir, mostra o exemplo de triplas RDF.

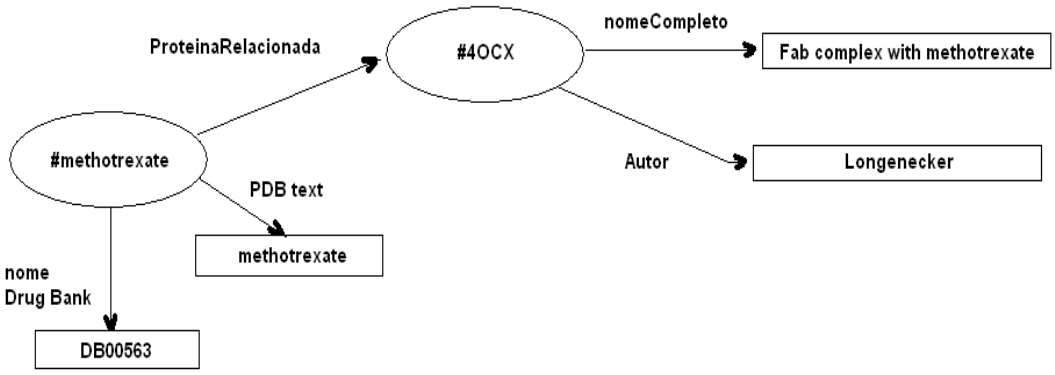


Figura 4 - Exemplo de triplas RDF.

O *pipeline* busca recuperar informação relevante por meio do *software* NCBO *Annotator* (MUSEN, 2012, NCBO, 2009) baseado nas associações do conteúdo textual e das palavras anotadas na ontologia *Protein Ontology* (PR).

A instituição *National Center for Biomedical Ontology* (NCBO) desenvolve e mantém uma Aplicação Web denominada *BioPortal Annotator* para acessar ontologias biomédicas. Esta biblioteca contém uma grande coleção de ontologias, como GO, NCIT, Classificação Internacional de Doenças (CID), disponibilizadas em diferentes formatos como OBO e OWL. Os usuários podem navegar e explorar este

repositório de ontologias, tanto on-line como por meio de uma *API* de serviços da *Web* (NCBO, 2009).

Um *corpus* textual eletrônico pode ser anotado semanticamente, seja por sua relevância em determinado trecho, seja pela necessidade de uma simples marcação. O objetivo é orientar o homem na compreensão do conteúdo textual ou permitir aos programas de computador, como é o caso do *pipeline* com os seus algoritmos, o processamento dos textos ou fragmentos de textos para que se alcancem os resultados desejados.

A anotação auxilia as relações e dá significado conciso ao conjunto de palavras. Demarca uma nova camada para o texto encarregada de prover os meios onde se descreve um conteúdo (UREN, 2005).

1.4 Bancos de dados biológicos

Nesta seção são descritas as bases de dados que são utilizadas durante o processamento do *pipeline*. De uma maneira geral, são os repositórios que contêm a informação biológica necessária para o processamento do *pipeline*, bem como as bases de dados com informação biomédica em formato texto disponibilizada para mineração de texto e o seu reuso no *pipeline*.

1.4.1 O repositório – UniProt

Formado por meio da união de três bases de dados: *Swiss-Prot Knowledge Database*, *TrEMBL* e *Protein Information Resource* (*PIR).* O *UniProt* oferece à comunidade científica livre acesso aos dados referentes à informação funcional e sequência de proteínas. O *UniProt* apresenta-se como um recurso universal de proteínas capaz de gerenciar serviços e acessos que estão divididos e disponibilizados em outras bases de dados ou repositórios: UniProt como base de conhecimento (UniProtKB), o UniProt Reference Clusters (UniRef), UniProt Archive (UniParc) e o repositório UniMES que é voltado para metagenômica e sequências ambientais do banco de dados UniProt (UNIPROT, 2014).

A anotação dos dados no UniProt é realizada por especialistas que atuam no registro de proteínas humanas, bacterianas, vegetais e outras. O repositório é curado e mantém um alto nível de anotações, como por exemplo, a descrição e a função da proteína em uma estrutura. Todas as anotações registradas no UniProt podem ser acessadas por meio de programas de computador. Portanto, no presente

projeto, o repositório UniProt tem a finalidade de validar os termos obtidos por meio da anotação semântica e fornecer os códigos de entrada denominados na estrutura interna do UniProt como <*accession*>. O código <*accession*> identifica entradas no UnitProt que possibilitam ao *pipeline* a busca e recuperação de informação sobre proteínas, ligantes e fármacos. O código <*accession*> funciona como uma palavra-chave e associados a ela uma série de dados importantes, no caso do *pipeline*, vão validar termos obtidos por meio da anotação semântica, recuperar todos os *accession* que estabelecem relação à palavra-chave e caracterizar proteínas.

1.4.2 A base de dados – PDB

O PDB é um repositório internacional de dados cujo conteúdo é composto por informação pertinente a estrutura 3-D de macromoléculas biológicas, incluindo proteínas e ácidos nucleicos. Foi projetado em 1971 no Brookhaven National Laboratory sob a liderança de Walter Hamilton, e continha na sua forma original apenas 7 estruturas. Após Hamilton, seguiram na liderança Tom Koetzle em 1993 e Joel Sussman no período de 1994 até 1998 quando o *Research Collaboratory for Structural Bioinformatics (RCSB)* tornou-se responsável pela gestão do PDB. No ano de 2003 foi fundada a wwPDB[26] para manter um único arquivo PDB de dados estruturais de macromoléculas que é livre e disponível publicamente para a comunidade global. É composto por organizações que atuam como centros de guarda, tratamento e distribuição de dados para os dados do PDB (Berman, 2000).

Segundo Zambenedetti (2002), na década de 80, o PDB apresentou um aumento nos registros que continua nos dias atuais com uma taxa exponencial. Atualmente, o PDB contém mais de 100.000 registros (BERMAN, 2000). Cada entrada do PDB contém uma lista de coordenadas para uma entidade molecular e informação adicional, tais como: procedimentos experimentais, referências bibliográficas, autores, comentários e anotações.

Originalmente, a informação começou a ser representada na forma de texto livre. Porém, em um formato não uniforme de dados, o que dificultava a recuperação da informação. Contudo, a busca por uma solução adequada ao problema tornou o PDB uma fonte de dados com grandes recursos para recuperar dados sobre proteínas e suas estruturas.

[26] *The Worldwide Protein Data Bank (wwPDB)* é formado por organizações que atuam como grandes centros para depósitos, tratamento e distribuição de dados do PDB. Os membros são: RCSB PDB (USA), PDBe (Europa), PDBj (Japão) e BMRB (USA). A missão é manter um únicoA arquivo PDB de dados estruturais de macromoléculas de acesso livre e de domínio póblico.

1.4.3 DrugBank

Segundo e DrugBank (2014) e Wishart (2006), a base de dados DrugBank contém 4100 entradas de fármacos, incluindo 800 pequenas moléculas aprovadas pela FDA e disponíveis na forma de um serviço de consulta *biotech drugs*[27], bem como 3.200 moléculas candidatas a fármacos experimentais. Além disso, 14.000 sequências de proteínas alvos ligadas a essas moléculas, como por exemplo: <*idaccession*> (chave de acesso a caracterização das proteínas ou moléculas); <*quimicalids*>/<*het_id*> (chave de acesso aos ligantes).

Cada entrada *DrugCard*[28] contém mais de 200 campos de dados com metade da informação que está sendo dedicada ao composto e dados químicos, tais como: composição química, uso terapêutico e outros. A outra metade dedicada a caracterização de alvos de fármacos ou dados de proteínas. Muitos campos de dados são *hiperlinks*[29] a outros bancos de dados (KEGG, PubChem, ChEBI, PDB, Swiss-Prot e GenBank) e uma variedade de programas de computador que permitem visualizar a estrutura.

No DrugBank é possível realizar busca com o conteúdo textual, sequência, estrutura química e por meio de consultas relacionais. O uso e o potencial das aplicações do DrugBank incluem a descoberta de medicamentos *in silico*, a concepção de medicamentos, descobertas de alvos e fármacos (*docking*) ou a previsão, interação, metabolismo e o uso farmacêutico (DRUGBANK, 2014).

1.4.4 BioMed Central

O BioMed Central (BMC, 2014) é um repositório que pertence a *Springer Science + Business Media*, e também hospeda a plataforma *SpringerOpen*. O BioMed Central reúne 268 periódicos nos segmentos da Ciência, Tecnologia e Medicina. São periódicos revisados por pares e está sob a política de acesso aberto.[30] O portfólio de revistas abrange todas as áreas da Biologia, Biomedicina e Medicina e inclui títulos de interesses amplos como *BMC Biology* e *BMC Medicine* ao lado de revistas especializadas, como *Retrovirology* e *BMC Genomics*. Todos os

[27] http://www.drugbank.ca/drugs?type=biotech
[28] Subdivisão de campos.

[29] *hiperlink* é um link presente em um arquivo de hipertexto, página *html* ou documento que direciona para um outro local ou outro arquivo. Normalmente o *hiperlink* é ativado ao acionar, com o ponteiro do mouse ou tecla, uma palavra ou imagem destacada na tela.
[30] Publicação de acesso aberto é aquela que permite o livre acesso e distribuição de artigos publicados em que o autor detém o copyright do seu trabalho por meio de uma licença Creative Commons Attribution. Portanto, não há impedimentos para o acesso.

artigos originais de pesquisa publicados pela BioMed Central são disponibilizados *online* e com livre acesso imediatamente após a publicação.

Publicando com o BioMed Central os autores mantêm os direitos autorais do seu trabalho e, ao licenciá-lo sob *Licence Creative Commons Attribution,* os autores permitem que os artigos sejam reutilizados e redistribuídos sem restrição, desde que a obra original seja corretamente citada.

O BioMed Central é a fonte para que o *pipeline* busque e recupere informação sob a forma de artigo completo e foi utilizada para a formação do *corpus* textual cujo conteúdo é composto pela seção resultados de cada arquivo recuperado.

1.5 Aspectos Teóricos

Nesta seção, é apresentada a base teórica utilizada para esclarecer conceitos, definir termos e como fonte de inspiração para a construção do método.

1.5.1 Reuso

O termo reuso ou reutilização do conhecimento, isto é, a transformação do que está registrado em bases de dados textuais para um ambiente diferente do qual esses textos foram originalmente concebidos foi proposto por Markus (2001). É o processamento da base de conhecimento necessária, que é composta pela matéria-prima dos processos inovadores. O reuso da informação é o que auxilia nas melhorias de todos os sistemas que dela dependem.

O trabalho de Swanson et al. (2006) prevê o potencial para novas descobertas por meio da análise da literatura cientifica. A proposta do artigo é um método que a partir da análise de conjuntos disjuntos formada por registros da base de dados MEDLINE (A, C); o método produz uma lista de palavras do título e respectivas frases (B) o índice permite que especialista possam identificar e estabelecer as relações tipo A-B e B-C e relações indiretas A-C. Descritores como o Medical Subject Headings (MeSH) são utilizados como apoio para coesão textual na medida que o número de relações entre os conjuntos aumentar.

As declarações de James Black (1988, p.23) e Wermuth (2003) são pontos de partida para a construção do presente trabalho. No caso das práticas laboratoriais de reposicionamento de fármacos buscou-se um medicamento existente que sirva para um tratamento de uma patologia diferente da qual o fármaco promove a cura ou tratamento. Tem-se o velho para o novo fármaco. Por outro lado, Wermuth (2003)

colabora no presente trabalho para resultados destinados à pesquisa básica. O pesquisador propõe a busca por moléculas cuja estrutura, propriedades terapêuticas, níveis de toxidade testados e eventos adversos são bem conhecidos. Os autores apontam os caminhos mais vantajosos para se chegar a um novo fármaco.

1.5.2 Ontologias

Spasic et al. (2005) fazem uso de ontologias para mineração de textos cuja aplicação se dá no segmento da biomedicina. O método agrupa e relaciona termos, analisa as variações e elimina ambiguidades. As ontologias fornecem as descrições e respectivos conceitos biomédicos permitindo que sejam processados por computadores. Há uma interpretação textual a partir dos conceitos biomédicos, as suas relações com os termos de domínio específico e as anotações semânticas (campos descritivos anotados semanticamente). A camada semântica e as ontologias permitem a extração e interpretação de informação sobre os conceitos biomédicos uma vez que pertencem a uma ontologia (domínio de alta complexidade). Somou-se a essa metodologia recursos estatísticos sobre coocorrência entre as classes específicas dos termos biomédicos identificados.

No projeto EPIWORK, o esforço de pesquisa multidisciplinar foi concentrado no desenvolvimento de um quadro de ferramentas e conhecimentos apropriados para projetos de infraestruturas de previsão epidemia, não explorou amplamente o uso de ontologias porque nunca houve uma grande quantidade de recursos anotados para este fim. Contudo, durante o projeto, foi criado o (NERO) *Network of Epidemiology-Related Ontologies* (FERREIRA et al., 2012) que é, de maneira exata, uma coletânea de ontologias que são úteis na representação do conhecimento epidemiológicos. O NERO especifica, respectivamente para representação de fármacos e doenças, as ontologias (ChEBI) *Chemical Entities of Biological Interest* e (DOID) *Human Disease Ontology*. Ambas estão disponíveis, por exemplo, no portal do *OBO Foundry* (http://www.obofoundry.org/). (FERREIRA, 2014).

CHEBI: (http://www.obofoundry.org/cgi-bin/detail.cgi?id=chebi);
DOID: (http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology)

O ChEBI representa muito mais do que medicamentos, contém informação detalhada sobre entidades químicas de interesse biológico. Por outro lado, o DOID é uma ontologia antiga que não é atualizada desde 2011. Contudo, não se pode afirmar a existência de nada mais detalhado que seja de uso genérico para doenças, portanto, torna-se necessária a sua construção ou busca mais detalhada no segmento de ontologias médicas ou biomédicas. O (CDO) *Cardiovascular Disease Ontology* e o (IDO) *Infectious Disease Ontology* que estão bem mais desenvolvidas, fazem uso de uma ontologia de topo comum à maioria das ontologias biomédicas (BFO) *Basic Formal Ontology.* No entanto, por serem mais específicas, não abrangem todas as categorias de doença (enquanto que DOID tem 8000 classes, CDO e IDO têm apenas cerca de 500 classes cada uma). (FERREIRA, 2012, 2014).

O NCBO Annotator é um serviço Web que anota metadados em formato texto - os resumos de periódicos do PubMed, por exemplo - com conceitos relevantes obtidos por meio de ontologias. Atualmente, o fluxo de trabalho de anotação é baseado em um conceito altamente eficiente de reconhecimento sintático (usando nomes de conceito e sinônimos). O motor de busca desenvolvido em colaboração com o (NCIBI) - *National Center for Integrative Biomedical Informatics* e o conjunto de algoritmos de expansão semântica alavancam a semântica em ontologias. A metodologia oferecida pelo BioPortal NCBO utiliza ontologias para criar anotações de texto simples e devolve-los usando padrões da web semântica. (MUSEN et al., 2012)

O NCBO faz uso da Biblioteca de Ontologias OBO – Open Biological Ontologies que é considerado por Campos (2009) um repositório com a finalidade de armazenar e compartilhar terminologias no segmento biomédico e em outros domínios de interesse. Contudo, OBO reúne vários tipos de vocabulários reconhecidos no âmbito da informática e ciência da informação, que são: os vocabulários controlados, os tesauros e as ontologias. A autora ressalta que tais repositórios podem adotar na sua construção características genéricas e, portanto, serem aplicados a outros organismos ou temas correlatos caracterizando assim a sua interdisciplinaridade.

1.5.3 Anotação Semântica

Segundo Oren et al. (2006), a Web Semântica permite às máquinas interpretar, combinar e usar os dados na Web. Considerando que a Web só é compreensível para o ser humano, a Web Semântica pode ter seus dados processados por computadores. A base para a Web Semântica são as descrições de recursos, classes, termos, objetos, etc. Tais descrições são realizadas por meio de softwares de computador e têm como suporte metadados. A combinação viabiliza para homens e máquinas a interpretação das anotações feitas para esses recursos.

A pesquisa de Fontes (2011) busca agregar mais informação aos documentos textuais digitais com anotações semânticas. Tais anotações são obtidas das ontologias de domínio acrescidas da possibilidade de explorar inferências ontológicas e o conceito de meta-anotação. O objetivo é guiar os usuários no sentido de promover recursos quanto ao uso das anotações semânticas inferidas por meio da informação e raciocínio no qual foram concebidas. O autor propõe a construção de uma ferramenta de anotação denominada AutôMeta (*Automatic Metadata annotation tool*).

1.5.4 Mineração de Textos em Bioinformática

Os autores Hearst (1999) e Zweigenbaum et al. (2007) afirmam que há inúmeras definições para mineração de textos na área biomédica. Contudo, a partir da definição que é a de identificar um conhecimento que está implícito em determinado conteúdo textual ou com um sentindo mais amplo, a mineração de textos pode ser entendida como um sistema que permita relacionar, identificar e extrair informação de documentos textuais digitais. Por outro lado, os primeiros experimentos ocorreram com a extração de dados explícitos registrados eletronicamente em textos científicos. Nesse aspecto, segundo Krallinger e Valencia (2005), o uso da mineração de textos estava orientado para artigos científicos no segmento da Biologia Molecular.

A mineração de textos na área biomédica é realizada por meio de técnicas e métodos em dados textuais cuja aplicação se faz necessária quando se deseja buscar, recuperar, analisar e extrair informação. Segundo Spasic et al. (2005), o volume que é registrado nas bases de dados da literatura biomédica é crescente e torna difícil a tarefa de extrair informação sem utilizar a mineração de textos.

Outra abordagem é o uso dos resumos para extrair informação. Os autores Chun, H. et al. (2005) e Zhou (2004) exploram o MEDLINE que é uma base de dados da literatura internacional da área médica. Por meio dos resumos, a informação é obtida com marcadores sob a forma de rótulos que identificam classes gramaticais. Por outro lado, Jezuz (2013) explora os resumos do PubMed para Mineração de textos científicos visando à identificação de componentes bioativos com potencial terapêutico para o tratamento de Dengue, Malária e Doença de Chagas.

Métodos de agrupamento são utilizados por Wives (1999). O autor, na sua tese, apresenta o método para agrupar documentos textuais digitais conforme a sua similaridade. O método permite obter um número reduzido de descritores que representam o conjunto similar. O resultado visa obter conceitos (tabela de descritores) que representem o(s) conjunto(s) similar(es), o que reduzirá, segundo o autor, as interferências decorrentes de problemas da própria linguagem e do vocabulário utilizado na construção dos documentos textuais digitais. Os descritores obtidos com o método auxiliam na exploração e análise dos dados, o que facilita para os homens e as máquinas estabelecerem relações e reconhecerem o conteúdo do conjunto textual cuja metodologia seja aplicada.

A próxima subseção descreve alguns softwares relacionados com o tema da pesquisa realizada na presente dissertação. São softwares e serviços disponíveis que fazem uso de recursos de mineração de textos e ontologias com objetivo de propor soluções para problemas específicos em vários segmentos da área médica e biomédica.

1.5.5 Softwares relacionados

BioPortal Annotator – fornece acesso a ontologias biomédicas de uso corrente e as ferramentas para manipular e trabalhar com essas ontologias. O BioPortal permite navegar-se pelas bibliotecas de ontologias, mapear termos em diferentes ontologias, buscar um termo em diferentes ontologias além de receber recomendações sobre quais ontologias, no domínio biomédico, são mais relevantes para determinado corpus textual (MUSEN et al., 2012).

AutôMeta (*Automatic Metadata annotation tool*) – construída como um mecanismo de anotação semântica multiplataforma e multi-interface. Segundo Fontes (2011), o software permite recuperar, a partir de um documento texto, anotações simples e múltiplas anotações (anotações em lote). A arquitetura do AutôMeta contempla mecanismos para anotação semântica em ambiente Web; anotação automática e semiautomática de um documento texto a partir de uma ontologia desenvolvida pelo usuário; visualizar estrutura da anotação em RDF; extrair informação de anotações semânticas de um documento textual digital na Web.

GOAnnotator – O principal objetivo da ferramenta é fornecer anotações semânticas Gene Ontology (GO) de alta qualidade para as proteínas registradas na base de conhecimento UniProt. A aplicação usa similaridade semântica para verificar as anotações de proteínas reconhecendo termos GO automaticamente extraídos da literatura (COUTO et al., 2006).

Terminizer – A proposta do Terminizer é ajudar o usuário da aplicação no uso de ontologias. O serviço detecta termos ontológicos em fragmentos de texto tipo publicações ou anotações experimentais, frases ou termos isolados. Uma ontologia é o passo acima de um vocabulário controlado, uma coleção de termos de determinado domínio como o da anatomia: osso, perna, etc.

O Terminizer faz uso de ontologias para anotar textos o que torna mais fácil para pessoas e softwares compreender e processar, respectivamente, o texto. Segundo os autores, o uso de ontologias colabora na melhoria dos processos de busca por documentos se as relações ontológicas são utilizadas. A busca começa por meio de um termo específico e a ontologia fica encarregada de estabelecer as possíveis relações. Isto é factível porque o software é capaz de descobrir que um documento contém o termo ou categoria desejada a partir do conhecimento contido na ontologia (NERC, 2015).

LAITOR – *Literature Assistant for Identification of Terms co-Occurrences and Relationships,* (BARBOSA-SILVA et al., 2010). Elaborado estrategicamente para mineração de textos o software LAITOR foi desenvolvido para normalizar nomes de bioentidades previamente definidas em um dicionário de dados proteicos. O algoritmo do software está baseado na coocorrência das entidades biológicas

(genes e proteínas) ao longo do texto. Tais coocorrências são extraídas levando em consideração a presença de termos em uma mesma frase contidas em resumos científicos. Há um conjunto de regras para filtrar pares de bioentidades que ocorrem em várias estruturas textuais e em cada sentença.

PESCADOR – Platform for Exploration of Significant Concepts Associated to co-Occurrences Relationships – A ferramenta web extrai uma rede de interações a partir de um conjunto de resumos do PubMed que são fornecidas por um usuário. O PESCADOR é capaz de filtrar a rede de interação em conformidade com os termos e/ou conceitos que foram previamente definidos pelo usuário (BARBOSA-SILVA, 2011; PESCADOR, 2011).

PubTator – PubTator é uma ferramenta baseada na Web para acelerar procedimentos manuais de manutenção, coleta e arquivamento de ativos digitais (por exemplo, a anotação de entidades biológicas e suas relações) por meio do uso de técnicas avançadas de text mining. PubTator fornece um serviço prático para anotar citações do PubMed, mantém um compromisso de sincronização com a base de referências do repositório PubMed e busca atualizar anotações de forma automática todos os dias (WEI et al., 2013).

1.6 Justificativa

1.6.1 Do ponto de vista científico e das políticas de saúde pública

A proposta apresentada nesta dissertação justifica-se por atender a um esforço das nações a partir de uma ação conjunta firmada em um acordo na Organização das Nações Unidas (ONU) denominado Objetivos do Milênio ONU 2000[1]. Dentre outras metas estabelecidas, destaca-se o combate a AIDS, malária e outras doenças que, no caso do presente trabalho, serão reconhecidas no âmbito da pesquisa como doenças negligenciadas.

Por outro lado, o aumento da produção científica registrada em bases de dados, repositórios de acesso livre e os serviços oferecidos por meio da rede mundial de computadores obriga o Estado, diante desse processo de transformação social, a traçar diretrizes conforme apontamentos do Ministério do Planejamento, Orçamento e Gestão (BRASIL, 2010).

> A rede mundial tornou-se um desafio para as empresas, instituições e organismos do governo em todo o mundo e não há como escapar desse processo de transformação da sociedade. Para todos aqueles que tiverem meios de acesso, as informações são diversas, públicas e gratuitas e, para os que não têm, o Estado assume um papel muito importante, voltado para a democratização do acesso à rede e a prestação eficiente de seus serviços aos cidadãos, usando as tecnologias de informação e comunicação (TIC's) (BRASIL, 2010).

Diante do contexto observado, há a possibilidade de empregar métodos e teorias já consolidadas no mundo científico, com objetivo de explorar bases de dados compostas por documentos textuais digitais e combinar outros métodos, reforçando o quadro interdisciplinar que compreende os estudos no campo da Biologia Computacional e Sistemas.

1.6.2 Do ponto de vista econômico financeiro

Encontrar proteínas com estruturas ou funções similares é um meio de oferecer alternativas de baixo custo para propor novas aplicações no uso de fármacos já existentes, reduzir os custos envolvidos na pesquisa básica e principalmente estimular laboratórios a produzir novos medicamentos.

O desenvolvimento de fármacos é um processo que demanda tempo e envolve quantias vultosas. As empresas farmacêuticas dedicam uma média de 12 a 15 anos (SLOAN, 2008) de pesquisa e cerca de US$500 milhões em investimentos para ir, a partir da descoberta de uma molécula promissora, para a comercialização de um medicamento (OLIVA, 2003).

O fluxo, na Figura 5, a seguir, mostra a cadeia de pesquisa e desenvolvimento de fármacos e medicamentos, a complexidade envolvida desde a pesquisa básica até o produto final e dá uma breve noção de tempo para a conclusão do processo.

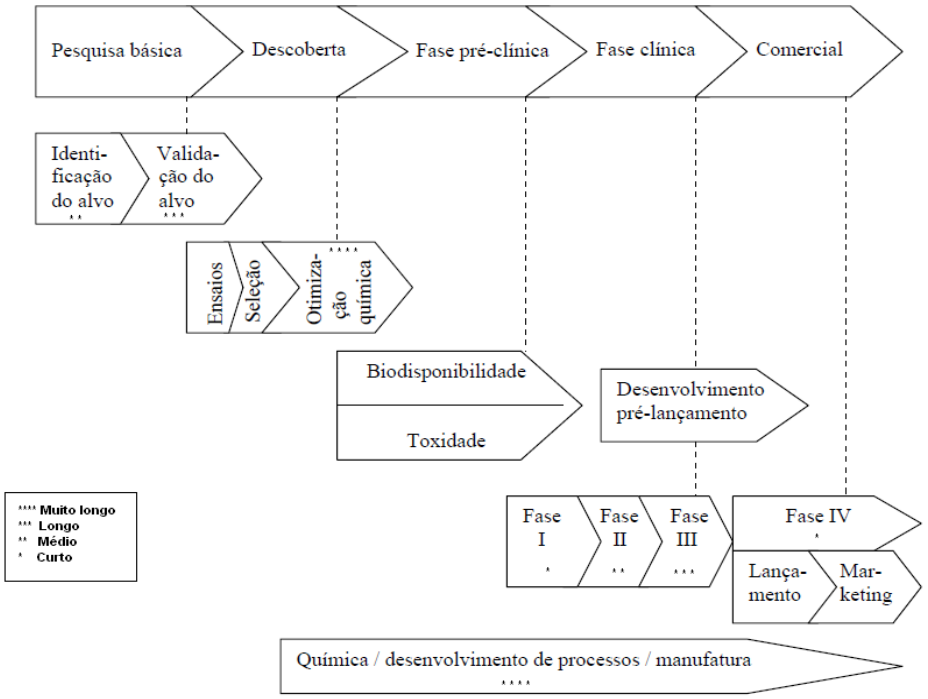


Figura 5 - Cadeia de pesquisa e desenvolvimento de fármacos e medicamentos.
Fonte: (Vieira, VMM, Ohayon P[31]. 2000)

Há uma dissociação entre a indústria farmacêutica, investidores empresários e grupos de pesquisa. No caso do Brasil, os investidores não possuem a cultura ou costume de empreender ou financiar em longo prazo (PALMEIRA, 2012); os resultados oferecidos pelos grupos de pesquisas não traduzem um produto rentável para o mercado de doenças negligenciadas, não dá lucro. Soma-se a isso a ausência de órgãos oficiais, a falta de políticas públicas, a falta de programas de prevenção de risco e doenças. Inclui-se nesse arrazoado o bom emprego e origem do termo negligenciada.

Muito mais grave, porém, é essa frágil articulação entre os agentes econômicos envolvidos no Sistema de Inovação e a falta de políticas públicas voltadas para mecanismos inovadores que possam atrair investimentos à Pesquisa e Desenvolvimento (P&D) (GADELHA, 2003). Isso faz do Brasil um país menos competitivo dadas as seguintes circunstâncias: limitações financeiras impostas pelo uso de insumos sob patentes; optar por uma posição econômica coadjuvante; postura tradicional na divisão internacional do trabalho; muito distante das atividades

[31] Teste pré-clínico de medicamentos em animais experimentais ou *in vitro*, para seus efeitos biológicos e tóxicos e aplicações clínicas potenciais.

inerentes à produção de alta tecnologia, patentes e insumos empregados em pesquisas futuras (QUEIROZ: GONZÁLEZ 2001 apud VIEIRA, 2014).

Portanto, o presente trabalho explorou a literatura científica indexada no PubMed buscando nomes de proteínas relacionadas a doenças negligenciadas, já que ligar essas proteínas a outros bancos de dados e identificar similares para reposicionamento, principalmente na pesquisa básica, é um caminho para a redução de custos envolvidos na produção de novos medicamentos, além de ser uma possibilidade para reposicionar um fármaco já existente com seus eventos adversos bem conhecidos.

O reuso de tecnologias e teorias consolidadas são pontos fortes no presente trabalho e um caminho para alcançar o objetivo desejado. Pode-se citar, como exemplo, o trabalho pioneiro de Swanson (1986, 1987) que, por meio de uma simples contagem de palavras que estabelecem um *ranking*, anteviu a importância dos métodos matemáticos e quantitativos nas relações entre conjuntos a partir dos dados obtidos nas bases de dados MEDLINE[32] X *Medical Subject Headings* (MeSH) (SWANSON, 2006)[33].

Reutilizar esse conhecimento para caracterizar proteínas e seus elementos, identificar a estrutura proteica e as relações entre elas e, principalmente, obter resultados e aplicá-los no reposicionamento de medicamentos é um dos fatores que estimula, motiva e torna relevante o presente trabalho.

Assim, de acordo com a abordagem discutida nesta dissertação, é importante investigar métodos para recuperar, organizar e extrair, no mínimo de forma semiautomática, a partir do texto digital, nomes de proteínas para reposicionamento de medicamentos em doenças negligenciadas.

1.7 Organização do trabalho

Esta dissertação de mestrado foi elaborada no formato de 5 capítulos que estão dispostos da seguinte maneira:

No Capítulo 2 estão definidos os objetivos do presente trabalho. A descrição detalhada dos materiais e dos métodos empregados na construção do *pipeline* é feita no Capítulo 3.

[32] MEDLINE contém citações de periódicos e resumos de literatura biomédica de todo o mundo.
[33] MeSH Natural Language MeSH NLM é um vocabulário controlado thesaurus utilizado para indexação de artigos para PubMed.

Com base nos resultados obtidos por meio do processamento dos artigos científicos do PubMed para o nome da doença malária, o Capítulo 4 corresponde à apresentação dos resultados, sob a forma de evidências para a confirmação das questões da pesquisa e dos objetivos alcançados.

No Capítulo 5 é detalhado as conclusões do trabalho e sugestões para a continuidade da pesquisa e para os novos trabalhos.

2 OBJETIVOS

2.1 Objetivo Geral

Criar um método que permita, a partir da literatura científica disponível no PubMed, a indicação de fármacos para o reposicionamento, por meio da identificação de nomes das doenças associadas aos nomes de proteínas e os ligantes aos fármacos correspondentes.

2.2 Objetivos Específicos

- Automatizar procedimentos para buscar e recuperar documentos textuais digitais no PubMeb;

- Explorar ontologias no domínio biomédico com objetivo de prover termos referentes a nomes de proteínas ou termos associados a esses nomes.

- Relatar proteínas que possam contribuir para procedimentos realizados na pesquisa básica, isto é, o *screening* de moléculas[34].

[34] Vieira, VMM, Ohayon P. Inovação em fármacos e medicamentos: estado-da-arte no Brasil e políticas de P&D adaptado do original The Pharma R&D Values Chain. In: Global Alliance for TB Drug Development. The Boston Consulting Group, 2000.

3 MATERIAL E MÉTODOS

O ponto inicial da metodologia consiste no entendimento do objeto de pesquisa que se dá por meio de uma revisão bibliográfica, além do conteúdo das disciplinas ministradas no curso de mestrado em Biologia Computacional e Sistemas – IOC/FIOCRUZ.

Os aspectos técnicos da metodologia serão apresentados na sequência desse texto. Em seguida, serão descritos os procedimentos executados no *pipeline* que se encontram divididos em quatro partes distintas: recuperar informação; reorganizar os dados com anotação semântica; processos de mineração de textos; categorizar e resultados.

O experimento para testes com este método consiste em executar o *pipeline* para buscar, recuperar e processar os dados relacionados ao nome da doença negligenciada denominada malária. Portanto, os resultados apresentados são referentes ao termo *malaria.*

A metodologia do presente trabalho é desenvolvida utilizando a linguagem de programação e o ambiente denominados R (R FOUNDATION, 2002).

Para os processos de mineração de textos (etapa 3) foram utilizadas como base os *scripts* escritos em linguagem R e a estrutura de tabelas do banco de dados PostgreSQL ambos disponíveis em Jezuz (2013). Algumas dessas rotinas foram ajustadas para se tornarem adequadas aos processos do *pipeline*.

A seguir, a Figura 6 representa o *pipeline* e procura demonstrar de forma sucinta o seu funcionamento.

