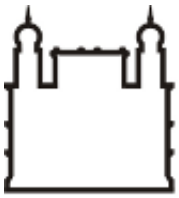


Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

**INSTITUTO OSWALDO CRUZ**
**Pós-Graduação em Biologia Computacional e Sistemas**

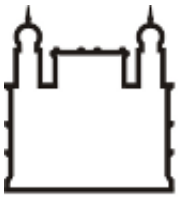
*RODRIGO JARDIM*

# Estudo de reposicionamento de fármacos para doenças negligenciadas causadas por protozoários através da integração de bases de dados biológicas usando Web Semântica

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

**Orientador(es):** Prof. Dra. Maria Luiza Machado Campos
Dr. Alberto Martín Rivera Dávila

Rio de Janeiro
2013



Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

**INSTITUTO OSWALDO CRUZ**
**Pós-Graduação em Biologia Computacional e Sistemas**

*AUTOR: RODRIGO JARDIM*

**ESTUDO DE REPOSICIONAMENTO DE FÁRMACOS PARA DOENÇAS NEGLIGENCIADAS CAUSADAS POR PROTOZOÁRIOS ATRAVÉS DA INTEGRAÇÃO DE BASES DE DADOS BIOLÓGICAS USANDO WEB SEMÂNTICA**

**Orientador(es):** Prof. Dra. Maria Luiza Machado Campos
Dr. Alberto Martín Rivera Dávila

**Aprovada em:**

**Examinadores:**

**Dr.** Floriano Paes Silva Junior - **Presidente**

**Prof$^a$. Dr$^a$.** Maria Claudia Reis Cavalcanti

**Prof$^a$. Dr$^a$.** Raquel Cardoso de Melo Minardi

**Suplentes:**

**Dr.** Ernesto Raúl Caffarena

**Prof$^a$. Dr$^a$.** Camila Silva de Magalhães

Rio de Janeiro, 01 de março de 2013.

*Dedico esta tese as mulheres da minha vida: minha mãe,*
*minha avó, minhas irmãs e a minha esposa.*

# Epígrafe

"É fazendo que se aprende a fazer aquilo que se deve aprender a fazer."

Aristóteles

# Agradecimentos

Agradeço ao Programa de Pós-Graduação em Biologia Computacional e Sistemas, extensivos a todos os funcionários que passaram pela Secretaria Acadêmica durante esses quase quatro anos, pela paciência e disponibilidade sempre que solicitada.

Ao Pablo Mendes e ao Dr. Amit Sheth por terem me ajudado a iniciar meu trabalho de tese.

Aos colegas do Laboratório de Biologia Computacional e Sistemas, Diogo, Rafael, Adriana, Fábio Bernardo, Fábio Motta, Joana e Gisele, pela ajuda e discussões salutares no laboratório e que me ajudaram a superar parte da minha "deficiência biológica" (Dávila, 2010).

Aos Professores-Doutores que tiveram a paciência de me aturar como aluno nas disciplinas do Programa.

À Prof$^a$. Dr$^a$. Maria Luiza que me orientou durante todo esse tempo, esclarecendo e ampliando meus horizontes.

Ao Dr. Alberto Dávila pelo apoio, orientação e firmeza com a qual conseguiu me conduzir ao seu laboratório e pela sua orientação neste trabalho.

Aos meus cachorros, Lilica (*in memoriam*), Brutus (*in memoriam*), Neo e João, por me proporcionarem momentos deliciosos na minha jornada.

Aos meus enteados, Renan e Mauro Silvino, por não me perturbarem durante o período desta tese e por terem ajudado nos momentos mais críticos.

À Barbara, Victória, Vanessa, Erika e Aline, mãe, avó e irmãs por terem me proporcionado oportunidades na vida pessoal, profissional e acadêmica.

A todos os deuses que povoam algum lugar do universo e que me trouxeram até aqui.

À minha esposa, Simone, pelo direcionamento, paciência, apoio, amizade, paciência, carinho, paciência, força, presença e por fim, paciência.

Obrigado!

## *Resumo*

A pesquisa e desenvolvimento de novos fármacos constituem um processo lento e oneroso. Novas técnicas têm sido propostas para agilizar esse processo. Uma dessas técnicas é chamada de reposicionamento de fármacos, cujo objetivo principal é utilizar fármacos já comercializados para tratamento de outras doenças. A proposição de um novo composto, fármaco ou nova utilização de um fármaco é um processo que necessita integrar informações de diversos campos do conhecimento. Na biologia, diversos dados têm sido gerados através das técnicas de larga escala e armazenados em bases de dados de diversos formatos, dificultando a integração desses dados e a consequente geração de novas informações. Neste estudo é proposta uma abordagem de integração de bases biológicas públicas, utilizando os preceitos do *Linked Open Data*, através da utilização de web semântica, RDF e URI, para propor uma lista de possíveis fármacos que possuem potencial para estudos mais aprofundados com a finalidade de serem reposicionados para tratamento de doenças causadas por protozoários.



**Palavras-chave:** reposicionamento, fármacos, integração, web, semântica

## *Abstract*

The research and development of new drugs constitute a slow and expensive process. New techniques have been proposed to accelerate this process. One of such technique is called drugs repositioning whose objective is to use drugs already marketed for the treatment of other diseases. The proposal of a new compound, drug or new use of a drug is a process that needs to integrate information from different fields of knowledge. In biology, many data have been generated through by means of high throughput techniques and stored in databases of different formats, making it difficult to integrate these data and the consequent generation of new information. In this study we propose an approach for integrating public biological databases, using the precepts of Linked Open Data, through the use of semantic web, RDF and URI, propose a list of possible drugs that have potential for further study in order to be repositioned for treatment of protozoal diseases.



**Keywords:** repositioning, drugs, integration, web, semantic

# Lista de Figuras

# Lista de Tabelas

# Sumário

# Capítulo 1

# Introdução

## 1.1 Fármacos e Reposicionamento de Fármacos

O processo de desenvolvimento de um novo fármaco é complexo e de elevado custo financeiro, demandando esforços integrados em vários aspectos relacionados com a inovação, tecnologias e gestão (Guido *et al.*, 2008). A forma tradicional de busca por novos fármacos tem se transformado em razão das pesquisas recentes em genômica e proteômica (Guido *et al.*, 2008).

O processo de Pesquisa e Desenvolvimento (P&D) de um fármaco pode ser dividido em duas grandes fases (Lombardino & Lowe, 2004): os estudos pré-clínicos e os estudos clínicos. Na primeira fase, as pesquisas estão voltadas para a identificação de moléculas com potencial de desenvolvimento clínico (Guido *et al.*, 2010). A segunda fase está voltada para testes clínicos, produção e aprovação do novo fármaco (Lombardino & Lowe, 2004) (Figura 1.1 ).

Estudo de Chong & Sullivan Jr. (2007) aponta que um novo fármaco, produzido pelo método convencional, denominado *de novo*, leva em média 15 anos para ser lançado no mercado, período que compreende o início das pesquisas até a autorização para

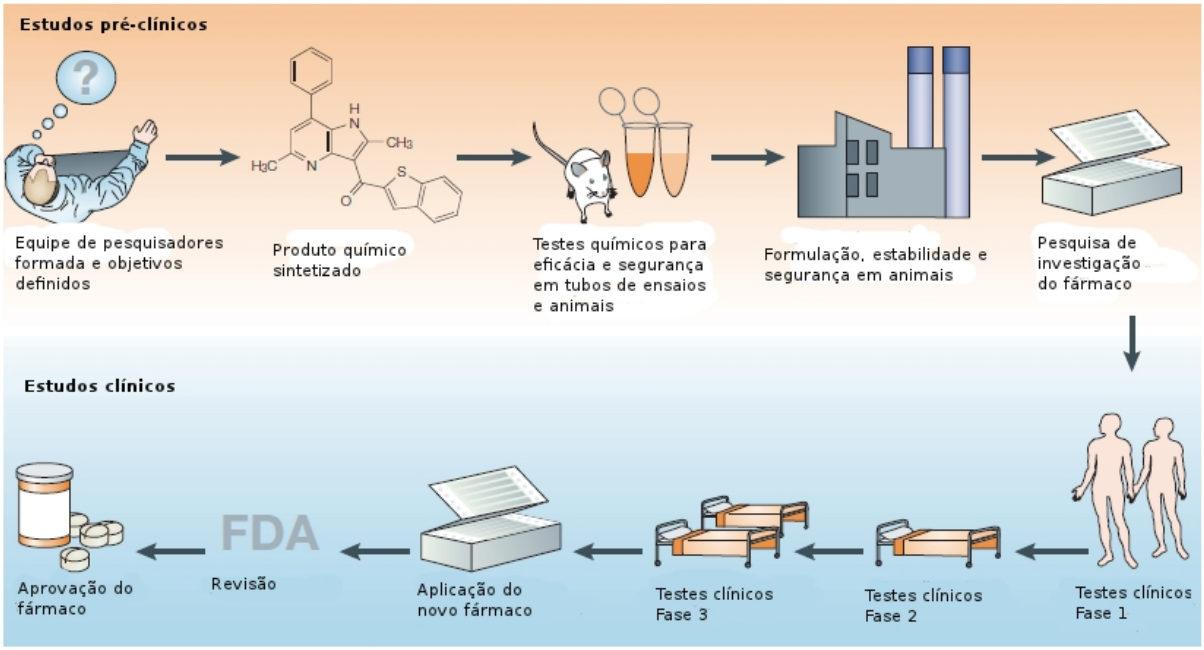



Figura 1.1: Etapas do processo de P&D para novos fármacos, através do processo denominado *denovo* (Fonte: Lombardino & Lowe (2004))

.

comercialização, a um custo financeiro de aproximadamente U$800 milhões. O risco de diminuir os investimentos em P&D de um fármaco é muito grande em virtude da pesquisa minuciosa que deve ser realizada em todas as etapas do desenvolvimento (Ashburn & Thor, 2004). Com a técnica de reposicionamento de fármacos essa questão é parcialmente resolvida pois boa parte das pesquisas farmacêuticas já foram realizadas (Ashburn & Thor, 2004) e, com isso, o tempo de P&D e os recursos envolvidos são diminuídos (Figura 1.2).

Reposicionamento de fármacos é o termo mais recente utilizado para novos usos de antigos fármacos, ou ainda, é o processo de encontrar novos usos fora do escopo da indicação médica original para fármacos existentes (Ashburn & Thor, 2004). Essa técnica é utilizada há muito tempo, havendo registro de seu uso desde meados do século passado (Aronson, 2007).

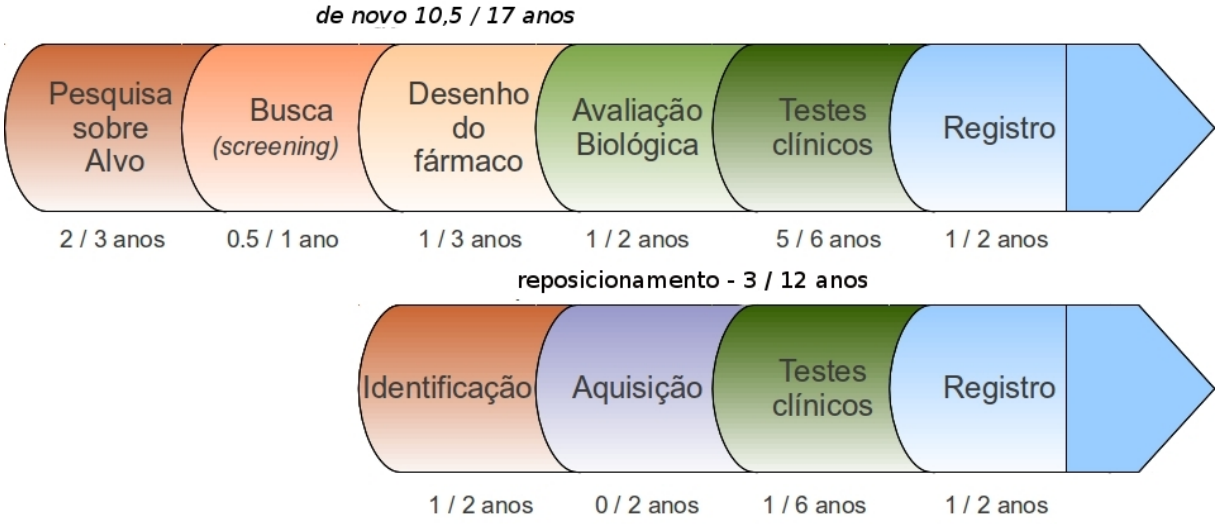



Figura 1.2: Comparação entre os métodos *de novo* e reposicionamento para novos fármacos (Fonte: adaptado de Ashburn & Thor (2004)

As pesquisas utilizando a abordagem de reposicionamento de fármacos vêm ganhando força com os recentes estudos sobre vias metabólicas e a grande capacidade dos equipamentos de sequenciamento de segunda e terceira gerações. Uma consulta realizada no PubMed[1] em fevereiro de 2012 com o termo “drug repositioning” retorna 87 artigos, sendo o primeiro em 2004 e 69 escritos após 2010. Além disso, a diminuição no tempo de P&D pode também reduzir o custo envolvido na sua produção.

A literatura descreve vários exemplos de sucesso do reposicionamento de fármacos. Um dos casos mais conhecido é da Talidomida que inicialmente foi prescrita como sedativo em 1946 e utilizada em 46 países para combater enjoos matinais em grávidas (Matthews & McCoy, 2003). Em pouco tempo, o uso desse medicamento foi associado com o nascimento de crianças com focomelia[2]. Em 1962 a talidomida foi retirada de comercialização em quase todo mundo e em 1965 Sheskin (1965) publicou um trabalho onde utilizou a talidomida para tratamento de hanseníase. Estudo recente revelou o alvo primário da talidomida que causa a teratogenicidade[3] (Ito *et al.*, 2010) abrindo

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] é uma anomalia congênita que impede a formação normal de braços e pernas.
[3] capacidade de produzir malformações congênitas.

caminho para o desenvolvimento de uma nova talidomida sem atividade teratogênica.

Outro caso bem conhecido é o do Citrato de sildenafila (Viagra) que foi desenvolvido para o tratamento da hipertensão arterial e durante os testes clínicos foi observado que sua função vasodilatadora agia primeiro no órgão genital masculino e não no coração, o que orientou sua comercialização para o tratamento da disfunção erétil (Terrett *et al.*, 1996).

A Miltefosina é outro exemplo de reposicionamento de fármaco. Inicialmente comercializada para o tratamento do câncer de mama (Clive *et al.*, 1999) e mais recentemente passou a ser utilizada para tratamento da leishmaniose visceral (Sundar *et al.*, 1998).

No trabalho de Aronson (2007) podemos encontrar uma breve revisão de diversos fármacos que já foram reposicionados com sucesso.

## 1.2 Doenças Negligenciadas

A Organização Mundial da Saúde (OMS) classifica as doenças como Tipo I, Tipo II e Tipo III, correspondendo a Doenças Globais, Negligenciadas e Muito Negligenciadas, respectivamente (Sachs, 2001).

As doenças negligenciadas ocorrem em regiões pobres e não são priorizadas pelas grandes indústrias farmacêuticas mundiais e de biotecnologia, responsáveis pela fabricação de vacinas, fármacos e kits de diagnósticos (Morel *et al.*, 2009). O fato de não serem priorizadas pelas indústrias farmacêuticas é conhecido como “falhas de mercado” (*market failures*), isto é, uma alocação ineficiente de produtos e serviços através dos mecanismos usuais do mercado livre. Apesar disso, nos últimos anos vêm

crescendo o número de pesquisas envolvendo as doenças negligenciadas, mostrando a relevância do tema para a saúde pública (Figura 1.3), sendo que cerca de 90% dos investimentos realizados em P&D provêm de recursos públicos ou filantrópicos (Moran *et al.*, 2009).

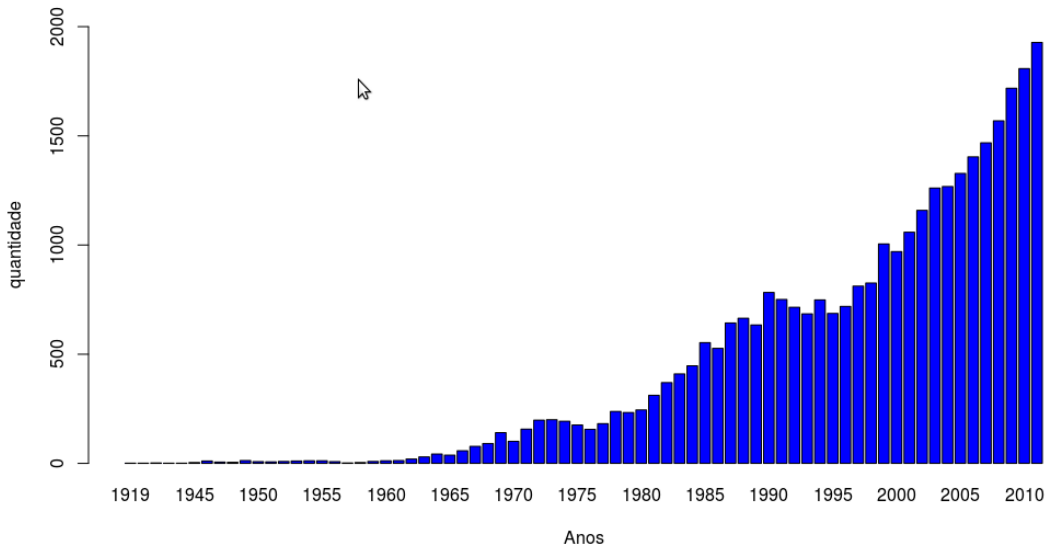



Figura 1.3: Quantidade de artigos disponíveis no PubMed com o termo "*neglected disease*"

De acordo com o último relatório do G-Finder (Moran *et al.*, 2009), atualmente esse percentual é de 80%. Neste mesmo relatório, o Brasil é apontado como o quarto país que mais investe em P&D para fármacos, vacinas e diagnósticos para as doenças negligenciadas, superado pelos Estados Unidos, Comunidade Europeia e Reino Unido, tendo sido realizado um investimento total de cerca de US$ 36,8 milhões. Apesar do volume financeiro apontado pelo relatório, esses investimentos são concentrados em diferentes tipos de doenças negligenciadas como dengue, pneumonia e tuberculose, representando 59,5%, enquanto que as doenças causadas por protozoários receberam apenas 13% (Figura 1.4).

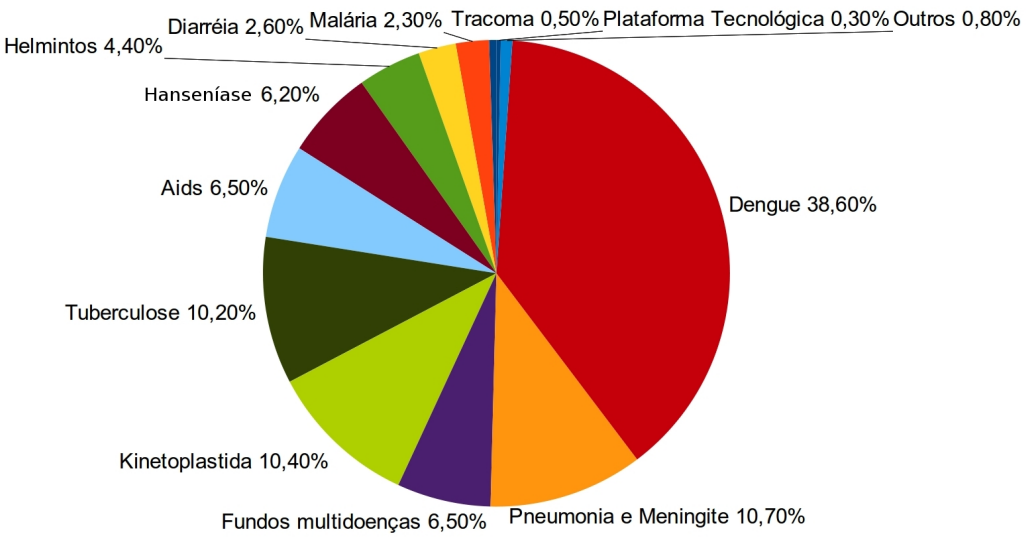



Figura 1.4: Distribuição dos investimentos em doenças negligenciadas no ano de 2009 no Brasil (Fonte: Moran *et al.* (2009))

Doenças causadas por protozoários ocorrem principalmente em áreas pobres de países em desenvolvimento e, prioritariamente, na zona tropical do globo terrestre, sendo também conhecidas como Doenças Tropicais. Tais doenças são classificadas como Tipo II ou III pela OMS e afetam cerca de 500 milhões de pessoas em todo o mundo (WHO, 2012a,b,c; Moran *et al.*, 2009).