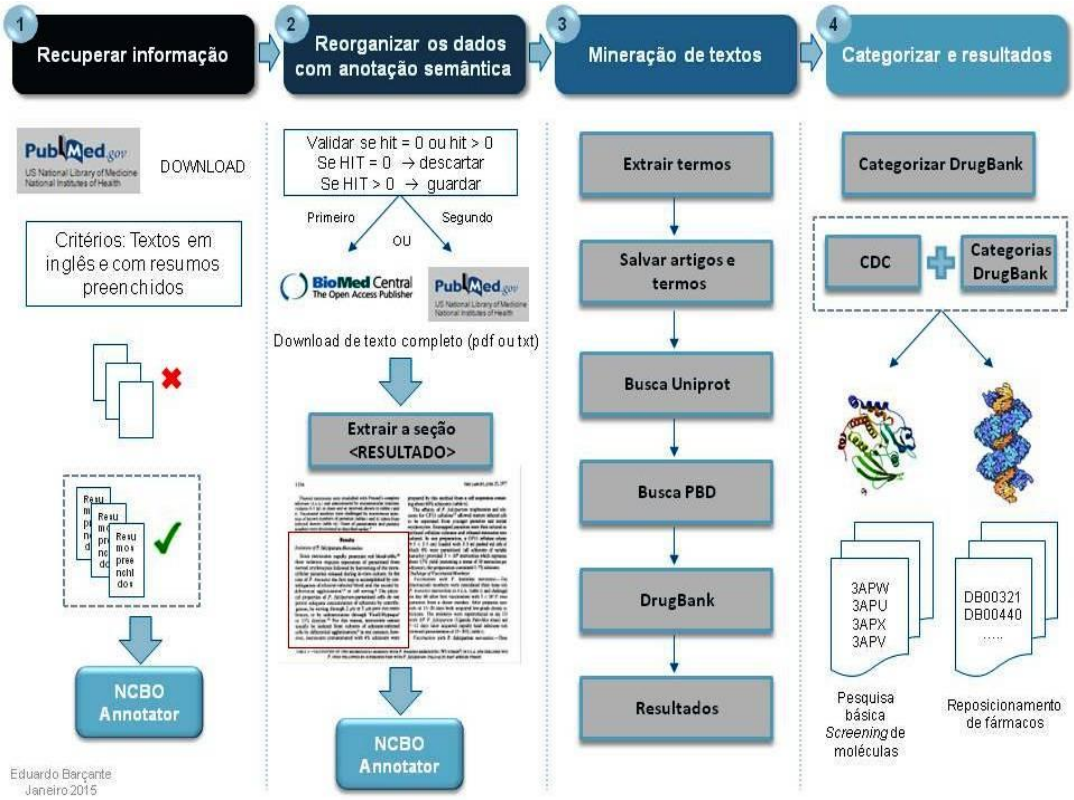


Figura 6 - Etapas da metodologia.
Fonte: Elaborado pelo autor.

3.1 Parte 1 – Recuperar informação

Por tradição, quando sistemas informatizados recuperam informação estão apoiados em um conjunto de palavras ou termos que estão à disposição para exprimir algo de forma oral ou escrita, isto é, o léxico. Contudo, o que é usado para uma consulta pode ser uma única palavra-chave ou um conjunto delas (CAMPOS; GOMES, 2008). Tais palavras são fornecidas pelo leitor que deseja encontrá-las nos metadados usados para descrever um objeto que, no caso do presente trabalho, é um texto digital.

O nível de complexidade está de acordo com a formalidade no qual o documento textual foi produzido, o que exige, de certa forma, um aporte de tecnologias que possam auxiliar na busca e recuperação da informação contida nesses documentos. No caso dessa dissertação, são os nomes de proteínas, e buscou-se auxílio na formalidade das relações e anotações semânticas obtidas por meio de uma ontologia.

Nas subdivisões deste tópico são descritos: os termos empregados na busca, origem do documento textual, o domínio, a ontologia e a abordagem para buscar e recuperar resumos e artigos científicos completos no PubMed.

3.1.1 Origem dos dados

A principal fonte de dados escolhida foi a base de dados PubMed, que é composta por mais de 24 milhões[35] de citações da literatura biomédica do MEDLINE, periódicos científicos e livros *online* (PUBMED, 2013). Quando possível, as citações podem incluir *links* para conteúdo de texto completo da PubMed Central e *sites* de editores.

Contudo, o acesso ao texto completo requer o redirecionamento para as corporações ou bancos de dados dos editores, mediante o pagamento de uma taxa, quando não há acordos financeiros para permitir acesso. Por isso, foi explorado automaticamente as bases de dados PubMed Central[36] e BioMed[37] que estão sob a iniciativa *Open Archives Initiative* (OAI)[38]. De outra maneira, buscou-se arquivos PDF em outros bancos de dados e *links*.

3.1.2 Termos

Os termos são as palavras-chave, palavras correlatas ao tema ou assunto relacionado ao objeto de busca, isto é, a palavra que é usada como parâmetro de busca aos resumos dos artigos[39] ou artigos completos[40] na base de dados PubMed. Baeza (1995) o define como:

> Um *termo de indexação* é uma palavra ou um grupo de palavras consecutivas em um documento. Em sua forma mais geral, um termo de indexação é qualquer palavra da coleção. Essa é a abordagem utilizada pelos projetistas de máquinas de busca. Em uma interpretação mais restrita, os termos de indexação são grupos de palavras pré-selecionadas que representam conceitos-chave (ou tópicos) em um documento. Essa é a abordagem adotada por bibliotecários e por cientistas da informação (BAEZA, 1995).

[35] No dia 15 de dezembro de 2014 o PubMed retorna um total de 24.240.347 resumos indexados.
[36] PubMed Central® (PMC) é um repositório de periódicos completos em formato texto contemplando Ciências Biomédicas e da Ciências da Vida. São artigos científicos mantidos por *U.S. National Institutes of Health's National Library of Medicine (NIH/NLM).*
[37] BioMed Central é um repositório para Ciência, Tecnologia e Medicina com 273 periódicos de acesso aberto e revistos por pares.
[38] Open Archives Initiative (OAI) desenvolve e promove padrões de interoperabilidade que visam facilitar a disseminação eficiente de conteúdos.
[39] Arquivos cujo conteúdo possuem um marcador denominado *abstract*.
[40] Arquivos em formato texto ou pdf cujo conteúdo traz todo o artigo publicado.

Nessa dissertação foram empregados os termos que dão nome às doenças: *chagas disease*, *dengue, leishmaniasis* e *malaria.* São esses quatro termos que orientaram a busca e a recuperação dos resumos publicados no PubMed. A seguir, na Figura 7, há um exemplo do termo denominado *malaria* e o resultado da busca na base de dados MeSH.

O Medical Subject Headings (MeSH) (NIH, 2014) é um vocabulário controlado utilizado para indexar os artigos científicos no PubMed. A base de dados MeSH permite identificar um termo adequado para uma buscar e recuperar artigos científicos no PubMed. Nessa dissertação, os termos empregados para buscar e recuperar resumos e artigos relacionados às doenças negligenciadas têm a sua origem do vocabulário controlado MeSH. No Anexo II, seguem os dados obtidos no MeSH para *chagas disease, dengue* e *leishmaniasis.*

Os termos empregados na busca, os artigos científicos recuperados (PubMed), os artigos processados e os resultados do *pipeline* foram escritos originalmente na língua inglesa. Portanto, existem relações estabelecidas entre as fases de busca até os resultados que também foram escritos automaticamente na língua inglesa. Por isso, traduzir o conjunto para o português não retrataria com fidelidade o que foi processado e tornaria impossível repetir o experimento realizado. Diante do exposto, o conteúdo de tabelas e figuras, em alguns casos, foi mantido na língua inglesa.

MeSH Heading	Malaria
Tree Number	C03.752.530
Annotation	GEN or unspecified; specify Plasmodium species IM if possible but note P. falciparum malaria = MALARIA, FALCIPARUM; P. vivax malaria = MALARIA, VIVAX; tertian malaria = MALARIA, VIVAX, quartan malaria: coord IM with PLASMODIUM MALARIAE (IM); malariotherapy = HYPERTHERMIA, INDUCED: do not confuse with MALARIA / ther; / drug ther: consider also ANTIMALARIALS
Scope Note	A protozoan disease caused in humans by four species of the PLASMODIUM genus: PLASMODIUM FALCIPARUM; PLASMODIUM VIVAX; PLASMODIUM OVALE; and PLASMODIUM MALARIAE; and transmitted by the bite of an infected female mosquito of the genus ANOPHELES. Malaria is endemic in parts of Asia, Africa, Central and South America, Oceania, and certain Caribbean islands. It is characterized by extreme exhaustion associated with paroxysms of high FEVER; SWEATING; shaking CHILLS; and ANEMIA. Malaria in ANIMALS is caused by other species of plasmodia.
Entry Term	Infections, Plasmodium
Entry Term	Marsh Fever
Entry Term	Paludism
Entry Term	Plasmodium Infections
Entry Term	Remittent Fever
See Also	Antimalarials
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH TM UR US VE VI
Online Note	use MALARIA to search MALARIA CONTROL 1966
History Note	MALARIA CONTROL was heading 1963-66
Date of Entry	19990101
Unique ID	D008288

Figura 7 - Dados associados a um termo recuperado numa consulta *online.*
Fonte: Base de dados MeSH.

3.1.3 A busca por resumos científicos

Uma vez definido(s) o(s) termo(s) de busca, foi feito uso da ferramenta de computação R (R FOUNDATION, 2002) e da biblioteca RisMed[41] com objetivo de obter, por meio de *download*, o maior número de resumos disponibilizados na base de dados PubMed. A Tabela 1 mostra as formas de consulta por meio da API[42] *entrez* e a Tabela 2 mostra o total de resumos indexados no PubMed.

[41] Kovalchik S. Package 'RISmed'. Biblioteca destinada, dentre outros procedimentos, a recuperar o conteúdo de bancos de dados do NCBI e a facilitar a análise de conteúdo do banco de dados do NCBI. [acesso em 22 abr 2014]. Disponível em: http://cran.r-project.org/web/packages/RISmed/index.html.
[42] Application Programming Interface (Interface de programação de aplicativos).

Tabela 1 - Consultas aos resumos por meio da API *entrez*.

Consulta
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=PubMed&term=(chagas disease)
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=PubMed&term=dengue
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=PubMed&term=leishmaniasis
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=PubMed&term=malaria

Tabela 2 - Total de resumos indexados no PubMed em 1/10/2014.

Termos	Resumos
Nome das doenças	Indexados (PubMed)
chagas disease	13.256
Dengue	12.739
Leishmaniasis	20.714
Malaria	71.220

3.2 Parte 2 – Reorganizar dados com anotação semântica

Definidos os termos e computados os totais dos resumos disponíveis no PubMed, é necessário trazer o conteúdo desses resumos para o ambiente de processamento em R.

Existem dois tipos de objetos ou serviços que retornam dados do PubMed no formato XML, que são: *eSearch* e *eFetch*. (SAYERS, 2008, 2010) O primeiro deles é um serviço de busca da *Web* denominado *eSearch* que essencialmente retorna os identificadores no PubMed a partir de um padrão de busca específica, documentado no *site* do NCBI eUtils. Como exemplo, tem-se a seguinte mensagem HTTP:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=PubMed&term=%28malaria%29+AND+%282001/01:2010/12%20[dp]%29&retmax=500

A instrução acima retorna os 500 primeiros registros com data de publicação compreendida no período de 2001/01 até 2010/12 no formato XML. Contudo, o resultado obtido com a instrução *eSearch* não é suficiente para o propósito desse trabalho. Além do código identificador do PubMed (PMID), é necessário o conjunto de dados ou o metadado do documento. O conteúdo desse metadado poderá ser recuperado individualmente ou em um conjunto de documentos por meio da instrução denominada *eFetch*.

O serviço *eFetch* retorna toda a informação armazenada para cada identificador, como: pmid, resumo e outros. No exemplo, a mensagem http retorna toda a informação relacionada ao PMID 24479197.

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=24479197&retmode=xml

No ambiente R, os serviços são combinados por meio da biblioteca RisMed, em uma única função. O resultado dessa função é o conteúdo do metadado disposto em vários objetos. Por exemplo:

query = "(chagas disease)+AND+(1800/01:1910/12 [dp])"

EUtilsSummary(query,type="esearch",db="pubmed",url=NULL,encoding="unknown",...)

chagas = EUtilsGet(x,type="efetch",db="pubmed")

Sob a forma de um vetor, a Tabela 3 contém a informação da nova estrutura adotada para o processamento do metadado. O resultado foi obtido por meio dos parâmetros *query* e das funções disponibilizadas na biblioteca RisMed e corresponde ao registro de número 1500 na lista de objetos.

Tabela 3 - Metadado reorganizado no ambiente R.

Lista de objetos e descrição	Conteúdo
MALARIA$PMID[1500] (Indentificador do PubMed)	"15328406"
MALARIA$TI[1500] (Título)	"Meager genetic variability of the human malaria agent Plasmodium vivax."
MALARIA$AB[1500] (Resumo)	"Malaria is a major human parasitic disease caused by four species of Plasmodium protozoa. Plasmodium vivax, the most widespread, affects millions of people across Africa, Asia, the Middle East, and Central and South America. We have studied the genetic variability of 13 microsatellite loci in 108 samples from 8 localities in Asia, Africa, South America, and New Guinea. Only one locus is polymorphic; nine are completely monomorphic, and the remaining three are monomorphic in all but one or two populations, which have a rare second allele. In contrast, Plasmodium falciparum displays extensive microsatellite polymorphism within and among populations. We further have analyzed, in 96 samples from the same 8 localities, 8 tandem repeats (TRs) located on a 100-kb contiguous chromosome segment described as highly polymorphic. Each locus exhibits 2-10 alleles in the whole sample but little intrapopulation polymorphism (1-5 alleles with a prevailing allele in most cases). Eight microsatellite loci monomorphic in P. vivax are polymorphic in three of five Plasmodium

	species related to P. vivax (two to seven individuals sampled). Plasmodium simium, a parasite of New World monkeys, is genetically indistinguishable from P. vivax. At 13 microsatellite loci and at 7 of the 8 TRs, both species share the same (or most common) allele. Scarce microsatellite polymorphism may reflect selective sweeps or population bottlenecks in recent evolutionary history of P. vivax; the differential variability of the TRs may reflect selective processes acting on particular regions of the genome. We infer that the world expansion of P. vivax as a human parasite occurred recently, perhaps <10,000 years ago."
MALARIA$LA[1500] (Idioma)	Eng

Uma lista de objetos no ambiente R é um objeto constituído por uma coleção ordenada de objetos conhecidos como os seus componentes. Não é necessário que os componentes sejam do mesmo modo ou tipo. Uma lista poderia consistir de um vetor numérico, um valor lógico, uma matriz, um vetor complexo, uma matriz de caracteres, uma função, e assim por diante. Os componentes de uma lista podem ser nomeados e o conteúdo do componente referido por: nomedalista[[1]] ou nomedalista$componente (VENABLES, 2013).

Transferido o metadado do PubMed para o ambiente R com o conteúdo semiestruturado, realizou-se a primeira triagem, uma vez que o recorte estabelecido para esse trabalho contempla apenas textos em inglês, MALARIA$LA igual a "Eng" e os resumos devidamente preenchidos MALARIA$AB diferente de um conteúdo vazio ou em branco (ver Tabela 3).

Essa regra se estende a todos os objetos de todos os nomes de doenças empregados no trabalho. São os seguintes objetos:[43] CHAGASDISEASE, para doença de chagas, DENGUE, para dengue; LEISHMANIASIS para leishmaniasis e MALARIA, para malária.

O próximo passo foi identificar quais são os resumos indexados que podem oferecer um *link* cujo conteúdo do texto completo contenha os nomes de proteínas desejados. A tarefa seguiu apoiada por uma ontologia, mais especificamente a ontologia *Protein Ontology* (PR) disponibilizada no portal *The National Center for Biomedical Ontology* (NCBO).

[43] Os objetos são variáveis no ambiente R que armazenarão o conteúdo recuperado no PubMed. No caso dos resumos obtidos na busca por doenças negligenciadas estarão em letras maiúsculas representando o respectivo nome da doença.

3.2.1 Resumos candidatos

Recuperados os resumos do PubMed, a continuidade do *pipeline* requer, nessa fase de processamento, uma pré-seleção dos resumos candidatos a fornecerem um texto completo que tenha no seu conteúdo, mais especificamente na seção resultados, nomes de proteínas.

Para a pré-seleção dos resumos utilizou-se o serviço de anotação semântica do *NCBO Annotator* (MUSEN et al., 2012) e a ontologia *Protein Ontology* (PR). O *annotator* é uma ferramenta que marca o texto livre com os termos anotados de uma ou várias ontologias. Disponibilizado para uso no BioPortal Annotator[44] ou por meio do serviço *online*, o *pipeline* para anotação semântica da Figura 8, a seguir, é fundamentado no conceito de reconhecimento sintático (usando um nome preferido pelo usuário e os sinônimos para os termos) e um conjunto de programas e algoritmos de expansão semântica (JONQUET, 2008) que aproveitam a estrutura da ontologia com relações tipo *is_a*[45], por exemplo.

O critério adotado para validar os resumos recuperados no PubMed foi: uma anotação simples ser retornada a partir da consulta à ontologia *Protein Ontology (PR)*. Os termos da consulta são palavras originárias do resumo e fornecidas como parâmetro ao serviço. Uma ocorrência é gerada se um termo anotado semanticamente (FILHO, 2010) for retornado (Anexo III), isto é, a ocorrência de um termo anotado é um *hit>0*.

[44] http://bioportal.bioontology.org/annotator#
[45] Transitividade da hierarquia Is_A toda classe deverá ser mapeada para todas as suas classes antecessoras. (Filho, 2010).

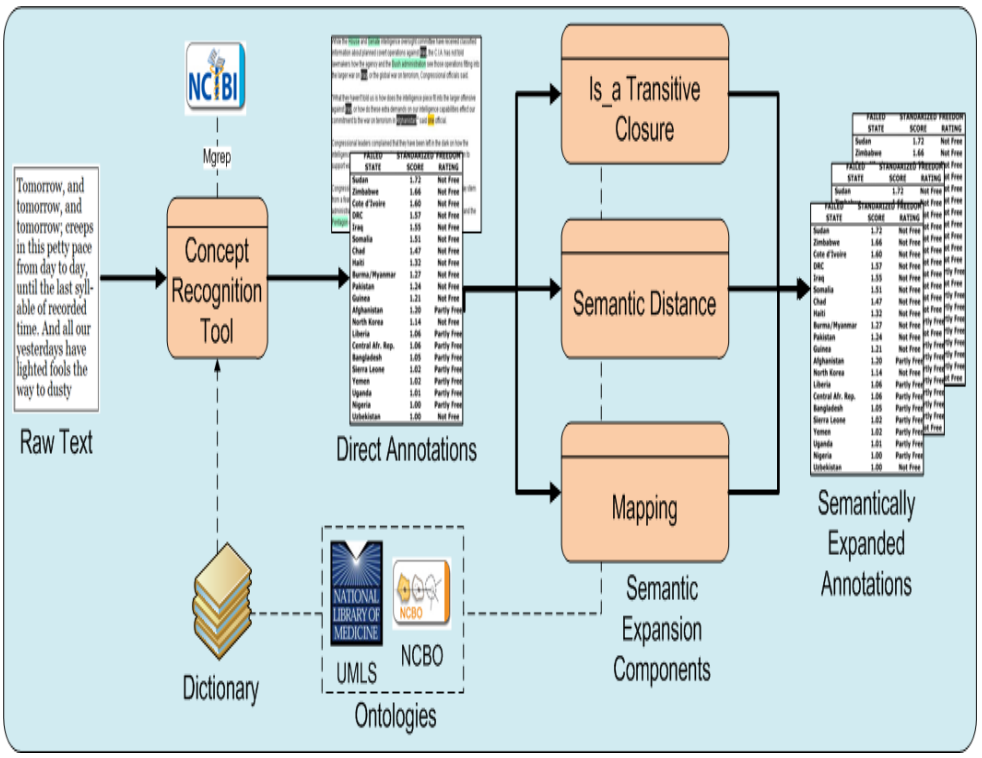


Figura 8 - *O pipeline para anotação semântica do serviço Web Annotator.*
Fonte: *NCBO Annotator (*Jonquet, 2008).

Um *hit>0* torna o resumo válido. No caso de um *hit=0* significa que não há anotação semântica para o termo ou conjunto de termos passado como um parâmetro ao serviço. Portanto, para um *hit=0* o resumo será descartado.

O procedimento é repetido para todos os resumos recuperados do PubMed até que não haja mais nenhum para validação.

Um novo conjunto está formado e contém resumos devidamente preenchidos, registrados com língua inglesa e validados por meio da ontologia *Protein Ontology* (PR). O *pipeline* passa à próxima etapa que é buscar e recuperar os artigos completos indexados pelo conjunto de resumos válidos.

3.2.2 Artigos completos

Essa seção trata da busca e recuperação dos artigos com o texto completo, uma vez que a formação de um novo *corpus* textual exige o conteúdo da seção resultados de cada artigo científico.

As principais fontes para o acesso aos arquivos estão indexadas no PubMed e, por meio dos *links*, o usuário chega ao repositório que detém a informação ou dá o acesso direto ao arquivo no formato texto ou pdf (PUBMED HELP, 2014). No entanto, os índices apontam para outros repositórios que detêm os artigos na sua forma completa. Então, eles serão explorados na seguinte ordem: o

BioMed Central (arquivos com acesso aberto[46]); arquivos recuperados no formato texto ou pdf (recuperados automaticamente pelo *pipeline*).

Uma vez recuperado o arquivo na sua forma completa, o *pipeline* verifica se é preciso converter o arquivo para o formato texto. Foi admitido o formato texto e o formato pdf para busca e recuperação de arquivos. Portanto, todos os arquivos no formato pdf estão alocados em um diretório e uma instrução é executada para convertê-los para o formato texto segue:

```
setwd("nome do diretório com arquivos no formato PDF")
system("for f in *.pdf; do pdftotext -enc ASCII7 -nopgbrk $f; done")
```

Todos os arquivos pdf localizados no diretório informado em *setwd* (“nome do diretório”) são convertidos automaticamente para o formato texto pelo procedimento fornecido em *system* ("for f in *.pdf; do pdftotext -enc ASCII7 -nopgbrk $f; done").

3.2.3 O texto na seção resultados

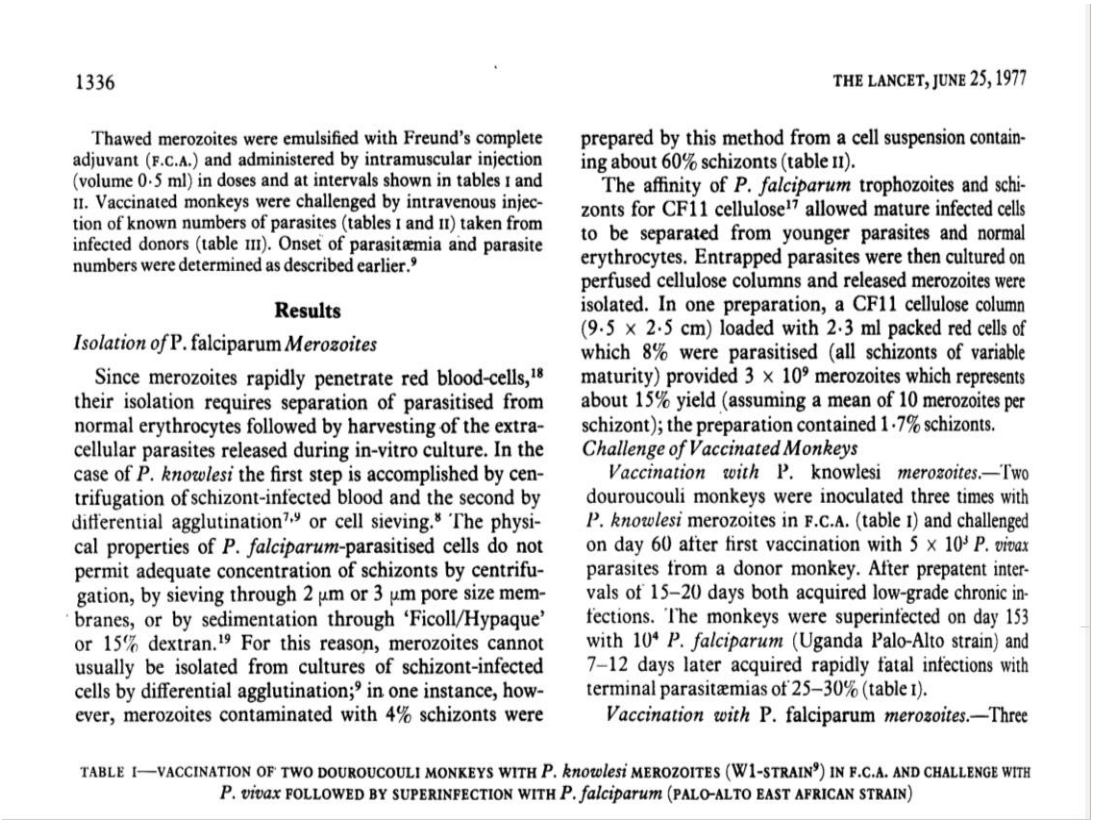
O *pipeline* agora precisa destacar a seção resultados de cada arquivo. O pressuposto é que a seção resultados contenha os termos necessários para que sejam extraídos nomes de proteínas válidos. Nomes de proteínas que levem a um composto bioativo com determinada função terapêutica, por exemplo.

Por outro lado, se o *pipeline* considerar todo o texto do artigo, a seção com a descrição dos métodos, por exemplo, contemplaria reações químicas associadas ao método empregado no artigo. Seriam considerados termos como catalisadores que visam acelerar ou parar uma reação química. São dados que podem não estar relacionados ao tema central do artigo e, portanto, não interessam ao propósito deste trabalho.

A seção resultados, como mostra a Figura 9, a seguir, é uma parte do texto completo, onde a identificação e extração dessa seção para posterior processamento permite uma redução no tempo de execução dos procedimentos subsequentes porque o fragmento de texto a ser processado é menor e traz termos relacionados ao tema central e as respostas aos problemas propostos.

[46] Publicação de acesso aberto permite o livre acesso e distribuição de artigos publicados em que o autor detém o copyright do seu trabalho através do emprego de uma licença Creative Commons Attribution, removendo Portanto, quaisquer barreiras ao acesso.

O critério adotado na construção do *pipeline* para extração da seção resultados é a identificação de um marcador e a palavra *results* com as suas possíveis variações. Caso não exista um marcador ou não exista a possibilidade de identificá-lo, todo o texto completo será considerado para o processamento.



1336 THE LANCET, JUNE 25, 1977

Thawed merozoites were emulsified with Freund's complete adjuvant (F.C.A.) and administered by intramuscular injection (volume 0·5 ml) in doses and at intervals shown in tables I and II. Vaccinated monkeys were challenged by intravenous injection of known numbers of parasites (tables I and II) taken from infected donors (table III). Onset of parasitæmia and parasite numbers were determined as described earlier.[9]

Results

Isolation of P. falciparum *Merozoites*

Since merozoites rapidly penetrate red blood-cells,[18] their isolation requires separation of parasitised from normal erythrocytes followed by harvesting of the extracellular parasites released during in-vitro culture. In the case of *P. knowlesi* the first step is accomplished by centrifugation of schizont-infected blood and the second by differential agglutination[7,9] or cell sieving.[8] The physical properties of *P. falciparum*-parasitised cells do not permit adequate concentration of schizonts by centrifugation, by sieving through 2 μm or 3 μm pore size membranes, or by sedimentation through 'Ficoll/Hypaque' or 15% dextran.[19] For this reason, merozoites cannot usually be isolated from cultures of schizont-infected cells by differential agglutination;[9] in one instance, however, merozoites contaminated with 4% schizonts were prepared by this method from a cell suspension containing about 60% schizonts (table II).

The affinity of *P. falciparum* trophozoites and schizonts for CF11 cellulose[17] allowed mature infected cells to be separated from younger parasites and normal erythrocytes. Entrapped parasites were then cultured on perfused cellulose columns and released merozoites were isolated. In one preparation, a CF11 cellulose column (9·5 × 2·5 cm) loaded with 2·3 ml packed red cells of which 8% were parasitised (all schizonts of variable maturity) provided 3×10^9 merozoites which represents about 15% yield (assuming a mean of 10 merozoites per schizont); the preparation contained 1·7% schizonts.

Challenge of Vaccinated Monkeys

Vaccination with P. knowlesi *merozoites.*—Two douroucouli monkeys were inoculated three times with *P. knowlesi* merozoites in F.C.A. (table I) and challenged on day 60 after first vaccination with 5×10^3 *P. vivax* parasites from a donor monkey. After prepatent intervals of 15–20 days both acquired low-grade chronic infections. The monkeys were superinfected on day 153 with 10^4 *P. falciparum* (Uganda Palo-Alto strain) and 7–12 days later acquired rapidly fatal infections with terminal parasitæmias of 25–30% (table I).

Vaccination with P. falciparum *merozoites.*—Three

TABLE I—VACCINATION OF TWO DOUROUCOULI MONKEYS WITH *P. knowlesi* MEROZOITES (W1-STRAIN[9]) IN F.C.A. AND CHALLENGE WITH *P. vivax* FOLLOWED BY SUPERINFECTION WITH *P. falciparum* (PALO-ALTO EAST AFRICAN STRAIN)

Figura 9 - Fragmento de um arquivo com a seção resultados em destaque.
Fonte: *http://www.sciencedirect.com/science/article/pii/S014067367792551X*.

Após a extração do fragmento de texto que representa a seção resultado o *pipeline* gera um arquivo chamado mapa que será descrito a seguir.

3.2.4 Indexar arquivos no pipeline

Indexar arquivos no *pipeline* significa criar um arquivo texto, denominado mapa, que servirá de índice para os diretórios e os arquivos que serão processados na seção de mineração de textos. A sua estrutura é simples e fornece o pmid do PubMed, o nome do diretório e o arquivo texto indicado para o processo de mineração.

O mapa é utilizado em dois momentos distintos: o primeiro indicará a posição dos arquivos cujo conteúdo é o texto extraído da seção resultados do texto completo e o segundo indicará a localização e arquivos com os termos anotados semanticamente.

A Figura 10, a seguir, mostra o uso do mapa e indica o momento em que o *pipeline* começa a processar a seção resultados.

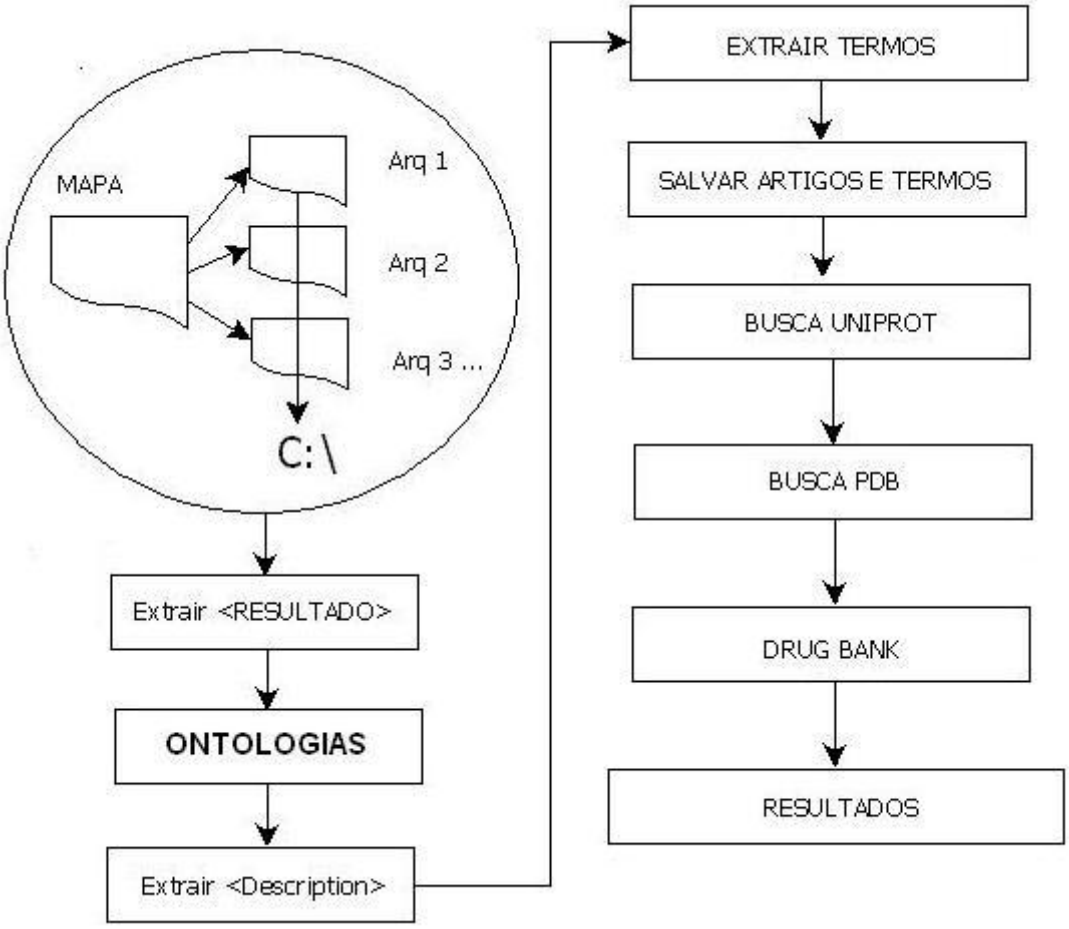


Figura 10 -Mapa como indexador no fluxo de processamento do *pipeline*.

O mapa tem um papel importante na construção e execução do *pipeline* porque é um meio de inserir dados oriundos de outras fontes. Essas fontes podem e normalmente empregam processamentos e ambientes diferentes do local onde o *pipeline* é executado. A inserção desses novos dados se dá na edição do mapa e na inclusão de novos registros e diretórios como mostra a Figura 11, a seguir.

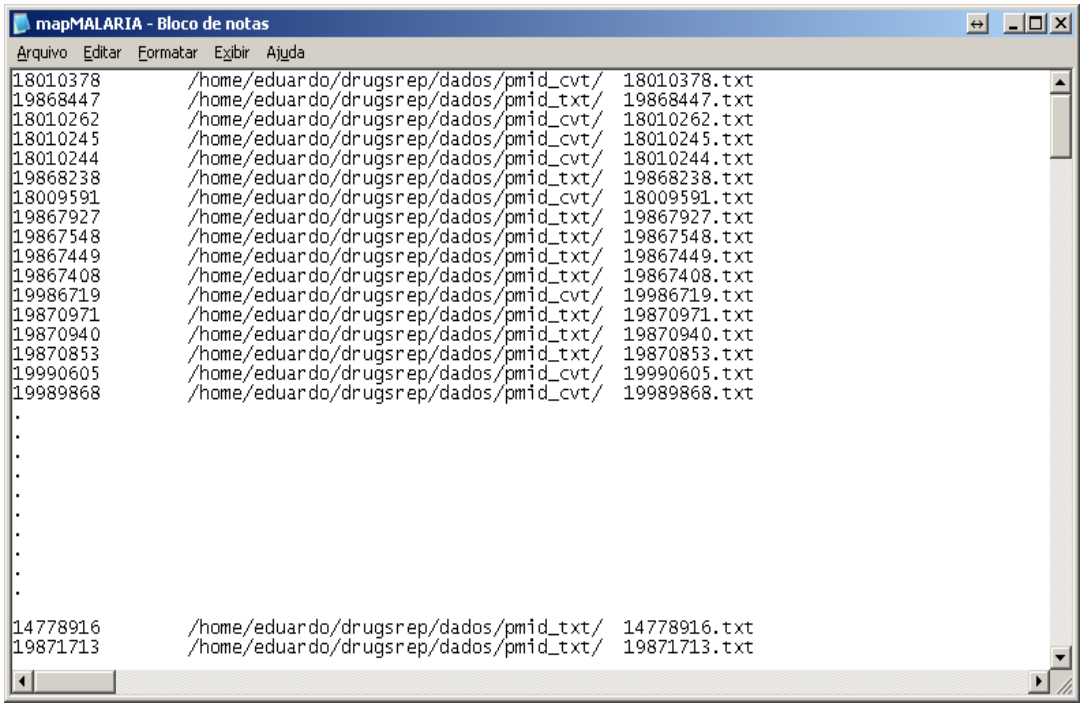


Figura 11 - Exemplo de um arquivo texto que representa um mapa para o *pipeline*.

3.3 Parte 3 – Processos para mineração de textos

O objetivo desta subseção é descrever as etapas que seguem os processos de mineração de textos adotados no *pipeline*. A Parte 3 é composta por onze subseções. A seguir, a breve descrição visa orientar um processo de extração de informação e estruturação de uma coleção de termos anotados semanticamente até a última fase que é um conjunto de fármacos recuperados do DrugBank. Como fases do processo para mineração são citados na ordem de processamento:

- **Anotação semântica** – a construção do corpus textual: consiste em formar um conjunto de termos anotados semanticamente para posterior mineração de textos, identificação de nomes de proteínas, ligantes e fármacos;

- **Pré-processamento**: eliminação de dados espúrios e transformação do corpus textual em uma matriz termo X documento;

- **Eliminação de espaços em branco**: reduz o tamanho do texto eliminando espaços em branco.

- **Conversão para letras minúsculas**: traz o texto para um padrão único de formatação em letras minúsculas.

- **Remoção de *stopwords***: eliminação de palavras do corpus textual a partir de uma lista pré-definida em um dicionário de palavras;

- **Indexar e normalizar**: eliminar ambiguidade ou trazer a um termo comum palavras com o mesmo significado;

- **Extração dos termos**: o resultado após o processamento desta fase do pipeline é um conjunto de termos que serão validados no UniProt e no PDB. São possíveis nomes de proteínas ou termos que podem conduzir o pipeline a esses nomes.

- **Salvar artigos e termos**: o registro das relações entre os termos encontrados e os respectivos artigos científicos;

- **Caracterizar dados – UNIPROT**: emprego dos termos identificados na busca por nomes de proteínas e campos chave válidos <*accession*> e seus respectivos dados;

- **Caracterizar dados – PDB**: uso dos <*accession*> para caracterizar proteínas com os dados de suas estruturas e ligantes;

- **Buscar fármacos no DrugBank**: buscar no DrugBank os fármacos relacionados aos ligantes identificados na caracterização de dados do PDB.

3.3.1 Anotação semântica – a construção do corpus textual

A construção do corpus textual com termos anotados semanticamente é uma etapa importante no processamento do *pipeline*. Esta é a fase do processamento, no qual o conteúdo recuperado, a partir das consultas realizadas no BioPortal Annotator, são termos da área biomédica porque a ontologia, Proteins Ontology (PR), utilizada para fornecer os dados anotados semanticamente, está intimamente relacionada a proteínas e a outras classes de interesse.

Após a execução dessa fase, o *pipeline* seguiu o processamento com os termos anotados semanticamente. Portanto, segue com um conjunto rico em termos biomédicos que forneceram os melhores parâmetros na busca por nomes de proteínas.

As anotações semânticas são recuperadas a partir dos fragmentos de textos colhidos na seção resultados de cada artigo científico ou do artigo na sua forma completa. Cada palavra submetida à consulta recupera termos relacionados à ontologia biomédica escolhida. No caso do *pipeline* a ontologia escolhida foi *Proteins Ontology (PR)*. Desta forma, o corpus textual foi formado por termos anotados semanticamente no domínio biomédico, termos que foram registrados e relacionados por especialistas e o resultado esperado é um conjunto especifico e com grandes chances de retornar nomes de proteínas válidos no âmbito das doenças negligenciadas. Este é o novo conjunto que foi processado nas próximas etapas do *pipeline*.

Os procedimentos dessa fase começam com a leitura e transferência dos arquivos endereçados no mapa para um vetor no ambiente R e todos os conteúdos dos arquivos no formato texto serão submetidos a uma consulta no formato *URL,* a ontologia *Proteins Ontology (PR),* disponível no *BioPortal Annotator.*

A URL permite acesso à interface API por meio da qual é realizada a captação do conteúdo semântico anotado. A API disponibilizada pelo BioPortal ou por uma interface *WEB* como mostra a Figura 12, a seguir, retornou vários marcadores e, no presente trabalho, são aplicados três deles, a saber:

- <*prefLabel*> o mesmo que um termo preferido. O marcador representa o conceito no domínio da ontologia. No BioPortal, termo é sinônimo de classe ou conceito;

- <*synonym*> Nome alternativo para um termo. Cada termo tem um único nome "preferido" e pode ter qualquer número de sinônimos.

- <*definition*> A descrição do marcador não é um campo obrigatório e não há tamanho definido.

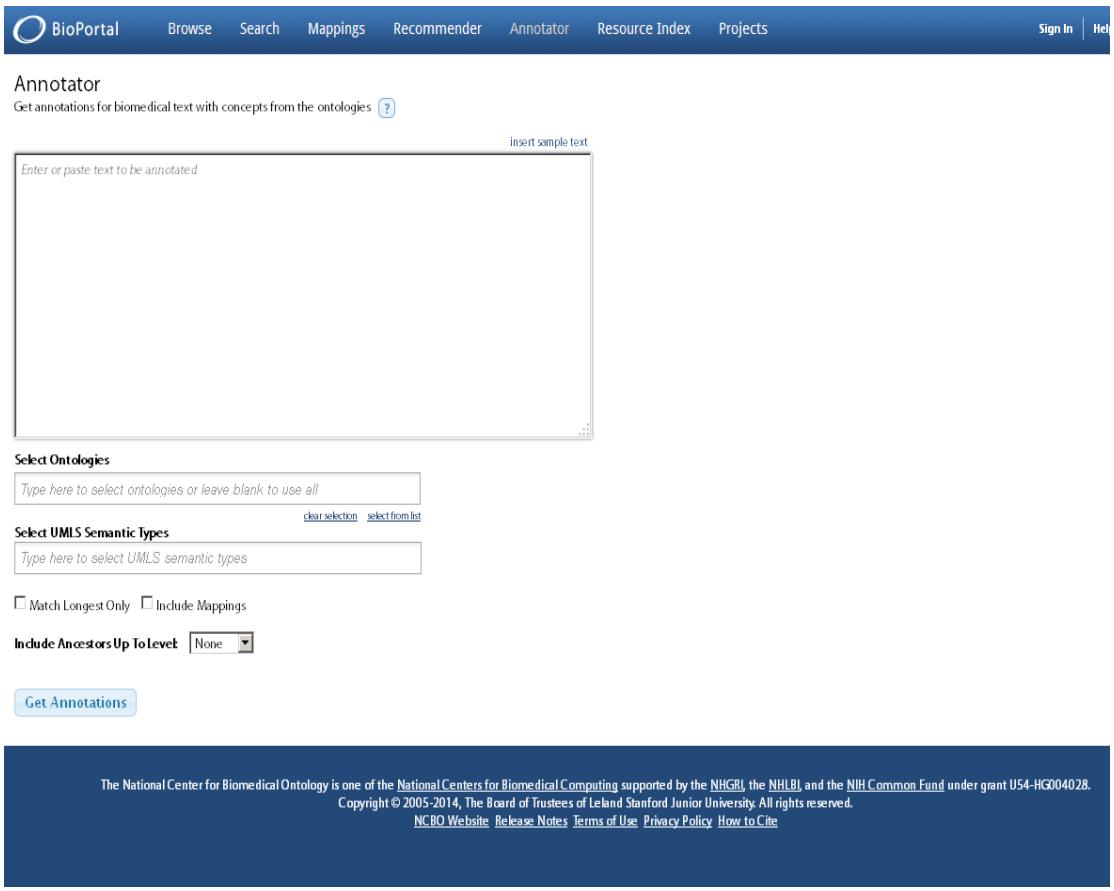


Figura 12 - BioPortal Annotator
Fonte: http://bioportal.bioontology.org/annotator

Na construção do *pipeline* empregou-se a consulta *Web* para identificar as ontologias disponíveis e a API para coletar os dados anotados semanticamente. O conteúdo do arquivo texto, registrado no mapa, é passado na forma de parâmetro, onde cada palavra é concatenada por meio de um sinal “+” juntamente com uma chave de acesso. Segue o exemplo, onde os parâmetros definidos são: palavra ou termo que se deseja obter a anotação semântica *beta-hydroxysteroid*. A Figura 13, a seguir, mostra o resultado após a execução da API (API, 2014) disponível no BioPortal Annotator, segue:

http://data.bioontology.org/search?q=beta-hydroxysteroid&apikey

```
{
   page: 1,
   pageCount: 1,
   prevPage: null,
   nextPage: null,
 - links: {
      nextPage: null,
      prevPage: null
   },
 - collection: [
    - {
         prefLabel: "3 beta-hydroxysteroid dehydrogenase type 7",
       - synonym: [
            "3 beta-hydroxysteroid dehydrogenase type VII",
            "C(27) 3-beta-HSD",
            "cholest-5-ene-3-beta,7-alpha-diol 3-beta-dehydrogenase",
            "3-beta-HSD VII",
            "HSD3B7",
            "3-beta-hydroxy-Delta(5)-C27 steroid oxidoreductase"
         ],
       - definition: [
            "A protein that is a translation product of the human HSD3B7 gene or a 1:1 ortholog thereof."
         ],
         obsolete: false,
         matchType: "prefLabel",
         @id: http://purl.obolibrary.org/obo/PR_000008787,
         @type: http://www.w3.org/2002/07/owl#Class,
       - links: {
```


Figura 13 - Fragmento de um arquivo com anotações semânticas para o termo *beta-hydroxysteroid*

O conteúdo dos marcadores <*prefLabel*>, <*synonym*> e <*definition*> são recuperados e armazenados em um vetor no ambiente R. No final do processamento de todos os arquivos do mapa, o vetor estará com todas as anotações semânticas obtidas por meio da seção resultados ou por um texto completo. A Figura 14, a seguir, mostra um exemplo do fragmento de termos anotados semanticamente.

Anotações Semânticas (fragmento)	protein Wnt-5b protein LMBR1L retinoblastoma-like protein protein FAM87A IgA-inducing
	protein protein HypA protein Wnt-10b protein CBFA2T2 protein S100-A8 protein FAM201A protein
	FAM205B lactase-like protein LIX1-like protein protein PPP5D1 hsc70-interacting protein protein
	FAM111A protein FAM106B protein Wnt-3a density-regulated protein neuritin-like protein protein AmpG
	protein SSX2 protein ZNF783 RRP12-like protein protein traV lnaD-like protein protein 2 one cut domain
	hFBXO41 hFBXO34 hFBXO28 LMO-7 LOMP LMO7 F-box only protein 20 hFBXO31 hFBXO22 hFBXO9 hFBXO33
	DAT1 LMO-3 hFBXO32 FCHO2 hFBXO24 hFBXO10 hFBXO2 hFBXO8 hFBXO44 hFBXO27 hFBXO3 hFBXO4
	hFBXO21 hFBXO6 hFBXO46 hFBXO39 hFBXO43 FCHO1 hFBXO7 hFBXO11 hFBXO17 hFBXO42 hFCHO1 hLMO7
	protein FAM220A protein hnr family member 3 (human) hFBXO5 hFBXO36 hFBXO16 LMO3 rhombotin-3 neuronal-specific
	hFBXO38 hNCCRP1 hFBXO30 hFBXO25 hFBXO40 hFBXO15
	hFCHO2 hLMO3 transmembrane and coiled-coil transcription factor

Figura 14 - Fragmento de um arquivo com um conjunto de anotações semânticas.

Como mencionado anteriormente, no ambiente R os dados ficam dispostos em um vetor. A Figura 15, a seguir, mostra a estrutura do vetor no ambiente R e os dados da posição número 10 desse vetor.

Descrição	Conteúdo empregado na API
Texto com parâmetros (fragmento)	used. This shows that this model using ion current measurement correctly describes the system under study. This value indicates that cationomycin is distributed mainly in the protein phase compared with the membrane phase. It must be noted that the K p value cannot be considered as an absolute measurement as many approximations were made, especially for the volume of proteins, but it may constitute a pertinent relative index to compare different compounds (provided they can transport ions or induce ion currents) studied under the same experimental conditions. We thus used K p to compare monensin and cationomycin partition between proteins and membranes. As previously described for cationomycin, the kinetic constants of transport (k Na2 and k K2) of monensin were measured in the absence of serum with increasing concentrations in the membrane (i ϭ k[mHA]) from the slopes of the straight lines in Figs. 4A and 4B, respectively. These k values were reported in Table II and they were used to calculate the partition coefficient of monensin (K p). The K p values obtained from sodium and potassium experiments are reported in Table III. The coefficients obtained with monensin for Na 2 and K 2 are similar (average value 0.15) and 30 times lower than K p of cationomycin. This indicates that monensin is mainly distributed in the membrane phase unlike cationomycin, which is mainly distributed in serum proteins.

Posição no Vetor	Conteúdo no vetor (resultado)
[linha, coluna]	Linha corresponde ao registro
Coluna 1	PMID
Coluna 2	Anotação Semântica obtida por meio da API
Coluna 3	No fim do processamento armazenará os termos para submissão ao UNIPROT
arrayIdCorpus[10,1] PMID	"10068460"
arrayIdCorpus[10,2] Anotação Semântica	"protein Wnt-5b protein LMBR1L retinoblastoma-like protein protein FAM87A IgA-inducing protein protein HypA protein Wnt-10b protein CBFA2T2 protein S100-A8 protein FAM201A protein FAM205B lactase-like protein LIX1-like protein protein PPP5D1 hsc70-interacting protein protein FAM111A protein FAM106B protein Wnt-3a density-regulated protein protein protein AmpG protein FAM220A protein hnr protein SSX2 protein ZNF783 RRP12-like protein protein traV InaD-like neuritin-like protein protein AmpG protein FAM220A protein hnr protein SSX2 protein ZNF783 RRP12-like protein protein traV InaD-like protein protein antigen FAM201A FAM205B klotho/lactase-phlorizin hydrolase-related protein LCTL LIX1L PPP5D1 PPP5 TPR repeat domain-containing protein 1 suppression of tumorigenicity 13 protein ST13 protein ST13 homolog protein FAM10A1 Hip putative tumor suppressor ST13 renal carcinoma antigen NY-REN-33 progesterone receptor-associated p48 protein FAM111A FAM106B WNT3A SMAP-3 DENR DRP protein DRP1 smooth muscle cell-associated protein 3 candidate plasticity gene 15-2 protein NRN1L ampG FAM220A acrosomal antigen OY-TES-1 proacrosin-binding protein sp32 CT23 ACRBP cancer/testis antigen 23 NOTUM protein X123 FAM189A2 FAM133B KHDC3L C21orf19-like protein mediator of ErbB2-driven cell motility 1 hepatitis C virus NS5A-transactivated protein 7 Memo-1 HCV NS5A-transactivated protein 7 MEMO1 pp110 RB1 Rb pp105 p105-Rb pRb CIAO1 WD repeat-containing protein 39 IWS1 IWS1-like protein elaB traube protein Rb-binding protein Che-1 AATF apoptosis-antagonizing transcription factor HECA BLOC1S5"

Figura 15 - Dados do vetor na posição 10.
Anotações semânticas obtidas com os termos descritos no quadro superior (fragmento do conteúdo).

3.3.2 Pré-processamento

Os documentos textuais digitais no seu formato bruto, originário, por exemplo, de metadados em formato XML, exigem um tratamento para a formação do *corpus* textual (FEINERER, 2008). O *corpus* texto precisa ser alterado de forma a manter

em seu conteúdo somente as palavras relevantes ao tema proposto. Isto é o tema desejado na escolha do termo da busca e recuperação dos artigos na base de dados PubMed, como é o caso do presente trabalho.

Um *corpus* composto por muitas palavras dificulta o processo de análise em mineração de textos porque está formado por palavras que não retratam o tema central do(s) resumo(s) ou artigo(s) completo(s).

Portanto, é oportuno que palavras como artigos, preposições, números, nomes de países, sinais de pontuação, siglas métricas, duplicações, dentre outras, sejam eliminadas. Segundo Feinerer (2008) “o pré-processamento, isto é, a aplicação de métodos de limpeza e estruturação do texto de entrada para uma análise mais aprofundada, é um componente central em estudos práticos de mineração de texto”.

O pré-processamento deve ser entendido como a fase inicial no processo de mineração de textos. Primeiramente, são removidas as palavras espúrias que não refletem o tema central. O objetivo é extrair um conjunto de palavras que represente todo o corpo textual que foi submetido às técnicas de processamento de linguagem natural; a seguir, é a transformação do *corpus* textual em uma matriz termo x documento. A matriz termo x documento segue o modelo de espaço vetorial (SALTON, 1975) e tem a finalidade de obter o conjunto de documentos, seus termos e respectiva frequência o que permite analisar e visualizar esses dados por meio de *clusters*, dendogramas e *wordclouds*, dentre outras técnicas ou funções.

3.3.3 Eliminação de espaços em branco

Essa eliminação consiste em suprimir caracteres em branco encontrados excessivamente ao longo do *corpus* textual. Reduzir o *corpus* textual, além de melhorar a visualização do texto, reduz o tempo gasto no processamento computacional para os *corpus* textuais excessivamente grandes.

A Figura 16, a seguir, mostra o resultado da função *stripWitespace* com apenas um documento textual para redução de espaços em branco.

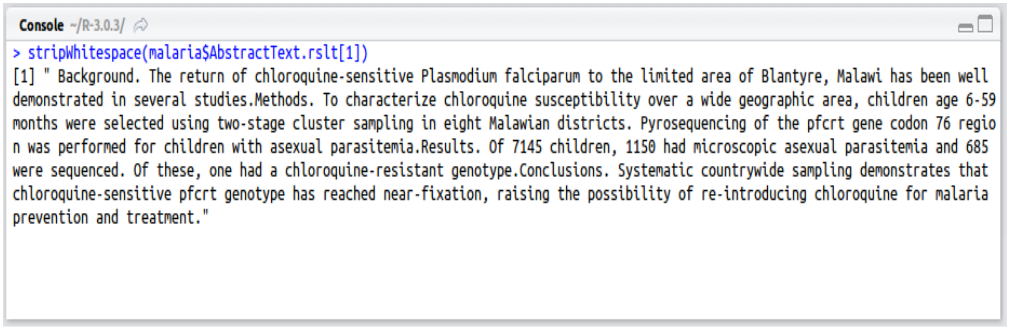


Figura 16 - *Corpus* textual após emprego da função *stripWhitespace().*

3.3.4 Conversão para letras minúsculas

A substituição de caracteres em caixa alta para caixa baixa tem a finalidade de tornar iguais termos com número de caracteres iguais com o mesmo significado. Para efeito de processamento computacional o termo *falciparum* é diferente de *Falciparum*. Sendo assim, a transformação permite que os termos permaneçam com a mesma quantidade de caracteres e o mesmo sentido.

A Figura 17, a seguir, mostra o resultado da função *tolower()* com apenas um documento textual para a substituição de caracteres em caixa alta para caixa baixa.

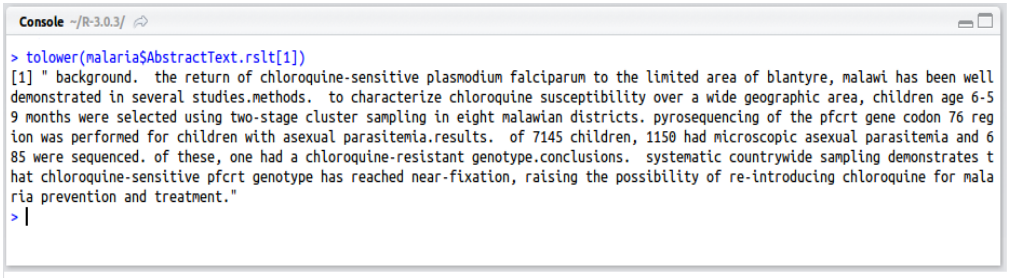


Figura 17 - Corpo textual após emprego da função *tolower().*

3.3.5 Remoção de stopwords

O termo *stopwords* (FEINERER, 2008) é empregado para designar palavras que aparecem com certa frequência no *corpus* textual, mas têm pouco ou nenhum significado ou relação com o tema proposto. São palavras consideradas irrelevantes, por isso é praxe removê-las permitindo reduzir ainda mais o *corpus* textual para futuro processamento.

Há inúmeros recursos de programação para a remoção de *stopwords.* Manter um arquivo de texto com as palavras que se deseja excluir é uma das alternativas ou simplesmente aplicar recursos oferecidos pelo ambiente de programação R.

O Anexo IV traz como exemplo um fragmento de arquivo cujo conteúdo é formado por *stopwords* previamente selecionadas. A Figura 18, a seguir, mostra um exemplo usando recursos da linguagem de programação R: a variável *set_mystopwords* contém as palavras listadas no Anexo V e a execução por meio da função *removeWords()*

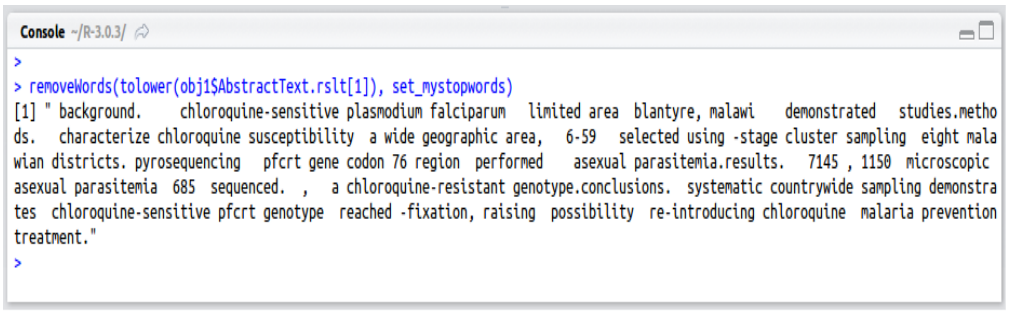


Figura 18 - *Corpus* textual após a função *removeWords().*

3.3.6 Indexar

No presente trabalho, indexar o *corpus* textual consiste em gerar um índice, isto é, estruturar e identificar as relações de um documento e seus respectivos termos. Logo, indexar os termos é uma tarefa que deverá ser executada de forma a obter um conjunto ordenado de termos de acordo com o tema central da busca.

3.3.7 Extração dos termos

A extração de termos determina um conjunto que é o resultado obtido após o *corpus* textual ser processado nas fases anteriores até a indexação. A Figura 19, a seguir, mostra a sequência de resultados de um corpus textual após ser processado por programas utilizados no *pipeline* que foram construídos na linguagem R.

PMID	17565676
Resumo (fragmento)	Malaria microscopy, while the gold standard for malaria diagnosis, has limitations.
	Efficacy estimates in drug and vaccine malaria trials are very sensitive to small errors in microscopy endpoints.
	This fact led to the establishment of a Malaria Diagnostics Centre of Excellence in Kisumu, Kenya.
	The primary objective was to ensure valid clinical trial and diagnostic test evaluations.
	Key secondary objectives were technology transfer to host countries, establishment of partnerships, and training of
	clinical microscopists
Seção Resultados (fragmento)	Results To date, 209 microscopists have participated from the
	following countries: Kenya (178), Uganda (10), Tanzania (5), Rwanda (1), Burundi (1), Malawi (6), Cameroon (3), Mali (2),
	Nigeria (2), and Thailand (1). One hundred and fifty eight participants (76%) have been primarily research and 51 (24%) have
	been primarily clinical. Leadership/expertise has participated from Kenya (4), Uganda (1), Nigeria (2), Mali (1), Cameroon (1),
	Zanzibar (2), Thailand (1), Peru (1), USA (1), and Indonesia (1). Data from the courses between July 2005 and January
	2006 are reported here. These data include 77 participants in the long course (including 10 who trained twice) and 23
	participants in the short course (including five who trained twice). Of these, 69% were conducting primarily research.
	Participants were from Kenya (81), Tanzania (5), Malawi (4), Rwanda (1), Burundi (1), Cameroon (2), Nigeria (2), Mali (2), and
	Uganda (2). On a questionnaire, participants reported a mean (median) of years of laboratory education 3 (3), years employed
	as a laboratory technician or technologist 9 (6), and years employed as a malaria microscopist 5 (3). Twelve percent reported
	that they were laboratory supervisors, 30% reported that they had a malaria proficiency program in their laboratory, and 84%
	reported that they used SOPs in their laboratory. Pre- and post-training scores and percent improvement by
	organizational group and course type from the most recent participants are presented in Additional File 1 .
	With a mean specificity of 80% and range in mean by organization on the
	pre-test of 6691%, these data clearly illustrate this is a serious problem in both the research and clinical setting.
	Specificity improved substantially with training (Additional File 1).
	Thirty-two percent of all true negative smears were spiked with artifact (mix Staphylococcus
	aureus with fungi at varying densities). On the pre-test, 26% of those spiked with artifact were read as false positives,
	while 18% of those not spiked were. On the post-test, 8% of those spiked with artifact were read as false positives, while 6% of
	those not spiked were. The slides with the highest density artifact had higher rates of false positives. Spiked slides did not
	appear to affect the species reported, except for a slightly higher rate reported as P. vivax in the post-test. The
	false positive smears reported as P. vivax and P. ovale increased overall in comparing the
	pre-test with the post-test, while the reporting of P. falciparum decreased. The effect of parasitaemia on
	the false negative rate was not as large as anticipated. The mean (median) densities of those with false negatives on the
	pre-test were 174 (128) parasites/l compared with 216 (142) parasites/l for the true positives.
Anotações Semânticas (fragmento)	protein Wnt-5b protein LMBR1L retinoblastoma-like protein protein FAM87A IgA-inducing
	protein protein HypA protein Wnt-10b protein CBFA2T2 protein S100-A8 protein FAM201A protein
	FAM205B lactase-like protein LIX1-like protein protein PPP5D1 hsc70-interacting protein protein
	FAM111A protein FAM106B protein Wnt-3a density-regulated protein neuritin-like protein protein AmpG
	protein SSX2 protein ZNF783 RRP12-like protein protein traV lnaD-like protein protein 2 one cut domain
	hFBXO41 hFBXO34 hFBXO28 LMO-7 LOMP LMO7 F-box only protein 20 hFBXO31 hFBXO22 hFBXO9 hFBXO33
	DAT1 LMO-3 hFBXO32 FCHO2 hFBXO24 hFBXO10 hFBXO2 hFBXO8 hFBXO44 hFBXO27 hFBXO3 hFBXO4
	hFBXO21 hFBXO6 hFBXO46 hFBXO39 hFBXO43 FCHO1 hFBXO7 hFBXO11 hFBXO17 hFBXO42 hFCHO1 hLMO7
	protein FAM220A protein hnr family member 3 (human) hFBXO5 hFBXO36 hFBXO16 LMO3 rhombotin-3 neuronal-specific
	hFBXO38 hNCCRP1 hFBXO30 hFBXO25 hFBXO40 hFBXO15
	hFCHO2 hLMO3 transmembrane and coiled-coil transcription factor
Termos	FAM87A HypA CBFA2T2 FAM201A FAM205B lactase-like LIX1-like PPP5D1 FAM111A FAM106B neuritin-like AmpG
	FAM220A hnr SSX2 ZNF783 RRP12-like traV hFBXO5 hFBXO36 hFBXO41 hFBXO34 hFBXO28 LMO-7 LOMP LMO7 hFBXO31
	hFBXO22 hFBXO9 hFBXO33 hFBXO16 LMO3 rhombotin-3 DAT1 LMO-3 hFBXO32 FCHO2 hFBXO24 hFBXO10 hFBXO2
	hFBXO8 hFBXO44 hFBXO27 hFBXO3 hFBXO4 hFBXO38 hNCCRP1 hFBXO30 hFBXO25 hFBXO40 hFBXO15 hFBXO21
	hFBXO6 hFBXO46 hFBXO39 hFBXO43 FCHO1 hFBXO7 hFBXO11 hFBXO17 hFBXO42 hFCHO1 hLMO7 hFCHO2 hLMO3
	ADAM-TS5 ADAMTS-5 ADAMTS-11 ADAM-TS ADAM-TS ADMP-2 ADAMTS5 hZKSCAN2 CBP/p300 ADAM-TS16
	ADAMTS-16 ADAMTS16 ADAM-TS ACAP3 centaurin-beta-5 hADAMTS18 ADAMTS-9 ADAM-TS ADAM-TS9 ADAMTS9
	hADAMTS7 hADAMTS9 CGRRF1 vti1-rp2 VTI1A v-SNARE

Figura 19 - Exemplo ilustra o conteúdo dos *arrays* retornado após a extração dos termos.

Na Figura 19 temos uma seção cujo título é “Termos” onde é apresentado um conjunto de palavras que são os termos relevantes. São os resultados obtidos nessa etapa do processamento que serão registrados na base de dados PostgreSQL e utilizados posteriormente na busca por nomes de proteínas.

Embora a API fornecida pelo BioPortal, descrita no item 3.3.1, detenha no seu código uma lista de *stopwords* que são eliminadas durante o processamento, o *pipeline* contém uma lista própria com termos previamente definidos e testados em projetos anteriores no laboratório no PROCC (Anexo V). Na medida em que o

pipeline é executado, essa lista pode ser ampliada e há um progresso relacionado à redução de termos espúrios no conjunto de termos que serão processados nas próximas fases do *pipeline*. A Figura 20, a seguir, mostra como um exemplo, o resultado do vetor na posição 10 apresentado anteriormente, na Figura 19 com 196 termos. Seguem os dados, após a execução dos passos 3.3.1 até 3.3.7 do *pipeline* com apenas 51 termos anotados semanticamente.

FAM87A HypA CBFA2T2 FAM201A FAM205B lactase-like LIX1-like PPP5D1
FAM111A FAM106B neuritin-like AmpG FAM220A hnr SSX2 ZNF783 RRP12-like
traV FAM201A FAM205B LCTL PPP5D1 PPP5 TPR FAM10A1 FAM111A FAM106B
WNT3A SMAP-3 DENR DRP DRP1 ampG FAM220A OY-TES-1 ACRBP FAM189A2
FAM133B HCV MEMO1 p105-Rb pRb CIAO1 IWS1 IWS1-like elaB traube
Che-1 AATF HECA BLOC1S5

Figura 20 - Termos após o processamento do pipeline.

3.3.8 Salvar artigos e termos

Antes de validar os termos que poderiam traduzir nomes de proteínas nas futuras consultas às bases do UNIPROT e PDB, são necessários alguns registros nas tabelas da base de dados do PostgreSQL. Na medida em que se avança com a descrição do *pipeline* neste trabalho, os dados foram registrados na base de dados PostgreSQL que oferece o suporte necessário para que esses dados sejam recuperados e processados posteriormente nas etapas de processamento do pipeline.

A etapa salvar artigos e termos faz o registro de dados em diversas tabelas definidas no PostgreSQL. São dados como: o número do *job*[47], código pmid, o resumo e outros. A Tabela 4, a seguir, mostra um fragmento que contém os 20 primeiros termos identificados no *pipeline* seguido de um número sequencial que individualiza o registro.

Tabela 4 - Os vinte primeiros termos registrados

Descrição	Sequencial
10-formyltetrahydrofolate	1
10-FTHFDH	2
13-acetate	3
16E1-BP	4
17-beta	5
17-beta-hydroxysteroid	6
1-alpha	7

[47] Na informática é o termo que designa um grupo, conjunto ou arranjo de tarefas executadas por um computador. Em alguns casos registra datas, nome do usuário que solcita tarefas, etc.

1-beta	8
1-epimerase	9
1-gamma	10
1-methylphosphonate	11
1-phosphatase	12
1-phosphohydrolase	13
1-type	14
20-HETE	15
26-hydroxylase	16
2-acylhydrolase	17
2-alpha	18
2-beta	19
2D-0014IV	20
2-epimerase	21

O próximo passo no *pipeline* é identificar nomes de proteínas. O processo de identificação requer uma consulta para todos os termos identificados e registrados. É o que será verificado em seguida.

3.3.9 Caracterizar dados – UNIPROT

O processo de caracterização envolve, basicamente, a busca por chaves de acesso a outras bases de dados. O objetivo é que a partir dessas chaves se estabeleça a ligação entre bases de dados e por meio dessa ligação obter informação complementar que caracterize nomes de proteínas. As chaves são elos de ligação com outros dados registrados em outras bases que permitem identificar, depois de uma série de consultas, os nomes dos fármacos que serão candidatos ao reposicionamento.