## 1.3 Protozoários e doenças negligenciadas

De acordo com Kotpal (2010) os protozoários formam um sub-reino do reino animal, constituídos por eucariotos unicelulares. Existem cerca de 50 mil espécies de protozoários, a maioria de vida livre, podendo ser encontrados na água doce, salobra ou salgada e na terra (Kotpal, 2010). O termo protozoário abrange protistas móveis unicelulares que apresentam uma grande variedade de complexidade estrutural, além de se adaptarem a diversas condições ambientais (Ruppert & Barnes, 1996).

Filogeneticamente, os protozoários estão espalhados pela árvore da vida, estando presente nas linhagens excavatas, discicristados, alveolatas e amebozoas (Figura 1.5).

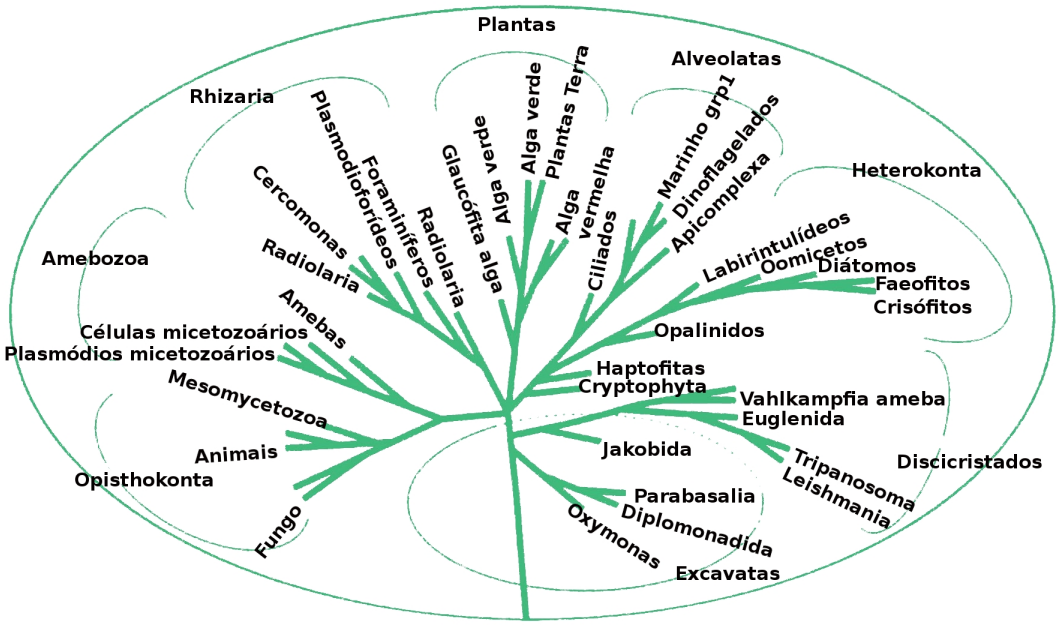



Figura 1.5: Filogenia de Eucariotos (Fonte: adaptado de Baldauf (2003)).

Várias espécies de protozoários são parasitas e causam doenças em seres humanos (Tabela 1.1). Dentre as doenças causadas por protozoários temos a doença de Chagas, malária e leishmaniose que afetam cerca de 500 milhões de pessoas no mundo, preferencialmente nas áreas tropicais e subtropicais do globo (WHO, 2012b,c).

Existem disponíveis atualmente em bases de dados públicas informações sobre o sequenciamento de genomas de diversos protozoários. Esses dados são de grande importância para os estudos de genômica comparativa, estudos de evolução e estudos de alvos para fármacos, vacinas e kits de diagnósticos.

A base de dados do ProtozoaDB (Dávila *et al.*, 2008) que também possui uma aplicação web contém informações sobre o proteoma e o genoma de protozoários patogênicos. Em sua nova versão, o ProtozoaDB disponibiliza informações sobre 22

Tabela 1.1: Doenças causadas pelos protozoários da versão atual do ProtozoaDB e a região predominante afetada.

| **Organismo** | **Doença** | **Área Predominante** |
|---|---|---|
| *Plasmodium falciparum* | Malária | África |
| *Plasmodium vivax* | Malária | América do Sul |
| *Plasmodium berghei* | Malária | Trópicos |
| *Plasmodium chabaudi* | Malária | Trópicos |
| *Plasmodium yoelii* | Malária | Trópicos |
| *Plasmodium knowlesi* | Malária | Ásia |
| *Babesia bovis* | Babesiose (Tristeza Bovina) | Trópicos |
| *Theileria annulata* | Febre da Costa Oriental da África | África |
| *Theileria parva* | Theileriose Tropical | América Latina |
| *Toxoplasma gondii* | Toxoplasmose | Cosmopolita |
| *Cryptosporidium muris* | Criptosporidiase | Cosmopolita |
| *Cryptosporidium parvum* | Criptosporidiase | Cosmopolita |
| *Cryptosporidium hominis* | Criptosporidiase | Cosmopolita |
| *Trypanosoma brucei* | Doença do Sono | África |
| *Trypanosoma cruzi* | Doença de Chagas | América Latina |
| *Leishmania braziliensis* | Leishmaniose | América Latina |
| *Leishmania infantum* | Leishmaniose | América Latina e Velho Mundo |
| *Leishmania major* | Leishmaniose | Velho Mundo |
| *Entamoeba histolytica* | Amebíase | Cosmopolita |
| *Entamoeba dispar* | Amebíase | Cosmopolita |
| *Giardia intestinalis* | Giardiase | Cosmopolita |
| *Trichomonas vaginalis* | Tricomoníase | Cosmopolita |

protozoários patogênicos das linhagens alveolata (*Plasmodium*, *Babesia*, *Theileria*, *Toxoplasma* e *Cryptosporidium*), amebozoa (*Entamoeba*), excavata (*Trichomonas*, *Giardia*) e discicristados (*Trypanosoma* e *Leishmania*) (Figura 1.5).

## 1.4 Fármacos para doenças negligenciadas causadas por protozoários

Entre 1975 e 2004, apenas 1,3% de todos os medicamentos desenvolvidos no mundo foram para tratamento de doenças negligenciadas e muito negligenciadas (Chirac & Torreele, 2006) e apesar das indústrias farmacêuticas ocuparem a terceira posição em investimento para P&D em doenças negligenciadas (Moran *et al.*, 2009), atrás dos investimentos governamentais e instituições filantrópicas, este valor pode ser considerado muito baixo, visto o quantitativo de acometidos por estas doenças, com o agravante de estar concentrado em apenas algumas doenças como HIV e tuberculose.

Estudo recente mostra que apenas 5% do financiamento mundial em inovação para doenças negligenciadas foram aplicados no grupo de doenças que inclui a doença do sono, leishmaniose visceral e doença de Chagas, caracterizadas como muito negligenciadas (Tipo III), ainda que essas doenças acometam cerca de 500 milhões de pessoas em todo o mundo (Moran *et al.*, 2009).

Em 1999 a organização “Médicos Sem Fronteiras” destinou os recursos recebidos pelo Prêmio Nobel da Paz à implantação de um novo modelo de P&D para doenças negligenciadas e muito negligenciadas. O resultado foi a criação em 2003 da iniciativa Medicamentos para Doenças Negligenciadas (*Drugs for Neglected Diseases initiative* - DNDi ) com a participação de sete organizações de diferentes países, entre elas a Fundação Oswaldo Cruz-Brasil (DNDi, 2010).

Atualmente, existem poucos fármacos disponíveis para tratamento de doenças negligenciadas e muito negligenciadas e esses ou são muito tóxicos ou são antigos e ineficazes (Tabela 1.2). Um exemplo é o fármaco Benznidazol para tratamento da doença de Chagas que apesar de eficaz para o tratamento da doença na fase aguda, apresenta diversos efeitos colaterais como depressão da medula óssea e dermatopatia alérgica (Cançado, 2002).

Tabela 1.2: Fármacos para doenças negligenciadas e muito negligenciadas disponíveis para tratamento - Fonte: (Coura, 2008; DNDi, 2010)

| **Doença** | **Fármaco** | **Desvantagens** |
|---|---|---|
| Doença de Chagas | Benznidazol, nas fases agudas e iniciais. De 60% a 80% de cura. | Tratamento de longa duração e alta toxicidade |
| Leishmaniose visceral | Antimoniais pentavalentes, Anfotericina B, Paromomicina e Miltefosina | Tóxicos e custo elevado |
| Doença do Sono | Suramina, Melarsoprol, Eflornitina e Pentamidina | Tóxicos e de difícil administração |
| Malária | Cloroquina e Primaquina | Já não é mais eficaz |
| Giardíase | Albendazol, Furazolidona, Metronidazol, Paramomicina, Quinacrina, Secnidazol e Tinidazol. De 60 a 97% de cura. | Não se aplica |
| Tricomoníase | Metronidazol, Tinidazol e Secnidazol | Tóxico |
| Amebíase | Não invasiva, Clorofenoxamida com 90% de cura e Cloroacetamida com 80% de cura. Na invasiva, Metronidazol ou Secnidazol com 97% de cura. | Não se aplica |
| Toxoplasmose | Pirimetamina, Sulfadiazina e Ácido folínico. Portadores de HIV, Sulfametoxazol | Tóxico |

## 1.5 Homologia

No estudo da genômica comparativao conceito de homologia é bastante explorado. Homologia designa a relação de descendência entre quaisquer entidades, sem especificar o cenário de evolução (Koonin, 2005). Dois genes são ditos homólogos se suas origens forem comuns. A relação de homologia pode ser subdividida em duas outras relações que revelam o cenário de evolução: a ortologia e a paralogia (Figura 1.6).

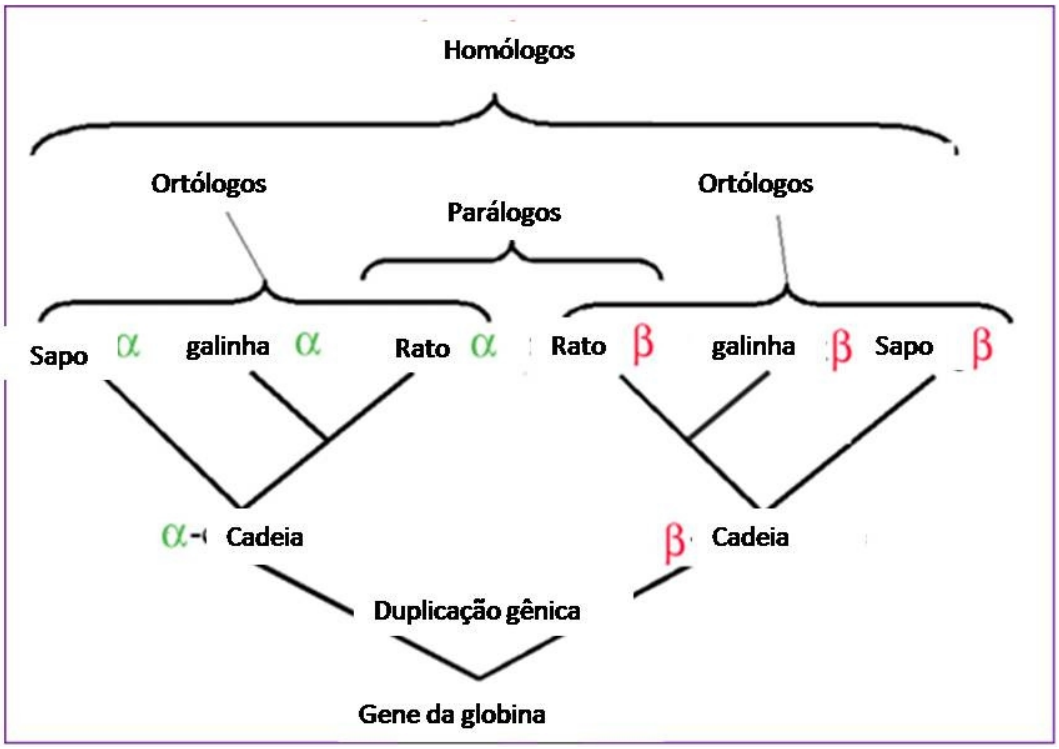



Figura 1.6: Exemplo de relações de homologia do gene da globina, demonstrando os eventos de duplicação (paralogia) e especiação (ortologia) (Fonte: adaptado de `http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html`).

Os eventos de evolução dos genes podem ser classificados como: especiação ou descendência vertical, duplicação, perda, transferência horizontal (HGT) e fusão, fissão e outros rearranjos (Koonin, 2005).

A ortologia é relacionada ao evento de especiação. Dois genes são ortólogos quando se especiaram na última relação de ancestralidade, isto é, possuem um an-

cestral comum antes do evento de especiação. A paralogia está relacionada ao evento de duplicação de um gene (Figura 1.6).

Os conceitos de homologia são fundamentais para a genômica comparativa e têm permitido avanços em termos de inferência de estruturas e funções hipotéticas de genes (Koonin, 2005). Métodos computacionais para a determinação da relação de homologia entre genes e até entre genomas podem ser realizados por árvores filogenéticas (Storm & Sonnhammer, 2002) ou métodos estatísticos (Remm *et al.*, 2001; Enright *et al.*, 2002; Li *et al.*, 2003), tendo todos estes métodos como ponto de partida a análise de similaridade.

## 1.6 Estrutura de Tecnologia da Informação em Bioinformática

O uso de tecnologia da informação (TI) em biologia molecular fez surgir o termo *in silico* para a área da saúde e que vem sendo utilizado há algumas décadas (León & Markel, 2006). A primeira sequência completa de um gene foi conseguida na década de 80 através de técnicas de biologia molecular (Richon, 2012). Nesta mesma década pesquisadores conseguiram publicar o mapa físico de *Escherichia coli* (Kohara *et al.*, 1987) e surgiram vários programas de computadores para tratar e analisar as sequências de nucleotídeos e proteínas (Richon, 2012).

Com a inserção da tecnologia à biologia molecular, ao longo dos últimos 30 anos, muitas ferramentas computacionais surgiram e se proliferaram em todo mundo através da Internet (León & Markel, 2006). Diversos grupos de pesquisa passaram a disponibilizar seus dados para a comunidade científica, de forma pública ou privada, através de arquivos do tipo texto ou de consultas diretas a suas bases de dados.

Com o advento dos equipamentos de sequenciamento e microarranjo, o volume de dados teve significativo crescimento, gerando problemas de armazenamento e escalabilidade (Kasprzyk & Smedley, 2006).

Se por um lado esse aporte tecnológico no campo da biologia molecular fez com que os repositórios de dados biológicos se proliferassem pelo mundo em diferentes formatos (principalmente heterogêneos), o que dificulta a integração, por outro lado, a interação entre esses conjuntos de dados é potencialmente interessante para análises dessas informações, podendo ser alcançada com iniciativas para sua integração (Kasprzyk & Smedley, 2006).

### 1.6.1 Integração de Bases de Dados Heterogêneas

Estruturas heterogêneas de bases de dados se referem às diferenças entre as bases, incluindo tanto a heterogeneidade no nível de sistema e estrutura, quanto no nível semântico (Kim & Seo, 1991; Liu *et al.*, 2010). O problema de integração de bases de dados heterogêneas vem sendo estudado desde o final do século passado (Litwin *et al.*, 1990; Sheth & Larson, 1990; Halevy, 2003) e diversas técnicas foram descritas na literatura para tentar solucionar o problema, tais como: o uso de bancos de dados federados, mediadores e *Data Warehousing* (Liu *et al.*, 2010; Cuadra *et al.*, 2011).

Bancos de dados federados pressupõem a criação de um esquema global que irá gerenciar a integração entre os esquemas locais sem que seja comprometida a autonomia das bases integrantes (Sheth & Larson, 1990).

A solução através de mediadores, também denominados *midlleware*, é feita através de desenvolvimento em camadas. A primeira camada é responsável pela tradução dos esquemas locais para um esquema global. A segunda camada mediará a comunicação entre a tradução e o usuário final, tornando invisível quais bases estão sendo

acessadas (Wiederhold, 1994).

O *Data Warehousing* é utilizado principalmente em organizações como suporte à tomada de decisão. Constitui um ambiente de dados composto por um ou mais bancos de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo (Inmon, 1997). Essa integração utiliza técnicas de Extração, Transformação e Carga (*Extraction, Transformation and Load* - ETL) para tratamento dos dados oriundos de diferentes fontes, permitindo que o processo de obtenção e integração dos dados seja monitorado e apoiado. Tais técnicas, normalmente, são específicas para cada solução, demandando tempo e dinheiro para elaboração, implementação, implantação e manutenção.

As bases de dados biológicas são naturalmente conjuntos de dados heterogêneos (Kasprzyk & Smedley, 2006) e são disponibilizadas em diversos formatos. As primeiras bases de dados biológicas eram arquivos em formato texto representando uma informação específica (Kasprzyk & Smedley, 2006). As bases de dados biológicas mais modernas representam dados segundo o modelo relacional, isto é, atraves de um esquema relacional e são administradas por Sistemas Gerenciadores de Bancos de Dados (SGBD) . Um exemplo é o esquema relacional do Esquema Genômico Unificado (*Genomic Unified Schema* - GUS) (Davidson *et al.*, 2000). O GUS é um esquema relacional e está associado a um conjunto de ferramentas e programas para desenvolvimento de rotinas de carga e análise de dados. Utiliza o dogma central da biologia como princípio de organização dos dados, contendo mais de 400 tabelas divididas categorias.

Embora as bases de dados mais recentes já busquem utilizar modelos de dados comuns, o esquema relacional utilizado por cada grupo de pesquisa pode ser diferente, além de ainda existirem diversas outras bases biológicas com formatos diversos, tais

como texto, HTML, entre outras. Isto torna o trabalho de integração entre essas bases uma tarefa difícil e custosa.

Com o uso da internet, o problema de integração passou a ter novas proposições uma vez que as informações passaram a ter uma organização semi-estruturada[4], além do uso do Protocolo de Transferência de Hipertexto (*HyperText Transfer Protocol* - HTTP) para comunicação, que permite transações sem estado (*stateless*[5]) (Barbosa *et al.*, 2001). Mais recentemente, procurou-se a utilização de linguagens com um poder maior de organização estrutural, como a Linguagem de Marcação Extensível (*eXtensible Markup Language* - XML)[6] para a produção de conteúdo na web.

Dois grupos de pesquisa (Kim & Seo, 1991; Kashyap & Sheth, 1996) propuseram integração de dados através do uso de semântica e não somente baseado na estrutura dos dados. Isso significa que um determinado conceito de um termo é preponderante para a integração de bases de dados e, de alguma forma, seu significado precisa estar explicitado.