O UNIPROT é a primeira base de dados e foi escolhida porque permite o emprego de termos em um de seus mecanismos de busca e recuperação de informação na forma de uma API. Esses termos processados e registrados pelo *pipeline* servem para a identificação da primeira chave de ligação que é o código *<accession>.*

O retorno do código *<accession>* por meio de uma API fornecida pelo UNIPROT permite validar o termo como uma proteína e retornar os códigos relacionados a outra base de dados que é o PDB ou retornar dados relacionados ao um termo identificado e os dados associados a este no PDB. O resultado que é exibido em seguida foi obtido com o emprego do termo *beta-hydroxysteroid* que é um substrato de uma enzima. A Figura 21, a seguir, mostra o fragmento do arquivo XML e os códigos listados a partir da execução da URL. E a Figura 22, mostra uma

outra forma de consulta normalmente empregada para uso na internet, contudo fornece o mesmo resultado.

Figura 21 - Fragmento de um arquivo com *accession codes* para o termo *beta-hydroxysteroid*.
Fonte:*http://www.uniprot.org/uniprot/?query=beta-hydroxysteroid&sort=score&format=xml&limit=50&*

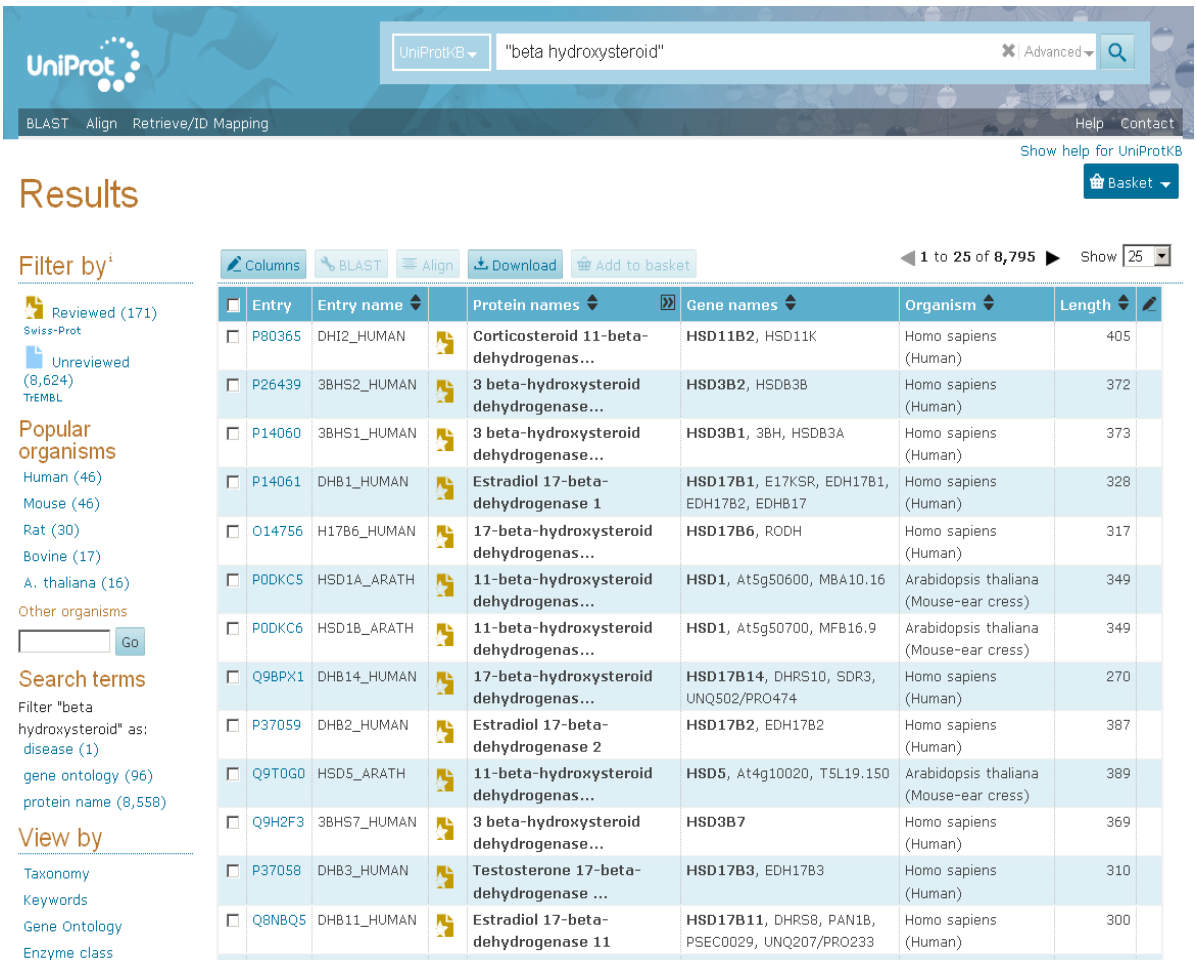


Figura 22 - Consulta no UniProt com *accession codes* para o termo *beta-hydroxysteroid*.
Fonte: *http://www.uniprot.org/uniprot/?query=%22beta-hydroxysteroid%22&sort=score*

O procedimento é executado para todos os termos registrados na base de dados PostgreSQL. Os resultados são armazenados e utilizados posteriormente na busca por dados no PDB. A seguir, a Tabela 5 contém os 50 registros identificados para o termo *beta-hydroxysteroid* com o respectivo código <*accession*> representado na tabela pelo campo <*iduniprot*>.

Tabela 5 - Nomes de proteínas relacionados ao termo *beta-hydroxysteroid*.

Nome UNIPROT	Nome Completo	Código UNIPROT
DHB11_HUMAN	Estradiol 17-beta-dehydrogenase 11	Q8NBQ5
DHB1_HUMAN	Estradiol 17-beta-dehydrogenase 1	P14061
DHB3_HUMAN	Testosterone 17-beta-dehydrogenase 3	P37058
DHB2_HUMAN	Estradiol 17-beta-dehydrogenase 2	P37059
DHB11_MOUSE	Estradiol 17-beta-dehydrogenase 11	Q9EQ06
AK1C3_HUMAN	Aldo-keto reductase family 1 member C3	P42330
DHB7_HUMAN	3-keto-steroid reductase	P56937
H17B6_HUMAN	17-beta-hydroxysteroid dehydrogenase type 6	O14756
DHB14_HUMAN	17-beta-hydroxysteroid dehydrogenase 14	Q9BPX1
DHB7_MOUSE	3-keto-steroid reductase	O88736
H17B6_MOUSE	17-beta-hydroxysteroid dehydrogenase type 6	Q9R092
DHB4_HUMAN	Peroxisomal multifunctional enzyme type 2	P51659

HCD2_HUMAN	3-hydroxyacyl-CoA dehydrogenase type-2	Q99714
H17B6_RAT	17-beta-hydroxysteroid dehydrogenase type 6	O54753
DHB13_HUMAN	17-beta-hydroxysteroid dehydrogenase 13	Q7Z5P4
DHB13_MOUSE	17-beta-hydroxysteroid dehydrogenase 13	Q8VCR2
DHB4_RAT	Peroxisomal multifunctional enzyme type 2	P97852
DHB4_MOUSE	Peroxisomal multifunctional enzyme type 2	P51660
HCD2_DROME	3-hydroxyacyl-CoA dehydrogenase type-2	O18404
H17B6_BOVIN	17-beta-hydroxysteroid dehydrogenase type 6	Q3T001
3BHS2_HUMAN	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 2	P26439
3BHS7_HUMAN	3 beta-hydroxysteroid dehydrogenase type 7	Q9H2F3
3BHS1_HUMAN	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 1	P14060
3BHS7_MOUSE	3 beta-hydroxysteroid dehydrogenase type 7	Q9EQC1
3BHS2_RAT	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 2	P22072
3BHS4_RAT	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 4	Q62878
3BHS1_MOUSE	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 1	P24815
3BHS2_MOUSE	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 2	P26149
3BHS3_MOUSE	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 3	P26150
3BHS4_MOUSE	3 beta-hydroxysteroid dehydrogenase type 4	Q61767
3BHS1_RAT	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 1	P22071
3BHS5_MOUSE	3 beta-hydroxysteroid dehydrogenase type 5	Q61694
3BHS5_RAT	3 beta-hydroxysteroid dehydrogenase type 5	P27364
3BHS6_MOUSE	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 6	O35469
3BHS7_RAT	3 beta-hydroxysteroid dehydrogenase type 7	O35048
DHI2_HUMAN	Corticosteroid 11-beta-dehydrogenase isozyme 2	P80365
HSD1A_ARATH	11-beta-hydroxysteroid dehydrogenase 1ª	P0DKC5
HSD1B_ARATH	11-beta-hydroxysteroid dehydrogenase 1B	P0DKC6
HSD5_ARATH	11-beta-hydroxysteroid dehydrogenase-like 5	Q9T0G0
HSD6_ARATH	11-beta-hydroxysteroid dehydrogenase-like 6	Q9LUE4
HSD4A_ARATH	11-beta-hydroxysteroid dehydrogenase-like 4A	P0DKC7
HSD4B_ARATH	11-beta-hydroxysteroid dehydrogenase-like 4B	Q9LUF2
HSD2_ARATH	11-beta-hydroxysteroid dehydrogenase-like 2	Q9STY8
HSD3_ARATH	11-beta-hydroxysteroid dehydrogenase-like 3	Q9STY7
3BHD_COMTE	3-beta-hydroxysteroid dehydrogenase	P19871
CBR1_RAT	Carbonyl reductase [NADPH] 1	P47727
DHB14_BOVIN	17-beta-hydroxysteroid dehydrogenase 14	Q9MYP6
3BHS3_MESAU	3 beta-hydroxysteroid dehydrogenase type 3	O35296
3BHS1_MACMU	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 1	P27365
3BHS1_MESAU	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 1	Q60555

3.3.10 Caracterizar dados – PDB

Neste trabalho, utilizando-se o PDB combinado com o UniProt, foi possível obter dados relacionados às proteínas, em quais organismos a proteína está presente e, principalmente, a sua estrutura e os códigos que identificam os ligantes. Os códigos dos ligantes permitiram ao *pipeline* relacionar ligantes aos fármacos registrados no DrugBank que são candidatos ao reposicionamento.

A lista de nomes de proteínas encontradas no procedimento anterior foi submetida a outro processo de validação que foi executado no banco de dados PDB. Nesse processo de validação, o pipeline confirma a existência de uma proteína

associada ao código <*accession*> obtido anteriormente e busca por meio da estrutura da proteína o código do ligante.

O PDB fornece uma API como serviço de interface de busca por informação relacionada à estrutura de proteínas (RCSB, 2015). O serviço está baseado no estilo denominado *RESTful Web Services* (RWS)[48]*.* A API baseada no RWS permite busca de acordo com o tipo de parâmetro informado pelo usuário ou programador. Assim há duas formas de se obter dados e são complementares para a execução contínua do *pipeline*, que são: a busca de registros pelo código <*accession*>; a busca por dados relacionados aos ligantes.

A busca pelo código <*accession*> consiste em um arquivo no formato XML onde o parâmetro <*accession*> é colocado no marcador <accessionIdList> conforme o exemplo descrito abaixo.

```
<orgPdbQuery>
<queryType>org.pdb.query.simple.UpAccessionIdQuery</queryType>
<description>Simple query for a list of UniprotKB Accession</description>
<accessionIdList> P14061 </accessionIdList>
</orgPdbQuery>
```

O exemplo usa o código P14061 que é o segundo item da Tabela 5. A execução da linha de instrução seguida de uma instrução POST (JAVA, 2013) em relação ao serviço *Web* apresenta o resultado exibido abaixo na Figura 23:

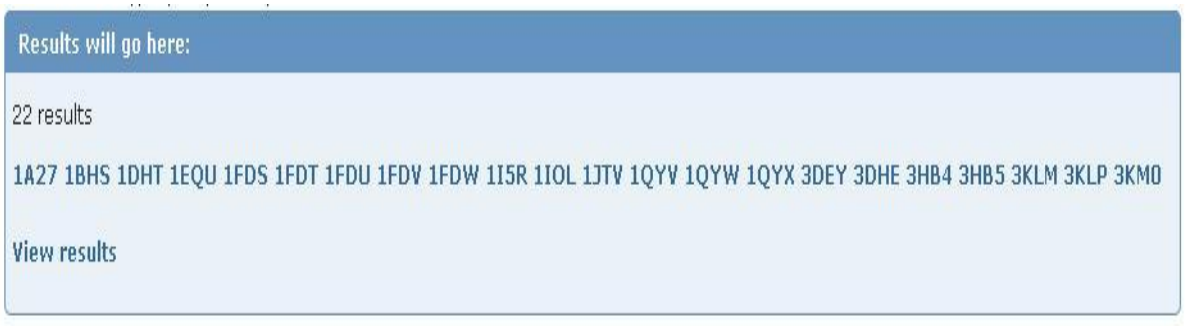


Figura 23 - Resultado da consulta com os códigos de identificação das estruturas contidas no PDB relacionada a P14061 usando (RWS).

O resultado são os códigos que identificam as estruturas contidas no banco de dados PDB relacionados a proteína passada como parâmetro que no caso desse

[48] *RESTful (Web Service Representational State Transfer)* é um estilo de arquitetura, uma abordagem de comunicação usado no desenvolvimento de serviçoes *WEB (Java , 2013)*.

exemplo é P14061. O próximo passo é empregar esses códigos para recuperar informação que permita ao *pipeline* identificar ligantes, classificação, anotações de domínio externas, enzimas homólogas e outros.

Os ligantes têm um papel fundamental para o *pipeline* porque eles são a chave para o acesso aos fármacos registrados no DrugBank.

Portanto, uma instrução URL é executada e recebe como parâmetro os códigos obtidos na Figura 23 e retorna para cada um desses códigos dentre outros dados os códigos que identificam os ligantes relacionados com a estrutura da proteína desejada.

Após a chamada da URL por meio de uma API ou navegador *Web* http://www.rcsb.org/pdb/rest/ligandInfo?structureId=1A27 o resultado é um arquivo XML com os códigos “EST” e “NAP” que identificam os ligantes. Esses códigos são precedidos de um parâmetro denominado <*chemicalID*> exatamente como é apresentado, a seguir, na Figura 24.

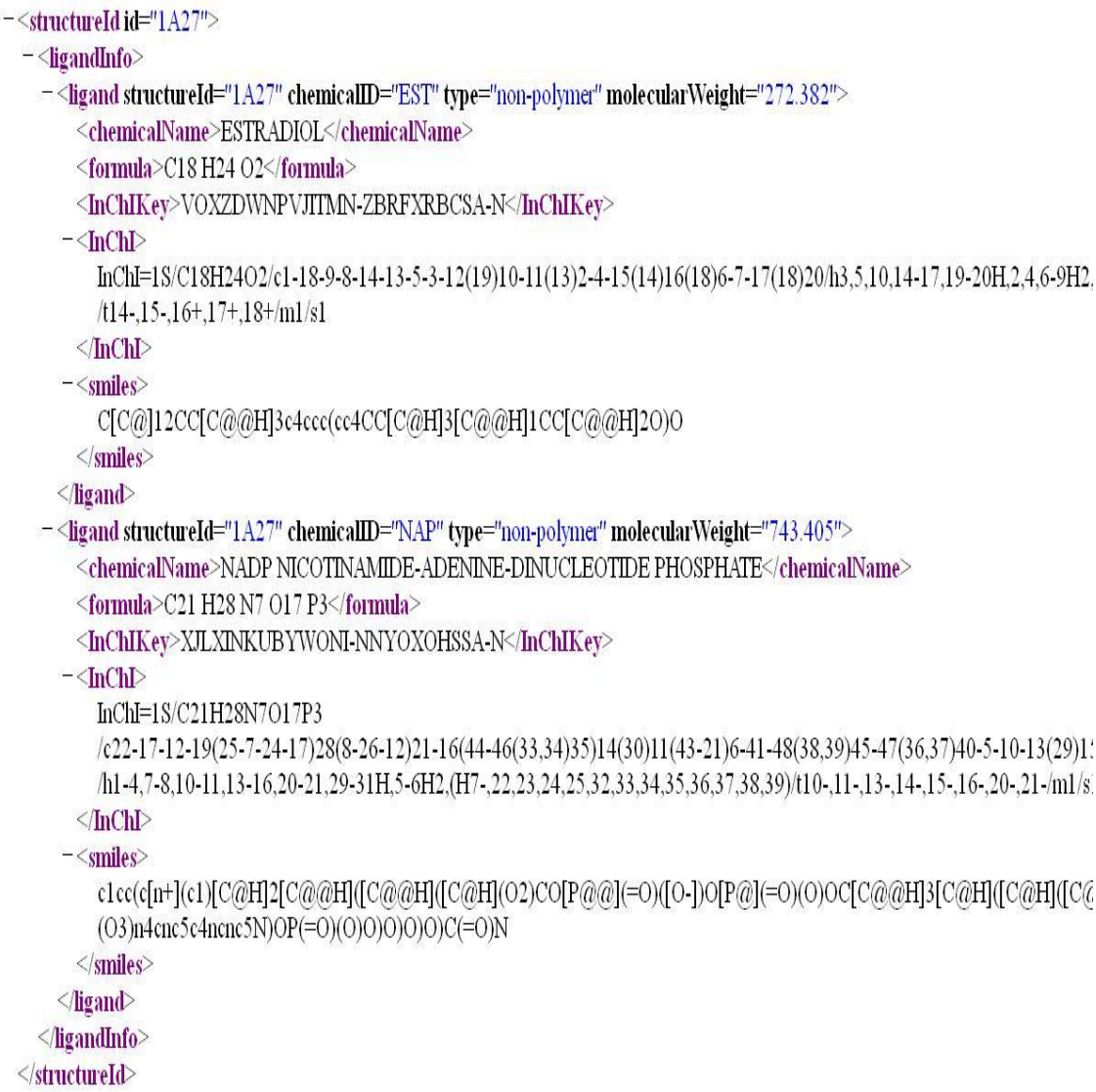


Figura 24 - Arquivo XML. Com os códigos dos ligantes para 1A27.
Fonte: http://www.rcsb.org/pdb/rest/ligandInfo?structureId=1A27

Todos os código listados na Figura 23 fizeram parte de um *loop* para obter os respectivos códigos dos ligantes, isto é, um *loop* para 22 códigos que recupera a informação referente ao ligante da estrutura especificada. O *pipeline* passa agora ao registro desses dados na base de dados PostgreSQL.

3.3.11 Buscar fármacos no DrugBank

Caracterizados os dados referentes aos nomes de proteínas, identificados respectivos ligantes a partir dos códigos de suas estruturas no PDB, o *pipeline* passa à busca de fármacos no DrugBank. A principal consulta ao DrugBank tem por finalidade associar os fármacos registrados no DrugBank aos códigos dos ligantes encontrados na caracterização de dados do PDB.

O procedimento emprega como chave de integração entre o PDB e o DrugBank o código do ligante denominado na base de dados do PDB como <*chemical_id*> e tem o seu equivalente na base de dados do DrugBank como <*het_id*>.

Contudo, o *pipeline* precisa, a partir do código do ligante recuperado no procedimento anterior, realizar consultas ao Drugbank e por meio de programas analisadores extrair a informação desejada que são os fármacos relacionados ao código do ligante e registrá-la na base de dados.

Portanto, o primeiro passo foi identificar os ligantes por meio de uma consulta em *SQL*, as tabelas registradas na base PostgreSQL. A consulta *SQL* empregada no *pipeline* é descrita a seguir:

SELECT DISTINCT tabela.ligante.chemicalid AS chemicalid, tabela.ligante.hetname AS hetname, tabela.ligante.idligante AS idligante FROM tabela.artigo AS a, tabela.artigo_termo AS at, tabela.termo_proteina AS tp, tabela.proteina AS p, tabela.pdb_ligante AS pdb, tabela.ligante AS ligante WHERE a.job = 1 AND a.idartigo = at.idartigo AND at.idtermo = tp.idtermo AND tp.idproteina = p.idproteina AND p.idproteina = pdb.idproteina AND pdb.idligante = li.idligante

A consulta em SQL descrita acima combina tabelas e dados de forma a permitir a identificação de todos os ligantes relacionados a todas as proteínas válidas no *job* em execução no *pipeline*. O objetivo é listar todos os ligantes por meio dos parâmetros recebidos. Como exemplo: o campo *a.job* com valor de parâmetro igual a 1 (um), atribuído automaticamente ao campo pelo *pipeline*, que representa o número do *job* para um lote referente ao termo *malaria*. A Tabela 6, a seguir, disponibiliza para o exemplo apenas o resultado dos 40 (quarenta) primeiros registros da consulta. Essa tabela mostra a lista de ligantes recuperados *chemicalid* e a sua respectiva descrição. Os valores em *chemicalid* são as chaves de acesso aos fármacos no Drugbank.

Tabela 6 - Os ligantes identificados no PDB.

chemicalid	hetname	idligante
001	1-[2,2-DIFLUORO-2-(3,4,5-TRIMETHOXY-PHENYL)-ACETYL]-PIPERIDINE-2-CARBOXYLIC	1029
002	N-[(2R)-2-BENZYL-4-(HYDROXYAMINO)-4-OXOBUTANOYL]-L-ISOLEUCYL-L-LEUCINE	48
004	(2S)-amino(phenyl)ethanoic acid	3208
004	(2S)-AMINO(PHENYL)ETHANOIC ACID	4822
007	1-METHYLAMINE-1-BENZYL-CYCLOPENTANE	5216
008	(S)-2-[(R)-3-AMINO-4-(2-FLUORO-PHENYL)-BUTYRYL]-1,2,3,4-TETRAHYDRO-ISOQUINOL	5214
009	(4S)-1,4-DIBENZYL-N-[(1S,2R)-1-BENZYL-3-{[3-(DIMETHYLAMINO)BENZYL]AMINO}-2-HYD	2887
00C	3-SULFO-D-ALANINE	9471
00G	5-[2-(1H-PYRROL-1-YL)ETHOXY]-1H-INDOLE	1683
00J	N,2-DIMETHYL-6-{[7-(2-MORPHOLIN-4-YLETHOXY)QUINOLIN-4-YL]OXY}-1-BENZOFURAN	5782
00S	4-(AMINOMETHYL)BENZENECARBOXIMIDAMIDE	5844
012	(4S)-N-[(1S,2R)-1-BENZYL-3-{[3-(DIMETHYLAMINO)BENZYL]AMINO}-2-HYDROXYPROPYL	2840
014	1-(5-CHLORO-6-FLUORO-1H-BENZIMIDAZOL-2-YL)-1H-PYRAZOLE-4-CARBOXYLIC ACID	5304
018	2-NITRO-N-(THIOPHEN-3-YLMETHYL)-4-(TRIFLUOROMETHYL)ANILINE	134
01E	(2S)-2-(3,3-DIMETHYLBUTANOYLAMINO)-N-[(2S)-1-[[(2S,3S)-3-HYDROXY-4-[(4-IODOPHE	8945
01P	N~2~-[3-METHOXY-4-(MORPHOLIN-4-YL)PHENYL]-N~4~-(QUINOLIN-3-YL)PYRIMIDINE-2,	5492
01S	N-[(2R)-2-(HYDROXYCARBAMOYL)-4-METHYLPENTANOYL]-L-ALANYLGLYCINAMIDE	7748
01T	[5-(AMINOMETHYL)-6-(2,2-DIMETHYLPROPYL)-2-ETHYL-4-(4-METHYLPHENYL)PYRIDIN-	5204
020	N-(FURAN-2-YLMETHYL)-2-NITRO-4-(TRIFLUOROMETHYL)ANILINE	130
023	N^2^-[(2R)-2-{(1S)-1-[FORMYL(HYDROXY)AMINO]ETHYL}-5-PHENYLPENTANOYL]-N,3-DI	7652
027	7-(4-METHYLPIPERAZIN-1-YL)-4-[(5-METHYL-1H-PYRAZOL-3-YL)AMINO]-2-(PROPAN-2-Y	3295
028	(1-HYDROXYHEPTANE-1,1-DIYL)BIS(PHOSPHONIC ACID)	5977
02K	1-aminocyclohexanecarboxylic acid	3267
02Y	6-DIAZONIO-5-OXO-L-NORLEUCINE	9296
02Z	4-AMINO-2-(PHENYLAMINO)-1,3-THIAZOLE-5-CARBOXAMIDE	8270
033	N-({4'-[(1-BENZOFURAN-2-YLCARBONYL)AMINO]-1,1'-BIPHENYL-4-YL}SULFONYL)-L-VAL	7880
034	5-(4-{[3-(2,6-DICHLOROPHENYL)-5-(PROPAN-2-YL)-1,2-OXAZOL-4-YL]METHOXY}PHENYL	8423
035	(2S)-2,7-BIS(PHOSPHONOOXY)HEPTANOIC ACID	8621
036	(2R)-2,7-BIS(PHOSPHONOOXY)HEPTANOIC ACID	8622
038	3-[[(2S)-2-[[[(2S)-2-[[(2S)-2-[[(2S)-2-AZANYL-3-(1H-1,2,3,4-TETRAZOL-5-YLCARBONYLAMI	2878
039	2-((9H-PURIN-6-YLTHIO)METHYL)-5-CHLORO-3-(2-METHOXYPHENYL)QUINAZOLIN-4(3H)-	4850
03B	5-(4-CHLOROPHENYL)-4-{3-[4-(4-{[(4-{[(2R)-4-(DIMETHYLAMINO)-1-(PHENYLSULFANYL)E	3320
03C	[5-AMINO-1-(2-METHYLPHENYL)-1H-PYRAZOL-4-YL]{3-[1-(METHYLSULFONYL)PIPERIDI	3301
03F	(9Z)-N-[(2S,3R,4E)-1-(BETA-D-GLUCOPYRANOSYLOXY)-3-HYDROXYOCTADEC-4-EN-2-Y	6051
03H	(2S)-2-CHLORO-4-METHYLPENTANOIC ACID	412
03K	N-(5-CYCLOPROPYL-1H-PYRAZOL-3-YL)BENZENE-1,4-DICARBOXAMIDE	8161
03M	(5Z)-5-[(6-CHLORO-7-METHYL-1H-INDOL-3-YL)METHYLIDENE]-3-(3,4-DIFLUOROBENZYL)	7614
03P	N-{2-[4-({3-CHLORO-4-[3-(TRIFLUOROMETHYL)PHENOXY]PHENYL}AMINO)-5H-PYRROLO	4197
03Q	2-{2-[4-({5-CHLORO-6-[3-(TRIFLUOROMETHYL)PHENOXY]PYRIDIN-3-YL}AMINO)-5H-PYRI	7711
03R	2-({6-{3-[AMINO(IMINO)METHYL]PHENOXY}-3,5-DIFLUORO-4-[(1-METHYL-3-PHENYLPRO	3976

A lista de ligantes e os respectivos valores no campo <*chemicalid*> serão usados para recuperar os fármacos na base de dados DrugBank. O acesso a esse tipo de informação é dado sob a forma de uma consulta no formato URL:

http://www.drugbank.ca/unearth/q?searcher=drugs&query=het_id:NAP

http://www.drugbank.ca/drugs/DB03461

Nesse caso, o DrugBank não disponibiliza um formato de saída tipo XML. Portanto, o *pipeline* precisa de programas analisadores específicos para interpretar o

resultado das duas consultas. Todos os fármacos recuperados são registrados com seus respectivos ligantes na base de dados PostgreSQL.

3.4 Parte 4 – Categorizar

Compilar o sistema de categorias e fármacos do DrugBank permite que o *pipeline* possa relacionar o conteúdo descritivo de cada fármaco identificado no *pipeline* com a sua respectiva categoria no DrugBank para propor o reposicionamento.

O critério adotado para que um fármaco seja candidato ao reposicionamento é que este não pertença a uma categoria *antimalarials.* Então, por exemplo, se o termo empregado na busca por resumos no PubMed faz referência à doença negligenciada malária, todos os fármacos que pertençam a categoria *antimalarials* não serão reposicionados.

Portanto, categorizar nessa fase do *pipeline* compreende: compilar todas os fármacos registrados no DrugBank; compilar todas as categorias da base de dados DrugBank; relacionar os fármacos com as respectivas categorias.

Para reposicionar, é necessária a leitura dos fármacos obtidos no processamento do *pipeline,* categorizá-los de acordo com as categorias compiladas no DrugBank, eliminar os fármacos relacionados à categorias ligadas ao nome da doença e verificar se não consta no relatório *Guidelines for Treatment of Malaria in the United* States (CDC, 2013) divulgado pelo *Center for Disease Control and Prevention* (CDC, 2014). Após esse procedimento, é formado um novo subconjunto que contém os fármacos sugeridos para o reposicionamento.

Nos próximos tópicos, são descritos como foram compiladas as categorias, fármacos e as relações para desfecho do *pipeline*.

3.4.1 Compilar fármacos no DrugBank

A compilação de fármacos consiste em identificar os códigos relacionados aos fármacos cadastrados na base de dados do DrugBank. O pipeline faz uma leitura sequencial de arquivos no padrão XML (Tabela 7) ou HTML e procura identificar os marcadores para extrair, se possível, o código desejado. Os marcadores são *tags* previamente definidas e no caso dos fármacos registrados no DrugBank, são: <drugbank-id primary="true"> e </drugbank-id>. A Tabela 7, a seguir, mostra como estão dispostos alguns marcadores e o conteúdo recuperado.

Tabela 7 - Fragmento do arquivo em formato XML do DrugBank.

<?xml version="1.0" encoding="UTF-8"?>
<drugbank>
<drug type="biotech" created="2005-06-13" updated="2013-05-12">
<drugbank-id primary="true">DB00001</drugbank-id>
<drugbank-id>BIOD00024</drugbank-id>
<drugbank-id>BTD00024</drugbank-id>
<name>Lepirudin</name>
<description>Lepirudin is identical to natural hirudin except for substitution of leucine for isoleucine at the N-terminal end of the molecule and the absence of a sulfate group on the tyrosine at position 63. It is produced via yeast cells.
</description>
<cas-number>120993-53-5</cas-number>
<groups>
<group>approved</group>
</groups>
<general-references># Smythe MA, Stephens JL, Koerber JM, Mattson JC: A comparison of lepirudin and argatroban outcomes. Clin Appl Thromb Hemost. 2005 Oct;11(4):371-4. "Pubmed":http://www.ncbi.nlm.nih.gov/pubmed/16244762
Tardy B, Lecompte T, Boelhen F, Tardy-Poncet B, Elalamy I, Morange P, Gruel Y, Wolf M, Francois D, Racadot E, Camarasa P, Blouch MT, Nguyen F, Doubine S, Dutrillaux F, Alhenc-Gelas M, Martin-Toutain I, Bauters A, Ffrench P, de Maistre E, Grunebaum L, Mouton C, Huisse MG, Gouault-Heilmann M, Lucke V: Predictive factors for thrombosis and major bleeding in an observational study in 181 patients with heparin-induced thrombocytopenia treated with lepirudin. Blood. 2006 Sep 1;108(5):1492-6. Epub 2006 May 11. "Pubmed":http://www.ncbi.nlm.nih.gov/pubmed/16690967
Lubenow N, Eichler P, Lietz T, Greinacher A: Lepirudin in patients with heparin-induced thrombocytopenia - results of the third prospective study (HAT-3) and a combined analysis of HAT-1, HAT-2, and HAT-3. J Thromb Haemost. 2005 Nov;3(11):2428-36. "Pubmed":http://www.ncbi.nlm.nih.gov/pubmed/16241940
Askari AT, Lincoff AM: Antithrombotic Drug Therapy in Cardiovascular Disease. 2009 Oct; pp. 440â€“. ISBN 9781603272346. "Google books":http://books.google.com/books?id=iadLoXoQkWEC&pg=PA440. </general-references>
<synthesis-reference/>
<indication>For the treatment of heparin-induced thrombocytopenia</indication>
<toxicity>In case of overdose (eg, suggested by excessively high aPTT values) the risk of bleeding is increased.</toxicity>
<metabolism>Lepirudin is thought to be metabolized by release of amino acids via catabolic hydrolysis of the parent drug. However, con-clusive data are not available. About 48% of the administration dose is excreted in the urine which consists of unchanged drug (35%) and other fragments of the parent drug.</metabolism>
<volume-of-distribution>* 12.2 L [Healthy young subjects (n = 18, age 18-60 years)]
* 18.7 L [Healthy elderly subjects (n = 10, age 65-80 years)]
* 18 L [Renally impaired patients (n = 16, creatinine clearance below 80 mL/min)]
<categories>
<category>
<category>Antithrombins</category>
<mesh-id/>
</category>
<category>

<category>Fibrinolytic Agents</category>
<mesh-id/>
</category>
</categories>
<categories>
<category>
<category>Antithrombins</category>

O resultado da compilação é uma lista de fármacos compilados. A seguir, a Tabela 8 apresenta um fragmento desses códigos.

Tabela 8 – Lista com alguns exemplos de fármacos compilados do DrugBank.

Fármacos				
DB00563	DB01082	DB00381	DB00570	DB03128
DB02764	DB00712	DB00471	DB01394	DB04214
DB00368	DB01050	DB01426	DB02342	..
DB04582	DB00624	DB00601	DB00235	..
DB00668	DB00753	DB00760	DB00350	..
DB01064	DB01159	DB00828	DB01783	..
DB00373	DB00818	DB01669	DB01022	..
DB01421	DB00907	DB01670	DB00152	..
DB00684	DB00730	DB01764	DB00280	..
DB00919	DB00115	DB00456	DB03966	..

3.4.2 Compilar as categorias do DrugBank

Para compilar categorias no DrugBank, o pipeline faz uma leitura sequencial do arquivo XML procurando identificar os marcadores que permitem o acesso ao conteúdo desejado. Os marcadores definidos são: <categories>, <category>, </category> e </categories>.

No momento da leitura e identificação das categorias, é possível associar uma categoria a seus respectivos códigos de fármacos. O objetivo aqui é sinalizar que existem medicamentos ou princípio ativo ligado a esses códigos. Então, ao final do processamento, uma lista de fármacos categorizados estará disponível na forma de um vetor para que o pipeline possa comparar os fármacos obtidos com os categorizados. A seguir, a Tabela 9 mostra uma lista com parte dos fármacos e suas respectivas categorias.

Tabela 9 - Lista com exemplos de fármacos e suas respectivas categorias.
Fonte: DrugBank.

Fármacos	Categorias
DB00115	Anti-anemic Agents
DB00290	Antibiotics, Antineoplastic
DB00321	Antidepressive Agents, Tricyclic

DB00336	Anti-Infective Agents
DB00194	Antimetabolites
DB00320	Anti-migraine Agents
DB00291	Antineoplastic Agents, Alkylating
DB00233	Antitubercular Agents
DB00184	Autonomic Agents
DB00311	Carbonic Anhydrase Inhibitors
DB00255	Estrogens, Non-Steroidal
DB00205	Folic Acid Antagonists
DB00259	Homeopathic Agents
DB00279	Hormone Replacement Agents
DB00227	Hydroxymethylglutaryl-CoA Reductase Inhibitors
DB00199	Macrolides
DB00207	Macrolides
DB00181	Muscle relaxant, Skeletal
DB00277	Muscle Relaxants, Respiratory
DB00295	Narcotics
DB00129	Non-Essential Amino Acids
DB00142	Non-Essential Amino Acids
DB00155	Non-Essential Amino Acids
DB00140	Photosensitizing Agents
DB00201	Purinergic P1 Receptor Antagonists
DB00238	Reverse Transcriptase Inhibitors
DB00328	Tocolytic Agents
DB00235	Vasodilator Agents
DB00152	Vitamins
DB00280	Voltage-Gated Sodium Channel Blockers

A lista completa com os 254 fármacos candidatos ao reposicionamento para malária que foram relatados após o processamento do pipeline está registrada no ANEXO VI.

3.4.3 Fármacos relatados X fármacos categorizados

O próximo passo é criar uma consulta para listar todos os dados obtidos durante o processamento do pipeline, descrito no item 3.3.11 deste documento. Os dados obtidos por meio dessa consulta são denominados fármacos relatados.

A lista de fármacos relatados é relacionada com a lista de fármacos categorizados. Os fármacos que não possuem categorias são descartados automaticamente. Deste modo, fármacos sem uma categoria definida não estarão presentes na próxima etapa porque para sugerir o reposicionamento de fármacos, isto é, dar novo uso terapêutico a um medicamento, é preciso que o *pipeline* identifique no DrugBank se existe o remédio disponível para consumo imediato. O relacionamento entre fármacos relatados e categorizados permite essa identificação.

A Tabela 10, a seguir, mostra o resultado parcial do processamento de uma consulta aos fármacos categorizados, isto é, uma lista com todos os fármacos caracterizados e serão processados na próxima etapa que determinará se é um candidato ao reposicionamento.

Tabela 10 - Lista com alguns fármacos relatados no pipeline e suas respectivas categorias.
Fonte: DrugBank.

Fármaco	Descrição	Categorias
DB00001	Lepirudin	Antithrombins
DB00001	Lepirudin	Fibrinolytic Agents
DB00002	Cetuximab	Antineoplastic Agents
DB00003	Dornase alfa	Enzymes
DB00004	Denileukin diftitox	Antineoplastic Agents
DB00005	Etanercept	Immunosuppressive Agents
DB00006	Bivalirudin	Antithrombins
DB00007	Leuprolide	Antineoplastic Agents, Hormonal
DB00007	Leuprolide	Fertility Agents, Female
DB00007	Leuprolide	Estrogen Antagonists
DB00008	Peginterferon alfa-2a	Immunosuppressive Agents
DB00009	Alteplase	Thrombolytic Agents
DB00010	Sermorelin	Hormone Replacement Agents
DB00011	Interferon alfa-n1	Antiviral Agents
DB00011	Interferon alfa-n1	Immunologic Factors
DB00011	Interferon alfa-n1	Immunosuppressive Agents
DB00012	Darbepoetin alfa	Hematinics
DB00012	Darbepoetin alfa	Anti-anemic Agents
DB00013	Urokinase	Thrombolytic Agents
DB00014	Goserelin	Component

3.4.4 Fármacos candidatos ao reposicionamento

Dois critérios foram adotados para determinar se um fármaco é candidato ao reposicionamento, que são: 1) o fármaco não pertencer a uma categoria relacionada ao nome da doença negligenciada *malaria*; 2) o fármaco não estar relatado no guia de tratamento e prevenção de doenças do *Centers For Disease Control And Prevention* (CDC).

Todos os fármacos relatados no item anterior são submetidos à verificação e, ao final do processamento, uma nova lista estará formada com todos os fármacos candidatos às práticas laboratoriais para reposicionamento de fármacos para a doença negligenciada malária.

4 RESULTADOS E DISCUSSÃO

O capítulo descreve resultados obtidos durante o desenvolvimento e execução do *pipeline*. A produção intelectual está registrada em dois eventos internacionais sob a forma de um pôster no *X-Meeting* (2014)[49] e um *short paper* no evento *KDIR* (2014)[50].

Nas próximas subseções são apresentados os dados observados seguidos dos comentários, das críticas ou das respostas aos problemas e questões da pesquisa.

4.1 O Método

Diante dos desafios impostos e as questões que envolvem a pesquisa, o método proposto neste trabalho é parte dos resultados obtidos e o objetivo determinado na presente dissertação. Portanto, o método é um resultado, uma vez que o recorte feito para a dissertação de mestrado retrata todos os passos que vão desde a coleta de documentos textuais digitais, processamento e, finalmente, o relato do conjunto de proteínas sugeridas para o screening de moléculas (pesquisa básica) e dos fármacos candidatos ao reposicionamento.

Uma das limitações apresentadas pela presente dissertação é o não estabelecimento de critérios de acurácia do pipeline enquanto instrumento de diagnóstico. Nesse sentido, seria de grande valia a determinação sensibilidade e da especificidade do mesmo para identificação de fármacos reposicionáveis.

Cabe ressaltar que, no tocante à acurácia do pipeline, a maior preocupação seria quanto à sua sensibilidade, uma vez que resultados falso-negativos seriam omitidos, impedindo o reconhecimento de fármacos potencialmente reposicionáveis como tais. Por outro lado, limitações quanto à especificidade do pipeline apenas aumentariam o número de fármacos a serem eliminados ao longo do screening.

[49] *ISCB-LATIN AMERICA X-Meeting in Bioinformatics with BSB and SoiBio OCTOBER 28 – 30 2014 – Belo Horizonte, BRAZIL.*

[50] KDIR 2014 6TH International Conference on Knowledge Discovery and Information Retrieval. Rome, Italy 21-24 October, 2014.

4.2 Verificação dos resumos coletados

O pressuposto de que o número de publicações é crescente nos últimos 10 anos está confirmado. Pode-se observar, a seguir, na Figura 25, que os termos empregados como palavras-chave na busca por publicações no PubMed retornam um número crescente de artigos quando acumulados por ano.

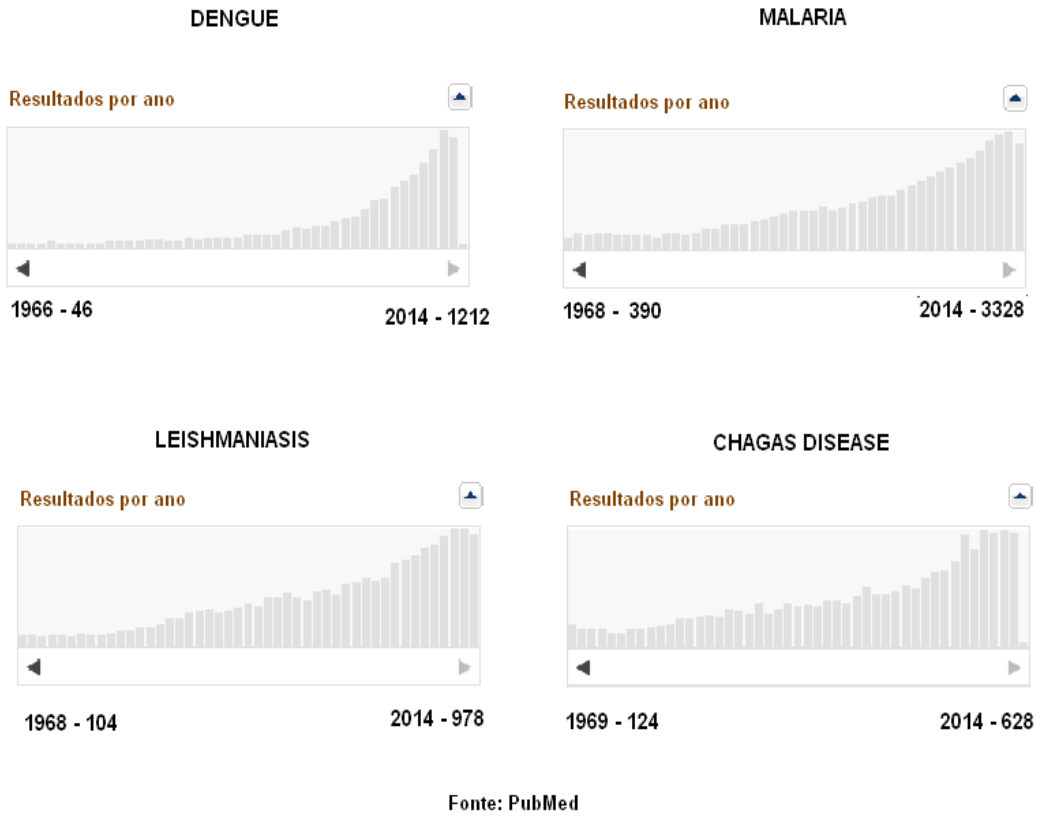


Figura 25 - Resumos publicados no PubMed.

4.3 Relatório Sintético dos dados processados no *pipeline*

Nesta seção, são descritos os dados registrados durante o processamento do pipeline, as versões dos bancos de dados biológicos e demais registros totalizados no processo no período de maio de 2014 até janeiro de 2015.

As fontes de dados para processamento no *pipeline* estão descritas abaixo, na Tabela 11. O campo <informação> expressa a versão, a quantidade de registros ou o período no qual a fonte foi consultada porque este lançamento depende da identificação da informação que não está sempre disponível no endereço do serviço.

Tabela 11 - Bases biológicas empregadas no *pipeline*.

BASES BIOLÓGICAS	INFORMAÇÃO
DrugBank	Versão 4.1
PDB	Informação para 106517 estruturas
UniProt	versão 2015_02 - 2015 UniProt Consortium
BioMed Central	Outubro 2014
PubMed	24.240.357 citações

A Tabela 12, a seguir, fornece os totais dos principais dados obtidos após a execução das quatro etapas definidas na metodologia da presente dissertação.

Tabela 12 - Registros contabilizados.

DESCRIÇÃO	TOTAL DE REGISTROS
Artigos PubMed	8444
Periódicos	2316
Termos	13815
PDB	55341
Proteínas	331756
Fármacos	2963
Ligantes	125202

Uma lista com os 50 periódicos com maior número de artigos processados no *pipeline* está disponível no ANEXO VII. A seguir, a Tabela 13 mostra os 10 periódicos que mais contribuíram na execução e processamento do *pipeline* quanto ao termo de busca da doença negligenciada malária.

Tabela 13 - Os periódicos com maior contribuição de artigos para processamento no *pipeline*.

PERIÓDICOS	REGISTROS
Malaria journal	**1534**
The American journal of tropical medicine and hygiene	**1116**
PloS one	**816**
Infection and immunity	**603**
Transactions of the Royal Society of Tropical Medicine and Hygiene	**578**
Molecular and biochemical parasitology	**494**
Tropical medicine & international health : TM & IH	**342**
Proceedings of the National Academy of Sciences of the USA	**331**
Vaccine	**319**
The Journal of infectious diseases	**318**

A lista completa contém 2316 periódicos e o *Malaria journal* figura com o maior número de contribuições com 1534 registros, seguido do *The American journal of tropical medicine and hygiene* com 1116 registros. O periódico *Memórias do Instituto Oswaldo Cruz* aparece na 25ª posição com 148 registros e o periódico *Nature* na 41ª posição com 96 artigos. Juntos, os 10 primeiros periódicos contribuem com cerca de 76% do total de artigos recuperados e processados no *pipeline.*

4.4 Resultado – Caracterização de Proteínas, Ligantes e Fármacos

Quanto à viabilidade de caracterizar proteínas segundo a sua função, caracterizar os ligantes segundo as suas propriedades terapêuticas e depois cruzar os bancos para propor reposicionamento de medicamentos, encontra-se o seguinte posicionamento: os resultados obtidos após o processamento do *pipeline* demonstram que é possível caracterizar ambos. Entretanto, o *pipeline* utilizou apenas os dados necessários para a construção de um elo entre proteínas, ligantes e os fármacos registrados no DrugBank.

A informação para caracterização de proteínas, como por exemplo, dados sobre as famílias das proteínas e organismos relacionados, foram registrados, porém não explorados (Tabela 14). Existem mais dados disponíveis quanto à função da proteína que não foram registrados porque é preciso identificar a fonte e a forma de acesso às respectivas bases biológicas. São necessários procedimentos mais sofisticados e mais tempo para implementação dessas rotinas. A Tabela 14, a seguir, mostra um registro das 55341 proteínas caracterizadas e registradas pelo *pipeline*.

Tabela 14 - Dados associados ao código da proteína 1XOA.
Fonte: PDB e UNIPROT.

Código PDB	Similares	Taxonomia	Organismo
1XOA	4OO4.A 2I4A.A 3M9J.A 4HUA.A 4WXT.A 2IFQ.A 2IFQ.B 2HSH.A 2YOI.A 2YN1.A 2J23.A 2I1U.A 4HU7.A 2VIM.A 2E0Q.A 1T00.A 3M9K.A	Homo sapiens,Acetobacter aceti,Homo sapiens,Escherichia coli K-12,Mycobacterium avium subsp. hominissuis 3388,Homo sapiens,Homo sapiens,Homo sapiens,synthetic construct,synthetic construct,Malassezia sympodialis,Mycobacterium tuberculosis,Escherichia coli K-12	Escherichia coli (strain K12)

A seguir, a Tabela 15 mostra 10 registros dos 125.202 correspondentes a caracterização dos ligantes identificados e registrados pelo *pipeline*.

Tabela 15 - Dados descritivos associados aos ligantes.
Fonte: PDB e UNIPROT.

Descrição	Ligante	Código PDB
DIETHYL ETHER	ETZ	4HP2
4-[2-(HYDROXYMETHYL)PYRIMIDIN-4-YL]-N, N-DIMETHYLPIPERAZINE-1-SULFONAMIDE	572	1PL6
(2S)-2-AMINO-2-METHYLOCTANOIC ACID	MKD	4N84
4-[(E)-{4-FORMYL-5-HYDROXY-6-METHYL-3-[(PHOSPHONOOXY)METHYL]PYRIDIN-2-YL}DIAZENYL]BENZOIC ACID	3RD 3RD	3RDH 3RDH
2-METHYL-D-NORLEUCINE	2JN	4N7Y
BETA-D-GALACTOFURANOSE	GZL	2VK2
MENADIONE	VK3	1TUV
6-DIAZENYL-5-OXO-L-NORLEUCINE	DON	2OSU
4-[(2S)-2-AMINO-3-HYDROXYPROPYL]PHENOL	TYE	1Q11
4-[(2S)-2-AMINO-3-HYDROXYPROPYL]PHENOL	TYE	2J5B

4.5 Resultado da compilação do DrugBank

Ao final do processamento do *pipeline,* para o nome da doença *malaria,* um conjunto com 2963 fármacos candidatos ao reposicionamento estava registrado. Isto significa que existem 2963 fármacos, componentes bioativos ou medicamentos, candidatos ao reposicionamento para a doença malária. Isto é, existem fármacos registrados no DrugBank destinados ao tratamento de outra patologia e, que após o processamento do *pipeline,* são sugeridos para o tratamento da doença malária. A confirmação desses registros somente poderá ser feita por meio de testes laboratoriais. O *pipeline* apenas sugere um conjunto candidato ao reposicionamento.

Nessa subseção, para melhor compreensão do que é apresentado como resultado e proposto pelo método descrito nesta dissertação, a figura 26 ilustra por meio da cadeia de produção de fármacos a fase onde estão localizados 3 dos 2963 candidatos ao reposicionamento. São fármacos destinados ao tratamento de doenças como câncer ou com a função de antidepressivo, por exemplo, que são indicados no tratamento de outra doença que, no caso da presente dissertação, é a malária.

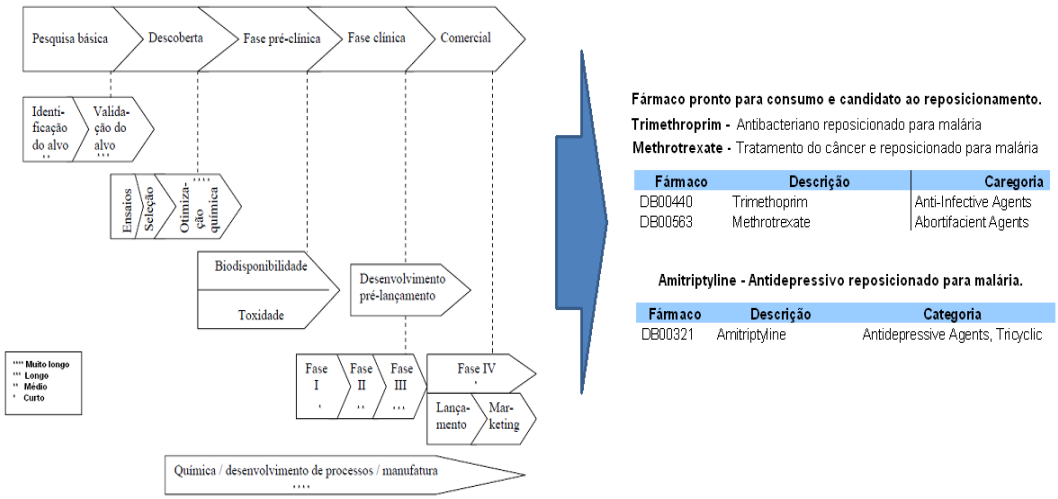


Figura 26 - Fármacos já desenvolvidos e reposicionados para tratamento da malária.

Um resultado relevante se for considerado todos os investimentos empregados na fase da pesquisa básica. Contudo, torna-se necessário nesse trabalho acadêmico um refinamento em busca de evidências de que no conjunto há fármacos reposicionáveis.

Portanto, alguns procedimentos e critérios foram adotados buscando aprimorar e refinar esse conjunto. Então, foram compiladas e relacionadas 8474 fármacos em 480 categorias do DrugBank. O resultado da compilação é uma tabela com 14849 registros de fármacos categorizados do DrugBank, isto é, o código do fármaco associado a uma ou mais categorias. A solução apresentada visa minimizar o efeito de fármacos registrados no DrugBank sem informação adicional para no mínimo identificar a existência de um medicamento no mercado de consumo.

O *pipeline* pode agora comparar os 2963 fármacos com os dados compilados do DrugBank fazendo uso de uma relação simples entre um fármaco sugerido pelo *pipeline* para o reposicionamento e uma tabela com fármacos categorizados do DrugBank. O resultado parcial dos fármacos sugeridos pelos *pipeline* e agora categorizados é apresentado, a seguir, na Tabela 16. Após a relação com categorias o número de candidatos ao reposicionamento foi reduzido para um total de 254 fármacos.

Uma vez categorizados,é possível reduzir ainda mais esse conjunto de 254 fármacos, e com esse objetivo, dois critérios foram adotados: 1) eliminar os fármacos categorizados como *antimalarials* porque são destinados ao tratamento da malária. Portanto, não podem ser reposicionados; 2) constar na lista Center for

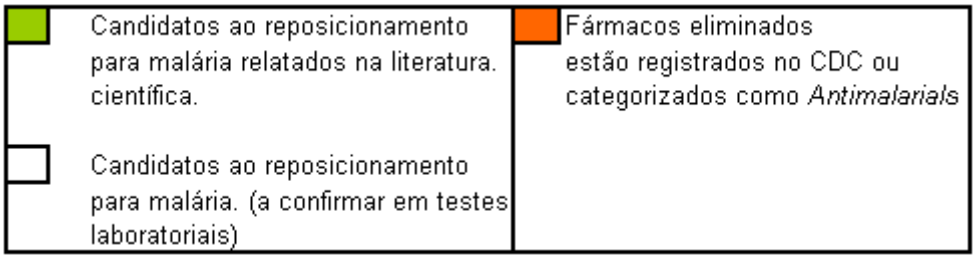
Disease Control and Prevention (CDC) (CDC, 2013) e com indicação para tratamento da doença malária.

A seguir, na Tabela 16, seguem marcados em vermelho, fármacos eliminados por estarem categorizados como *antimalarials* ou constarem na lista Center for Disease Control and Prevention (CDC) (CDC, 2013). Seguem marcados em verde, fármacos candidatos ao reposicionamento assinalados na literatura científica. Os 249 fármacos restantes são candidatos ao reposicionamento para a doença malária e esta confirmação deve ocorrer por meio de testes em práticas laboratoriais.

As subseções 4.6 e 4.7 descrevem com mais detalhes os fármacos eliminados e os fármacos confirmados na literatura científica como candidatos ao reposicionamento para a doença malária.

Tabela 16 - Fármacos candidatos ao reposicionamento após a categorização do DrugBank.

Fármaco	Descrição	Categoria
DB00563	Methotrexate	Abortifacient Agents
DB01157	Trimetrexate	Antibiotics
DB01190	Clindamycin	Antibiotics
DB01103	Quinacrine	Anticestodal Agents
DB00715	Paroxetine	Antidepressive Agents
DB04599	Aniracetam	Antidepressive Agents
DB00321	Amitriptyline	Antidepressive Agents, Tricyclic
DB01910	Adenosyl-Ornithine	Antifungal Agents
DB03600	Decanoic Acid	Antifungal Agents
DB03601	5-deoxyflavanone	Antifungal Agents
DB00440	Trimethoprim	Anti-Infective Agents
DB00741	Hydrocortisone	Anti-Inflammatory Agents
DB00693	Fluorescein	Fluorescent Dyes
DB00205	Pyrimethamine	Folic Acid Antagonists
DB08878	Aminopterin	Folic Acid Antagonists
DB00974	Edetic Acid	Food Additives
DB03793	Benzoic Acid	Food Preservatives
DB00437	Allopurinol	Free Radical Scavengers



4.6 Resultado observando critérios de eliminação

Da lista de fármacos, foram descartados do conjunto, os fármacos categorizados como *antimalarials* listados na Tabela 17 ou registrados pelo guia de tratamento e prevenção de malária do *CDC* (CDC, 2013) apresentados na Tabela 18. São eles:

Fármacos classificados como *Antimalarials*

Tabela 17 - Fármacos classificados como *antimalarials.*

Fármaco	Descrição	Categoria
DB01103	Quinacrine	Anticestodal Agents
DB00205	Pyrimethamine	Folic Acid Antagonists

Fármacos listados no CDC

Tabela 18 - Fármaco listado no CDC.

Fármaco	Descrição	Categoria
DB01190	Clindamycin	Antibiotics

4.7 Resultado observando literatura científica – O Reposicionamento de Fármacos

Os resultados obtidos após a execução do *pipeline* e apresentados nesta seção são evidências de que o método é capaz de sugerir o reposicionamento de fármacos ao explorar a literatura científica e relacionar ontologias às bases de dados do segmento biomédico.

No presente trabalho não foram realizados os testes laboratoriais para comprovar que alguns fármacos relatados no *pipeline* são reposicionáveis. Portanto, foram realizadas consultas à literatura científica biomédica em busca de relatos que comprovem o uso de alguma substância obtida no processamento do *pipeline* candidata ao reposicionamento de fármaco. Isto é, foi realizada uma busca não exaustiva na literatura científica por estudos que relatem fármacos empregados originalmente no tratamento de alguma outra doença e que este tenha sido reutilizado no tratamento da malária. Após o procedimento da busca foram comprovados três fármacos que são utilizados nas práticas laboratoriais para reposicionamento de fármacos. Os três fármacos constam nos relatos de Nzila et al., (2011) e no artigo de Partha et al. (2013) juntamente com as respectivas referências e dados caracterizados.

Os dois fármacos (DB00440 e DB00563) sugeridos para o reposicionamento foram encontrados no artigo *Drug repositioning in the treatment of malaria and* TB (NZILA et al., 2011) cuja referência é destacada, a seguir. A Figura 27 exibe os fármacos citados no trabalho de Nzila e a Tabela 19 contém os fármacos obtidos no *pipeline.*

Artigo: Drug repositioning in the treatment of malaria and TB. Nzila A, Ma Z, Chibale K. Drug repositioning in the treatment of malaria and TB. Future Medicinal Chemotherapy,; n. 3, p. 1413–1426, 2011. [acesso em 15 dez 2014]. Disponível em*:* http://tballiance.org/newscenter/research_papers/2011_FMC%20Drug%20reposistioning%20in%20the%20treatment%20of%20malaria%20and%20TB.pdf

Segundo Nzila et al. (2011) o antibacteriano trimethoprim, destinado ao tratamento de Infecção do Trato Urinário, foi utilizado no passado para tratamento da malária e segue atualmente sendo avaliada em combinações com derivados de artemisinina. Por outro lado, o *antifolates methotrexate* é um agente usado como uma droga anticâncer e testes demonstram que pode bloquear o crescimento do parasita da malária. Segue a figura 27 com outros fármacos relatados nos estudos de Nzila.

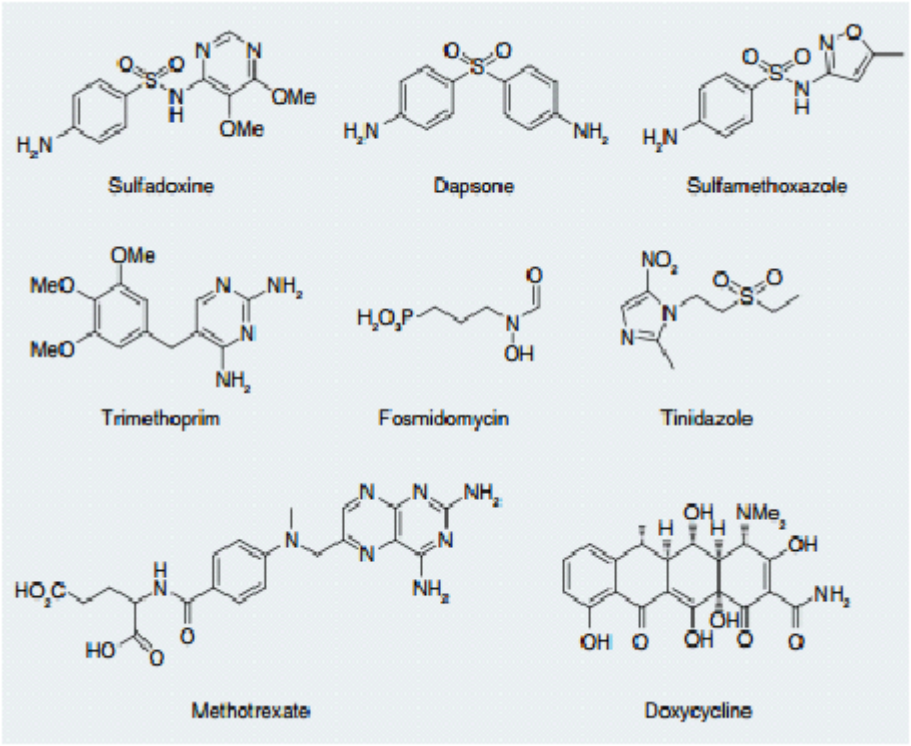


Figura 27 - Fármacos selecionados que foram reposicionados para o tratamento de malária.
Fonte: (Nzila et. al., 2011)

Tabela 19 - Trimethoprim e Methrotrexate relatados no artigo de Nzila e os códigos do DrugBank.
Fonte: (Nzila et. Al, 2011)

Fármaco	Descrição	Categoria
DB00440	Trimethoprim	Anti-Infective Agents
DB00563	Methrotrexate	Abortifacient Agents

Na figura 28, a seguir, há mais detalhes obtidos no DrugBank sobre o fármaco DB00440 denominado Trimethoprim. A consulta com todos os detalhes e a descrição completa do fármaco pode ser realizada por meio da URL: *http://www.drugbank.ca/drugs/DB00440.*

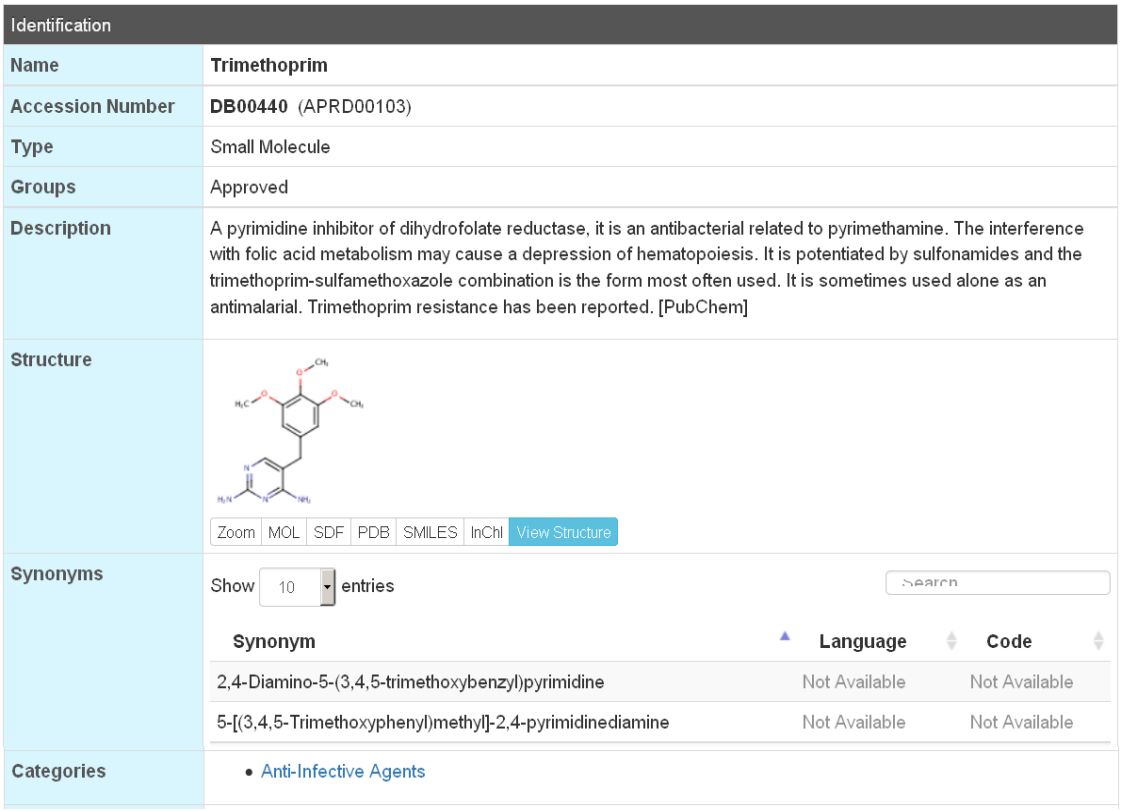


Figura 28 - Descrição para o fármaco DB00440 - Trimethoprim
Fonte: DrugBank

A figura 29, a seguir, exibe os dados referentes ao fármaco DB00563 denominado methotrexate. Assim como o DB00440 uma descrição mais detalhada também pode ser obtida no DrugBank com a URL: *http://www.drugbank.ca/drugs/DB00563.*

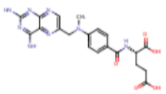
Identification	
Name	**Methotrexate**
Accession Number	**DB00563** (APRD00353)
Type	Small Molecule
Groups	Approved
Description	An antineoplastic antimetabolite with immunosuppressant properties. It is an inhibitor of tetrahydrofolate dehydrogenase and prevents the formation of tetrahydrofolate, necessary for synthesis of thymidylate, an essential component of DNA. [PubChem]
Structure	 Zoom MOL SDF PDB SMILES InChI View Structure
Categories	• Antirheumatic Agents • Dermatologic Agents • Immunosuppressive Agents • Enzyme Inhibitors • Folic Acid Antagonists • Nucleic Acid Synthesis Inhibitors • Antimetabolites, Antineoplastic • Abortifacient Agents, Nonsteroidal • Abortifacient Agents

Figura 29 - Descrição para o fármaco DB00563 – methotrexate
Fonte: DrugBank

O terceiro fármaco foi relatado nos estudos de Partha et al. (2013) cuja descrição é *Amitiptyline* (Tabela 20). Portanto, também faz parte do resultado do *pipeline* porque não foi categorizado como um *antimalarials* no DrugBank e não está listado no CDC.

É descrito, a seguir, a referência do artigo de Partha e os dados caracterizados para DB00321 correspondente a *Amitiptyline.*

Artigo "*Reuse of Old, Existing, Marketed Non-antibiotic Drugs as Antimicrobial Agents: a New Emerging Therapeutic Approach "* Partha Palit, Subhash C. Mandal and Nirup Bikash Mandal. 2013. Microbial pathogens and strategies for combating them: science, technology and education. Vol.3 [acesso em 15 dez. 2014] Disponível em: http://www.formatex.info/microbiology4/vol3/1883-1892.pdf

O artigo de Partha (2013) destaca a importância do reposicionamento de fármacos quanto à redução de custos proporcionada pela vantagem de reposicionar um medicamento existente. Partha, também considera emergencial a fase de testes em laboratórios com fármacos, moléculas ou proteínas reposicionáveis dado o

problema mundial causado por doenças negligenciadas e fornece informação relevante sobre fármacos reposicionados que já foram relatados em revistas científicas como modelo pré-clínico.