### 1.6.2 Web Semântica

O aumento de expressividade na Web foi proposto em 1999 (Berners-Lee *et al.*, 1999) através de um modelo de representação que facilita a automatização na interpretação de páginas publicadas. O RDF (*Resource Description Framework*)[7], um padrão expresso em XML, é um modelo para descrição e integração de informações, dando suporte à interoperabilidade entre diferentes padrões de descritores utilizados.

[4]Principalmente com a utilização da linguagem HTML, que possui uma estrutura basal, permitindo extensões de qualquer tipo, forma e conteúdo

[5]Transações "sem estado", ou seja, a comunicação entre cliente e servidor não é mantida ao longo de toda a transação, assim como ocorre com bancos de dados (*statefull*)

[6]http://www.w3. org/XML/

[7]http://www.w3.org/RDF/

Em 2001, Berners-Lee *et al.* (2001) sugeriram o uso de semântica para automatizar a recuperação de conteúdos na Web, cunhando o termo *Web Semântica*. A ideia básica é trazer uma estrutura para o conteúdo da Web onde a informação é oferecida com um significado bem definido, com a finalidade de prover programas (agentes de *software*) que possam realizar tarefas sofisticadas para os usuários, permitindo um trabalho cooperativo entre pessoas e computadores (Berners-Lee *et al.*, 2001).

Posteriormente dois novos termos surgiram neste contexto: *web de documentos* (*Web of documents*) e *web de dados* (*Web of data*). Berners-Lee (2006) preconiza que assim como a web de documentos, a web de dados é construída sobre documentos com marcações específicas na Web. A diferença é que na web de documentos os conteúdos estão disponibilizados na Linguagem de Marcação de Hipertexto (*Hyper-Text Markup Language* - HTML), enquanto que na web de dados os documentos são elaborados no modelo RDF.

Dentre as tecnologias adotadas pela Web Semântica, uma em especial é fundamental para o modelo adotado: o uso de Identificador Uniforme de Recurso (*Uniform Resource Identifier* - URI)(Berners-Lee, 1998). O URI é uma cadeia de caracteres que identifica um recurso físico ou abstrato e é semelhante ao utilizado pelos endereços de páginas na Web. Um exemplo de URI seria **http://biowebdb.org/protozoadb/protein/** para identificar o significado do termo proteína (*protein*).

### 1.6.3 Dados em RDF

RDF é um modelo padrão para intercâmbio de dados, provendo uma linguagem assertiva para expressar proposições com um vocabulário formal, particularmente o RDF *Schema* (RDFS), para acesso e uso de dados na internet.

Documentos RDF são representados no formato de triplas de informação (Figura 1.7), utilizando o URI e formam um grafo[8].


Figura 1.7: Exemplo de Tripla de informação em um formato de grafo. Os nós representam sujeito e objeto e a aresta, o predicado. (Fonte: `http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#section-Concepts`)

Cada informação é representada como: sujeito, predicado e objeto. O sujeito pode ser um URI ou um nó vazio, o predicado é sempre um URI e o objeto pode ser um URI, um valor literal ou nó vazio. Uma sentença em tripla RDF fornece a informação na qual dois recursos (sujeito e objeto) estão ligados entre si através de uma propriedade (predicado). Assim, por exemplo, pode-se designar o termo gene como sujeito, proteína como objeto e codifica como predicado, tendo assim a tripla: gene codifica proteína (Figura 1.8).


Figura 1.8: Exemplo de Grafo de RDF para gene e proteína.

A *World Wide Web Consortium* (W3C)[9] recomenda o uso de RDF e disponibiliza documentação para a padronização da linguagem [10].

[8]Grafo é uma representação gráfica que consiste em um conjunto de dados interligados par-a-par através de arestas (Figuras 1.7, 1.8, 1.13 e 1.14).

[9]W3C é uma comunidade internacional que desenvolve e recomenda o uso de padrões abertos para o uso da internet (http://w3.org)

[10]`http://www.w3.org/TR/2004/REC-rdf-mt-20040210/` e `http://www.w3.org/TR/2004/REC-rdf-schema-20040210/`

### 1.6.4 Formatos para informação em modelos RDF

Existem alguns formatos de arquivos para expressar o conhecimento em formato de triplas de informação, conforme veremos a seguir:

1. Formato RDF/XML: utilizando a linguagem XML é possível expressar um grafo de RDF em um documento do tipo texto para processamento automatizado (Figura 1.9). Esse formato é recomendado pela W3C e seus padrões podem ser encontrados em http://www.w3.org/TR/REC-rdf-syntax/.

```
<rdf:Description>
  <ex:editor>
    <rdf:Description>
      <ex:homePage>
        <rdf:Description>
        </rdf:Description>
      </ex:homePage>
    </rdf:Description>
  </ex:editor>
</rdf:Description>
```

Figura 1.9: Modelo de Grafo de RDF expresso em formato RDF/XML. (Fonte: `http://www.w3.org/TR/REC-rdf-syntax/`)

2. Formato Notation3 (N3): outro formato bastante utilizado para expressar grafos de RDF é com a linguagem N3 (Figura 1.10), também recomendada e padronizada pela W3C[11]. Este formato além de expressar RDF, permite também expressar regras aplicáveis às triplas.

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.

<http://en.wikipedia.org/wiki/Tony_Benn>
  dc:title "Tony Benn";
  dc:publisher "Wikipedia".
```

Figura 1.10: Modelo de Grafo de RDF expresso em formato Notation 3 (N3). (Fonte: `http://en.wikipedia.org/wiki/Notation3`)

3. Formato N-Triples: é um formato texto baseado em linha de fácil leitura e mais

[11] `http://www.w3.org/DesignIssues/Notation3.html`

compacto que os outros dois formatos (Figura 1.11). Foi baseado em um subconjunto da N3 e também é recomendado e padronizado pela W3C[12].