Tabela 20 - Amitriptyline relatado no artigo de Partha.
Fonte: (Partha, 2013).

Fármaco	Descrição	Categoria
DB00321	Amitriptyline	Antidepressive Agents, Tricyclic

Dentre outros fármacos listados por Partha, a Figura 30, abaixo, exibe o Amitriptyline, que é descrito com a sua aplicação original (Antidepressivo) e o novo uso (Antimalarial e Agente anti-leshmania).

Table 1 Additional use of drugs discovered during or after clinical usage

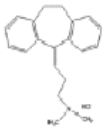
Drug (Classification)	Structure	Initial Use	Additional or New Primary Use
Amitriptyline		Tricyclic antidepressants	Antimalarial and anti leishmanial agent

Figura 30 - Classificação da droga Amitriptyline e nova aplicação.
Fonte: (Partha, 2013).

Na figura 31, a seguir, há dados referentes ao fármaco DB00321 denominado Amitriptyline. A descrição detalhada do fármaco DB00440 pode ser obtida no DrugBank com a URL: *http://www.drugbank.ca/drugs/DB00312.*

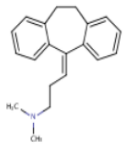
Identification	
Name	**Amitriptyline**
Accession Number	**DB00321** (APRD00227)
Type	Small Molecule
Groups	Approved
Description	Amitriptyline hydrochloride is a dibenzocycloheptene-derivative tricyclic antidepressant (TCA). TCAs are structurally similar to phenothiazines. They contain a tricyclic ring system with an alkyl amine substituent on the central ring. In non-depressed individuals, amitriptyline does not affect mood or arousal, but may cause sedation. In depressed individuals, amitriptyline exerts a positive effect on mood. TCAs are potent inhibitors of serotonin and norepinephrine reuptake. Tertiary amine TCAs, such as amitriptyline, are more potent inhibitors of serotonin reuptake than secondary amine TCAs, such as nortriptyline. TCAs also down-regulate cerebral cortical β-adrenergic receptors and sensitize post-synaptic serotonergic receptors with chronic use. The antidepressant effects of TCAs are thought to be due to an overall increase in serotonergic neurotransmission. TCAs also block histamine-H_1 receptors, α_1-adrenergic receptors and muscarinic receptors, which accounts for their sedative, hypotensive and anticholinergic effects (e.g. blurred vision, dry mouth, constipation, urinary retention), respectively. See toxicity section below for a complete listing of side effects. Amitriptyline may be used to treat depression, chronic pain (unlabeled use), irritable bowel syndrome (unlabeled use), diabetic neuropathy (unlabeled use), post-traumatic stress disorder (unlabeled use), and for migraine prophylaxis (unlabeled use).
Structure	 Zoom | MOL | SDF | PDB | SMILES | InChI | View Structure
Categories	• Adrenergic Uptake Inhibitors • Analgesics, Non-Narcotic • Antidepressive Agents, Tricyclic

Figura 31 - Descrição para o fármaco DB00321 – Amitriptyline
Fonte: DrugBank

Contudo, há um conjunto de 249 fármacos candidatos ao reposicionamento e que precisam de uma investigação clínica mais detalhada. A investigação clínica compreende alguns dos procedimentos científicos adotados na fase da pesquisa básica que demandam por muito tempo e um elevado custo financeiro, tais como: screening de moléculas, insumos sob registro de patentes e outros.

O método permite sugerir uma nova forma de colaborar junto a esses procedimentos científicos, reduzindo tempo e os custos da pesquisa básica em projetos que envolvem o fabrico de novos medicamentos.

O relato feito nesta seção atende aos dois últimos objetivos específicos descritos na seção 2.2 do presente trabalho.

4.8 Resultados para a pesquisa básica

A pesquisa básica consiste na análise de novos compostos que se mostrem promissores no combate a alguma patologia. A busca por moléculas similares em estrutura ou função que possam contribuir para a síntese ou análise de novos compostos demanda por tempo e muito dinheiro. Se o pipeline retorna um conjunto

de fármacos candidatos ao reposicionamento, é possível propor que o conjunto de moléculas ou proteínas que traduzem esse conjunto de fármacos são candidatos ao *screening* de moléculas. O pressuposto é que exista a possibilidade de reduzir a busca tão dispendiosa a um primeiro conjunto que satisfaça a determinada demanda na pesquisa básica, no caso do presente trabalho, moléculas ou proteínas que tenham potencial como alvo tratamento terapêutico para doenças negligenciadas.

Para melhor compreensão do que é apresentado como resultado, nessa subseção e proposto nos objetivos específicos dessa dissertação, a figura 32 ilustra por meio da cadeia de produção de fármacos a fase onde o pesquisador demanda por moléculas similares na sua estrutura ou função, ou seja, são as moléculas candidatas ao *screening* durante a fase inicial da pesquisa básica.

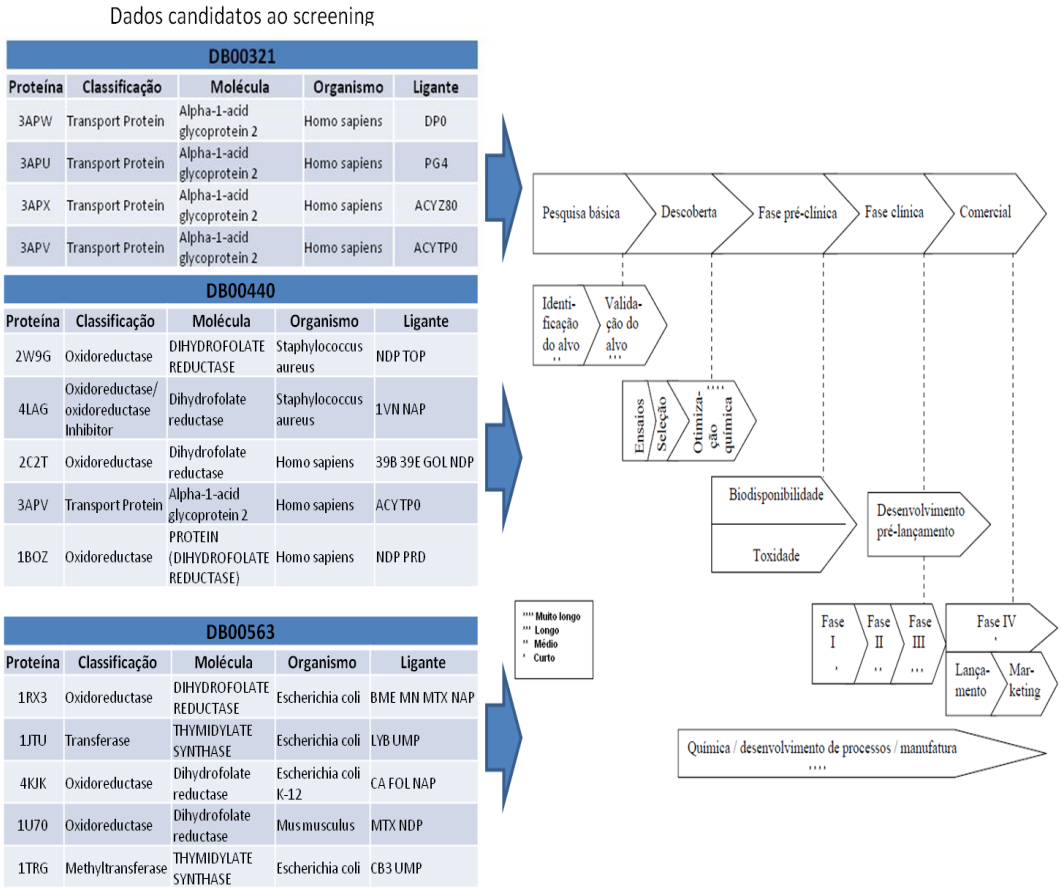


Figura 32 - Candidatos ao screening de moléculas

Os dados candidatos ao screening apresentados no pipeline para reposicionamento de fármacos e, confirmados na literatura científica, são: DB00321, DB00440 e DB00563. Portanto, existe um conjunto de moléculas de interesse à pesquisa básica relacionadas a esses fármacos que foram registradas durante o

processamento do *pipeline*. Nas Tabelas 21, 22 e 23, a seguir, são relatados dados caracterizados e elas contêm informação sobre proteínas, classificação, organismo e respectivo(s) ligante(s). Tais dados estão associados aos fármacos listados após o processamento e validação por meio da literatura científica.

Tabela 21 - Fármaco DB00321 – Amitriptyline – Dados para screening de moléculas.

DB00321				
Proteína	**Classificação**	**Molécula**	**Organismo**	**Ligante**
3APW	Transport Protein	Alpha-1-acid glycoprotein 2	Homo sapiens	DP0
3APU	Transport Protein	Alpha-1-acid glycoprotein 2	Homo sapiens	PG4
3APX	Transport Protein	Alpha-1-acid glycoprotein 2	Homo sapiens	ACY Z80
3APV	Transport Protein	Alpha-1-acid glycoprotein 2	Homo sapiens	ACY TP0

Tabela 22 - Fármaco DB00440 – Trimethoprim – Dados para screening de moléculas.

DB00440				
Código PDB	**Classificação**	**Molecula**	**Organismo**	**Ligante**
2W9G	Oxidoreductase	DIHYDROFOLATE REDUCTASE	Staphylococcus aureus	NDP TOP
4LAG	Oxidoreductase/ oxidoreductase Inhibitor	Dihydrofolate reductase	Staphylococcus aureus	1VN NAP
2C2T	Oxidoreductase	Dihydrofolate reductase	Homo sapiens	39B 39E GOL NDP
3APV	Transport Protein	Alpha-1-acid glycoprotein 2	Homo sapiens	ACY TP0
1BOZ	Oxidoreductase	PROTEIN (DIHYDROFOLATE REDUCTASE)	Homo sapiens	NDP PRD

Tabela 23 - Fármaco DB00563 – Methrotrexate – Dados para screening de moléculas

DB00563				
Código PDB	**Classificação**	**Molecula**	**Organismo**	**Ligante**
1RX3	Oxidoreductase	DIHYDROFOLATE REDUCTASE	Escherichia coli	BME MN MTX NAP
1JTU	Transferase	THYMIDYLATE SYNTHASE	Escherichia coli	LYB UMP
4KJK	Oxidoreductase	Dihydrofolate reductase	Escherichia coli K-12	CA FOL NAP
1U70	Oxidoreductase	Dihydrofolate reductase	Mus musculus	MTX NDP
1TRG	Methyltransferase	THYMIDYLATE SYNTHASE	Escherichia coli	CB3 UMP

Há evidências que apontam a viabilidade de contribuir com informação para procedimentos realizados na pesquisa básica, isto é, o *screening* de moléculas. Contudo, o conjunto apenas sugere nomes de proteínas candidatas ao

reposicionamento e passível de verificação por laboratório. O intuito desta subseção é suprir pesquisadores com informação para um “primeiro” conjunto para testes na fase inicial e tão dispendiosa que é a fase de pesquisa básica.

Os conjuntos apresentados anteriormente nas Tabelas 21, 22 e 23 apresentam uma parcela das respostas obtidas pelo *pipeline.* O ANEXO VIII apresenta um conjunto completo de dados para pesquisa básica a partir dos fármacos reposicionados para malária. O conjunto de fármacos no final do processamento registra um total de 2963 candidatos. Se considerar apenas os resultados para fármacos categorizados, o número é reduzido para 249 fármacos e cada um com suas respectivas proteínas destinadas a suprir uma demanda originada na pesquisa básica. Devem ser caracterizados como as tabelas descritas acima de forma a permitir, na pesquisa básica, a reutilização de qualquer desses dados na fase preliminar da cadeia de produção de fármacos.

Os fármacos identificados foram categorizados de acordo com suas atividades farmacológicas primárias ou maior relevância, conforme demonstrado no ANEXO VI. Após a categorização, realizou-se a distribuição desses fármacos conforme as atividades farmacológicas, apresentado na figura 33, a seguir:

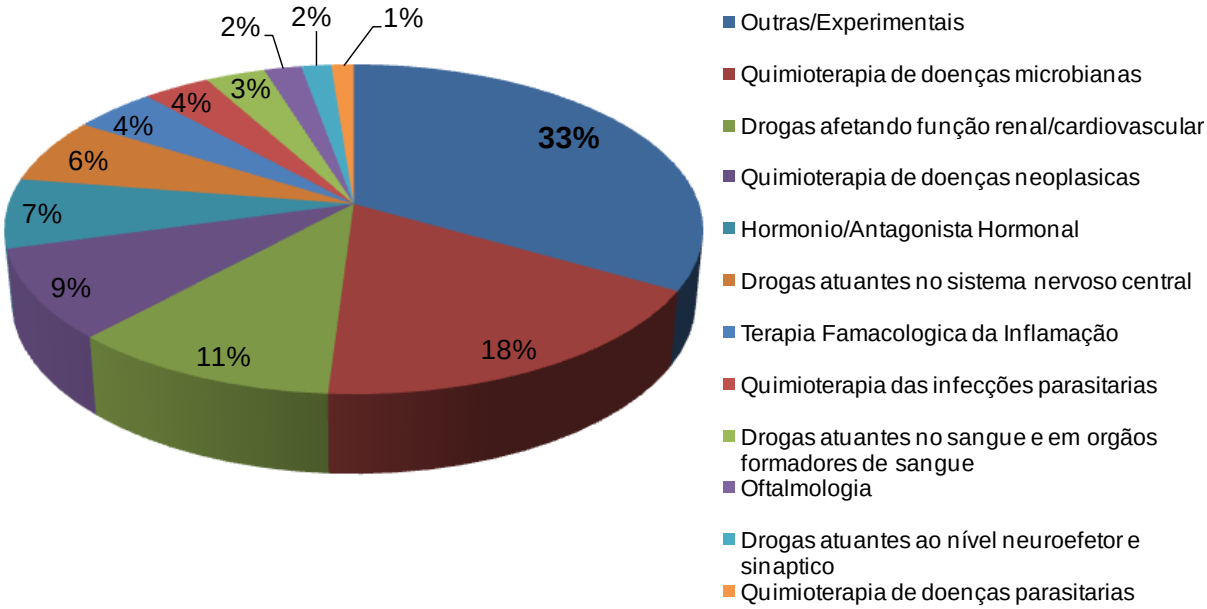


Figura 33 - Classificação dos resultados conforme atividades farmacológicas (BRUNTON et al., 2011)

Em relação às categorias de atividade farmacológica dos candidatos identificados, aquelas de maior frequência foram: Outras/experimentais (33%), Quimioterapia de doenças microbianas (18%), Drogas afetando a função

renal/cardiovascular (11%), Quimioterapia de doenças neoplásicas (9%) e Hormônios/antagonistas hormonais (7%).

A maior alocação de candidatos na categoria Outras/experimentais é justificada pelo fato de diversos fármacos classificados como candidatos atuarem mais à jusante em vias de sinalização celular, determinando menor especificidade de efeitos e consequentemente, suas alocações em mecanismos em uma miríade de classes. Em adição, os fármacos experimentais são assim denominados dada sua classificação no DrugBank. Por experimental, entende-se que são as drogas mostradas experimentalmente para se ligar a proteínas específicas em mamíferos, bactérias, vírus, fungos ou parasitas. Um medicamento experimental não está necessariamente sendo, em caráter formal, investigado.

A segunda categoria mais representada (Quimioterapia de doenças microbianas) refere-se a um termo guarda-chuva, encampando desde moléculas antirretroviais (como inibidores de protease) a fármacos antibacterianos e antibióticos, estes em maior proporção.

Dentre a categoria farmacológica “Drogas afetando a função renal/cardiovascular”, encontram-se moléculas inibidoras do transporte de sódio (a exemplo da anidrase carbônica) e potássio, diuréticos e antiarrítmicos. Fármacos utilizados na quiomioterapia de neoplasias e hormônios/antagonistas hormonais contribuem em conjunto para 1 em cada 5 dos candidatos identificados.

5 CONCLUSÃO E PERSPECTIVAS

O trabalho realizado consistiu no desenvolvimento de uma metodologia que visa a busca de nomes de proteínas cuja combinação das estruturas proteicas, ligantes e fármacos proporcionem um conjunto de fármacos candidatos ao reposicionamento e um conjunto de proteínas para a fase de pesquisa básica, o *screening* de moléculas. A metodologia partiu de documentos textuais digitais registrados no PubMed, submetidos a um *pipeline* que explora bases de dados da área biomédica como termos anotados semanticamente, mineração de textos e de relacionamentos entre dados.

A iniciativa do *OBO Foundry* e as ontologias ampliadas e disponibilizadas pelo BioPortal NCBO abrem um novo caminho que possibilita aos pesquisadores meios de construir ferramentas que delineiam um novo arranjo e interpretações diferentes entre o que é produzido e registrado em documentos textuais digitais no domínio biomédico. Portanto, viabilizam o reuso da informação que processada, gera resultados em um novo ambiente diferente do qual textos e documentos digitais foram originalmente concebidos.

O *pipeline* cumpre a sua função de reunir um conjunto de moléculas e fármacos otimizados para pesquisa básica e reposicionamento de fármacos. Um trabalho realizado por meio de métodos computacionais para explorar bases de documentos textuais digitais, relacionar termos às bases de dados e repositórios da área biomédica. Contudo, além da verificação da literatura para certificação dos conjuntos, um teste e exploração *in-silico* com especialistas é necessário para que o método seja funcional e útil.

Os fármacos reposicionáveis, confirmados na literatura científica, utilizados em experimentos destinados ao tratamento da malária e o respectivo código do DrugBank são: DB00321, Amitriptyline; DB00563, Methrotrexate; DB00440 – Trimethoprim. Há mais dois conjuntos de dados, que são: 1) um conjunto de 249 fármacos restantes candidatos ao reposicionamento para a doença malária e deverão ser confirmados posteriormente em práticas laboratoriais; 2) um conjunto que sugere moléculas para o *screening* (fase inicial da pesquisa básica) e que foram

obtidos a partir das relações existentes entre os fármacos identificados na literatura como reposicionáveis e a base de dados utilizada (ANEXO IX), com os dados processados, pelo *pipeline*.

O trabalho realizado abre diversas frentes para investigação e pesquisa:

- Informação - No âmbito da informação *online,* explorar recursos computacionais para identificar fontes de resumos, artigos completos e outros dados que possam contribuir para o que fomenta o mecanismo propulsor do *pipeline* que são os artigos científicos. Por outro lado, explorar e tornar o *pipeline* um serviço público, com um processo automático e contínuo capaz de identificar textos e processá-los até as indicações dos fármacos que poderão servir às práticas laboratoriais voltadas para o reposicionamento de fármacos.

- Ontologias - Compilação de classes, termos, descrições e anotação semântica que respondam a uma tupla {proteínas, fármacos, nomes de doenças}. A criação de uma ontologia que por meio de uma API forneça termos, sinônimos e um padrão de resposta inteligível para homens e máquinas.

- Similaridade de Proteínas – Reuso de técnicas que permitem caracterizar proteínas no PDB com a construção de clusters em R, por exemplo. A ideia é buscar dados relacionados às proteínas registradas com intuito de refinar o *pipeline* e evitar um grande número de registros e fármacos para processamento.

- Bancos de Dados Públicos - Ampliar o uso de bases de dados disponíveis atualmente na Internet. Explorar, além do DrugBank, PDB, UniProt outras bases de domínio público tais como KEGG, *Drugs.com* e *U.S. Food and Drug Administration*. Ampliar as possibilidades de comparar, categorizar e reposicionar fármacos.

Algumas dificuldades durante a investigação e pesquisa são discutidas a seguir:

- Qual ontologia usar? Dado que existe um grande número de ontologias na área biomédica, como por exemplo, o BioPortal NCBO que oferece 405 ontologias, além de reorganizar e estabelecer relações entre elas. Determinar a ontologia empregada para os primeiros testes do *pipeline* reservou um tempo considerável na bateria de testes. Sendo importante considerar, para a escolha da ontologia adequada, os termos anotados semanticamente, os termos que traduzem nomes de proteínas e a quantidade de termos anotados. Além de considerar diretamente os recursos disponíveis para processar, relacionar e registrar uma grande quantidade de dados.

- Qual serviço de anotação semântica que retorna dados anotados semanticamente a partir de um termo e uma consulta? A busca por um portal de ontologias, que disponibilizasse o recurso desejado, requereu uma investigação não somente para a escolha do software como por qual formato a ontologia estaria disponível. Assim, foram identificados os seguintes portais: 1) *Terminizer* (NERC, 2015), o qual detecta termos ontológicos em pedaços de texto, tais como publicações ou anotações experimentais. As ontologias compiladas no serviço *web* têm suas origens do *OBO Foundry* e está incompleta. Situação: descontinuado; 2) AutôMeta (FONTES, 2011), mecanismo de anotação semântica multiplataforma e *multi-interface* que permite desde uma anotação simples até múltiplas. O mecanismo está operacional, porém, exige formatos OWL, RDF ou N-Triples. No período de desenvolvimento do *pipeline,* a ontologia *Proteins Ontology* (PO) disponibilizada no portal *OBO Foundry,* em formato OWL, não estava ativa para download. Situação: descartado; 3) Owltools (OWLTOOLS, 2015), um conjunto de ferramentas projetadas para lidar com arquivos de ontologia. Owltools possui grande número de recursos, métodos para propriedades OBO, classes, definições textuais, etc. Extensa documentação e muito complexo para a atual fase e o propósito deste *pipeline.* Situação: descartado.

- Modificações nas interfaces HTML para exibição dos dados sobre proteínas, ligantes ou fármacos. Qualquer modificação no nome do

marcador ou a troca de posição na página implica na revisão do programa analisador.

- Restrições impostas pelo PubMed para *download* de arquivos digitais por meio de processamentos automáticos. Há um controle para evitar um grande número de downloads no PubMed. O objetivo é de não sobrecarregar os servidores que desempenham tarefas específicas e não são dedicados somente a um único usuário. Portanto, há recomendações para que um lote de processamento que demanda um grande número de downloads de arquivos ou maior tempo nos servidores seja realizado nos finais de semana, feriados ou horários de pouco uso dos sistemas.

Perspectivas e trabalhos futuros:

- Estruturas Similares: O *pipeline* registra informação específica sobre as proteínas identificadas durante a sua execução. Logo, complementar esses dados com mais informação sobre: estrutura, família e organismos.

- Uma das funções do *pipeline* é coletar dados sobre a estrutura das proteínas obtidas do UniProtKB. A informação sobre estruturas é obtida do próprio PDB. Portanto, explorar outras bases de dados por meio dos nomes das proteínas e registrar informação adicional de acordo com o propósito do *pipeline*.

- Empregar recursos de programação com intuito de reduzir os custos de processamento principalmente nas busca por técnicas e métodos que visem reduzir o tempo de processamento do *pipeline*; recursos estatísticos e matemáticos com objetivo de construir uma camada de confiança estatística que poderia reduzir os falsos positivos no processos de busca e recuperação da informação.

- Comparar a lista de candidatos à lista dos compostos de outras fontes que sejam provedores de dados testados para atividade anti-malarial com os compostos da GSK, por exemplo.

Todos os códigos e bibliotecas empregados na construção do pipeline serão disponibilizados por meio de um portal ou repositório de acesso livre ficando a cargo da FIOCRUZ a reprodução e divulgação do material.

6 REFERÊNCIAS BIBLIOGRÁFICAS

Adamic LA, Wilkinson D, Huberman BA, Adar E. 2002. A literature based method for identifying gene-disease connections. In Proceedings of the IEEE Computer Society Conference on Bioinformatics, Stanford, CA, pp. 109–117.

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Biologia molecular da célula. 3ª ed. Porto Alegre: Artes Médicas, 1997. 128-129 p.

Alecrim MGC, Carvalho LM, Andrade SD, Arcanjo An+ RL, Alexandre MA, Alecrim WD. Tratamento de crianças com malária pelo *Plasmodium Falciparum* com derivados da artemisinina. Rev Soc Bras Med Trop [periódico na Internet]. 2003 [acesso em nov 2013]. 36(2):223-226. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0037-86822003000200005&lng=en. http://dx.doi.org/10.1590/S0037-86822003000200005.

Al-Mubaid H, Singh RK. 2005. A new text mining approach for finding protein-to-disease associations. Am J Biochem Biotechnol, n.1, p. 145–152. [acesso em 3 out 2014]. Disponível em http://thescipub.com/PDF/ajbbsp.2005.145.152.pdf

Andrade M, Valencia A. 1998. Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. Bioinformatics, v.14. [acesso em 5 out 2014]. Disponível em: http://bioinformatics.oxfordjournals.org/content/14/7/600.full.pdf+html

API Documentation [internet]. 2014. [acesso em 8 dez 2015]. Disponível em: http://data.bioontology.org/documentation

Arzucan Özgür, Thuy Vu, Güneş Erkan, and Dragomir R. Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network Bioinformatics (2008) 24 (13): i277-i285 doi:10.1093/bioinformatics/btn182 [acesso em 5 out 2014]. Disponível em: http://bioinformatics.oxfordjournals.org/content/24/13/i277.full

Attwood TK, Kell DB, Mcdermott P, Marsh J, Pettifer SR, Thorne D. Calling international rescue: Knowledge lost in literature and data landslide! Biochem. J., v. 424, p. 317–333, 2009. [acesso em 18 jan 2011]. Disponível em: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2805925/pdf/bj4240317.pdf.

Barbosa-Silva A, Fontaine J, Donnard ER, Stussi F, Ortega JM, Andrade-Navarro, MA. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. BMC Bioinformatics , v. 12, p. 435, 2011.

Baeza-Yates AR, Ribeiro-Neto B. Modern information retrieval [Internet]. New York: ACM Press, 1995. [acesso em 21 abr 2014]. Disponível em www.cs.ucla.edu/~miodrag/cs259-security/baeza-yates99modern.pdf

Barçante E. Propostas e metodologias de processamento automático de documentos textuais digitais: uma análise da literatura, Dissertação (Mestrado em Ciência da Informação) – Universidade Federal Fluminense / Programa de Pós-

Graduação em Ciência da Informação, 2011. [acesso em 1 out 2014]. Disponível em: http://www.ci.uff.br/ppgci/arquivos/Dissert/Diss_EduardoBarcante.pdf

Belloze KT. 2013. Priorização de alvos para fármacos no combate a Doenças Tropicais Negligenciadas causadas por protozoários. FIOCRUZ/IOC.

Berners-Lee T, Hendler J, Lassila O. The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American Magazine. p. 34-43, May, 2001.

Berners-Lee T, Hendler J, Lassila O. The semantic web, scientific American, May 2001. [acesso em 27 set 2014]. Disponível em: http://www.w3.org/2001/sw

Berman HM. Westbrook J, Zukang Feng, Gilliland G, Bhat Weissig TN, H, Shindyalov IN, Bourne PE. The Protein Data Bank Nucl. Acids Res. *(2000) 28 (1): 235-242 doi:10.1093/nar/28.1.235* [acesso em 20 out 2014]. Disponível em: http://nar.oxfordjournals.org/content/28/1/235.full.pdf+html

Black J. Drug pioneers win Nobel Laureate. This Week. 1988. New Scientist 22 October 1988. [intermet] [acesso em 1 mar. 2015] disponível em https://books.google.com.br/books?id=8yAeP2tY6wwC&pg=PA26&lpg=PA26&dq=james+black+fruitful+basis&source=bl&ots=0T4oggQuSy&sig=TIOuMkiTPOyhGARmyogyeHROMKE&hl=en&sa=X&ei=OY3yVOiAHY-TsQT-1YLYDA&ved=0CCoQ6AEwAg#v=onepage&q=james%20black%20fruitful%20basis&f=false

BMC. BioMed Central The Open Archive Publisher. 2014. [Internet] [acesso em 14 dez 2014]. Disponível em: http://www.biomedcentral.com/about

BRASIL. ANVISA. Agência Nacional de Vigilância Sanitária. O que você deve saber sobre MEDICAMENTOS. 2010. [Acesso em 1 de mar. 2015] Disponível em: http://portal.anvisa.gov.br/wps/wcm/connect/92aa8c00474586ea9089d43fbc4c6735/Cartilha%2BBAIXA%2Brevis%C3%A3o%2B24_08.pdf?MOD=AJPERES

BRASIL. Ministério do Planejamento, Orçamento e Gestão. 2010. Princípios e Diretrizes [portal na Internet]. 2010. [acesso em 6 out 2013]. Disponível em: https://www.governoeletronico.gov.br/o-gov.br/principios.

BRASIL. Ministério da Saúde. Departamento de Ciência e Tecnologia. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. Doenças negligenciadas: estratégias do Ministério da Saúde. Rev Saúde Pública. 2010;44(1):200-202.

BRASIL. Ministério da Saúde. Portaria nº 3.916/GM, de 30 de outubro de 1998. *Diário Oficial da União*, Brasília, nº 215-E, 10 nov. 1998a. Seção 1, p. 18-22. [acesso em 30 set 2014]. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/gm/1998/prt3916_30_10_1998.html

BRASIL. Presidência da República. Casa Civil. LEI Nº 5.991, DE 17 DE DEZEMBRO DE 1973. [acesso em 30 set 2014]. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/L5991.htm

Braganholo VP, Heuser CA. 2001. .XML Schema, RDF(S) e UML: uma Comparação. Universidade Federal do Rio Grande do Sul – UFRGS. Insituto de Informática. Porto Alegre – RS – Brasil. [acesso em 8 jan 2015] disponível em: http://www.inf.ufrgs.br/~heuser/papers/ideas2001.pdf

Breitman K. *Web* Semântica: a internet do futuro. Rio de Janeiro: LTC, 2005.

Brickley D, Guha RV. Resource Description Framework (RDF) Schema Specification 1.0, 2000. World Wide Web Consortium.

Brunton LL, Chabner BA, Knollmann BC, editors. Goodman and Gilman's The Pharmacological Basis of Therapeutics. 12th ed. New York, NY: McGraw-Hill Companies Inc;; 2011

CAMPOS, M. L. A. ; GOMES, H. E. Taxonomia e Classificação: princípios de categorização. Datagramazero (Rio de Janeiro), v. 9, p. 01, 2008. [acesso em 8 jan 2015] disponível em: http://www.dgz.org.br/ago08/Art_01.htm

Campos M, Campos M, D'Avila A, Gomes H, Campos L, Lira L. Aspectos metodológicos no reuso de ontologias: um estudo a partir das anotações genômicas no domínio dos tripanosomatídeos - DOI: 10.3395/reciis.v3i1.243pt. Revista Eletrônica de Comunicação, Informação & Inovação em Saúde, Brasil, 3, mar. 2009. [acesso em 27 set 2014]. Disponível em: http://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/243.

Carvalho HF, RECCO-PIMENTEL, SM. RECEPTORES [Internet], 2001. Departamento de Biologia Celular – UNICAMP. [acesso em 15 jan. 2015] disponível em http://www.nucleodeaprendizagem.com.br/receptores.pdf

CDC. Guidelines for Treatment of Malaria in the United States. 2013. CDC Center for Disease Control and Prevention [acesso em 23 dez 2014]. Disponível em: http://www.cdc.gov/malaria/resources/pdf/treatmenttable.pdf

CDC. Center for Disease Control and Prevention. 2014. [acesso em 22 dez 2014]. Disponível em: http://www.cdc.gov/malaria/diagnosis_treatment/treatment.html

Coimbra RS, Vanderwall DE, Oliveira GC. Disclosing ambiguous gene aliases by automatic literature profiling. BMC Genomics. 2010 Dec 22;11 Suppl 5:S3. doi: 10.1186/1471-2164-11-S5-S3. [acesso em 16 de fev. 2015] disponível em http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3045796/pdf/1471-2164-11-S5-S3.pdf

Cox MM, Doudna JÁ, O'Donnell M. (2012) Biologia Molecular – Princípios e Técnicas;Artmed. ISBN: 9788536327402

Creighton TE. Proteins, Structures and Molecular Principles, 2nd. Edition. WH Freeman and Co, New York NY, 1993

DE ROBERTIS EM, HIB JDR. bases da biologia celular e molecular. 4. ed. rev., e atual. Rio de Janeiro: Guanabara Koogan, 2006.

DECS. Descritores em ciências da saúde – Biblioteca Virtual de Saúde (BVS). 2014. [acesso em 14 abr 2014]. Disponível em: http://decs.bvsalud.org/cgi-

bin/wxis1660.exe/decsserver/?IsisScript=../cgi-bin/decsserver/decsserver.xis&search_language=p&interface_language=p&previous_page=homepage&task=exact_term&search_exp=Reposicionamento%20de%20Medicamentos

DNDi. Drugs for neglected diseases initiative [Internet]. 2010. [acesso em 12 nov 2013]. Disponível em: http://www.dndi.org.br/pt/doencas-negligenciadas.

DrugBanK. 2014. [acesso em 15 dez 2014]. Disponível em: http://www.drugbank.ca/.

Feinerer I. A text mining framework in R and its applications. [PhD thesis]. Vienna. Department of Statistics and Mathematics, Vienna University of Economics and Business Administration; October 2008. [acesso em 1 abr 2014]. Disponível em: http://epub.wu.ac.at/1923/1/document.pdf

Feinerer I, Hornik K. (2014). tm: Text Mining Package. R package version 0.5-10. [acesso em 6 out 2014] Disponível em : http://CRAN.R-project.org/package=tm

Feinerer I, Hornik K, Meyer D. (2008). Text Mining Infrastructure in R. Journal of Statistical Software 25(5): 1-54. [acesso em 6 out 2014] Disponível em: http://www.jstatsoft.org/v25/i05/.

Feinerer I. (2014) Introduction to the tm Package Text Mining in R. [Internet]. [acesso em 10 dez 2014]. Disponível em: http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Feldman R, Aumann Y, Zilberstein A, Ben-Yehuda Y. Mining biomedical literature using information extraction. 1998. [Internet]. [acesso em 21 nov 2013]. Disponível em: http://pluto.huji.ac.il/~rfeldman/papers/CDD.pdf

FERREIRA RS, GLAUCIUS O, ANDRICOPULO AD. Integração das técnicas de triagem virtual e triagem biológica automatizada em alta escala: oportunidades e desafios em P&D de fármacos. Quím. Nova, São Paulo , v. 34, n. 10, 2011 . [Accesso em 13 Jan. 2015]. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-40422011001000010&lng=en&nrm=iso>.