```
<http://www.w3.org/2001/sw/RDFCore/ntriples/> <http://purl.org/dc/elements/1.1/creator> "Dave Beckett" .
<http://www.w3.org/2001/sw/RDFCore/ntriples/> <http://purl.org/dc/elements/1.1/creator> "Art Barstow" .
<http://www.w3.org/2001/sw/RDFCore/ntriples/> <http://purl.org/dc/elements/1.1/publisher> <http://www.w3.org/> .
```

Figura 1.11: Modelo de Grafo de RDF expresso em formato N-Triples (Fonte: `http://www.w3.org/2001/sw/RDFCore/ntriples/`).

A proposta da Web Semântica, desde o início, inclui a utilização de ontologias, como forma de representar o conhecimento consensual sobre um domínio para determinada comunidade. Esse conhecimento, uma vez representado, pode ser utilizado para a anotação e referência de recursos na Web, fazendo com que possam ser processados considerando-se um contexto em particular.

### 1.6.5 Ontologias

O termo ontologia surgiu na filosofia associada ao estudo da essência do ser e sua existência. Na computação, uma ontologia é uma especificação formal da conceitualização de uma representação de conhecimento sobre um determinado domínio (Gruber, 1993). Uma ontologia tenta capturar o conhecimento de um domínio de uma comunidade como uma coleção estruturada de termos e definições (Uschold & Gruninger, 1996).

Para representar ontologias, formalmente foram criadas diversas linguagens estruturadas. Em 2004, a W3C aprovou a *Web Ontology Language* (OWL)[13] como linguagem para representação de ontologias (McEntire & Stevens, 2006). Como fruto de estudos e avanços em pesquisas na área de Inteligência Artificial, a OWL surgiu para estender as facilidades do RDF, como forma de representar conhecimento sobre

[12] `http://www.w3.org/2001/sw/RDFCore/ntriples/`
[13] http://www.w3.org/2004/OWL/

um domínio, através de um formalismo que associa termos entre si, utilizando-se de um conjunto de propriedades e relações mais ricas (McEntire & Stevens, 2006). Em 2009 a W3C disponibilizou os padrões para a OWL, versão 2 (OWL 2), estendendo as características e resolvendo alguns problemas identificados na primeira versão.

A OWL 2 permite representar classes, propriedades, tipos de dados e suas propriedades, objetos, operações, restrições, hierarquias, entre outras características e pode ser representada em diversos formatos de sintaxe, como o padrão RDF/XML (Figura 1.12) ou a sintaxe Manchester. Cada uma das sintaxes possui sua finalidade, mas representam a mesma linguagem. Por exemplo, o RDF/XML é a sintaxe obrigatória para as ferramentas da OWL 2, enquanto que a Manchester é facilmente entendida por pessoas que não possuem conhecimento avançado sobre ontologias.

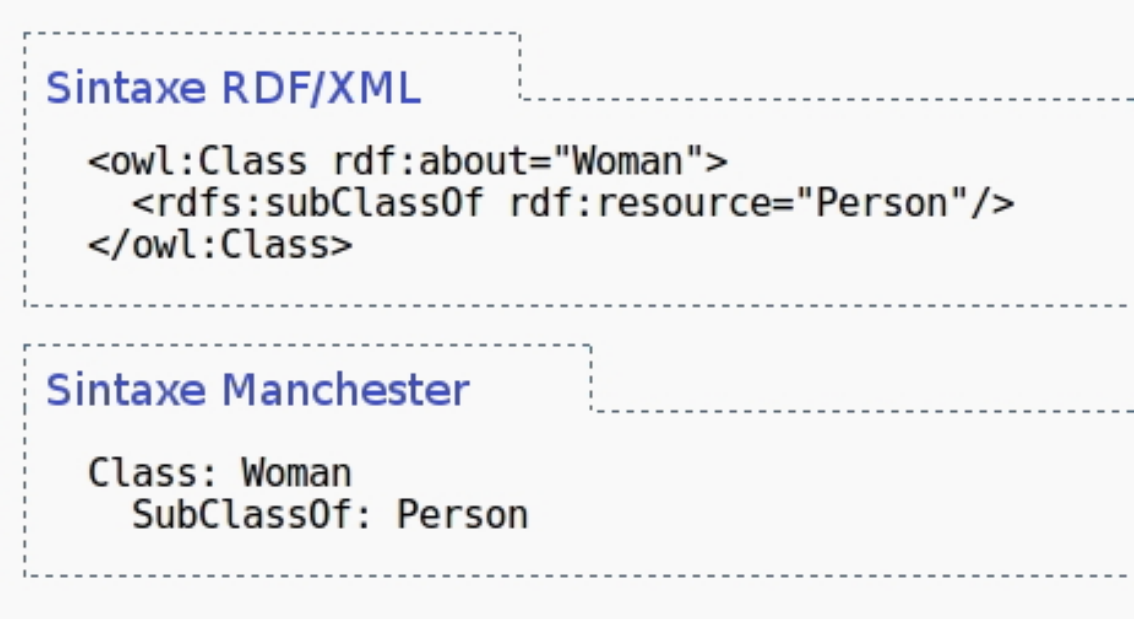



Figura 1.12: Exemplo de sintaxes para OWL demontrando a diferença entre a sintaxe RDF/XML e Manchester (Fonte: `http://www.w3.org/TR/2009/REC-owl2/primer-20091027/`).

### 1.6.6 Bases de dados enriquecidas semanticamente

Bases de dados enriquecidas semanticamente podem ser armazenadas de diversas maneiras, entretanto, ainda há uma preferência pelas bases relacionais (Hausenblas & Karnstedt, 2010). Esquemas de armazenamento de RDF expressos em modelos relacionais podem ser eficientes, agregando os benefícios de ambas as tecnologias. Berners-Lee (1998) destacou a conexão direta entre o modelo de dados da web semântica com o modelo relacional.

Diversas soluções têm sido propostas para trabalhar com grande volume de dados em RDF, como o JENA, Virtuoso e SESAME, descritos abaixo.

O projeto Apache JENA[14] abrange uma biblioteca em Java para montar aplicações de Web Semântica contendo interpretadores SPARQL)[15], inferências baseadas em regras de ontologias e uma variedade de estratégias de armazenamento de triplas RDF, incluindo um modelo relacional.

O Virtuoso é um servidor multimodelos que permite armazenar dados em formato relacional, triplas de RDF, XML e texto livre. Funciona ainda como um servidor Web com suporte aos protocolos HTTP e SPARQL. Possui interface amigável para administração das bases de dados.

De forma semelhante ao JENA, o projeto SESAME[16] provê uma biblioteca para trabalhar com triplas RDF, incluído armazenamento, inferências e consulta aos dados. Suporta os principais tipos de dados que expressam RDF: RDF/XML, N3 e N-triples.

[14] http://jena.apache.org/index.html

[15] SPARQL é um acrônimo recursivo para *SPARQL Protocol And RDF Query Language* e é divido em duas tecnologias: um protocolo e uma linguagem de consulta à RDF.

[16] http://www.openrdf.org/about.jsp

Várias organizações já disponibilizam suas bases de dados na web permitindo consultas e até mesmo que sejam baixadas para uso público. Com o número crescente de bases disponibilizadas na web foi proposta a interligação desses dados, dando origem aos Dados Abertos Ligados (Linked Open Data - LOD).

### 1.6.7 Dados abertos ligados

Bizer *et al.* (2007) propuseram a interligação de dados na Web com o uso de RDF. Este foi o primeiro trabalho que deu origem aos Dados Abertos Ligados (*Linked Open Data* - LOD). O LOD é uma iniciativa para interligar dados relacionados na Web, expondo, compartilhando e ligando informações através de Web Semântica, com o uso de URI e RDF.

O projeto inicial contemplou 8 conjuntos de dados com cerca de 1,5 bilhão de triplas e aproximadamente 150 mil ligações de RDF (Figura 1.13).

No último grafo disponibilizado pelo LOD, em 2010, existem 295 conjuntos de dados com 31 bilhões de triplas e cerca de 504 milhões de ligações RDF[17], demonstrando um crescimento rápido, tanto em números de bases disponibilizadas, quanto em qualidade de ligações entre essas bases (Figura 1.14).

Com a finalidade de padronizar os dados disponibilizados na nuvem de LOD, foram definidos quatro princípios (Berners-Lee, 2006):

- Todos os itens em um conjunto de dados precisam ser identificados utilizando URI;
- Todos os URI precisam ser reais e apontar para um endereço real;

[17]Fonte: `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`, acessado em novembro de 2012.

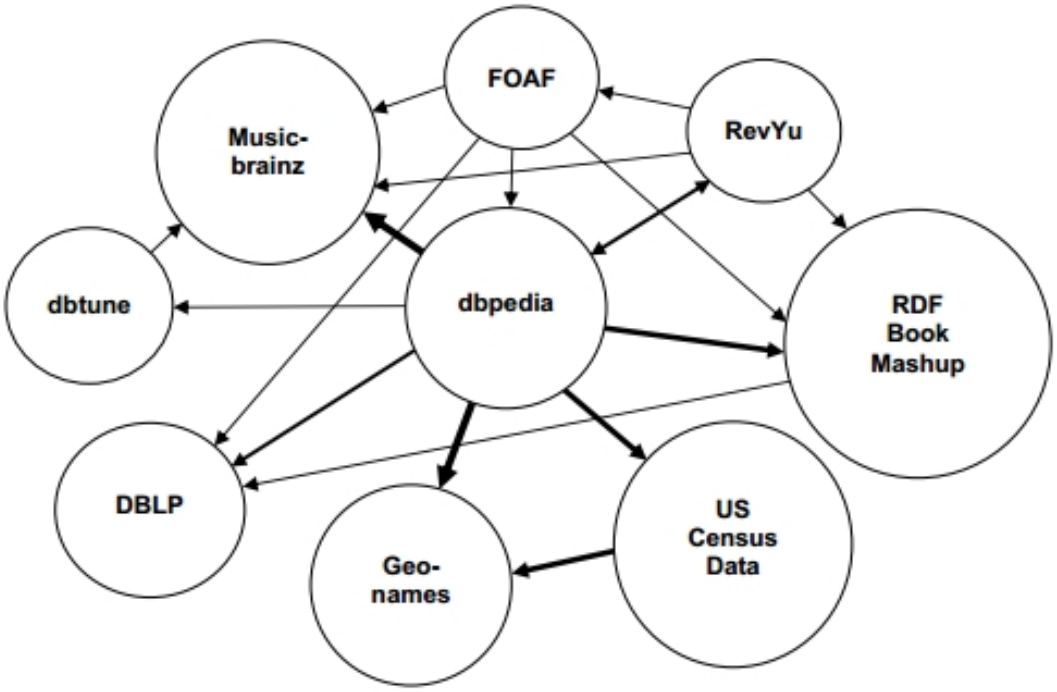



Figura 1.13: Grafo do projeto LOD em sua versão de 2007. Os nós representam as bases de dados em formado RDF e as arestas são ligações entre as bases (Fonte: Bizer *et al.* (2007)).

- O URI precisa expressar informação;
- Incluir ligações para outros URI para que novas informações possam ser descobertas.

Este modelo proposto é flexível, permitindo interoperabilidade entre as bases e extensibilidade da nuvem (Hors *et al.*, 2012). Embora não seja uma premissa para o LOD, o uso de vocabulários e ontologias é fortemente recomendado e de grande importância para a ligação dos dados, sendo definida como um dos principais passos para a publicação de dados no LOD (Bauer & Kaltenböck, 2012). Já existem diversos vocabulários e ontologias para várias áreas do conhecimento, tais como: FOAF[18], SIOC[19], SKOS[20], etc. A escolha pela reutilização de algum vocabulário ou ontolo-

[18]Vocabulário para descrever projetos, organizações e pessoas (`http://xmlns.com/foaf/spec/`).

[19]Ontologia que provê conceitos e propriedades para descrever informações sobre comunidades *online* (`http://rdfs.org/sioc/spec/`).

[20]Vocabulário para expressar sistemas de ogranização, tais como taxonomias, tesauros, etc. (`http://www.w3.org/TR/swbp-skos-core-spec`).

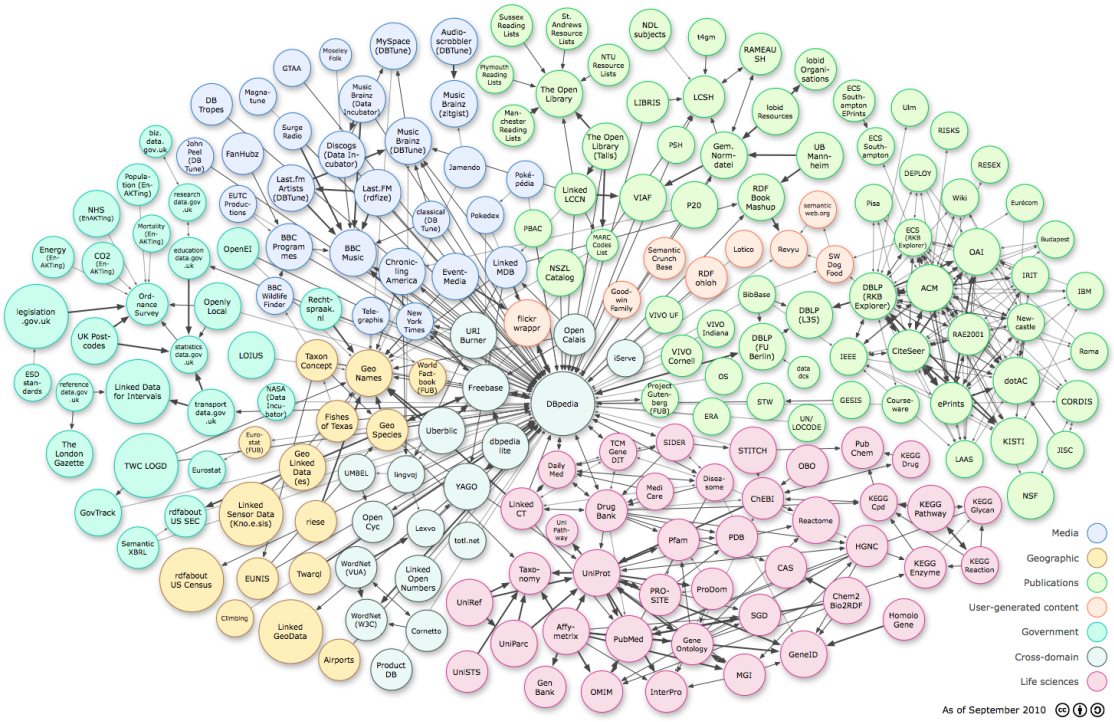



Figura 1.14: Nuvem do LOD em 2010. Os conjuntos de dados na cor rosa são dados da área das Ciências da Vida (Fonte: `http://richard.cyganiak.de/2007/10/lod/`).

gia pode garantir um número maior de ligações com outras bases de conhecimento também disponíveis na nuvem do LOD.

Os dados em RDF para serem disponibilizados na Web precisam de um servidor com suporte aos protocolos HTTP e SPARQL, o que é denominado pela comunidade computacional de *EndPoint*, isto é, um endereço web onde é possível utilizar a linguagem SPARQL para recuperar informações em formato de triplas.

### 1.6.8 O uso de dados ligados na biologia

Diversos trabalhos demonstram as vantagens de se utilizar semântica para integrar dados biológicos (Neumann *et al.*, 2004; Neumann, 2005; Pasquier, 2008; Chen *et al.*, 2010), cujo grande desafio atual é extrair informações de bases com enorme volume de dados gerados pelos novos equipamentos de sequenciamento.

Apesar de diversas bases biológicas estarem no formato relacional, a semântica só se expressa no nível de esquema (tabelas, colunas, etc.), sendo fundamental enriquecer a descrição desses dados para que esquemas heterogêneos possam ser integrados ou inter-relacionados (Cuadra *et al.*, 2011).

Com a finalidade de enriquecer semanticamente as bases biológicas foram criados diversos vocabulários e ontologias por diversos grupos de pesquisa diferentes. O Gene Ontology (GO) (Ashburner *et al.*, 2000) é uma iniciativa de se utilizar ontologia para anotação funcional de genes e proteínas. Outra iniciativa é a do BioPortal (Musen *et al.*, 2012)[21] que reune mais de 270 ontologias e vocabulários controlados da área biomédica, incluindo o próprio GO e o tesauros[22] do NCI (National Cancer Institute).

Bases de dados enriquecidas semanticamente podem ser armazenadas de diversas maneiras, entretanto, ainda há uma preferência pelo modelo relacional (Hausenblas & Karnstedt, 2010). Esquemas de armazenamento de triplas expressos em modelos relacionais podem ser eficientes, agregando os benefícios de ambas as tecnologias.

O LOD pode ser analisado como uma grande base de dados na Web, como proposto por Hausenblas & Karnstedt (2010). Desta forma, é possível realizar consultas em diversas bases do LOD, obtendo informações de diferentes origens. Como cada informação é expressa em um URI, é possível navegar por essas ligações (*links*), partindo de um domínio da informação para outro domínio. Para a biologia isso se torna fundamental pois, muitas vezes, as informações estão disponíveis em diversas áreas das ciências da vida.

Para Slater *et al.* (2008) não há dúvidas sobre os benefícios que a integração de informações biológicas pode trazer para o processo de descoberta de fármacos, pois

[21] http://bioportal.bioontology.org/
[22] Tesauro é uma compilação do léxico de uma língua ou de uma área do saber.

a indústria de biotecnologia necessita de várias disciplinas especializadas trabalhando juntas (interdisciplinariedade), não sendo possível para o processo de P&D separar os problemas para serem resolvidos independentemente.

Embora na nuvem do LOD existam diversas bases de dados biológicas no contexto das ciências da vida, poucos grupos de pesquisa converteram suas bases em triplas RDF segundo as premissas do LOD e disponibilizam esses dados em um endereço de internet para consultas SPARQL. O Drugbank (Knox *et al.*, 2011), por exemplo, disponibiliza seus dados em um *endpoint*[23], enquanto que o Uniprot (The UniProt Consortium, 2012) disponibiliza em sua página a sua base de dados convertida para triplas RDF. Outras bases de dados biológicas disponibilizadas na nuvem do LOD (PDB, Gene Ontology, Kegg Pathway, Kegg Drug, HomoloGene, etc) foram disponibilizadas pelo projeto Bio2RDF (Belleau *et al.*, 2008) que utilizou uma ontologia e um vocabulário próprios para converter dados públicos em triplas de informação.

## 1.7 Perguntas Biológicas e termos computacionalmente complexos

Responder a determinadas perguntas biológicas pode ser uma tarefa computacionalmente difícil e que pode requerer grande capacidade de processamento. As questões biológicas apresentam um alto nível de complexidade computacional e de riqueza semântica. Questões que são aparentemente fáceis para o entendimento humano são de alta complexidade computacional e precisam ser tratadas de forma similar à perspectiva da Teoria Geral de Sistemas (TGS) proposta por Ludwig von Bertalanffy no século passado, que preconiza, entre outras, a divisão de um grande sistema em vários sistemas menores. Analogamente, uma questão biológica complexa computacional-

[23] `http://www4.wiwiss.fu-berlin.de/drugbank/`

mente pode ser dividida em várias questões menos complexas computacionalmente e processadas ao mesmo tempo para responder a questão maior (Figura 1.15).

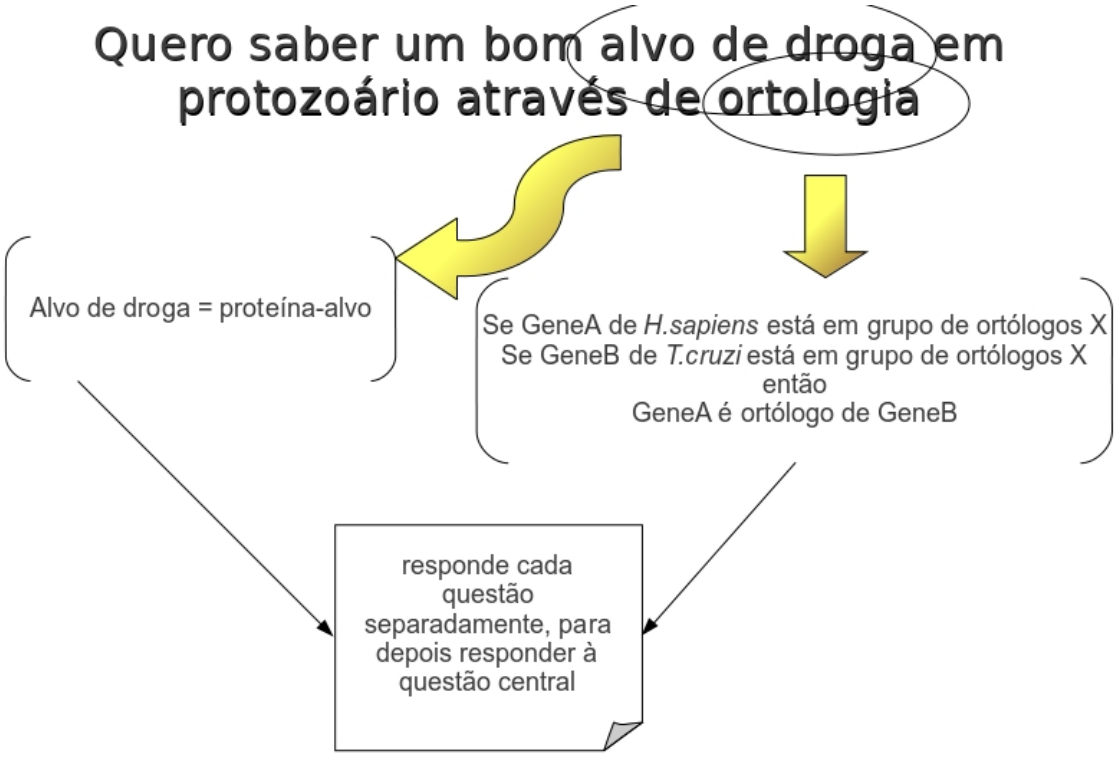



Figura 1.15: Exemplo da aplicação da TGS em uma pergunta biológica complexa.

Neste trabalho é proposta a integração de bases de dados biológicas, com domínios específicos da informação, através de semântica para integrar à nuvem do LOD. Uma vez disponibilizadas em formatos RDF, essas bases podem ser interoperadas para a busca de genes/proteínas, utilizando-se o conceito de homologia, que possam ser utilizados como alvos para fármacos já comercializados.

# Capítulo 2

# Objetivos

## 2.1 Objetivo Geral

Estudar o reposicionamento de fármacos para tratamento de doenças negligenciadas causadas por protozoários no contexto da biologia computacional, utilizando o conceito de homologia.

## 2.2 Objetivos Específicos

- Verificar a viabilidade da utilização do conceito de homologia para propor fármacos que possam ser reposicionados para tratamento de doenças causadas por protozoários;
- Propor uma lista de fármacos que possuam potencial para serem reposicionados para tratamento de doenças negligenciadas causadas por protozoários;
- Avaliar os fenótipos associados às proteínas de protozoários, através da integração de bases de dados e utilizando o conceito de homologia;
- Avaliar as vias metabólicas que possuem proteínas de protozoários que são ortólogas a alvos de fármacos;

- Avaliar a alternativa do uso de semântica para integração de bases de dados biológicas;
- Propor uma solução de conversão de dados em modelos relacionais para o padrão RDF;
- Disponibilizar as informações das bases de dados integradas para consultas, de forma pública;
- Disponibilizar a base de dados do ProtozoaDB em padrão RDF, segundo as premissas do LOD, para a publicação na nuvem do LOD.

# Capítulo 3

# Materiais e Métodos

## 3.1 Abordagem de ortologia para reposicionamento de fármacos

O trabalho foi realizado utilizando os conceitos de homologia para buscar fármacos que pudessem ser reutilizados para tratamento de doenças causadas por protozoários.

A Figura 3.1 representa uma visão biológica minimalista da solução adotada, permitindo que o problema complexo de reposicionamento de fármacos fosse ser apoiado computacionalmente.

## 3.2 Arquitetura da Solução

A solução adotada pode ser esquematizada por camadas, onde cada uma possui o foco em um problema específico (Figura 3.2). A primeira camada foi denominada de camada de bases externas, onde são obtidas as informações das bases de dados públicas que foram baixadas via internet ou conseguidas através de serviços web

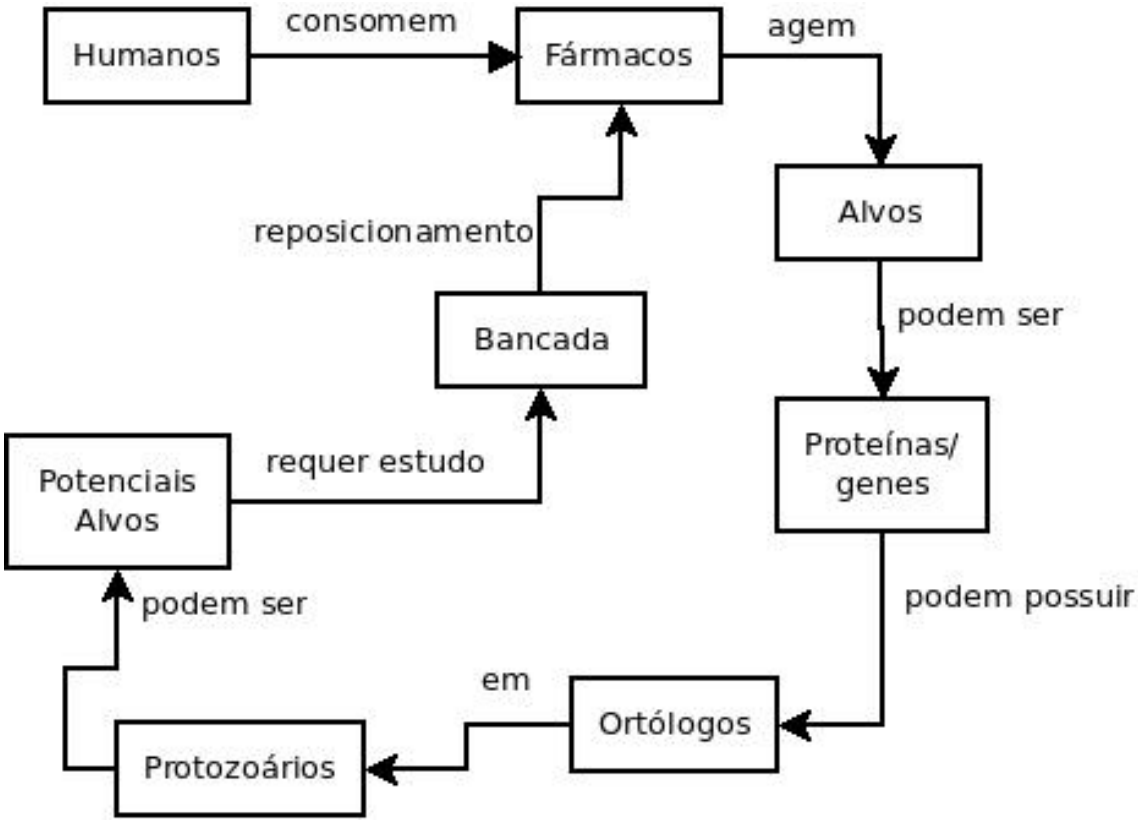



Figura 3.1: Esquema da solução apresentada na tese utilizando a abordagem de ortologia para tentar solucionar o problema complexo de reposicionamento de fármacos.

disponibilizados. A segunda camada foi denominada camada do servidor e é responsável pela carga dos dados no modelo relacional, a conversão para o padrão RDF e a disponibilização dos dados para consultas com tratamento para perguntas complexas. A terceira camada foi denominada de camada cliente e permite consultas SPARQL locais ou remotas nas bases de dados convertidas e integradas.

## 3.3 Bases de Dados utilizadas

Foram selecionadas oito bases de dados biológicas públicas (Tabela 3.1) que foram baixadas diretamente da internet ou através da utilização de Serviços Web. As bases de dados que foram selecionadas representam um determinado domínio da informação biológica, contribuindo, cada uma, com a solução de uma parte do problema.

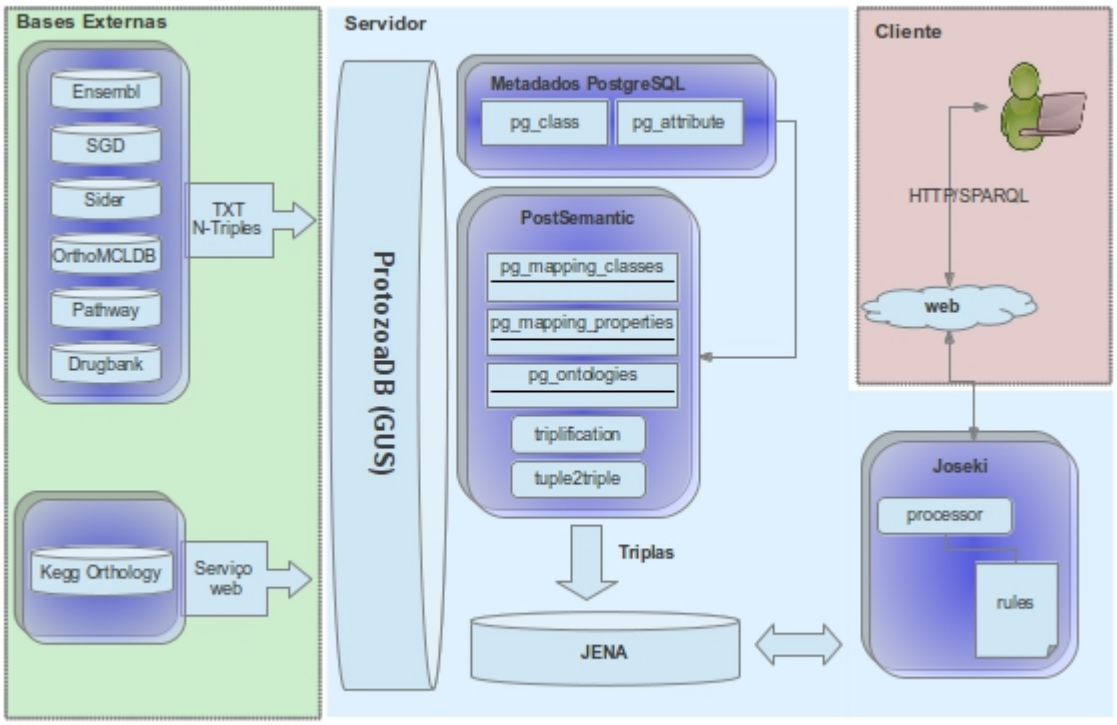



Figura 3.2: Arquitetura em camadas da solução para o problema de reposicionamento de fármacos. O diagrama contempla a carga das bases externas (na cor verde), o processamento de conversão de tuplas para triplas (na cor azul) e o processamento de consultas complexas pelos usuários (na cor rosa).

Tabela 3.1: Bases de Dados escolhidas para integração

| Domínio | Base de Dados | Conteúdo | Formato Tamanho | Endereço Internet | Versão |
|---|---|---|---|---|---|
| Protozoários | ProtozoaDB | Genoma e proteoma de protozoários | Relacional (64Gb) | http://protozoadb.biowebdb.org/home/download | Outubro de 2010 (release 180) |
| Fármacos | Drugbank | Fármacos e alvos de fármacos | RDF (145Mb) | http://www4.wiwiss.fu-berlin.de/drugbank/drugbank_dump.nt.bz2 | Agosto de 2010 |
| | Sider | Efeitos adversos | Texto (2.3Mb) | http://sideeffects.embl.de/media/download/meddra_adverse_effects.tsv.gz | Novembro de 2012 |
| Homologia | OrthoMCLDB | Grupos de ortólogos | Texto (5.8Mb) | http://orthomcl.org/common/downloads/release-5/groups_OrthoMCL-5.txt.gz | Versão 5 |
| | Kegg Orthology | Grupos de ortólogos | Texto (5.9Mb) | http://kineto2.biowebdb.org/services/phenotypes/mappingGI_KO.dat | Maio de 2012 |
| Vias metabólicas | Kegg Pathway | Mapas de vias metabólicas | Texto (169Kb) | http://www.biowebdb.org/pub/kegg/pathway/ko/ko_map.tab | Abril de 2011 |

*Continua na próxima página*

Tabela 3.1 – *Continuação da página anterior*

| **Domínio** | **Base de Dados** | **Conteúdo** | **Formato Tamanho** | **Endereço Internet** | **Versão** |
|---|---|---|---|---|---|
| Proteoma humano | Ensembl | Proteoma humano | Texto (25Mb) | `ftp://ftp.ensembl.org/pub/release-69/gtf/homo_sapiens/Homo_sapiens.GRCh37.69.gtf.gz` | Outubro de 2012 |
| Organismos modelos | SGD | Genoma e proteoma de Saccharomyces cerevisiae | Texto (14Mb) | `http://downloads.yeastgenome.org/curation/literature/phenotype_data.tab` | Novembro de 2012 |

## 3.4 Estratégia de conversão das bases de dados

As bases de dados em formato texto foram carregadas para um modelo relacional através de programas específicos (*wrappers*) e armazenadas juntamente com as informações do ProtozoaDB. Essas bases foram então convertidas pelo módulo desenvolvido nesta tese denominado PostSemantic, para o formato N-Triples. A base de dados em formato N-Triples foi carregada diretamente para a base, por não precisar tratamento prévio (Figura 3.3).

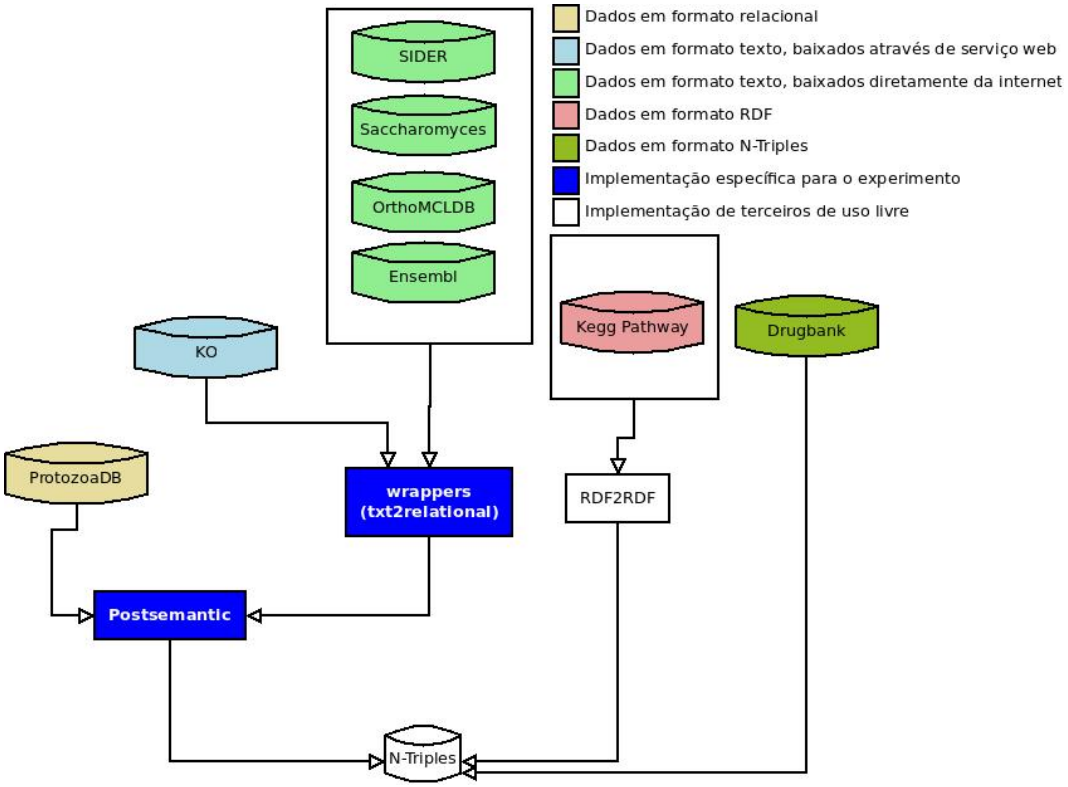



Figura 3.3: Diagrama com o esquema utilizado para a conversão de cada base de dados em formato N-Triples.

### 3.4.1 Escolha dos termos das URI

Para a escolha das URI adotadas em cada domínio da informação afim de converter os dados para triplas, foram adotados os seguintes critérios:

- Preferencialmente adotar o próprio termo do modelo relacional ou da fonte de origem;
- Utilizar o mesmo termo para os campos identificados como chave estrangeira;
- A URI deverá identificar a proveniência dos dados.

### 3.4.2 Dominio de conhecimento - Protozoários

A base de dados ProtozaDB (Dávila *et al.*, 2008), criada no contexto do Consórcio Biowebdb[1], armazena dados de genoma e proteoma de 22 protozoários patogênicos. Informações sobre homologia entre os organismos também estão disponíveis e foram utilizadas neste trabalho. A base de dados possui cerca de 64Gb[2], com 218.100 proteínas armazenadas em um modelo relacional com esquema GUS. Sete tabelas do esquema foram selecionadas para serem convertidas para o formato N-Triples: **nafeatureimp**, com informações sobre nucleotídeos; **nasequeceimp**, que armazena as sequências de nucleotídeo; **aafeatureimp**, que possui informações sobre aminoácidos; **aasequeceimp**, com sequências de aminoácidos; **dbrefnafeature** e **dbref**, com os identificadores do Genbank (Benson *et al.*, 2011); **nagene**, com o identificador do gene; e, **taxonname**, com dados sobre taxonomia dos organismos.

Os dados utilizados nesta tese foram gerados através do resultado de uma consulta em Linguagem de Consulta Estruturada (*Structured Query Language* - SQL) englobando essas tabelas (Apêndice A). Como resultado, foram retornados 218.100 registros. Para cada campo foi atribuído um URI (Tabela 3.2) para a conversão em triplas.

[1] http://biowebdb.org
[2] Informação acessada em setembro de 2012.

Tabela 3.2: Campos escolhidos do domínio de protozoários para a conversão em triplas com os respectivos URI associados.

| **Campo** | **Informação** | **Tipo** | **URI** |
|---|---|---|---|
| aa_sequence_id | Identificador da proteína | Classe | `http://biowebdb.org/protozoadb/aa_sequence_id` |
| na_feature_id | Identificador do gene | Classe | `http://biowebdb.org/protozoadb/na_feature_id` |
| source_id | Número de acesso da proteína | Classe | `http://biowebdb.org/protozoadb/accession_number` |
| description | Anotação Funcional da proteína | Propriedade | `http://biowebdb.org/protozoadb/annotation` |
| gene | Número de acesso do gene | Classe | `http://biowebdb.org/protozoadb/accession_number` |
| taxonname | Descrição da taxonomia | Propriedade | `http://biowebdb.org/protozoadb/taxonname` |
| gi | Indentificador da proteína no Genbank | Propriedade | `http://biowebdb.org/protozoadb/gi` |

### 3.4.3 Dominio de Conhecimento - Fármacos

A base de dados do Drugbank (Knox *et al.*, 2011) contém 6.711 fármacos e 4.227 proteínas distribuídas em cartões, denominados *DrugCard*, com mais de 150 campos[3]. Está disponível para consulta online e para baixar em diversos formatos de arquivos, inclusive em formato N-Triples e, por essa razão, foi carregada diretamente para a base, sem necessidade de conversão.

A base de dados do SIDER (Kuhn *et al.*, 2010) armazena informações sobre efeitos adversos de fármacos comercializados. A informação foi baixada em arquivo do tipo texto, em formato tabular e para ser convertida para o formato N-Triples foi necessário convertê-la inicialmente para uma tabela do esquema relacional do GUS, através de um programa (Apêndice B), para posterior conversão em formato RDF/N-Triples. Os campos dessa tabela foram associados à URI para o processo de conversão para triplas (Tabela 3.3).

Tabela 3.3: Campos escolhidos do domínio de fármacos para a conversão em triplas com os respectivos URI associados.

| **Campo** | **Informação** | **Tipo** | **URI** |
|---|---|---|---|
| conceptid | Identificador do efeito colateral | Classe | `http://sideeffects.embl.de/conceptid` |
| drugname | Nome do fármaco | Propriedade | `http://sideeffects.embl.de/genericName` |
| side_effect | Efeito Colateral | Propriedade | `http://sideeffects.embl.de/side_effect` |

### 3.4.4 Domínio de conhecimento - Homologia

Duas bases de dados foram selecionadas para prover a informação sobre ortologias: OrthoMCLDB (Chen *et al.*, 2006) e KEGG Orthology (KO) (Goto *et al.*, 1997).

[3]Informação acessada em setembro de 2012.

A base de dados do OrthoMCLDB está em sua versão (*release*) de número 5 e contém 150 genomas com 124.740 grupos de ortólogos. Está disponibilizada em arquivo do tipo texto em formato tabular. Foi necessário converter o formato tabular para uma tabela em um esquema relacional e depois convertê-la para N-Triples, através de um programa (Apêndice B), conforme os URI selecionados para cada coluna (Tabela 3.4).

Tabela 3.4: Campos escolhidos do domínio de homologia (OrthoMCLDB) para a conversão em triplas com os respectivos URI associados.

| **Campo** | **Informação** | **Tipo** | **URI** |
|---|---|---|---|
| orthologous | Identificador do grupo de ortólogo | Classe | `http://www.orthomcl.org/orthologous_group` |
| organism | Sigla do organismo | Propriedade | `http://www.orthomcl.org/organism` |
| accession_number | Número de acesso da proteína | Propriedade | `http://www.orthomcl.org/accession_number` |

O KEGG Orthology (KO) possui dados de ortologia de genes de diversos organismos, incluindo organismos modelos. O Instituto KEGG suspendeu o serviço de disponibilização de suas bases de dados através de arquivos para serem baixados diretamente da Internet. Contudo, ainda é possível acessar e recuperar seus dados através de Serviços Web[4]. Desta forma os identificadores das proteínas dos 22 protozoários do ProtozoaDB foram mapeados com os grupos de ortólogos do KO, através de um programa escrito para esta tese (Apêndice B). Esse mapeamento gerou um arquivo do tipo texto, no formato tabular. De forma similar ao OrthoMCLDB, uma tabela foi criada em um esquema relacional e os dados foram armazenados nessa tabela, através de um programa (Apêndice B), antes de serem convertidos para o formato N-Triples, de acordo com os URI escolhidos para essa base (Tabela 3.5).

[4]http://www.genome.jp/kegg/soap/, acessado em setembro de 2012.

Tabela 3.5: Campos escolhidos do domínio de homologia (KO) para a conversão em triplas com os respectivos URI associados.

| Campo | Informação | Tipo | URI |
|---|---|---|---|
| keggid | Identificador do organismo | Classe | `http://www.genome.jp/kegg/keggid` |
| ko | Identificador do grupo de ortólogo | Classe | `http://www.genome.jp/kegg/ko` |
| genbankidprotein | Identificador da proteína no Genbank | Propriedade | `http://www.genome.jp/kegg/gi` |

### 3.4.5 Domínio do conhecimento - Vias Metabólicas

As vias metabólicas foram obtidas em formato texto e armazenadas em uma tabela em um esquema relacional, através de um programa (Apêndice B). Foram associados URI para as colunas antes do processo de conversão para triplas (Tabela 3.6).

Tabela 3.6: Campos escolhidos do domínio de vias metabólicas para a conversão em triplas com os respectivos URI associados.

| Campo | Informação | Tipo | URI |
|---|---|---|---|
| ko | Identificador do grupo de ortólogo | Classe | `http://www.genome.jp/kegg/ko` |
| mapid | Identificador do mapa | Classe | `http://www.genome.jp/kegg/mapid` |
| description | Descrição do mapa | Propriedade | `http://www.genome.jp/kegg/description` |

### 3.4.6 Domínio do conhecimento - Proteoma humano

A base de dados do Ensembl (Flicek *et al.*, 2011) contém informações sobre o genoma e proteoma humano e está disponível para baixar em arquivos com formato texto. Foi incluído em um esquema relacional (Apêndice B) e convertido para o formato N-Triples, de acordo com os URI escolhidos (Tabela 3.7).

Tabela 3.7: Campos escolhidos do domínio de proteoma humano para a conversão em triplas com os respectivos URI associados.

| Campo | Informação | Tipo | URI |
|---|---|---|---|
| accession_number | Número de acesso do ENSEMBL | Classe | `http://www.ensembl.org/accession_number` |
| gene_name | Número de acesso da proteína | Classe | `http://www.ncbi.nlm.nih.gov/accession_number` |

### 3.4.7 Domínio do conhecimento - Organismos modelos

O projeto Saccharomyces Genome Database (SGD) (Cherry *et al.*, 2012) possui informações sobre o genoma e proteoma do organismo modelo *Saccharomyces cerevisiae*, incluindo informação sobre fenótipos. A base de dados está disponível em arquivo texto em formato tabular. Após baixar a base de dados, as informações foram armazenadas em uma tabela em um esquema relacional (Apêndice B), antes de serem convertidas para o formato N-Triples, de acordo com os URI escolhidos (Tabela 3.8).

Tabela 3.8: Campos escolhidos do domínio de organismos modelos para a conversão em triplas com os respectivos URI associados.

| Campo | Informação | Tipo | URI |
|---|---|---|---|
| accession_number | Número de acesso da proteína | Classe | `http://www.yeastgenome.org/accession_number` |
| phenotype | Descrição do fenótipo | Propriedade | `http://www.yeastgenome.org/phenotype` |

## 3.5 Sistema Gerenciador de Banco de Dados

O SGBD PostgreSQL v9.1[5] foi utilizado para armazenar as bases de dados em formato N-Triples e o esquema GUS (Davidson *et al.*, 2000)[6]. As bases de dados em

[5]http://www.postgresql.org/
[6]http://gusdb.org

formato texto que foram inicialmente convertidas para o formato relacional também foram armazenadas no SGBD PostgreSQL.

## 3.6 Modelo de dados para armazenamento no padrão RDF

O esquema relacional do projeto Apache JENA(TM)[7] (Figura 3.4) foi escolhido para armazenar os dados das bases em formato N-Triples, sendo composto por quatro tabelas que armazenam grafos, nós, triplas e prefixos em um modelo relacional.

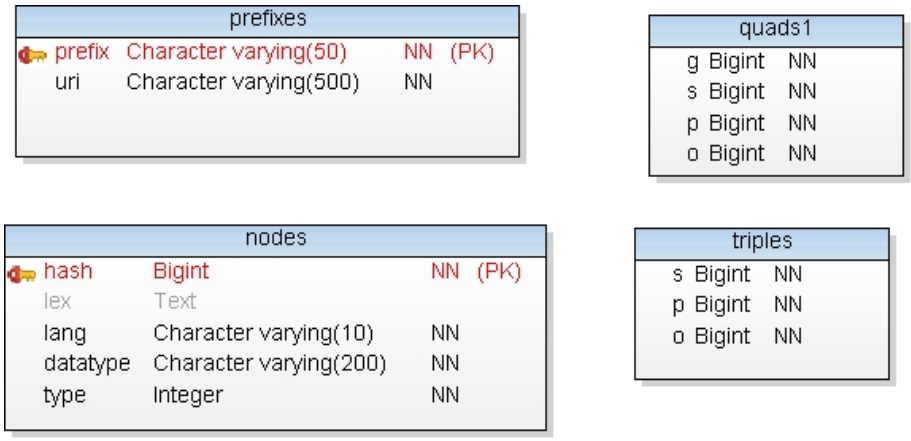



Figura 3.4: Modelo de dados do projeto Apache JENA

O campo chave primária da tabela **nodes**, chamado de *hash*, é um campo inteiro que armazena um valor baseado em MD5[8]. A tabela **nodes** armazena os dados que formam as triplas e a tabela **triples** armazena o valor do *hash* de cada dado. Esse esquema garante que o dado esteja armazenado apenas uma única vez na tabela **nodes** mas que possa ser utilizado diversas vezes na tabela **triples**.

[7] http://jena.apache.org/index.html

[8] O MD5 (Message-Digest algorithm 5) é um algoritmo de hash de 128 bits unidirecional desenvolvido pela RSA Data Security, Inc., descrito na RFC 1321 (Fonte: http://pt.wikipedia.org/wiki/MD5, acessado em setembro de 2012).

De forma semelhante, a tabela **prefixes** pode armazenar os prefixos dos URI (exemplo: bio:http://biowebdb.org) uma única vez e esse prefixo (bio) pode ser utilizado diversas vezes na tabela **nodes** em substituição ao URI original.

## 3.7 Conversão das bases relacionais para o formato RDF

### 3.7.1 Visão Geral

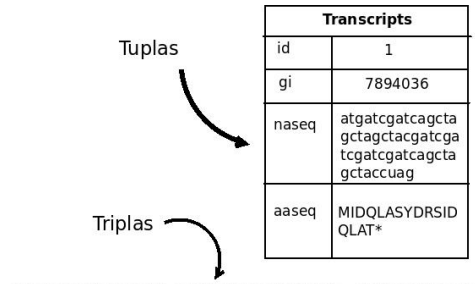
Um modelo relacional é representado por tabelas que são compostas por colunas (tabela) e linhas (campos). As informações armazenadas neste formato são chamadas de **tuplas**.

Uma **tripla** é composta de sujeito (s), predicado (p) e objeto (o). Para a conversão de tuplas em triplas, seguimos os seguintes requisitos (Berners-Lee, 1998)(Figura 3.5):

- Toda tabela é considerada um sujeito (classe) e toda chave primária sequencial também faz parte do sujeito;
- Toda coluna é considerada um predicado (propriedade);
- Toda tupla é considerada um objeto, que pode ser do tipo literal ou classe.

### 3.7.2 Diagrama de conversão do formato relacional para o formato RDF

O modelo de dados do catálogo do SGBD PostgreSQL foi estendido, criando novas tabelas que armazenaram a informação das tabelas a serem convertidas (Figura 3.6). Diversos programas de computador foram escritos para serem executados diretamente de dentro do SGBD PostgreSQL com o objetivo de alimentar as tabelas




**<http://biowebdb.org/transcripts/1> <http://biowebdb.org/gi> "7894036"**

**<http://biowebdb.org/transcripts/1> <http://biowebdb.org/naseq> "atgatcgatcagctagctagctacgatcgatcgatcgatcagctagctaccuag"**

**<http://biowebdb.org/transcripts/1> <http://biowebdb.org/aaseq> "MIDQLASYDRSIDQLAT*"**

Figura 3.5: Exemplo de conversão de Tupla para Tripla.

de conversão. Os programas foram escritos em linguagens procedurais, próprias do SGBD (PL/PGSQL, PL/Ruby e PL/Java).

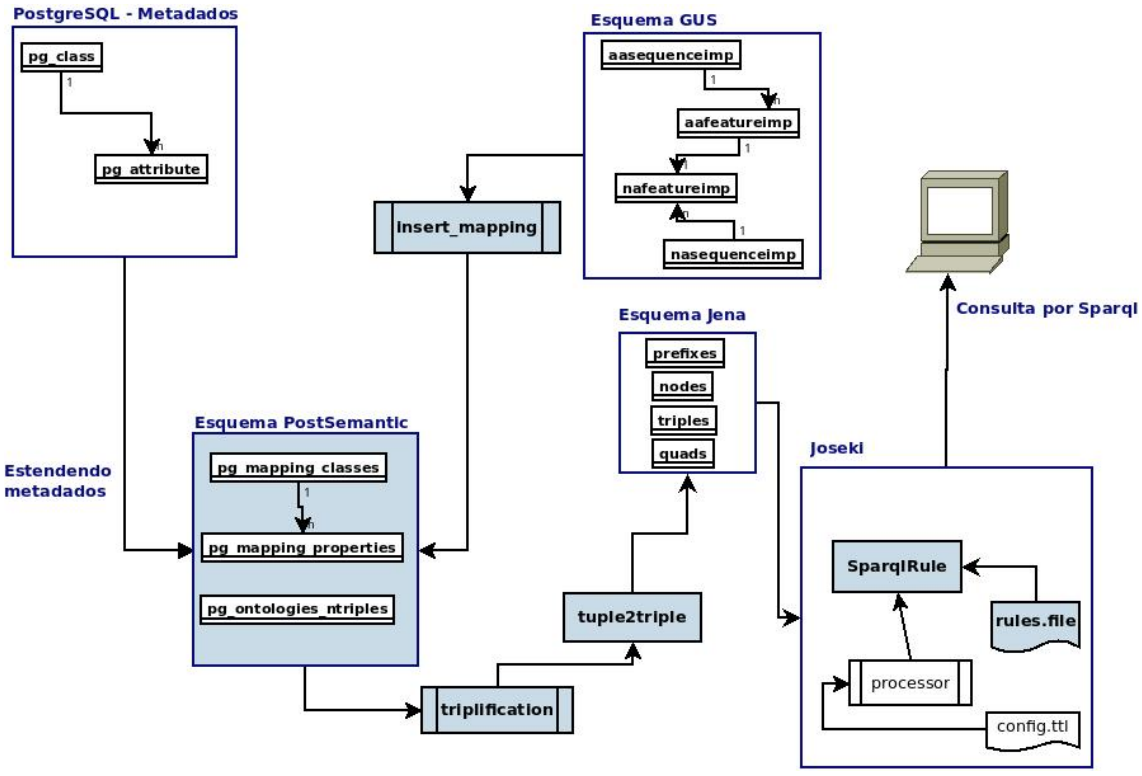



Figura 3.6: Diagrama para a conversão de dados relacionais para N-Triples e disponibilização de consultas via SPARQL.

#### 3.7.2.1 Modelo de Dados

As tabelas do projeto PostSemantic foram estendidas do metadados do SGBD PostgreSQL para permitir o armazenamento das informações sobre a conversão do formato relacional para o padrão RDF, no formato N-Triples. Três tabelas foram criadas (Figura 3.6 e Figura 3.7): **pg_mapping_classes**, **pg_mapping_properties** e **pg_ontologies_ntriples**. A tabela **pg_mapping_classes** armazena informações de tabelas do esquema do GUS que foram convertidas para o formato N-Triples. A tabela **pg_mapping_properties** contém informações das colunas de cada tabela que foi convertida e a tabela **pg_ontologies_ntriples** contém ontologias em formato N-Triples.

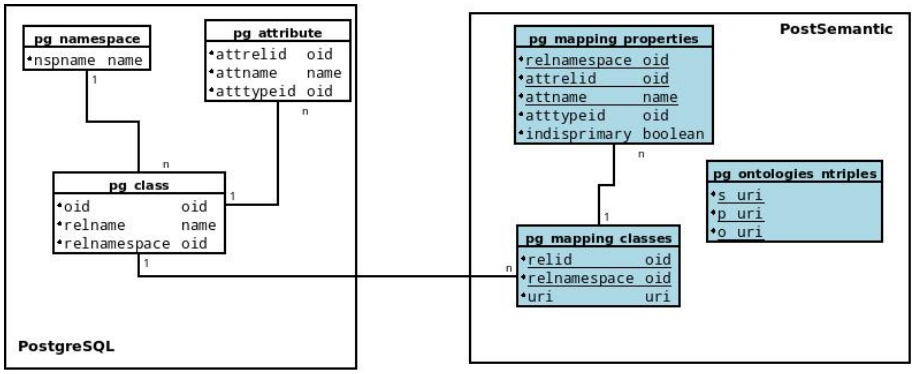



Figura 3.7: Metadados do PostgreSQL estendido para adequar à solução de triplificação de bases relacionais. As tabelas com fundo em azul claro foram implementadas neste projeto.

#### 3.7.2.2 Função para mapear classes e propriedades

A função denominada **insert_mapping_classes** é responsável por inserir dados nas tabelas **pg_mapping_classes** e **pg_mapping_properties**. Seu código-fonte (Apêndice B) foi escrito em linguagem PL/PGSQL e recebe três parâmetros: nome do esquema, nome da tabela a ser convertida e URI.

#### 3.7.2.3 Função para criar triplas

A função **triplification** realiza duas operações: criação das triplas mapeadas pela função **insert_mapping_classes** e execução da função **tuple2triple** que irá converter as tuplas em triplas. Foi escrita em PL/PGSQL (Apêndice B) e não possui parâmetro.

#### 3.7.2.4 Função para converter tuplas em triplas

A conversão de tuplas para triplas é realizada pela função **tuple2triple** escrita em PL/Ruby (Apêndice B). Recebe como parâmetro o nome da tabela e insere registros no esquema relacional do projeto JENA.

#### 3.7.2.5 Função Hash

A função **hash** foi escrita em linguagem PL/Java (Apêndice B) e é responsável por calcular o campo hash (baseado em MD5) da tabela nodes do projeto JENA. Recebe como parâmetros todos os campos da tabela **nodes**, exceto o campo hash.

#### 3.7.2.6 Função para inserir nós

A função **insert_node** verifica a existência ou não de um nó na tabela nodes. Caso o nó já exista, a função apenas retorna o identificador do nó (campo hash). Caso contrário, a função insere o nó no modelo de dados do projeto JENA e retorna o identificador do nó. O código-fonte dessa função foi escrito em PL/PGSQL (Apêndice B) e recebe como parâmetros todos os campos da tabela nodes, exceto o campo hash.

## 3.8 Disponibilização dos dados para consultas

As tabelas do projeto JENA foram acessadas através de consultas SPARQL realizadas em um servidor web. Foi utilizado o servidor web Apache Tomcat 6.0 [9] com JOSEKI [10] que possui suporte ao protocolo SPARQL.

Foi criado um programa para traduzir termos computacionalmente complexos em consultas SPARQL de forma a permitir a quebra de uma pergunta biológica complexa em outras perguntas menores (Figura 3.8).

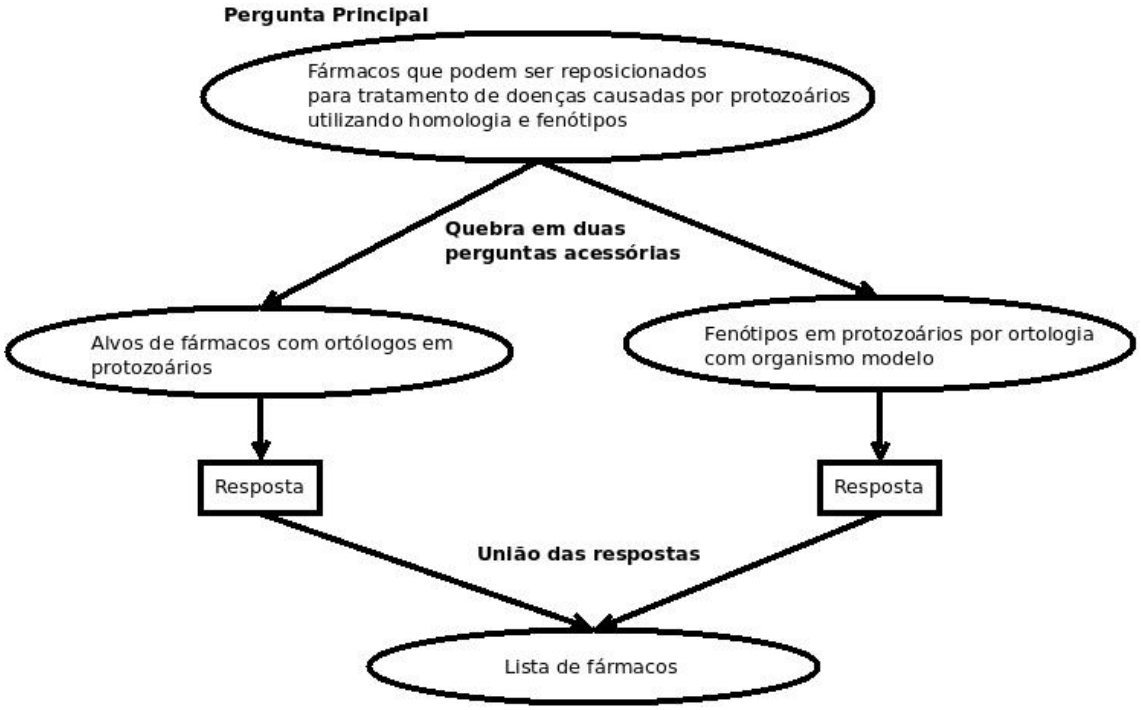



Figura 3.8: Esquema da pergunta central do projeto realizada através de consulta SPARQL às bases convertidas para o modelo de triplas. A pergunta principal é quebrada em duas outras perguntas que são processadas separadamente. As respostas dessas duas perguntas são unificadas para responder a pergunta principal.

A classe do núcleo do processamento do JOSEKI (classe constructor implementada pelo programa org.joseki.SPARQL) foi alterada para permitir a tradução, dinami-

[9] http://tomcat.apache.org
[10] http://joseki.sourceforge.net/

camente, de termos complexos durante uma consulta SPARQL (Figura 3.9). Esses termos complexos ficam armazenados em um arquivo do tipo texto no servidor e são traduzidos quando incluídos em alguma consulta SPARQL (Figura 3.10).

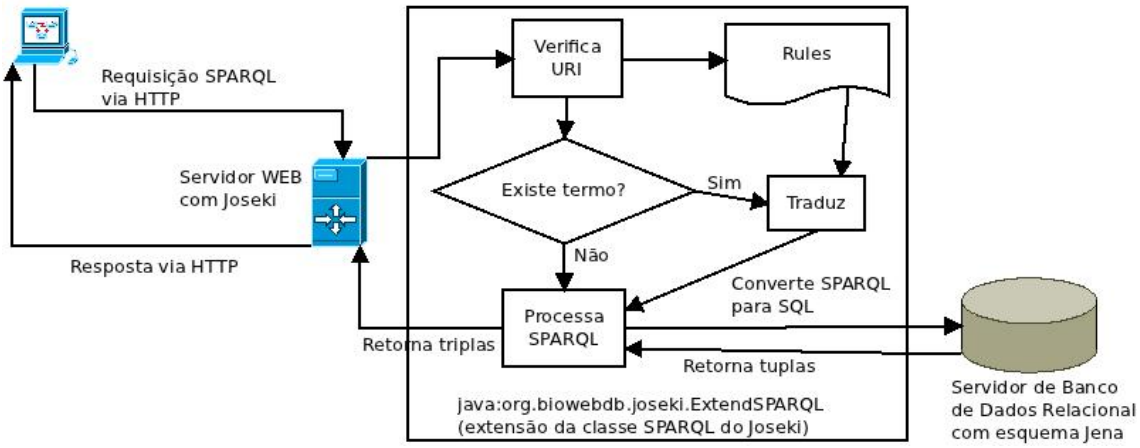



Figura 3.9: Esquema de tradução dinâmica de termos complexos. Após a consulta realizada, um processo verifica se algum termo da consulta SPARQL encontra-se no arquivo **Rules**. Caso exista, esse termo é traduzido e processado antes do processamento da consulta inicial. Após o processamento do termo, é realizado o processamento da consulta principal, retornando o resultado em formato de triplas.