Filho SNV. Tratamento semântico de documentos em sistemas gerenciadores de banco de dados relacionais. 2010. Rio de Janeiro, Instituto Militar de Engenharia. [acesso em 6 out 2014]. Disponível em: http://www.comp.ime.eb.br/dissertacoes/2010-Sidney_Venturi.pdf

Fontes CA. 2011. Explorando Inferência em um Sistema de Anotação Semântica. Dissertação (Mestrado em Sistemas e Computação) - Instituto Militar de Engenharia. [Internet] 2015. [acesso em 18 fev. 2015]. Disponível em http://www.comp.ime.eb.br/dissertacoes/2011-Celso_Fontes.pdfb

Gadelha CAG, Quental C, Fialho BC. Saúde e inovação: uma abordagem sistêmica das indústrias da saúde. [Internet]. Cadernos de Saúde Pública. 2003 jan/fev. p 47 [acesso em 14 abr 2014]. Disponível em: http://www.scielo.br/scielo.php?pid=S0102-311X2003000100006&script=sci_arttext

Golan, DE et al. Princípios de farmacologia: interações fármaco-receptor. Rio de Janeiro: Guanabara Koogan, 2009. [Internet]. Cap. 1 p. 9. [acesso 15 abr 2014]. Disponível em: http://www.ufpi.br/subsiteFiles/lapnex/arquivos/files/Interacoes%20farmaco-receptor.PDF

Ginter F, Boberg J, J¨arvinen J, Salakoski T. 2004. New techniques for disambiguation in natural language and their application to biological text. JMLR, 5. [acesso em 5 out 2014]. Disponível em: http://www.jmlr.org/papers/volume5/ginter04a/ginter04a.pdf

Gruber T. What is an ontology? 1993. [acesso em 22 dez. 2014]. Disponível em: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.

Gruber T. Ontology. Entry in the Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2008. [acesso em 22 dez 2014]. Disponível em: http://tomgruber.org/writing/ontology-in-encyclopedia-of-dbs.pdf

Haupt VJ, Schroeder M. Old friends in new guise: repositioning of known drugs with structural bioinformatics Brief Bioinform first published online March 26, 2011 doi:10.1093/bib/bbr011

Hotho A, Nürnberger A, Paaß G. A brief survey of text mining. Journal for Computational Linguistics and Language Technology, v. 20, n.1, p. 19-62, 2005. [acesso em 22 dez 2014]. Disponível em: http://www.cin.ufpe.br/~rbcp/artigos/2005-Hotho-A%20Brief%20Survey%20of%20Text%20Mining.pdf

Hristovski D, Stare J, Peterlin B, Dzeroski S. 2001. Supporting discovery in medicine by association rule mining in Medline and UMLS. Proc. MedInfo Conf., London, England, Sep. 2-5, n.10, p.1344-1348. [acesso em 4 out 2014]. Diponível em: http://www.google.com.br/url?sa=t&rct=j&q=&esrc=s&source=web&cd=6&ved=0CDUQFjAF&url=http%3A%2F%2Fnl.ijs.si%2Fet%2Ftalks%2Ftsujiilab%2Fsaso%2Fhristovski.doc&ei=qHAwVKG7FoTCggSgkYGwDQ&usg=AFQjCNFX-b-JgQCgI4n5B_AlCuOHh7ayTQ&sig2=vfWonB4IcUUVDayUOnWpHw&bvm=bv.76802529,d.eXY&cad=rja

Hunter L., Hunter L. Molecular biology for computer scientists. Artificial Intelligence & Molecular Biology (1993) (MIT Press, Cambridge, MA) 1–46 [acesso em 14 jan 2015] disponível em https://www.cs.princeton.edu/~mona/IntroMaterials/hunter-bio-for-CS.pdf

ICSP. Institute of Chemistry of São Carlos (USP) – IQSC [internet]. 2015. CATÁLISE. [acesso em 15 jan. 2015] disponível em www.iqsc.usp.br/cursos/quimicageral/catalise-dwt2.htm

INCT. Instituto Nacional de Ciência e Tecnologia de Inovação em Doenças Negligenciadas (INCT). 2014. Doenças Negligenciadas [homepage na Internet]. [acesso em 24 nov de 2013]. Disponível em: http://www.cdts.fiocruz.br/inct-idn/index.php?option=com_k2&view=item&layout=item&id=112&Itemid=61

IVF. Instituto Virtual de Fármacos do Estado do Rio de Janeiro. 2006. IVFRJ *On Line* – 13ª Edição [INTERNER] acesso em 28 de fev. 2015] disponível em http://www.ivfrj.ccsdecania.ufrj.br/ivfonline/edicao_0013/terminologia.html

Jardim R. Estudo de reposicionamento de fármacos para doenças negligenciadas causadas por protozoários através da integração de bases de dados biológicas usando Web Semântica (2013). FIOCRUZ/IOC

Java. The Java EE 6 Tutorial. 2013 What Are RESTful Web Services? [acesso em 4 out 2014]. Disponível em : http://docs.oracle.com/javaee/6/tutorial/doc/gijqy.html

Jezuz MPG. Mineração de textos científicos visando à identificação de componentes bioativos com potencial terapêutico para o tratamento de dengue, malária e doença de chagas (2013). FIOCRUZ/IOC.

Jonquet C, Musen MA, Shah NH. A System for Ontology-Based Annotation of Biomedical. 2008. [acesso em 22 dez 2014]. Disponível em: http://www.lirmm.fr/~jonquet/publications/documents/Article-DILS08_Jonquet_Musen_Shah_published.pdf

Jonquet C, Musen MA, Shah NH. A System for ontology-based annotation of biomedical data. In International Workshop on Data Integration in The Life Sciences (DILS) Evry, France. 2008. THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY NCBO Annotator : Semantic Annotation of Biomedical Data. [acesso em 27 set 2014]. Disponível em: http://www.bioontology.org/sites/default/files/NCBO_Annotator_poster_0.pdf

Lancaster FW. Indexação e resumos: teoria e prática. Brasília: Briquet de Lemos/Livros, 1993.

LASSILA O, SWICK RR. Resource Description Framework (RDF) Model and Syntax Specification. World Wide Web Consortium, 1999

Leal A, Mateus B, Nunes D, Malta G. 2012. Síntese Proteica – 11º ano. [Internet] [acesso em 6 jan 2015] Disponível em: https://www.youtube.com/watch?v=IWkBlsEE1Cw

Lima G. Modelo hipertextual -MHTX: um modelo para organização hipertextual de documentos. DataGramaZero - Revista de ciência da informação Rio de Janeiro, v. 8, n. 4, ago. 2007. [acesso em 6 out 2014] . Disponível em: http://www.brapci.ufpr.br/download.php?dd0=7550.

Lima JC, Carvalho CL. 2005. Ontologias - OWL (*Web Ontology Language*). Instituto de Informática Universidade Federal de Goiás. [acesso em 2 jan 2015]. Disponível em: http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-05.pdf

Maia MAR. Gramática e Parser. In: II Congresso Internacional da Abralin, 2001, Fortaleza, CE. Anais do II Congresso Internacional da Abralin, Boletim 26. Fortaleza. Imprensa Universitária UFC, 2001. v. I. p. 188-192.

Markus LM. Toward a theory of knowledge reuse: Types of knowledge reuse

situations and factors in reuse success. Journal of management information systems, 2001;18(1):57-93.

Moura AMC. A *web* semântica: fundamentos e tecnologias. Rio de Janeiro:IME, [2002].

Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, Smith B; NCBO team. The National Center for Biomedical Ontology. J Am Med Inform Assoc. v.19, n.2, p.190-5, Mar-Apr 2012. Epub 2011 Nov 10. [acesso em 12 dez 2014]. Disponível em: http://www.bioontology.org/annotator-service

NCBO Annotator Web Service. 2009. [acesso em 26 set 2014]. Disponível em: .http://www.bioontology.org/wiki/index.php/Annotator_Web_service

NERC Environmental Bioinformatics Cente. Enabling environmental science in the molecular age. [Internet]. 2015. [acesso em 14 jan. 2015]. Disponível em: http://nebc.nerc.ac.uk/nebc_website_frozen/nebc.nerc.ac.uk//tools/terminizer/overview

NIH U.S. National Library of Medicine. 2014. [internet] [acesso em 7 jan 2015]. Disponivel em: http://nnlm.gov/training/resources/meshtri.pdf

Nzila A, Ma Z, Chibale K. Drug repositioning in the treatment of malaria and TB. Future Medicinal Chemotherapy, n.3, p. 1413-1426, 2011. [acesso em 15 dez 2014]. Disponível em: http://tballiance.org/newscenter/research_papers/2011_FMC%20Drug%20reposistioning%20in%20the%20treatment%20of%20malaria%20and%20TB.pdf

OBO. Open Biomedical Ontologies, 2005. [acesso em 27 set 2014]. Disponível em: http://obo.sourceforge.net.

OBO. The Open Biological and Biomedical Ontologies. 2014. OBO Foundry Principles. [Internet] [acesso em 7 jan 2015]. Disponível em: http://www.obofoundry.org/crit.shtml

OBO. The Open Biological and Biomedical Ontologies. 2006. Archive of original principles. [Internet] [acesso em 7 jan 2015]. Disponível em: http://www.obofoundry.org/crit_2006.shtml

O'Hara K, Hall W. Semantic Web. In: Bates MJ, Maack MN, Drake M. (Ed.). Encyclopedia of library and information science. 2009. 2. ed. Taylor & Francis. [acesso em 18 fev 2015] Disponível em: http://eprints.ecs.soton.ac.uk/17126/.

Oliva G. Universidades públicas e instituições de pesquisa: experiência com empresas. In: I Seminário Nacional sobre o Complexo Industria l da Saúde. 2003, Rio de Janeiro. Trabalho apresentado. Rio de Janeiro: BNDES, 5 a 7 de maio de 2003. [acesso em 3 fev 2004]. Disponível em: http://www.bndes.gov.br/SiteBNDES/export/sites/default/bndes_pt/Galerias/Arquivos/conhecimento/seminario/Saude07_4_1.pdf

Oliveira C, Garrão M, Amaral L. Recognizing complex prepositions Prep+N+Prep as negative patterns in automatic term extraction from texts. In: Proceedings of 1st

workshop em tecnologia da informação e da linguagem humana (TIL2003). São Carlos - SP. 2003. [acesso em 6 out 2014]. Disponível em: http://www.nilc.icmc.usp.br/til/til2003/oral/oliveira_garrao_amaral25.pdf.

ONU. Organização das Nações Unidas 2000. Millennium development goals and beyond 2015 [homepage na Internet]. [acesso em 5 out 2013]. Disponível em: http://www.un.org/millenniumgoals/.

OWL - Web Ontology Language, 2008. [acesso em 27 set 2014]. Disponível em: http://www.w3.org/TR/owl-ref/.

owltools. Wrapper fpr OWL API [Internet]. 2015. [acesso em 15 jan. 2015] disponível em: https://code.google.com/p/owltools/

Palakal M, Stephens M, Mukhopadhyay S, Raje R, Rhodes S. 2002. A multi-level text mining method to extract biological relationships. Proc. IEEE Computer Soc. Bioinformatics (CSB) Conf., p. 97-108. [acesso em 7 out 2014]. Disponível em: http://ai.arizona.edu/mis596A/book_chapters/medinfo/Chapter_16.pdf

Palit P, Mandal SC, Mandal NB. Reuse of old, existing, marketed non-antibiotic drugs as antimicrobial agents: a new emerging therapeutic approach. 2013. Microbial pathogens and strategies for combating them: science, technology and education, v.3 [acesso em 15 dez 2014]. Disponível em: http://www.formatex.info/microbiology4/vol3/1883-1892.pdf

Palmeira FPL, Pieroni JP, Antunes A, Bomtempo JV. O desafio do financiamento à inovação farmacêutica no Brasil: a experiência do BNDES. 2012. Profarma. Revista do BNDES. [acesso em 22 mar 2014]. Disponível em: http://www.bndes.gov.br/SiteBNDES/export/sites/default/bndes_pt/Galerias/Arquivos/conhecimento/revista/rev3703.pdf 2004.

Pisanelli DM, Gangemi A, Steve G. An ontological analysis of the UMLS metathesaurus. Journal of American medical informatics association, v. 5 (symposium supplement), 1998.

PostgreSQL. The PostgreSQL Global Development Group, 2015. [acesso em 18 fev 2015]. Disponível em: http://www.postgresql.org/.

PubMed Help [Internet]. National Center for Biotechnology Information. Disponível em: http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.

PubMed. National Center for Biotechnology Information. 2013. [Internet]. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed.

RDF, Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation, 2004. [acesso em 6 out 2014]. Disponível em: http://www.w3.org/RDF/.

R Foundation 2002. The R project for statistical computing. [Internet]. [acesso em 16 mar 2014]. Disponível em: http://www.r-project.org/

RSCB PDB PROTEIN DATA BANK An Information Portal to 106082 Biological Macromolecular Structures [Internet] 2015. [acesso em 22 jan. 2015]. Disponível em: http://www.rcsb.org/pdb/software/rest.do

Salton G, Wang A, Yang CS. A vector space model for information retrieval. Communications of the ACM. [Internet]. v.18, n.11, p.613–620, 1975. [acesso em 28 mar 2014]. Disponível em: http://www.cs.uiuc.edu/class/fa05/cs511/Spring05/other_papers/p613-salton.pdf

Sayers E. A General Introduction to the E-utilities. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. [acesso em 7 jan 2015] Dispnível em: http://www.ncbi.nlm.nih.gov/books/NBK25497/

Sayers E. E-utilities Quick Start. 2008 Dec 12 [Updated 2013 Aug 9]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. [acesso em 7 jan 2015] Disponível em: http://www.ncbi.nlm.nih.gov/books/NBK25500/

Sayers E. The E-utilities In-Depth: Parameters, Syntax and More. 2009 May 29 [Updated 2014 Oct 23]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. [acesso em 7 jan 2015] Disponível em: http://www.ncbi.nlm.nih.gov/books/NBK25499/

SBC. Stockholm Bioinformatic Centre. OrthoXML & SeqXML. 2013. [Internet]. [acesso em 10 out 2013]. Disponível em: http://seqxml.org./xml/Main.html.

SciTable by Nature Education. Protein Structure [Internet]. 2014. [acesso em 15 jan. 2015] disponível em: http://www.nature.com/scitable/topicpage/protein-structure-14122136

Schimitt T, Messina David N, Schreiber F, Sonnhammer Erik LL. Letter to the Editor: SeqXML and OrthoXML: standards for sequence and orthology information. 2011. [Internet]. [acesso em 10 nov 2013]. Disponível em: http://bib.oxfordjournals.org/content/12/5/485.long.

Smith D, Binet L, Bonnevie L, Hakokongas L, Meybaum J. Fatal imbalance: the crisis in research and development for drugs for neglected diseases. Geneva; Sept. 2001. [Internet] [acesso em 2 abr 2014]. Disponível em: http://www.msfaccess.org/sites/default/files/MSF_assets/NegDis/Docs/NEGDIS_report_FatalImbalance_CrisisInR&D_ENG_2001.pdf

Sloan FA, Hsieh C. “The Effects of Incentives on Pharmaceutical Innovation,” in Frank A. Sloan (ed.), Incentives and Choice in Health Care, MIT Press, 2008, p. 238. DOI:10.7551/mitpress/9780262195775.003.0009 [acesso em 22 may 2014] Disponível em http://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262195775.001.0001/upso-9780262195775-chapter-9

Srinivasan P, Libbus B. 2004. Mining MEDLINE for implicit links between dietary substances and diseases. ISMB 2004 and in Bioinformatics (Supplement). [acesso

em 6 out 2014]. Disponível em http://homepage.cs.uiowa.edu/~psriniva/Papers/TurmericBio.pdf

Souza W. Doenças negligenciadas. Ciência e tecnologia para o desenvolvimento nacional. Estudos Estratégicos [Internet]. 2010. [acesso em 11 nov 2013]. Disponível em: http://www.abc.org.br/IMG/pdf/doc-199.pdf.

Swanson DR. Fish-oil, Raynaud's syndrome and undiscovered public knowledge. Perspectives in biology and medicine, v.30, n.1, p.7-18, 1986.

Swanson DR. Two medical literatures that are logically but not bibliographically connected. Journal of the American society for information science, v.38, n.4, p.228-333, 1987.

Swanson DR, Smalheiser NR, Torvik VI. Ranking indirect connections in literature based discovery. The role of Medical Subject Headings. Journal of the American Society for Information Science and Technology, v.57, n.11, p.1427–39, 2006.

TechTarget. 2014. [Internet] [acesso em 14 dez 2014]. Disponível em http://searchsoa.techtarget.com/definition/

UFRJ. Universidade Federal do Rio de Janeiro. 2014. [Internet Portal dos Fármacos]. [acesso em 14 abr 2014]. Disponível em: http://www.portaldosfarmacos.ccs.ufrj.br/atualidades_profwermuth.html

UNICODE. 2014. About the Unicode Standard. [Internet] [acesso em 8 jan 2015] disponível em: http://www.unicode.org/standard/standard.html
65. Hearst M. (1999). Untangling Text Data Mining. [Internet] School of Information Management & Systems. University of California, Berkeley. [acesso em 10 dez 2014]. Disponível em: http://people.ischool.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html

UniProt. Universal Protein Resource (UniProt). 2014. [acesso em 15 dez 2014]. Disponível em: http://www.uniprot.org/.

Uramoto N, Matsuzawa H, Nagano T, Murami A, Takeuchi H. 2004. A text-mining system for knowledge discovery from biomedical documents. [acesso em 4 out 2014]. Disponível em: http://www.ccs.neu.edu/home/futrelle/bionlp/papers/Uramoto2004IBMSysJ.pdf

Uren V, Cimiano P, Iria J, Handschuh S, Vargas-Vera M, Motta E, Ciravegna F. 2005. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. [acesso em 26 dez 2014]. Disponível em: http://www.google.com.br/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CB0QFjAA&url=http%3A%2F%2Fwww.researchgate.net%2Fpublication%2F42790034_Semantic_annotation_for_knowledge_management_requirements_and_a_survey_of_the_state_of_the_art%2Flinks%2F09e415088547bad003000000&ei=oz6dVKOyMIPmggTikYRo&usg=AFQjCNFKHThY4DxUSA3T4ve43O5p74-nhg&sig2=c1WZ-DmYPCCo6OYtjL0OGw&bvm=bv.82001339,d.eXY

Venables WN, Smith DM. R Core Team. 2013. Notes on R: A Programming Environment for Data Analysis and Graphics
Version 3.0.1 (2013-05-16) [acesso em 7 jan 2015] Disponível em: http://cran.r-project.org/doc/manuals/R-intro.pdf

Wermuth, CG. The Practice of Medicinal Chemistry. London : Elsevier Academic Press, 2003.

Vieira VMM, Ohayon P. Inovação em fármacos e medicamentos: estado-da-arte no Brasil e políticas de P&D. Revista Economia & Gestão, v.6, n.13, p.60-82, 2006. [acesso em 4 abr 2014]. Disponível em: http://www.ufpi.br/subsiteFiles/ppgcf/arquivos/files/PI%20e%20TT%202.pdf

W3C. 1999. Namespaces in XML. World Wide Web Consortium 14-January-1999 [acesso em 5 jan 2015]. Disponível em: http://www.w3.org/TR/1999/REC-xml-names-19990114/

W3C. 1998. A. The Internal Structure of XML Namespaces [internet]. 1998. [acesso em 14 jan. 2015]. Disponível em http://www.w3.org/TR/1998/WD-xml-names-19980916#Philosophy

W3C 2014. Extensible Markup Language (XML) . W3C Ubiquitous Web domain. [acesso em 5 out 2014]. Disponível em: http://www.w3.org/XML/.

W3C. 2014. W3C Brasil *Web* Semântica. [acesso em 27 set 2014]. Disponível em: http:// http://www.w3c.br/Padroes/WebSemantica/.

W3C (2014). XML Schema. W3C Ubiquitous Web domain. [acesso em 5 jan 2015]. Disponível em: http://www.w3.org/XML/Schema

W3C Semantic Web. 2014. RDF. Resource Description Framework (RDF). [internet] [acesso em 8 jan 2015] disponível em http://www.w3.org/RDF/
33. Witten IH, Don KJ, Dewsnip M, Tablan V. Text mining in a digital library. Int J Digit Libr Journal. 2004;4:56-9 [Internet]. [acesso em 8 de dez. 2014] disponível em http://www.cs.waikato.ac.nz/~ihw/papers/03-IHW-KJD-MD-VT-Textminingina.pdf. http://dx.doi.org/ 10.1007/s00799-003-0066-4

W3C Semantic Web. 2014. RDFS. RDF Vocabulary Description Language 1.0: RDF Schema (RDFS). [internet] [acesso em 8 jan 2015] disponível em http://www.w3.org/2001/sw/wiki/RDFS

W3C (2004). OWL Web Ontology Language Overview. W3C Recommendation, World Wide Web Consortium. [acesso em 4 out 2014]. Disponível em: http://www.w3.org/TR/2004/REC-owl-features-20040210/.

W3C (2004). OWL Web Ontology Language Guide. W3C Recommendation, 10 February 2004. [acesso em 4 dez 2014]. Disponível em: http://www.w3.org/TR/owl-guide/.

W3C (2004). OWL Web Ontology Language Overview. W3C Recommendation, World Wide Web Consortium. [acesso em 24 jul 2010]. Disponível em: http://www.w3.org/TR/owl-features/.

W3C Semantic Web. (2012). OWL Web Ontology Language (OWL). [acesso em 4 jan 2015]. Disponível em: http://www.w3.org/2001/sw/wiki/OWL

W3C Semantic Web. 2014. URI. Uniform Resource Identifier. [internet] [acesso em 8 jan 2015] disponível em http://semanticweb.org/wiki/Uniform_Resource_Identifier

Weeber M et al. 2003. Generating hypothesis by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses of thalidomide. J. Am. Med. Inform. Assoc., n.10, p.252-259. [acesso em 5 out 2014]. Disponível em: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC342048/

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration *Nucl. Acids Res. (2006) 34 (suppl 1): D668-D672 doi:10.1093/nar/gkj067*

Wren J, Bekeredjian R, Stewart JA, Shohet RV, Garner HR. 2004. Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics, n.20, p.3. [acesso em 5 out 2014]. Disponível em: http://bioinformatics.oxfordjournals.org/content/20/3/389.full.pdf

WHO. World Health Organization. Commission on macroeconomics and health. Macroeconomics and health: investing in health for economic development. Geneva: 2001. p. 78.

WHO. World Health Organization. First WHO report on neglected tropical diseases: working to overcome the global impact of neglected tropical diseases. Geneva; 2010. [acesso em 13 abr 2014]. Disponível em: http://whqlibdoc.who.int/publications/2010/9789241564090_eng.pdf?ua=1

WHO. World Health Organization. Malaria 2013. 14/08/2013. [homepage na Internet]. [acesso em 21 abr 2014]. Disponível em: http://www.who.int/mediacentre/factsheets/fs094/en/.

Zambenedetti C. Extração de Informação sobre Bases de Dados Textuais. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Rio Grande do Sul / Programa de Pós-Graduação em Computação, 2002.
[acesso em 10 out 2014]. Disponível em
http://www.lume.ufrgs.br/bitstream/handle/10183/1628/000353940.pdf?sequence=1

Zweigenbaum P, Demner-Fushman D, Hong Yu, Cohen KB. Frontiers of biomedical text mining: current progress. Briefings in bioinformatics**,** v. 8, n. 5, p. 358-375, 2007. [acesso em 22 dez 2014] Disponível em: <http://bib.oxfordjournals.org/content/8/5/358.full.pdf+html>.

RELAÇÃO BIBLIOGRAFICA ANALISADA

Barbosa-Silva A, Fontaine J, Donnard ER, Stussi F, Ortega JM, Andrade-Navarro, MA. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. BMC Bioinformatics , v. 12, p. 435, 2011.

Barbosa-Silva A, Soldatos T, Magalhães I, Pavlopoulos G, Fontaine J-F, Andrade-Navarro M, et al. LAITOR - Literature Assistant for Identification of Terms co-

Occurrences and Relationships. BMC bioinformatics. 2010;11(1):70. acesso em 22 fev. 2015. Disponível em http://www.biomedcentral.com/1471-2105/11/70

Campos MLA. INTEGRAÇÃO DE ONTOLOGIAS: O DOMÍNIO DA BIOINFORMÁTICA E A PROBLEMÁTICA DA COMPATIBILIZAÇÃO TERMINOLÓGICA. 2009. [Internet] [Acesso em 27 fev. 2015] disponível em http://www.uff.br/ppgci/editais/mlconto.pdf

Chun H. et al. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In: PACIFIC SYMPOS IUM ON BIOCOMPUTING, 13., October 2005. p. 4-15. [acesso em 26. fev. 2015] Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17094223>

Couto F, Silva MJ, Lee V, Dimmer E, Camon E, Apweiler R, Kirsch H, Rebholz-Schuhmann D. GOAnnotator: linking electronic protein GO annotation to evidence text. Journal of Biomedical Discovery and Collaboration 1(19), 2006.

Ferreira JD, Pesquita C, Couto FM, Silva MJ. "Bringing epidemiology into the Semantic Web." In ICBO. 2012.

Ferreira JD. EPIWORK (Doenças x Fármacos). [mensagem pessoal]. eduardo.barcante@gmail.com. 14 dez. 2014

Fontes CA. 2011. Explorando Inferência em um Sistema de Anotação Semântica. Dissertação (Mestrado em Sistemas e Computação) - Instituto Militar de Engenharia. [Internet] 2015. [acesso em 18 fev. 2015]. Disponível em http://www.comp.ime.eb.br/dissertacoes/2011-Celso_Fontes.pdfb

HEARST MA. Untangling text data mining. In: Annual meeting of the association for computational linguistics, University of Maryland, 1999.

Jezuz MPG. Mineração de textos científicos visando à identificação de componentes bioativos com potencial terapêutico para o tratamento de dengue, malária e doença de chagas (2013). FIOCRUZ/IOC.

Jonquet C, Musen MA, Shah NH. A System for Ontology-Based Annotation of Biomedical. 2008. [acesso em 22 dez 2014]. Disponível em: http://www.lirmm.fr/~jonquet/publications/documents/Article-DILS08_Jonquet_Musen_Shah_published.pdf

KRALLINGER M, VALENCIA,A.; Text-mining and information-retrieval services for molecular biology. Genome biology, London, v. 6, p. 224, 2005.

Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, Smith B; NCBO team. The National Center for Biomedical Ontology. J Am Med Inform Assoc. v.19, n.2, p.190-5, Mar-Apr 2012. Epub 2011 Nov 10. [acesso em 12 dez 2014]. Disponível em: http://www.bioontology.org/annotator-service

NERC Environmental Bioinformatics Cente. Enabling environmental science in the molecular age. [Internet]. 2015. [acesso em 14 jan. 2015]. Disponível em: http://nebc.nerc.ac.uk/nebc_website_frozen/nebc.nerc.ac.uk//tools/terminizer/overview

Oren E, Moller KH, Scerri S, Handschuh S, Sintek M. What are Semantic Annotations? 2006. [internet] [acesso em 26 fev 2015] disponível em http://www.siegfried-handschuh.net/pub/2006/whatissemannot2006.pdf

PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. BMC Bioinformatics. 2011 Nov 9;12(1):435

SPASIC, I; ANANIADOU, S; McNAUGHT, J.; KUMAR, A. Text mining and ontologies in biomedicine: making sense of raw text. **Briefings in bioinformatics**, v. 6, n. 3, p. 239-251, 2005. Disponível em: <http://personalpages.manchester.ac.uk/staff/sophia.ananiadou/BIB.pdf>. Acesso em: 2 jan. 2011.

SWANSON, DR; SMALLHEISER, NR; TORVIK, VI. Ranking indirect connections in literature based discovery. The role of Medical Subject Headings. Journal of the American society for information science and technology, v. 57, n. 11, p. 1427-1439, 2006

Wei CH et. al., PubTator: a Web-based text mining tool for assisting Biocuration, Nucleic acids research, 2013, 41 (W1): W518-W522. doi: 10.1093/nar/gkt44

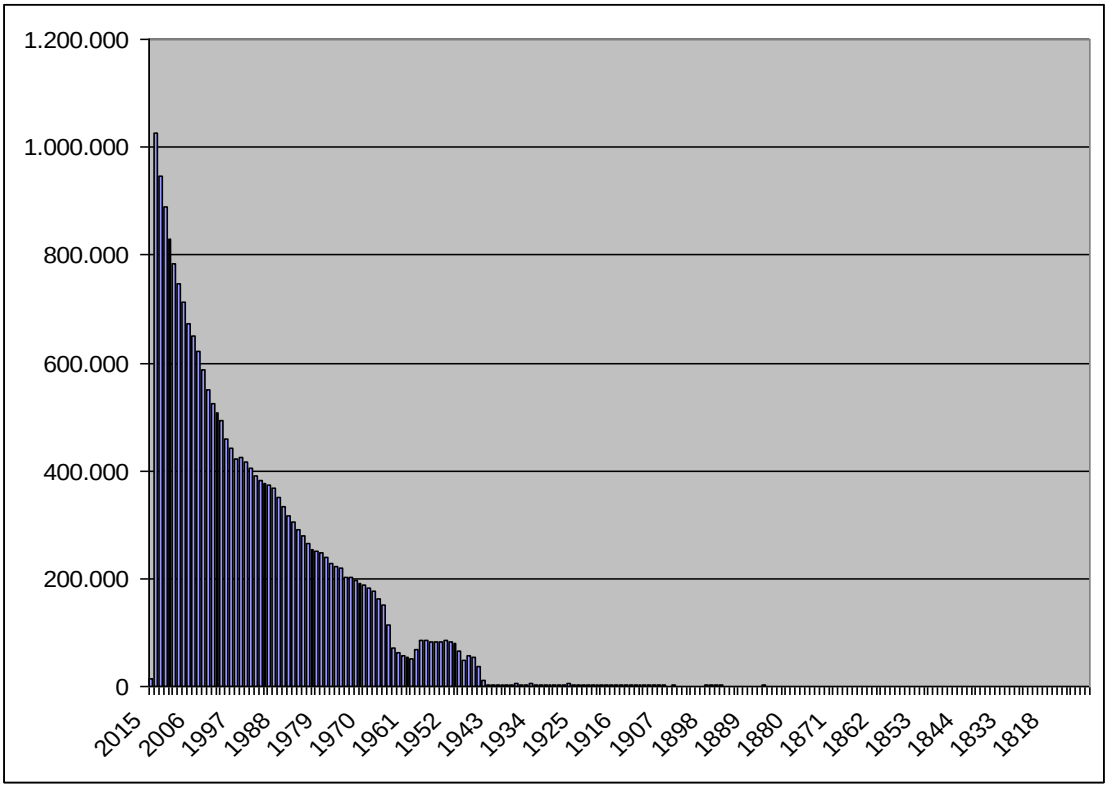
WIVES, L. K. Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos. 2004. 136 f. 88 Tese (Doutorado) - Universidade Federal do Rio Grande do Sul, Porto Alegre: Programa de Pós-Graduação em Computação. [Acesso em: 26 jan. 2015.] Disponível em: <http://www.lume.ufrgs.br/handle/10183/4576 >

Zhou, G.; Zhang, J.; Su, J.; Shen, D.; Tan, C. et al. Recognizing names in biomedical texts: a machine learning approach. Bioinformatics, v. 20, n. 7, p. 1178-1190, 2004.

ZWEIGENBAUM P, DEMNER-FUSHMAN D, HONG YU; COHEN, KB. Frontiers of biomedical text mining: current progress. Briefings in bioinformatics, v. 8, n. 5, p. 358-375, 2007. [acesso em 26 fev. 2015] Disponível em: <http://bib.oxfordjournals.org/content/8/5/358.full.pdf+html>.

7 ANEXOS

ANEXO I – Frequência de publicações por ano no PubMed



Fonte: PubMed. Consulta em dezembro 2014.

ano	publicações		ano	publicações		ano	Publicações
2015	13.742		1997	422.591		1979	248.517
2014	1.025.740		1996	425.391		1978	238.887
2013	946.507		1995	414.898		1977	229.036
2012	888.560		1994	404.277		1976	223.348
2011	830.675		1993	391.914		1975	218.146
2010	783.375		1992	383.304		1974	202.591
2009	746.586		1991	377.164		1973	201.623
2008	713.644		1990	373.788		1972	197.829
2007	674.070		1989	368.129		1971	191.625
2006	648.721		1988	349.189		1970	187.166
2005	621.168		1987	333.554		1969	183.377
2004	585.998		1986	316.075		1968	176.308
2003	549.822		1985	304.502		1967	162.017
2002	524.882		1984	289.692		1966	151.011
2001	508.324		1983	278.921		1965	114.789
2000	494.124		1982	264.399		1964	72.476
1999	458.196		1981	253.798		1963	62.340
1998	440.879		1980	251.481		1962	58.280

Fonte: PubMed. Consulta em dezembro 2014.