```
## Property renaming
http://biowebdb.org/protozoadb/reuseTo<-CONSTRUCT {?farmacoId <http://biowebdb.org/protozoadb/reuseTo>
?protozoa}WHERE{ ?protozoa <http://biowebdb.org/protozoadb/accession_number> ?an .    ?an
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.orthomcl.org/> .    ?protozoa
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://biowebdb.org/protozoadb/> .    ?protozoa
<http://biowebdb.org/protozoadb/gi> ?gi .    ?an <http://www.orthomcl.org/gi> ?gi .    ?grupo
<http://biowebdb.org/protozoadb/accession_number> ?an .    ?grupo <http://www.orthomcl.org/orthologous_group>
?string_grupo .    ?grupo <http://www.ensembl.org/accession_number> ?codigoensembl .    ?ncbiAccession
<http://www.ensembl.org/accession_number> ?codigoensembl .    ?ncbiAccession
<http://www.ncbi.nlm.nih.gov/accession_number> ?accession_number_string .    ?alvo
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/geneName> ?accession_number_string .    ?farmacoId
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/target> ?alvo .    ?farmacoId <http:
//www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/genericName> ?farmaco .    ?protozoa
<http://biowebdb.org/protozoadb/taxonname> ?taxon  }

http://biowebdb.org/protozoadb/phenotype<-CONSTRUCT {?protozoa <http://biowebdb.org/protozoadb/phenotype>
?phenotype} WHERE { ?protozoa <http://biowebdb.org/protozoadb/gi> ?gi . ?ko <http://www.genome.jp/kegg/gi> ?gi
. ?ko <http://www.yeastgenome.org/accession_number> ?yeast . ?yeast <http://www.yeastgenome.org/phenotype>
?phenotype}

http://biowebdb.org/protozoadb/pathway<-CONSTRUCT { ?protozoa <http://biowebdb.org/protozoadb/pathway>
?mapid } WHERE { ?protozoa <http://biowebdb.org/protozoadb/gi> ?gi . ?ko <http://www.genome.jp/kegg/gi> ?gi .
?ko <http://www.genome.jp/kegg/mapid> ?mapid}
```

Figura 3.10: Exemplo de termo complexo e sua tradução para consulta SPARQL.

## 3.9 Lista de fármacos

Foram considerados como fármacos potencialmente reposicionáveis para tratamento de doenças negligenciadas causadas por protozoários apenas aqueles que possuem alvos com ortólogos em protozoários e que esses ortólogos estejam em pelo menos uma via metabólica e que possuam pelo menos um fenótipo associado (Figura 3.11).

Com essa lista de fármacos foi realizada uma análise de similaridade com o programa BL2SEQ (versão 2.2.26)[11] (Tatusova & Madden, 1999) entre cada par de informação: alvo do fármaco / alvo no protozoário. Foram selecionadas para cada gênero de protozoário o par com menor similaridade e foi realizado um alinhamento par-a-par (*pairwise*) com o programa Jalview (Waterhouse *et al.*, 2009). Em seguida foram capturados do ProtozoaDB as proteínas ortólogas ao alvo do protozoário com menor similaridade com o alvo do fármaco e construído um novo alinhamento. Uma ficha com informações sobre o fármaco, alvo no protozoário, fenótipos do alvo no protozoário e mapas de vias metabólicas em que o alvo no protozoário participa foi elaborada para análises.

## 3.10 Validação dos resultados

Para validar os resultados obtidos, um programa foi elaborado (Apêndice B) para fazer uma pesquisa no PubMed e retornar os resumos que contivessem os seguintes termos: espécie do protozoário, anotação funcional do alvo no protozoário e os termos *drug* e *target* (Ex: ("Trypanosoma cruzi"AND "transketolase"AND (drug OR target))). Em seguida foi realizada uma análise com os artigos encontrados para avaliar a relevância dos termos encontrados para este estudo. Apenas o primeiro resumo retornado por cada consulta foi considerado no estudo (Figura 3.11).

[11] Este programa faz uma análise de similaridade entre duas sequências.

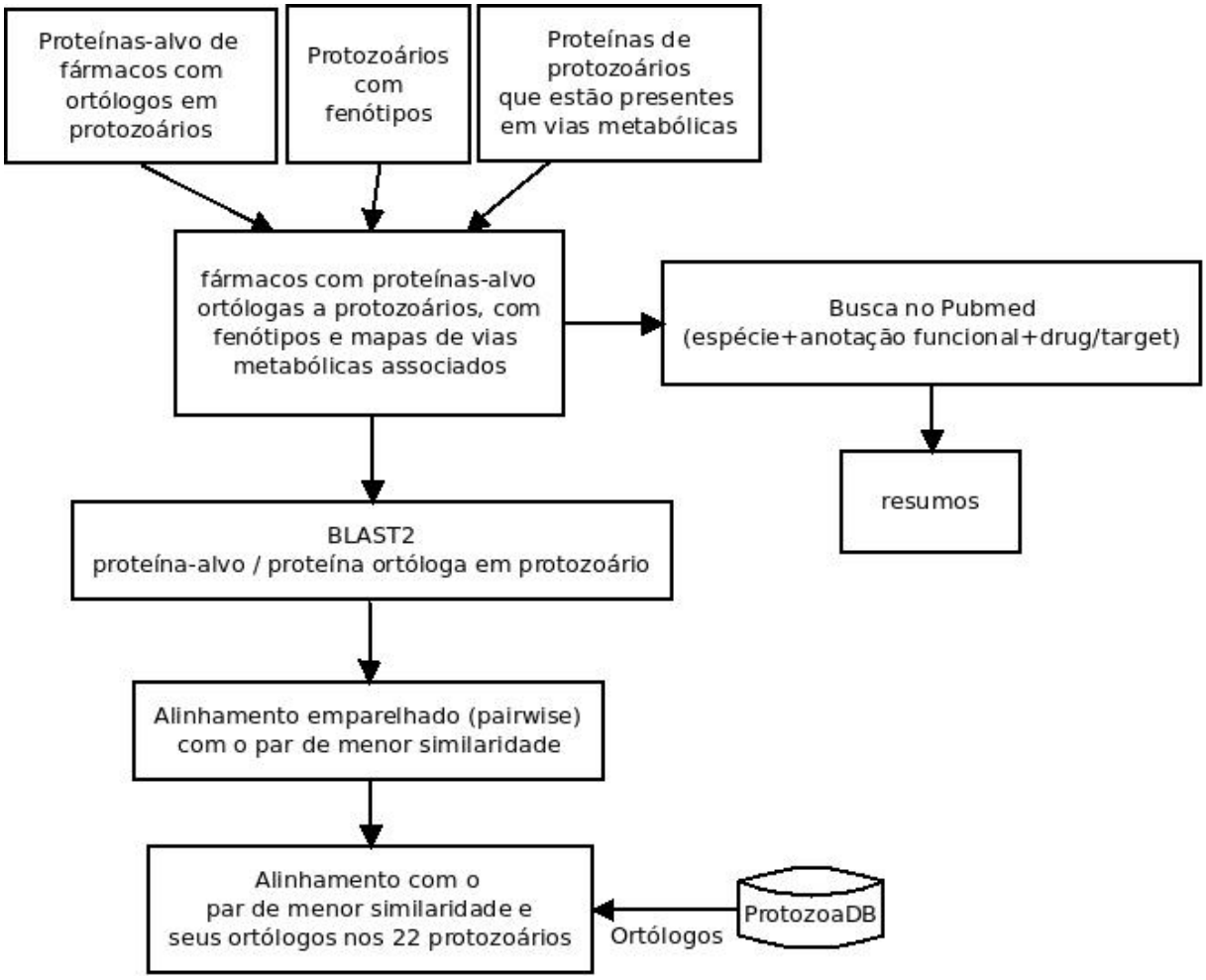



Figura 3.11: Validação dos resultados. Com o resultado dos fármacos cujas proteínas-alvo possuem ortólogos em protozoários e esses ortólogos possuem fenótipos e mapas de vias metabólicas associadas, foi realizada uma consulta no PubMed para encontrar artigos com relevância para a validação do estudo. Também foi realizada uma BLAST2SEQ com as sequências ortólogas por gênero de protozoário e um alinhamento par-a-par para encontrar o par com menor similaridade.

# Capítulo 4

# Resultados

## 4.1 Módulo PostSemantic

O módulo PostSemantic foi desenvolvido neste estudo para permitir que bases de dados armazenadas em formato relacional possam ser convertidas para o padrão RDF. Este módulo é composto por objetos de banco de dados armazenados em um SGDB. Nesta primeira versão, apenas o código para PostgreSQL foi disponibilizado[1]. Ao todo são 3 tabelas, 8 funções e 1 programa escrito em linguagem Java (Tabela 4.1).

Tabela 4.1: Objetos que compõem o módulo PostSemantic

| **Nome do objeto** | **Tipo de objeto** | **Descrição** |
|---|---|---|
| pg_mapping_classes | Tabela | Armazena informações sobre as tabelas que serão convertidas para triplas |

*Continua na próxima página*

[1] O código fonte está disponível em `http://sourceforge.org/p/postsemantic`

Tabela 4.1 – *Continuação da página anterior*

| **Nome do objeto** | **Tipo de objeto** | **Descrição** |
|---|---|---|
| pg_mapping_properties | Tabela | Armazena informações sobre os campos das tabelas que serão convertidas para triplas |
| pg_ontologies_ntriples | Tabela | Armazena ontologias no formato N-Triples |
| Getcount | Função em plpgsql | Calcula a quantidade de linhas a serem convertidas para triplas |
| hash | Função em pljava | Calcula o campo chave primária de acordo com o projeto Apache JENA |
| insert_mapping_classes | Função em plpgsql | Insere o mapeamento de uma tabela para posterior conversão para triplas |
| insert_node | Função em plpgsql | Insere um nó da tripla de informação |
| triplification | Função em plpgsql | Converte o metadado da tabela para triplas |
| triplification2 | Função em plpgsql | Converte o metadado da tabela para triplas com controle de transação |
| tuple2triple | Função em plruby | Converte os dados relacionais para triplas |
| tuple2triple2 | Função em plruby | Converte os dados relacionais para triplas com controle de transação |
| Hash.java | Função em Java | Código-fonte para calcular o campo hash utilizado pela função hash. |

## 4.2 Dados em formato RDF

Foram geradas 4.091.239 triplas depois da conversão das 7 bases de dados. A Tabela 4.2 mostra o número de triplas resultante da conversão de cada base de dados

Tabela 4.2: Quantidade de triplas geradas por base de dados convertida.

| **Base de Dados** | **Número de triplas** |
|---|---|
| ProtozoaDB | 2.715.917 |
| OrthoMCLDB | 379.314 |
| Kegg | 86.045 |
| Sider | 65.745 |
| SGD | 62.204 |
| Ensembl | 263.503 |
| DrugBank | 518.511 |
| **Total** | 4.091.239 |

## 4.3 Nuvem de dados do projeto

Com as 7 bases de dados convertidas em triplas foi possível disponibilizar uma nuvem de dados que poderia ser agregada à nuvem do LOD (Figura 4.1). Entretanto, somente os dados relativos ao ProtozoaDB serão disponibilizados pois pertencem ao grupo de pesquisa deste estudo. Os demais conjuntos de dados estão disponíveis apenas para consultas SPARQL[2].

## 4.4 Termos dinamicamente criados pelo arquivo de regras

Considerando que para responder a pergunta central do projeto havia a necessidade de se utilizar termos computacionalmente complexos, 3 novas propriedades (Tabela

[2]Disponível em `http://biowebdb.org/drugrepositioning/sparql-rules`

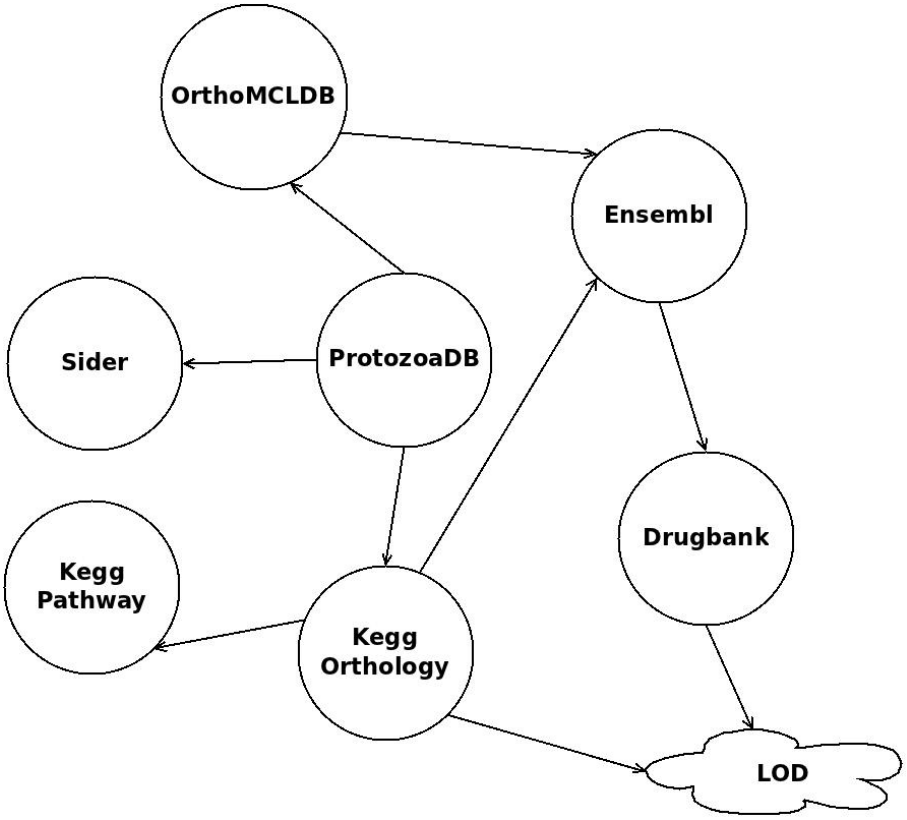



Figura 4.1: Nuvem de dados criada pelo estudo após a conversão das bases de dados para o padrão RDF.

4.3) foram criadas com a finalidade de pré-processar esses termos, permitindo assim que a pergunta fosse respondida utilizando-se os termos complexos. No Apêndice C encontram-se as consultas SPARQL utilizadas para a criação desses novos termos.

Tabela 4.3: Termos dinamicamente criados pelo arquivo de regras.

| **Nova Propriedade** | **Informação** |
|---|---|
| reuseTo (`http://biowebdb.org/protozoadb/reuseTo`) | Fármacos que possuem como alvos proteínas de *Homo sapiens* com ortólogos em protozoários. |
| phenotype (`http://biowebdb.org/protozoadb/phenotype`) | Fenótipos de genes de *Sacharomyces cerevisae* com ortólogos em protozoários. |
| pathway (`http://biowebdb.org/protozoadb/pathway`) | Mapas de vias metabólicas dos genes de protozoários. |

## 4.5 Consultas realizadas na base convertida para triplas

Com as 4.091.239 triplas disponibilizadas para consultas, 19 consultas foram realizadas (Apêndice C) para explorar o potencial da nuvem de dados, gerando informações quantitativas e qualitativas.

### 4.5.1 Informações quantitativas

Foram montadas e executadas 9 consultas para capturar informações quantitativas da nuvem de dados gerada. A Tabela 4.4 mostra as informações gerais extraídas dessa nuvem.

Tabela 4.4: Informações gerais extraídas da nuvem de dados gerada.

| **Informação** | **Qtd** |
|---|---|
| Classes | 18 |
| Propriedades | 145 |
| Fármacos | 4.408 |
| Alvos | 4.553 |
| Efeitos colaterais | 3.494 |
| Grupos de Ortólogos do OrthoMCLDB | 37.439 |