ANEXO II – Nomes das doenças no MeSH

MeSH Heading	Chagas Disease
Tree Number	C03.752.300.900.200
Annotation	coordinate with NEGLECTED DISEASES if pertinent; CHAGAS CARDIOMYOPATHY is also available.
Scope Note	Infection with the protozoan parasite TRYPANOSOMA CRUZI, a form of TRYPANOSOMIASIS endemic in Central and South America. It is named after the Brazilian physician Carlos Chagas, who discovered the parasite. Infection by the parasite (positive serologic result only) is distinguished from the clinical manifestations that develop years later, such as destruction of PARASYMPATHETIC GANGLIA; CHAGAS CARDIOMYOPATHY; and dysfunction of the ESOPHAGUS or COLON.
Entry Term	American Trypanosomiasis
Entry Term	Chagas' Disease
Entry Term	Trypanosomiasis, South American
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH TM UR US VE VI
Entry Version	CHAGAS DIS
Online Note	use CHAGAS DISEASE to search TRYPANOSOMIASIS, SOUTH AMERICAN 1966-91
History Note	92; was TRYPANOSOMIASIS, SOUTH AMERICAN 1966-91; CHAGAS DISEASE was see TRYPANOSOMIASIS, SOUTH AMERICAN 1981-91
Date of Entry	19990101
Unique ID	D014355

MeSH Heading	Dengue
Tree Number	C02.081.270
Tree Number	C02.782.350.250.214
Tree Number	C02.782.417.214
Annotation	coordinate with NEGLECTED DISEASES if pertinent; SEVERE DENGUE is also available
Scope Note	An acute febrile disease transmitted by the bite of AEDES mosquitoes infected with DENGUE VIRUS. It is self-limiting and characterized by fever, myalgia, headache, and rash. SEVERE DENGUE is a more virulent form of dengue.
Entry Term	Classical Dengue
Entry Term	Classical Dengue Fever
Entry Term	Dengue Fever
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH TM UR US VE VI
Date of Entry	20140626
Unique ID	D003715

MeSH Heading	Leishmaniasis
Tree Number	C03.752.300.500
Tree Number	C03.858.560
Tree Number	C17.800.838.775.560
Annotation	general or unspecified; prefer specifics; coordinate with NEGLECTED DISEASES if pertinent; LEISHMANIASIS, AMERICAN see LEISHMANIASIS, CUTANEOUS is also available; tegumentary leishmaniasis = LEISHMANIASIS, CUTANEOUS
Scope Note	A disease caused by any of a number of species of protozoa in the genus LEISHMANIA. There are four major clinical types of this infection: cutaneous (Old and New World) (LEISHMANIASIS, CUTANEOUS), diffuse cutaneous (LEISHMANIASIS, DIFFUSE CUTANEOUS), mucocutaneous (LEISHMANIASIS, MUCOCUTANEOUS), and visceral (LEISHMANIASIS, VISCERAL).
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH TM UR US VE VI
Date of Entry	19990101
Unique ID	D007896

Fonte: MeSH. Dezembro 2014.

ANEXO III – *Web Annotator online* (hit>0)


```
- <class>
    <page>1</page>
    <pageCount>1</pageCount>
    <prevPage/>
    <nextPage/>
  - <links>
      <nextPage/>
      <prevPage/>
    </links>
  - <collection>
    - <class>
        <prefLabel>molybdenum transport system permease protein ModB</prefLabel>
      - <synonymCollection>
          <synonym>modB</synonym>
        </synonymCollection>
      - <definitionCollection>
          <definition>Category=gene.</definition>
        </definitionCollection>
        <obsolete>false</obsolete>
        <matchType>prefLabel</matchType>
        <id>http://purl.obolibrary.org/obo/PR_000023268</id>
        <type>http://www.w3.org/2002/07/owl#Class</type>
      - <linksCollection>
        - <links>
            <self href="http://data.bioontology.org/ontologies/PR/classes/http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FP
          </links>
        - <links>
            <ontology href="http://data.bioontology.org/ontologies/PR" rel="http://data.bioontology.org/metadata/Ontology"/>
          </links>
        - <links>
            <children href="http://data.bioontology.org/ontologies/PR/classes/http%3A%2F%2Fpurl.obolibrary.org%2Fobo%
          </links>
        - <links>
            <parents href="http://data.bioontology.org/ontologies/PR/classes/http%3A%2F%2Fpurl.obolibrary.org%2Fobo%
          </links>
        - <links>
            <descendants href="http://data.bioontology.org/ontologies/PR/classes/http%3A%2F%2Fpurl.obolibrary.org%2Fc
            /owl#Class"/>
          </links>
        - <links>
```


Fonte: http://data.bioontology.org. Dezembo 2014.

http://data.bioontology.org/search?q=*termo*&ontologies=PR&format=xml&apikey=*apikey*

ANEXO IV – Stopwords registradas na API fornecida pelo BioPortal.

I, a, above, after, against, all, alone, always, am, amount, an, and,
any, are, around, as, at, back, be, before, behind, below, between,
bill, both, bottom, by, call, can, co, con, de, detail, do, done,
down, due, during, each, eg, eight, eleven, empty, ever, every, few,
fill, find, fire, first, five, for, former, four, from, front, full,
further, get, give, go, had, has, hasnt, he, her, hers, him, his, i,
ie, if, in, into, is, it, last, less, ltd, many, may, me, mill, mine,
more, most, mostly, must, my, name, next, nine, no, none, nor, not,
nothing, now, of, off, often, on, once, one, only, or, other, others,
out, over, part, per, put, re, same, see, serious, several, she, show,
side, since, six, so, some, sometimes, still, take, ten, the, then, third,
this, thick, thin, three, through, to, together, top, toward, towards,
twelve, two, un, under, until, up, upon, us, very, via, was, we, well,
when, while, who, whole, will, with, within, without, you, yourself,
yourselves

Fonte: http://data.bioontology.org/documentation

ANEXO V – Lista parcial de *stopwords* empregadas no *pipeline.*

Países Meses (fragmento)	Algeria Angola Benin Botswana Burkina Faso Burundi Cameroon Canary Islands Cape Verde Islands Central African Republic
	Comoros Islands Côte d'Ivoire Democratic R. of Congo Djibouti Egypt Equatorial Guinea Eritrea Ethiopia Gabon Gambia Ghana
	Guinea Guinea-Bissau Kenya Lesotho Liberia Libya Madagascar Malawi Mali Mauritania Mauritius Mayotte Morocco Mozambique
	American Samoa Australia Cocos Islands Cook Islands Christmas Island Easter Island Fiji French Polynesia Guam Hawaiian
	Islands Kiribati Marshall Islands Micronesia Midway Islands Nauru New Caledonia New Zealand Niue Norfolk Island Northern
	Vietnam January February March April May June July August September October November December
Outras Palavras (fragmento)	A AA AAH AAHED AAHING AAHS AAL AALII AALIIS AALS
	AARDVARK AARDVARKS AARDWOLF AARDWOLVES AARGH AARRGH AARRGHH AAS
	AASVOGEL AASVOGELS AB ABA ABACA ABACAS ABACI ABACK ABACTERIAL ABACUS
	ABACUSES ABAFT ABAKA ABAKAS ABALONE ABALONES ABAMP ABAMPERE ABAMPERES
	ABAMPS ABANDON ABANDONED ABANDONER ABANDONERS ABANDONING ABANDONMENT
	ABANDONMENTS ABANDONS ABAPICAL ABAS ABASE ABASED ABASEDLY ABASEMENT
	ABASEMENTS ABASER ABASERS ABASES ABASH ABASHED ABASHES ABASHING
	ABASHMENT ABASHMENTS ABASIA ABASIAS ABASING ABATABLE ABATE ABATED
	ABATEMENT ABATEMENTS ABATER ABATERS ABATES ABATING ABATIS ABATISES
	ABATOR ABATORS ABATTIS ABATTISES ABATTOIR ABATTOIRS ABAXIAL ABAXILE
	ABBA ABBACIES ABBACY ABBAS ABBATIAL ABBE ABBES ABBESS ABBESSES ABBEY
	ABBEYS ABBOT ABBOTCIES ABBOTCY ABBOTS ABBREVIATE ABBREVIATED
	ABBREVIATES ABBREVIATING ABBREVIATION ABBREVIATIONS ABBREVIATOR
	ABBREVIATORS ABDICABLE ABDICATE ABDICATED ABDICATES ABDICATING
	ABDICATION ABDICATIONS ABDICATOR ABDICATORS ABDOMEN ABDOMENS ABDOMINA
	ABDOMINAL ABDOMINALLY ABDUCE ABDUCED ABDUCENS ABDUCENT ABDUCENTES
	ABDUCES ABDUCING ABDUCT ABDUCTED ABDUCTING ABDUCTION ABDUCTIONS
	ABDUCTOR ABDUCTORES ABDUCTORS ABDUCTS ABEAM ABECEDARIAN ABECEDARIANS
	ABED ABELE ABELES ABELIA ABELIAN ABELIAS ABELMOSK ABELMOSKS ABERRANCE
	ABERRANCES ABERRANCIES ABERRANCY ABERRANT ABERRANTLY ABERRANTS
	ABERRATED ABERRATION ABERRATIONAL ABERRATIONS ABET ABETMENT ABETMENTS
	ABETS ABETTAL ABETTALS ABETTED ABETTER ABETTERS ABETTING ABETTOR
	ABETTORS ABEYANCE ABEYANCES ABEYANCIES ABEYANCY ABEYANT ABFARAD
	ABFARADS ABHENRIES ABHENRY ABHENRYS ABHOR ABHORRED ABHORRENCE
	ABHORRENCES ABHORRENT ABHORRENTLY ABHORRER ABHORRERS ABHORRING ABHORS

Fonte: Laboratório PROCC/FIOCRUZ

ANEXO VI – Lista de fármacos candidatos ao reposicionamento para malária.

FÁRMACO	Atividades Farmacológicas	Categorias DrugBank (Adaptado do DeCS)
DB00563	Quimioterapia de doenças neoplásicas	Abortivo
DB02764	Outros/Experimental	Adesivos
DB00368	Drogas afetando função renal/cardiovascular	Agonistas alfa-Adrenérgicos
DB04582	Outros/Experimental	Antagonistas Adrenérgicos alfa
DB00668	Drogas afetando função renal/cardiovascular	Agonistas Adrenérgicos beta
DB01064	Drogas afetando função renal/cardiovascular	Agonistas Adrenérgicos beta
DB00373	Drogas afetando função renal/cardiovascular	Antagonistas Adrenérgicos beta
DB01421	Quimioterapia das infecções parasitárias	Amebicidas
DB00919	Quimioterapia de doenças microbianas	Aminoglicosídeos
DB01082	Quimioterapia de doenças microbianas	Aminoglicosídeos
DB00684	Quimioterapia de doenças microbianas	Aminoglicosídeos
DB00712	Quimioterapia de doenças microbianas	Analgésicos
DB01050	Terapia Famacologica da Inflamação	Analgésicos não Entorpecentes
DB00624	Hormonio/Antagonista Hormonal	Androgênios
DB00753	Drogas atuantes no SNC	Anestésicos
DB01159	Drogas atuantes no SNC	Anestésicos
DB00818	Drogas atuantes no SNC	Anestésicos Intravenosos
DB00907	Drogas atuantes no SNC	Anestésicos Locais
DB00730	Quimioterapia das infecções parasitárias	Anti-Helmínticos
DB00115	Drogas atuantes no sangue e em órgãos formadores de sangue	AntiAnemia
DB00381	Drogas afetando função renal/cardiovascular	Antianginosos
DB00471	Drogas afetando função renal/cardiovascular	Antiarrítmicos
DB01426	Drogas afetando função renal/cardiovascular	Antiarrítmicos
DB00828	Quimioterapia de doenças microbianas	Antibacterianos
DB01764	Quimioterapia de doenças microbianas	Antibacterianos
DB03166	Quimioterapia de doenças microbianas	Antibacterianos
DB00601	Quimioterapia de doenças microbianas	Antibacterianos
DB00760	Quimioterapia de doenças microbianas	Antibacterianos
DB01669	Outros/Experimental	Antibacterianos
DB01670	Outros/Experimental	Antibacterianos
DB01942	Outros/Experimental	Antibacterianos
DB03966	Outros/Experimental	Antibacterianos
DB03967	Outros/Experimental	Antibacterianos
DB04124	Outros/Experimental	Antibacterianos
DB04626	Outros/Experimental	Antibacterianos
DB04741	Outros/Experimental	Antibacterianos
DB04785	Outros/Experimental	Antibacterianos
DB04797	Outros/Experimental	Antibacterianos
DB00994	Quimioterapia de doenças microbianas	Antibióticos
DB01045	Quimioterapia de doenças microbianas	Antibióticos
DB01190	Quimioterapia de doenças microbianas	Antibióticos
DB00997	Quimioterapia de doenças microbianas	Antibióticos

FÁRMACO	Atividades Farmacológicas	Categorias DrugBank (Adaptado do DeCS)
DB01157	Quimioterapia de doenças microbianas	Antibióticos
DB00290	Quimioterapia de doenças microbianas	Antibióticos Antineoplásicos
DB00615	Quimioterapia de doenças microbianas	Antibióticos Antituberculose
DB01103	Quimioterapia de doenças parasitárias	Anticestoides
DB00715	Drogas atuantes no SNC	Antidepressivos
DB04599	Outros/Experimental	Antidepressivos
DB02959	Drogas atuantes no SNC	Antidepressivos de Segunda Geração
DB02960	Outros/Experimental	Antidepressivos de Segunda Geração
DB00321	Drogas atuantes no SNC	Antidepressivos Tricíclicos
DB01235	Drogas atuantes no SNC	Antidiscinético
DB01234	Terapia Famacologica da Inflamação	Antieméticos
DB04794	Quimioterapia de doenças microbianas	Antifúngicos
DB01910	Outros/Experimental	Antifúngicos
DB03600	Outros/Experimental	Antifúngicos
DB03601	Outros/Experimental	Antifúngicos
DB03758	Outros/Experimental	Antifúngicos
DB04297	Outros/Experimental	Antifúngicos
DB00905	Oftalmologia	Antiglaucoma
DB00481	Hormonio/Antagonista Hormonal	Antihipocalcimiante
DB00336	Quimioterapia de doenças microbianas	Anti-Infecciosos
DB00440	Quimioterapia de doenças microbianas	Anti-Infecciosos
DB00741	Terapia Famacologica da Inflamação	Anti-Inflamatórios
DB00812	Terapia Famacologica da Inflamação	Anti-Inflamatórios não Esteroides
DB00861	Terapia Famacologica da Inflamação	Anti-Inflamatórios não Esteroides
DB08797	Terapia Famacologica da Inflamação	Anti-Inflamatórios não Esteroides
DB00580	Terapia Famacologica da Inflamação	Anti-Inflamatórios não Esteroides
DB04743	Terapia Famacologica da Inflamação	Anti-Inflamatórios não Esteroides
DB00783	Hormonio/Antagonista Hormonal	Antimenopausa
DB00194	Quimioterapia de doenças neoplásicas	Antimetabólitos
DB00811	Quimioterapia de doenças neoplásicas	Antimetabólitos
DB01011	Quimioterapia de doenças neoplásicas	Antimetabólitos
DB00987	Quimioterapia de doenças neoplásicas	Antimetabólitos Antineoplásicos
DB01667	Outros/Experimental	Antimetabólitos Antineoplásicos
DB03172	Outros/Experimental	Antimetabólitos Antineoplásicos
DB00320	Drogas atuantes no SNC	Tratamendo de enxaqueca
DB02709	Outros/Experimental	Antimutagênico
DB01201	Quimioterapia de doenças microbianas	Antimicrobiantes
DB00592	Quimioterapia de doenças parasitárias	Antinematódeos
DB00398	Quimioterapia de doenças neoplásicas	Antinematódeos
DB04868	Quimioterapia de doenças neoplásicas	Antinematódeos
DB02877	Outros/Experimental	Antinematódeos
DB00291	Quimioterapia de doenças neoplásicas	Antineoplásicos Alquilantes
DB04216	Outros/Experimental	Antioxidantes
DB02300	Hormonio/Antagonista Hormonal	Terapia da psoríase
DB03255	Outros/Experimental	Terapia da psoríase

FÁRMACO	Atividades Farmacológicas	Categorias DrugBank (Adaptado do DeCS)
DB02731	Outros/Experimental	Antiséptico
DB00763	Hormonio/Antagonista Hormonal	Antitireóideos
DB00233	Quimioterapia de doenças microbianas	Antituberculosos
DB01004	Quimioterapia de doenças microbianas	Antivirais
DB00184	Drogas atuantes ao nível neuroefetor e sinaptico	Fármacos do Sistema Nervoso Autônomo
DB00399	Outros/Experimental	Disfosfanatos
DB01244	Drogas afetando função renal/cardiovascular	Bloqueadores dos Canais de Cálcio
DB00819	Drogas afetando função renal/cardiovascular	Inibidores da Anidrase Carbônica
DB00311	Drogas afetando função renal/cardiovascular	Inibidores da Anidrase Carbônica
DB06728	Outros/Experimental	Carcinógenos
DB00988	Drogas afetando função renal/cardiovascular	Cardiotônicos
DB00640	Drogas afetando função renal/cardiovascular	Fármacos Cardiovasculares
DB00380	Drogas afetando função renal/cardiovascular	Fármacos Cardiovasculares
DB02348	Outros/Experimental	Cariostáticos
DB02349	Outros/Experimental	Cariostáticos
DB06777	Outros/Experimental	Catárticos
DB03329	Outros/Experimental	Catepsina B
DB04436	Outros/Experimental	Estimulantes do Sistema Nervoso Central
DB00456	Quimioterapia de doenças microbianas	Cefalosporinas
DB04795	Outros/Experimental	Quelantes
DB05088	Outros/Experimental	Quelantes
DB04272	Outros/Experimental	Quelantes
DB01586	Outros/Experimental	Colagogos e Coleréticos
DB03128	Outros/Experimental	Colinérgicos
DB04214	Outros/Experimental	Compostos Cromogênicos
DB00396	Hormonio/Antagonista Hormonal	Anticoncepcionais
DB00717	Hormonio/Antagonista Hormonal	Anticoncepcionais Orais Sintéticos
DB01093	Terapia Famacologica da Inflamação	Crioprotetores
DB00461	Terapia Famacologica da Inflamação	Inibidores de Ciclo-Oxigenase 2
DB00586	Terapia Famacologica da Inflamação	Inibidores de Ciclo-Oxigenase
DB01023	Drogas afetando função renal/cardiovascular	Di-Hidropiridinas
DB01261	Hormonio/Antagonista Hormonal	Inibidores da Dipeptidil Peptidase IV
DB00606	Drogas afetando função renal/cardiovascular	Diuréticos
DB00903	Drogas afetando função renal/cardiovascular	Diuréticos
DB00742	Drogas afetando função renal/cardiovascular	Diuréticos Osmóticos
DB00551	Drogas afetando função renal/cardiovascular	Inibidores Enzimáticos
DB00558	Quimioterapia de doenças microbianas	Inibidores Enzimáticos
DB00786	Quimioterapia de doenças microbianas	Inibidores Enzimáticos
DB01686	Outros/Experimental	Inibidores Enzimáticos
DB03206	Outros/Experimental	Inibidores Enzimáticos
DB03796	Outros/Experimental	Inibidores Enzimáticos
DB03797	Outros/Experimental	Inibidores Enzimáticos
DB04223	Outros/Experimental	Inibidores Enzimáticos
DB00255	Hormonio/Antagonista Hormonal	Estrogênios não Esteroides
DB02999	Outros/Experimental	Agonistas de Aminoácidos Excitatórios

FÁRMACO	Atividades Farmacológicas	Categorias DrugBank (Adaptado do DeCS)
DB00996	Drogas atuantes no SNC	Antagonistas de Aminoácidos Excitatórios
DB06151	Outros/Experimental	Expectorantes
DB00951	Quimioterapia de doenças microbianas	Inibidores da Síntese de Ácidos Graxos
DB00693	Oftalmologia	Corantes Fluorescentes
DB00205	Quimioterapia das infecções parasitárias	Antagonistas do Ácido Fólico
DB08878	Quimioterapia de doenças neoplásicas	Antagonistas do Ácido Fólico
DB00974	Drogas atuantes no sangue e em órgãos formadores de sangue	Aditivos Alimentares
DB03793	Outros/Experimental	Conservantes de Alimentos
DB00437	Outros/Experimental	Depuradores de Radicais Livres
DB03644	Outros/Experimental	Depuradores de Radicais Livres
DB03645	Outros/Experimental	Depuradores de Radicais Livres
DB03646	Outros/Experimental	Depuradores de Radicais Livres
DB04657	Outros/Experimental	Fungicidas Industriais
DB03497	Outros/Experimental	Inibidores do Crescimento
DB03496	Outros/Experimental	Inibidores do Crescimento
DB04539	Outros/Experimental	Herbicidas
DB02546	Quimioterapia de doenças neoplásicas	Inibidores de Histona Desacetilases
DB00259	Quimioterapia de doenças microbianas	Agentes Homeopáticos
DB01169	Quimioterapia de doenças neoplásicas	Agentes Homeopáticos
DB00279	Hormonio/Antagonista Hormonal	Reposição Hormonal
DB00227	Outros/Experimental	Inibidores de Hidroximetilglutaril-CoA Redutases
DB01393	Drogas afetando função renal/cardiovascular	Hipolipemiantes
DB00627	Drogas afetando função renal/cardiovascular	Hipolipemiantes
DB03424	Outros/Experimental	Hipolipemiantes
DB04258	Outros/Experimental	Imunossupressores
DB04259	Outros/Experimental	Imunossupressores
DB00936	Outros/Experimental	Ceratolíticos
DB03085	Outros/Experimental	Ceratolíticos
DB00976	Quimioterapia de doenças microbianas	Cetolídeos
DB02386	Outros/Experimental	Leucil Aminopeptidase
DB00199	Quimioterapia de doenças microbianas	Macrolídeos
DB00207	Quimioterapia de doenças microbianas	Macrolídeos
DB01211	Quimioterapia de doenças microbianas	Macrolídeos
DB01361	Quimioterapia de doenças microbianas	Macrolídeos
DB00877	Quimioterapia de doenças microbianas	Macrolídeos
DB00778	Quimioterapia de doenças microbianas	Macrolídeos
DB00834	Hormonio/Antagonista Hormonal	Indutores da Menstruação
DB00411	Oftalmologia	Mióticos
DB03775	Outros/Experimental	Mitógenos
DB03776	Outros/Experimental	Mitógenos
DB00614	Drogas atuantes no SNC	Inibidores da Monoaminoxidase
DB00181	Drogas atuantes ao nível neuroefetor e sinaptico	Relaxates musculares esqueléticos
DB00356	Drogas atuantes no SNC	Relaxantes Musculares Centrais
DB00277	Drogas atuantes no SNC	Relaxantes Musculares Respiratório

FÁRMACO	Atividades Farmacológicas	Categorias DrugBank (Adaptado do DeCS)
DB00675	Hormonio/Antagonista Hormonal	Moduladores Seletivos de Receptor Estrogênico
DB01216	Outros/Experimental	Agentes atuantes na pele mucosa
DB00594	Drogas afetando função renal/cardiovascular	Bloqueadores dos Canais de Sódio
DB00695	Drogas afetando função renal/cardiovascular	Inibidores de Simportadores de Cloreto de Sódio e Potássio
DB00898	Outros/Experimental	Solventes
DB04699	Outros/Experimental	Solventes
DB04682	Outros/Experimental	Espermicidas
DB02901	Outros/Experimental	Glicosídeos Cardíacos
DB01692	Outros/Experimental	Reagentes de Sulfidrila
DB00795	Quimioterapia de doenças microbianas	Sulfonamidas
DB00560	Quimioterapia de doenças microbianas	Tetraciclinas
DB00759	Quimioterapia de doenças microbianas	Tetraciclinas
DB02579	Outros/Experimental	Adesivos Teciduais
DB00328	Outros/Experimental	Tocolíticos
DB01030	Quimioterapia de doenças neoplásicas	Inibidores da Topoisomerase I
DB04690	Outros/Experimental	Inibidores da Topoisomerase I
DB04691	Outros/Experimental	Inibidores da Topoisomerase I
DB04692	Outros/Experimental	Inibidores da Topoisomerase I
DB00738	Quimioterapia das infecções parasitárias	Tripanossomicidas
DB04786	Quimioterapia das infecções parasitárias	Tripanossomicidas
DB03608	Outros/Experimental	Tripanossomicidas
DB00570	Quimioterapia de doenças neoplásicas	Moduladores de Tubulina
DB01394	Quimioterapia de doenças neoplásicas	Moduladores de Tubulina
DB02342	Quimioterapia de doenças neoplásicas	Moduladores de Tubulina
DB00235	Drogas afetando função renal/cardiovascular	Vasodilatadores
DB00350	Drogas afetando função renal/cardiovascular	Vasodilatadores
DB01783	Drogas atuantes no sangue e em órgãos formadores de sangue	Complexo Vitamínico B
DB01022	Drogas atuantes no sangue e em órgãos formadores de sangue	Vitamina K
DB00152	Drogas atuantes no sangue e em órgãos formadores de sangue	Vitaminas
DB00280	Drogas afetando função renal/cardiovascular	Bloqueadores do Canal de Sódio Disparado por Voltagem

ANEXO VII – Lista de periódicos com maior número de artigos processados no *pipeline.*

PERIÓDICOS	REGISTROS
Malaria journal	1534
The American journal of tropical medicine and hygiene	1116
PloS one	816
Infection and immunity	603
Transactions of the Royal Society of Tropical Medicine and Hygiene	578
Molecular and biochemical parasitology	494
Tropical medicine & international health : TM & IH	342
Proceedings of the National Academy of Sciences of the United States of America	331
Vaccine	319
The Journal of infectious diseases	318
Acta tropica	297
The Journal of biological chemistry	293
Bulletin of the World Health Organization	268
The Southeast Asian journal of tropical medicine and public health	266
Annals of tropical medicine and parasitology	259
Antimicrobial agents and chemotherapy	259
Journal of immunology (Baltimore, Md. : 1950)	255
Parasitology research	255
Experimental parasitology	246
Parasitology	201
Parasite immunology	194
International journal for parasitology	187
Lancet	162
Zhongguo ji sheng chong xue yu ji sheng chong bing za zhi = Chinese journal of parasitology & parasitic diseases	159
Memorias do Instituto Oswaldo Cruz	148
PLoS pathogens	142
Molecular microbiology	139
Medecine tropicale : revue du Corps de sante colonial	138
Journal of medicinal chemistry	134
Blood	131
Journal of travel medicine	130
Bulletin de la Societe de pathologie exotique (1990)	121
Parasites & vectors	120
The Journal of experimental medicine	120
Parassitologia	118
Trends in parasitology	116
Medical and veterinary entomology	114
The Journal of parasitology	111
Journal of medical entomology	107
Clinical infectious diseases : an official publication of the Infectious Diseases Society of America	102
Nature	96
Biochemical and biophysical research communications	95
Science (New York, N.Y.)	95
Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases	94
East African medical journal	89
Journal of vector borne diseases	88
PLoS neglected tropical diseases	88
European journal of immunology	85
Journal of the American Mosquito Control Association	80
Cellular microbiology	77

Fonte: Elaborado pelo autor.

ANEXO VIII – Moléculas destinadas à pesquisa básica.

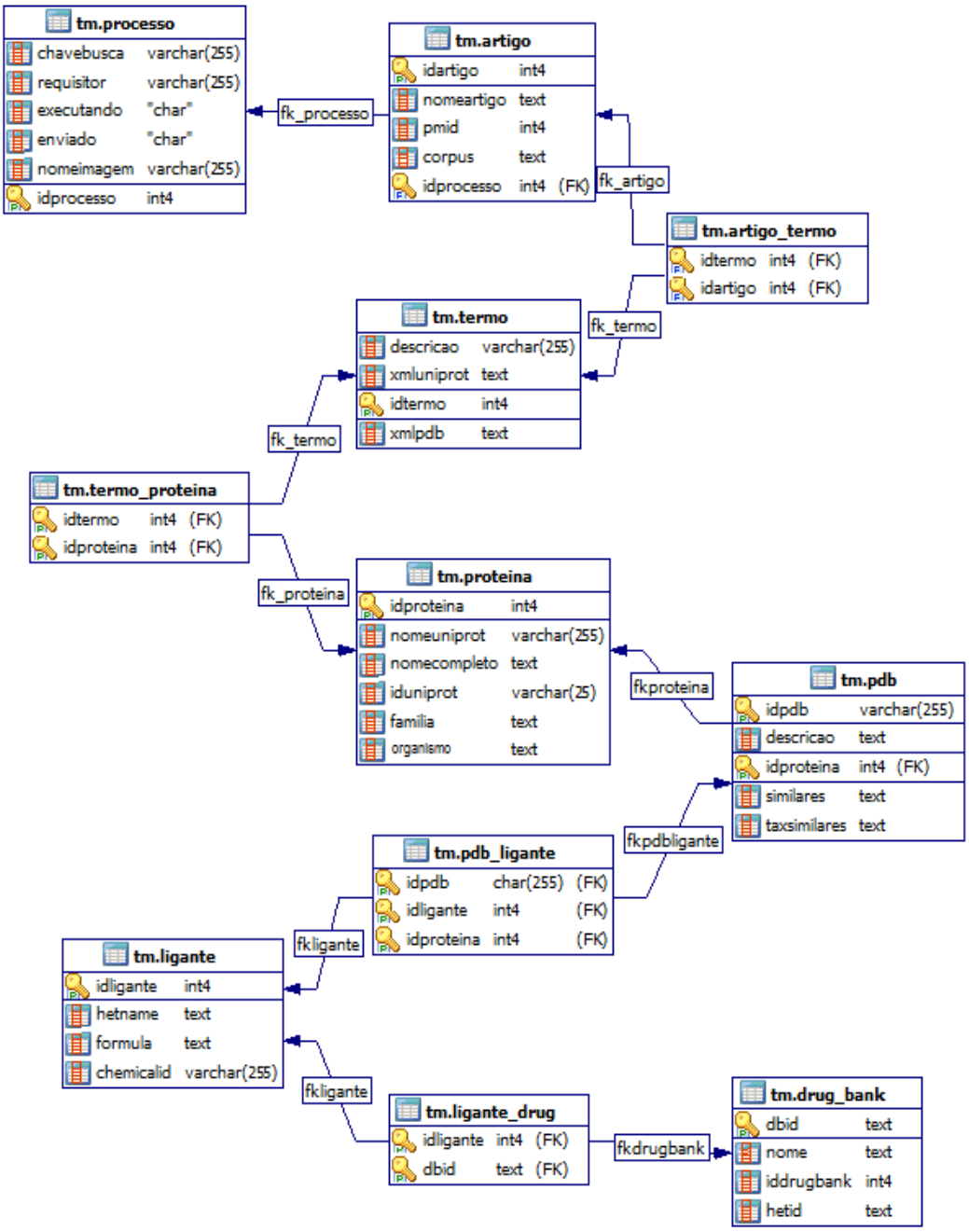
Tabela 24 - Proteínas associadas aos fármacos DB00321 e DB00440.

DB00321	DB00440				
3APW	1PD8	2W9H	3S3V	4KBN	3SR5
3APU	4LAG	3SRQ	4KEB	1MVT	3FS6
3APX	4DDR	4LAE	3LG4	4KFJ	1KMS
3APV	3GHC	1DHF	2C2S	3NXX	3NU0
	1HFP	1HFR	3EIG	1PD9	4M6J
	4LAH	3F8Y	3GHW	3FRD	3SRU
	3GYF	1DRF	1OHK	2W9G	3FRE
	4G95	3FY9	2C2T	3NXY	3FY8
	3GHV	3N0H	3FRA	3SRS	3M08
	3L3R	3NZD	1KMV	2DHF	1OHJ
	3OAF	3SRR	3FRF	3NXV	1HFQ
	1U72	1U71	3I8A	4M6L	3F8Z
	4FGG	3FRB	3NXO	1S3V	4KAK
	1PDB	1BOZ	4M6K	2W3M	3SQY
	3NTZ	3GI2	1DLS	3FYW	3F91
	4FGH	4LEK	1S3U	3SRW	3SGY
	2W3B	1DLR	2W3A	3NXT	3SH2
	3NXR	1YHO	3M09	1MVS	1S3W
	3S7A	3FYV	4KD7		

Tabela 25 - Proteínas associadas ao fármaco DB00563.

DB0563						
3NXY	1RB2	1DRE	1BJG	4EIZ	1RX4	1PD8
4KNZ	2DHF	2KCE	1F4D	4DFR	3OCH	1HFR
1BQ1	3K47	2VET	1F4G	1DYH	1DNA	1TSD
2D0K	3NXV	3QYL	1KMV	1DHF	3QL3	1DRB
3QYO	4KJJ	2FTQ	3BHR	4KAK	1HFQ	1JOM
1TDR	4M6L	1DLR	3GHW	4M6J	2DRC	3BFI
1PD9	1RE7	1LUD	1AJM	4GH8	4EJ1	4KEB
2ANQ	1RH3	2BBQ	1DRH	4NX7	1NCE	3L3R
1AIQ	1RX8	1BOZ	2C2S	1D1G	1KCE	1TJS
4KFJ	4KMZ	3N0H	3EIG	2W3B	1OHJ	1YHO
3NXX	1TYS	1EV5	4KN0	3S7A	1RF7	3D80
1JTQ	1DHI	1RA1	4G95	3DFR	1F4E	3GI2
1RA9	1TSN	1DDU	3R33	4NX6	2H2Q	1MVT
4KBN	2INQ	1DRF	1RX1	3IRN	1JOL	1EVG
2G8X	2ANO	4I1N	2FZJ	1U72	3HBB	1KMS
3K74	1JUT	1MVS	1HFP	3NU0	1U71	2A9W
2HQP	3F8Z	4I13	4DDR	3IRM	3B9H	4EIG
1AXW	1RX2	3INV	1DIS	1TLC	3CLB	1RG7
4KMY	1DRA	4F2V	2HM9	1JTU	3F8Y	1RX5
1DDS	2C2T	1JG0	3NZD	3B5B	1ZPR	1RX9
1KZJ	6DFR	1AO8	1QQQ	3GHV	1AOB	3S3V
1RA8	1OHK	4GEV	3FS6	3GYF	1EV8	1BZF
3NXT	1RA2	2FTO	4KN2	4KJK	1TDU	2LF1
1CZ3	4FHB	1TRG	1DHJ	2VF0	3NTZ	1BID
2L28	3DRC	3CL9	1RX7	4KN1	1PDB	1F4F
1FFL	3TMS	1DYI	1S3U	1RX3	3OAF	1DLS
1RB3	1BDU	4KJL	3K45	3IRO	1F4B	1BQ2

ANEXO IX – Estrutura do banco de dados



Fonte: Jezuz (2013)