| Grupos de Ortólogos do KEGG | 8.804 |
| Fenótipos | 583 |
| Mapas de Vias Metabólicas | 393 |
| Proteínas de protozoários | 218.100 |
| Fármacos com alvos ortólogos a protozoários | 394 |
| Proteínas de protozoários ortólogas a alvos de fármacos | 386 |
| Fármacos com alvos ortólogos a protozoários e fenótipos associados | 216 |
| Fármacos com alvos ortólogos a protozoários, com fenótipos associados e mapas de vias metabólicas associados | 150 |

### 4.5.2 Informações qualitativas

Foram realizadas 10 consultas para obtenção de dados qualitativos, mostrados nas próximas seções.

#### 4.5.2.1 Fármacos

Foram encontrados 394 fármacos cujos alvos são ortólogos às proteínas de protozoários (Apêndice D). Em relação aos protozoários, foram encontradas 386 proteínas que são ortólogas a alvos de fármacos. As análises seguintes são baseadas nas classificações dos fármacos segundo o DrugBank.

#### 4.5.2.2 Organismos afetados

Na base de dados do DrugBank há a informação de quais organismos são afetados pelos fármacos. Dos 394 fármacos com alvos com ortólogos em protozoários, 255 (64,72%) não possuem organismos afetados cadastrados e 3 fármacos (0,76%) estão cadastrados como fármacos para protozoários (Figura 4.2)

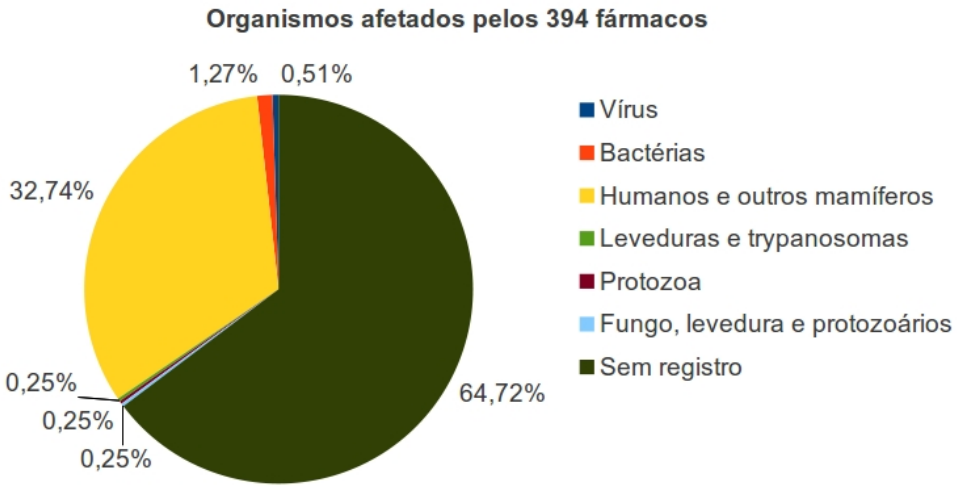



Figura 4.2: Classificação do DrugBank sobre qual organismo é afetado pelo fármaco. O gráfico exibe informações sobre os 394 fármacos cujos alvos possuem ortólogos com proteínas de protozoários.

#### 4.5.2.3 Categoria dos fármacos

Outra informação existente na base do DrugBank é a classificação dos fármacos por categorias. Dos 394 fármacos encontrados, 212 (53,80%) não possuem categoria cadastrada (Figura 4.3).

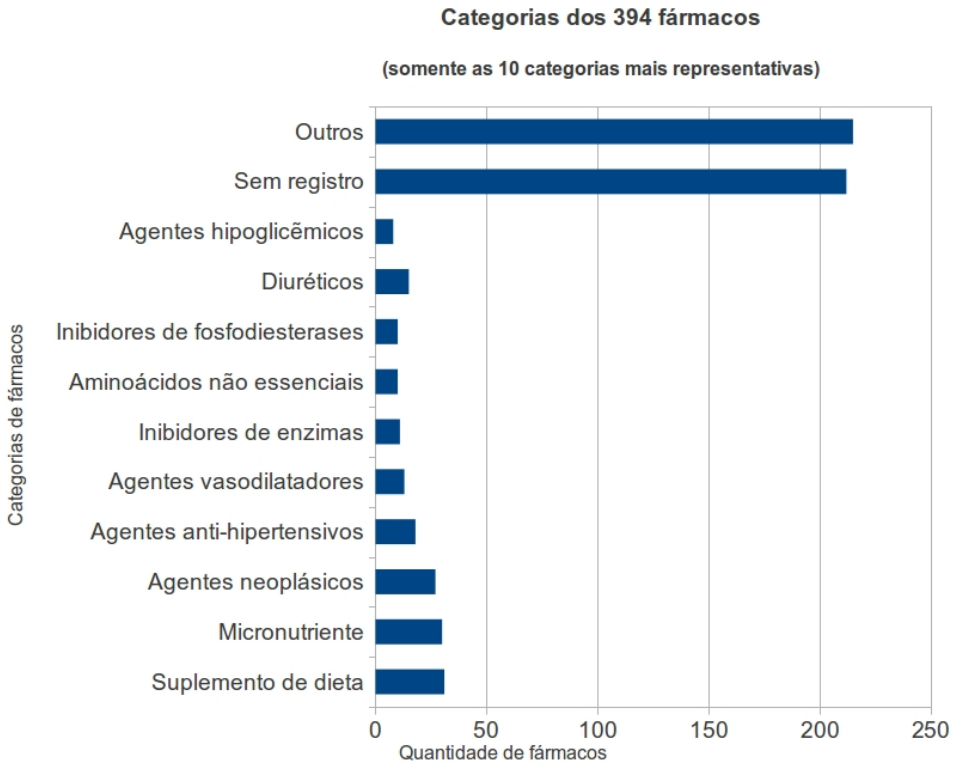



Figura 4.3: Classificação do DrugBank quanto à categoria dos fármacos. O gráfico exibe informações das categorias mais representativas dos 394 fármacos cujos alvos possuem ortólogos com proteínas de protozoários.

#### 4.5.2.4 Protozoários identificados

Os 394 fármacos cujos alvos possuem ortólogos com protozoários representam 9 espécies de protozoários (Figura 4.4).

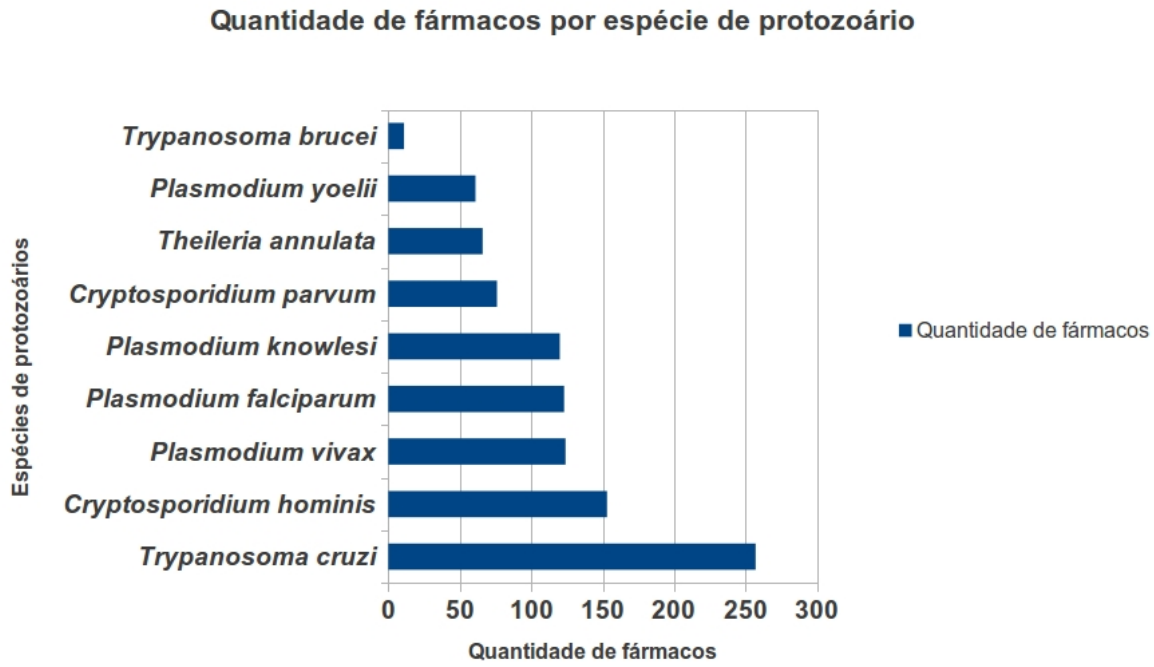



Figura 4.4: Distribuição dos 394 fármacos identificados no estudo que possuem proteínas-alvo ortólogas a proteínas de protozoários. O gráfico apresenta a quantidade de fármacos por cada espécie de protozoário.

#### 4.5.2.5 Fenótipos

Foram encontrados 430 fenótipos em 10.852 proteínas de protozoários. A Figura 4.5 apresenta os 10 fenótipos associados com o maior número de proteínas. Na Figura 4.6 está representada a quantidade de fenótipos por espécie de protozoário.

Nos 394 fármacos com alvos com ortólogos em protozoários, foram encontrados 157 (39,85%) fenótipos associados às proteínas de protozoários (Figura 4.7).

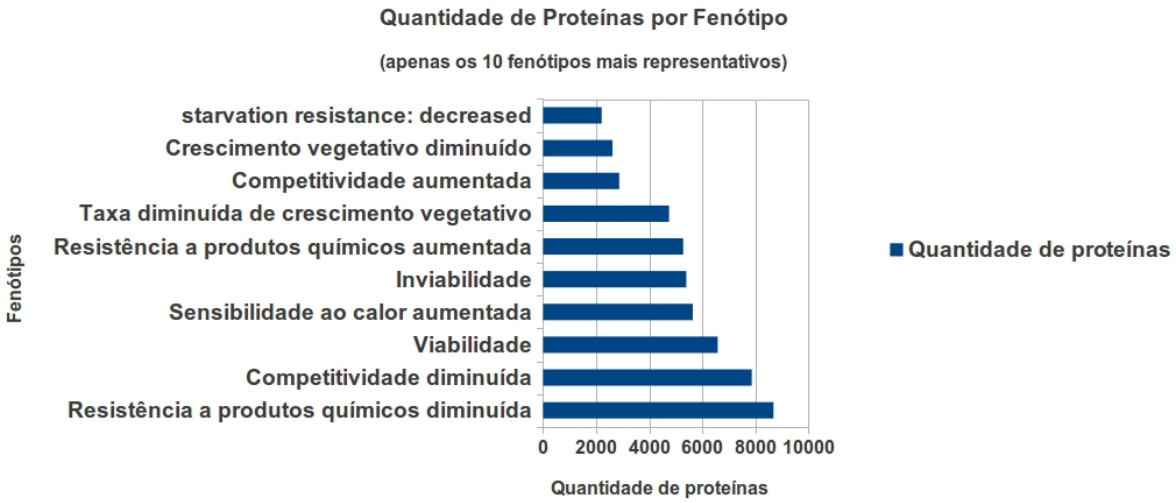



Figura 4.5: Fenótipos associados à proteínas de protozoários. O gráfico mostra os 10 fenótipos com maior número de proteínas associadas. Foram encontradas 10.852 proteínas de protozoários associadas a 430 fenótipos.

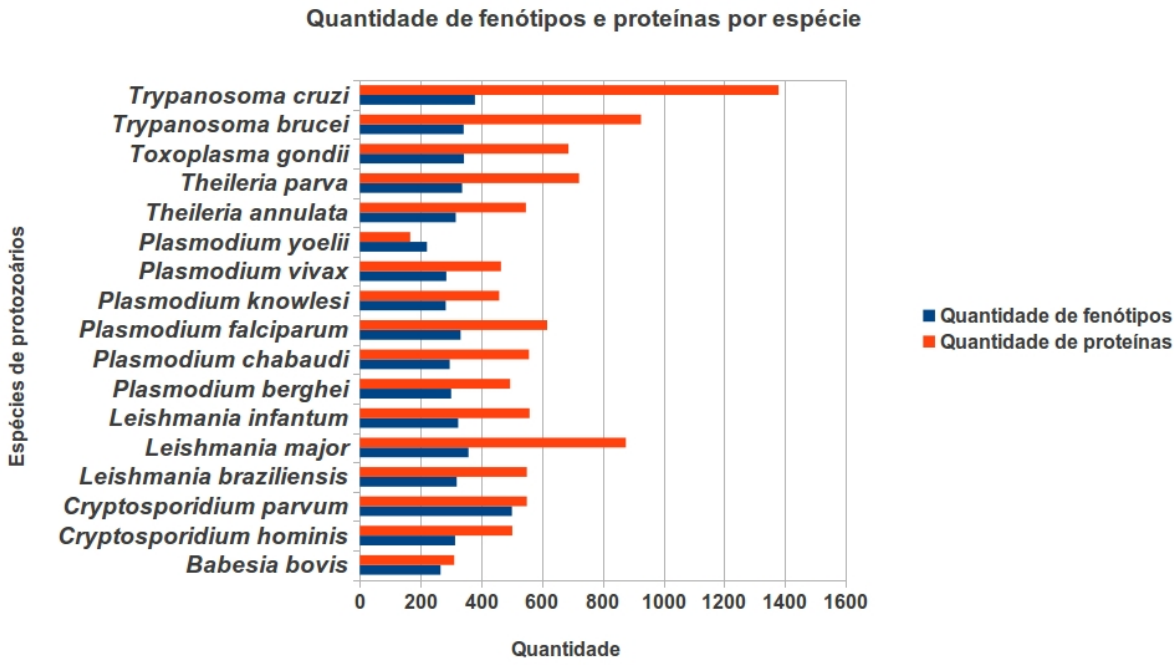



Figura 4.6: Quantidade de fenótipos e número de proteínas associadas a fenótipos por espécie de protozoário. A barra na cor azul exibe o número de fenótipos associados à proteínas de protozoários. A barra na cor vermelha mostra o número de proteínas de protozoários associados a pelo menos 1 fenótipo.

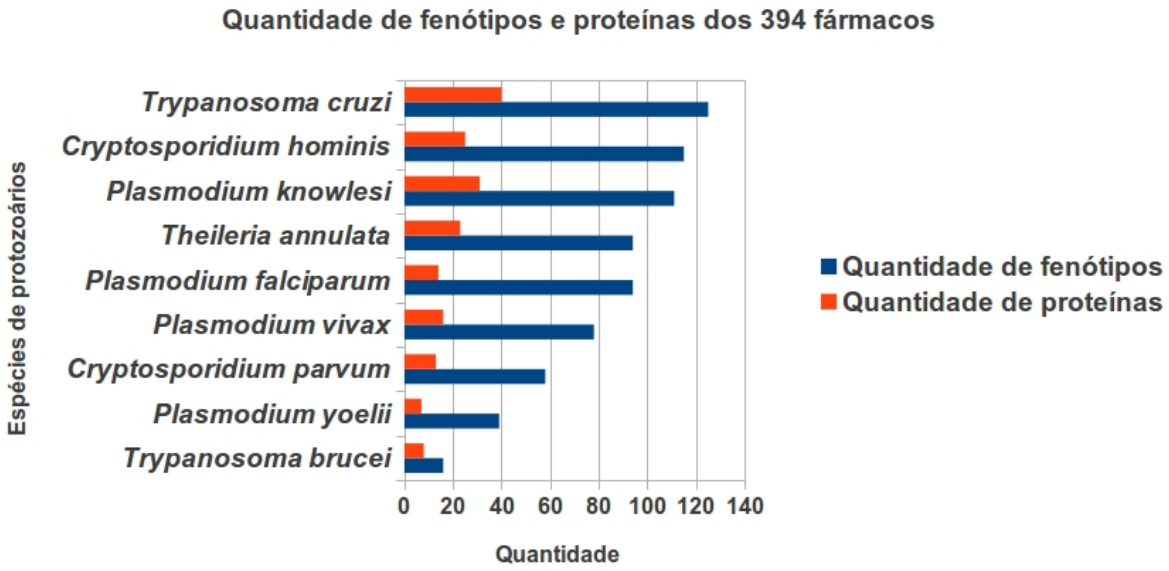



Figura 4.7: Quantidade de fenótipos associados às proteínas de protozoários ortólogas a alvos de fármacos. O gráfico exibe as proteínas de protozoários ortólogas aos 394 fármacos.

#### 4.5.2.6 Vias Metabólicas

Dos 393 mapas de vias metabólicas convertidas para triplas, 255 (64,89%) mapas possuem proteína(s) de protozoários (Figura 4.8) e 9.590 proteínas de protozoários estão presentes em pelo menos 1 mapa (Figura 4.9).

#### 4.5.2.7 Fármacos com fenótipos anotados

Dos 394 fármacos cujos alvos possuem ortólogos com protozoários, 216 (54,82%) possuem fenótipos associados (Figura 4.10).

#### 4.5.2.8 Fármacos com fenótipos e vias metabólicas

Dos 394 fármacos com alvos em ortólogos em protozoários, 150 (38,07%) possuem fenótipos e vias metabólicas associadas (Figura 4.11).

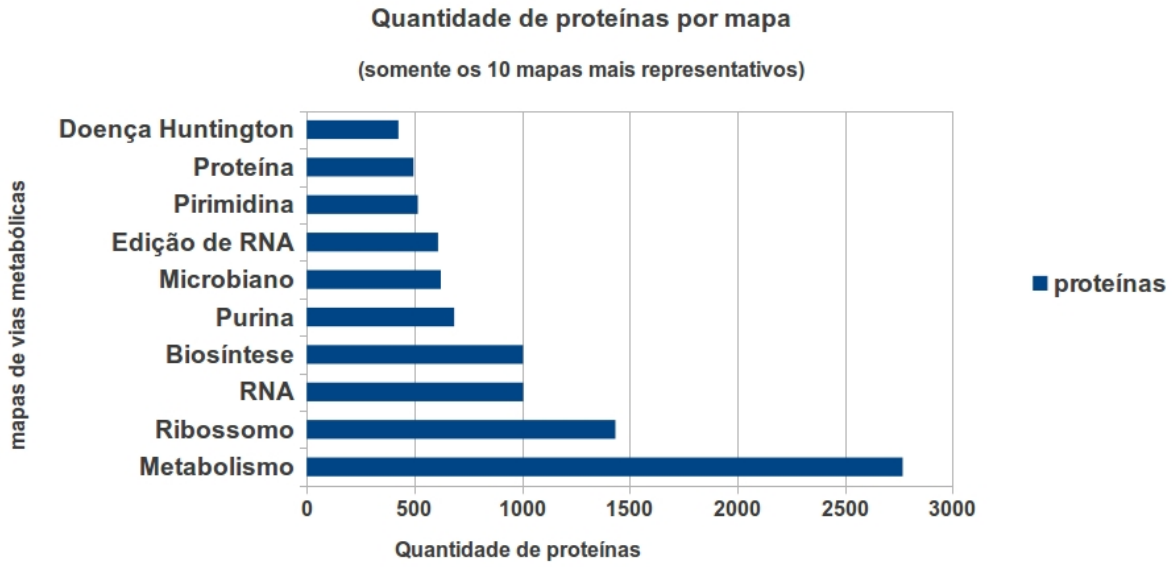



Figura 4.8: Quantidade de proteínas de protozoários por mapas de vias metabólicas. O gráfico mostra os 10 mapas de vias metabólicas com maior número de proteínas de protozoários identificadas. Foram encontradas 9.590 proteínas de protozoários associadas a 255 mapas de vias metabólicas.

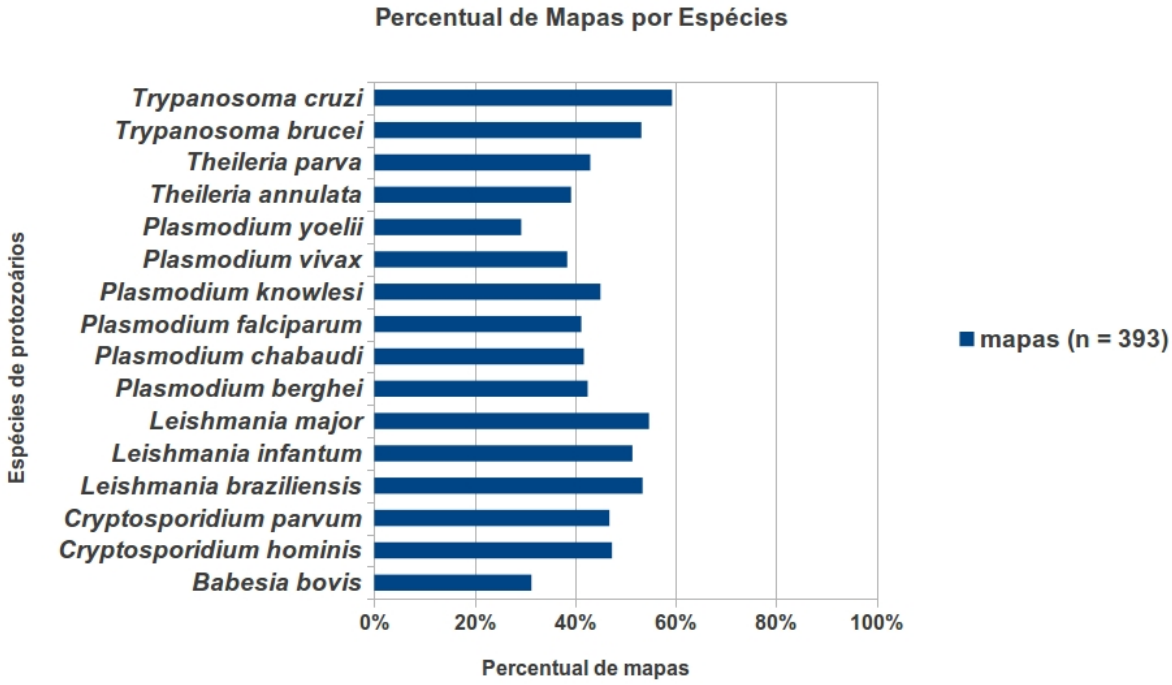



Figura 4.9: Percentual de mapas por espécie de protozoário identificados no estudo, onde pelo menos uma proteína do protozoário foi identificada.

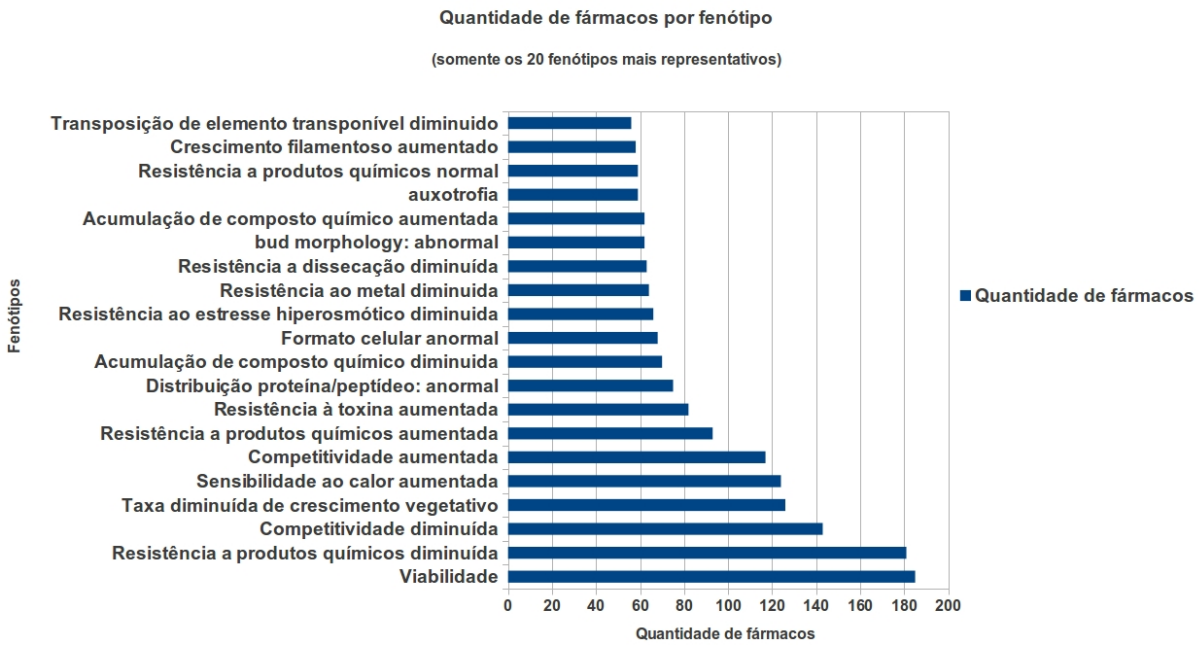



Figura 4.10: Quantidade de fármacos com alvos com ortólogos em protozoários com fenótipos anotados. O gráfico mostra os 20 fenótipos com maior quantidade de fármacos.

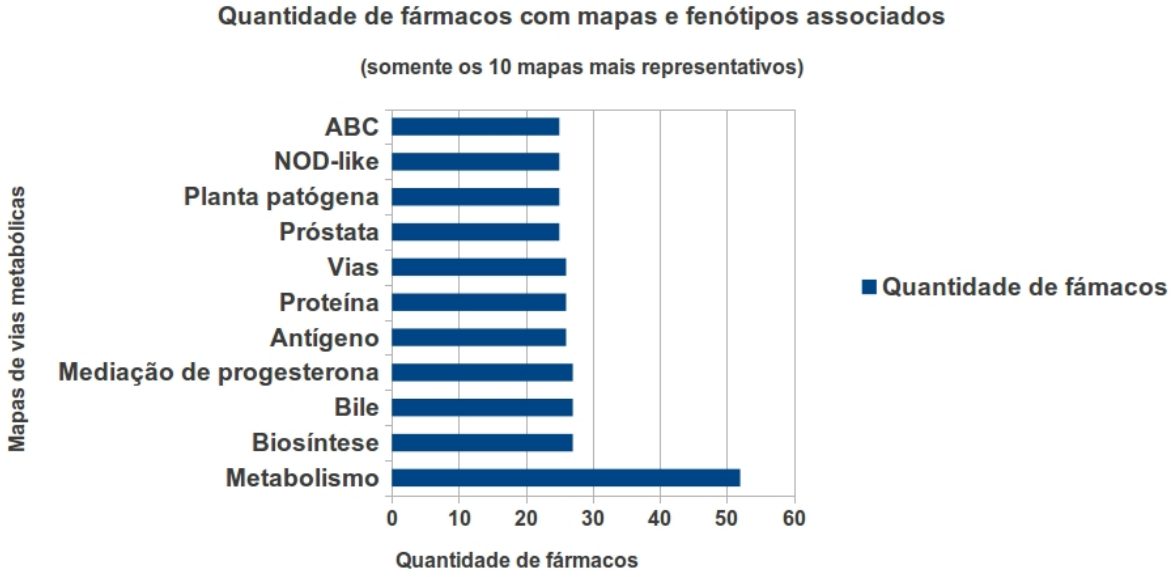



Figura 4.11: Quantidade de fármacos com fenótipos e vias metabólicas associados. O gráfico apresenta as 10 vias metabólicas com maior número de fármacos identificados.

#### 4.5.2.9 Alvos em Protozoários

A Figura 4.12 mostra a anotação funcional das proteínas de protozoários identificadas no estudo como ortólogas a alvos de fármacos, que possuem fenótipos associados e participam de pelo menos uma via metabólica.

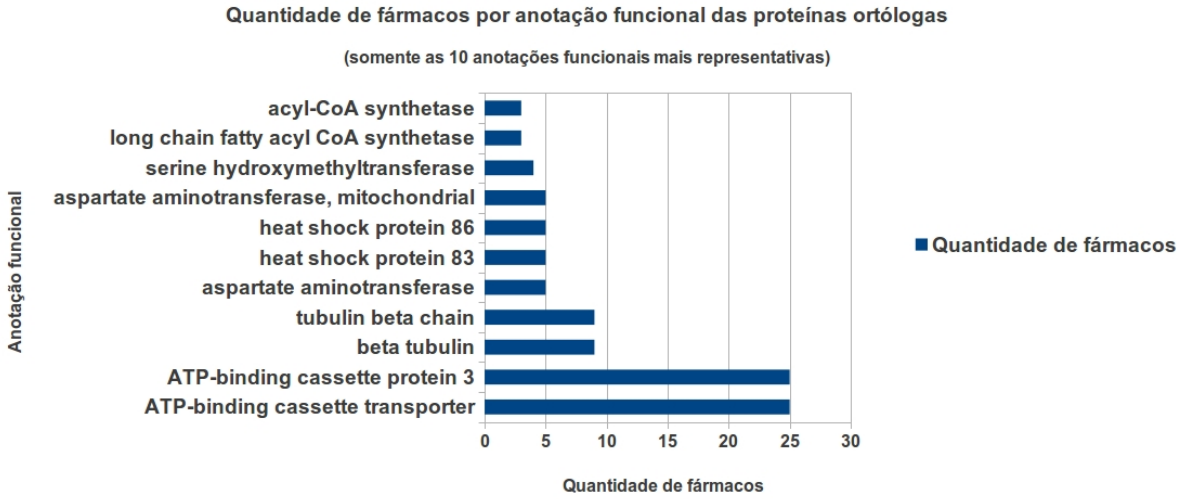



Figura 4.12: Anotação funcional das proteínas de protozoários que são ortólogas aos alvos de fármacos e que possuem fenótipos e participam de via metabólica. O gráfico mostra a anotação funcional e a quantidade de fármacos das 10 anotações funcionais com mais fármacos associados.

### 4.5.3 Lista de fármacos com potencial para reposicionamento

A partir dos resultados da integração das bases e das consultas realizadas, foi elaborada uma lista com 150 fármacos com potencialidade de serem reposicionados para tratamento de doenças negligenciadas causadas por protozoários (Apêndice D).

### 4.5.4 Análise dos alinhamentos das proteínas-alvo de fármacos com suas ortólogas em protozoários

As 9 espécies de protozoários que possuem ortólogos à alvos de fármacos representam 4 gêneros: *Cryptosporidium*, *Plasmodium*, *Theileria* e *Trypanosoma*. Para

cada um desses gêneros foi realizada uma análise de similaridade utilizando o par proteína-protozoário/proteína-alvo.

#### 4.5.4.1 Gênero *Cryptosporidium*

Foram encontradas 70 proteínas-alvo de fármacos com proteínas ortólogas do gênero *Cryptosporidium*. A proteína de protozoário com menor similaridade com sua proteína-alvo foi entre a proteína de protozoário com número de acesso (accession number) XP_668169 - Histona deacetilase e a proteína-alvo HDAC2 - Histona deacetilase 2. O alinhamento par-a-par (*pairwise*) (Figura 4.13) entre as duas mostrou uma identidade de 52,78%.

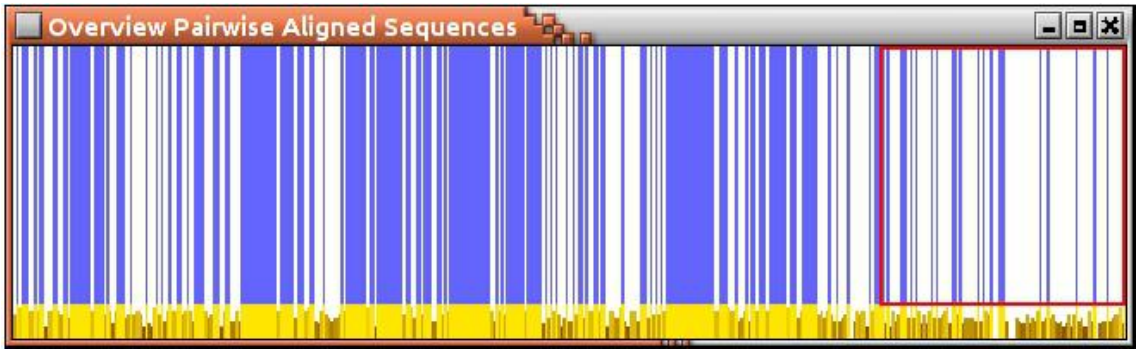



Figura 4.13: Alinhamento par-a-par entre a proteína do gênero *Cryptosporidium* com a proteína-alvo ortóloga do fármaco que apresentou menor similaridade.

Foram recuperadas no ProtozoaDB mais 3 proteínas ortólogas à proteína do gênero *Cryptosporidium*. Um novo alinhamento foi realizado entre essas proteínas (Figura 4.14).

O par com menor similaridade para este gênero está relacionado ao fármaco **Vorinostat**. A Tabela 4.5 mostra as características extraídas da nuvem de dados gerada para esse fármaco.

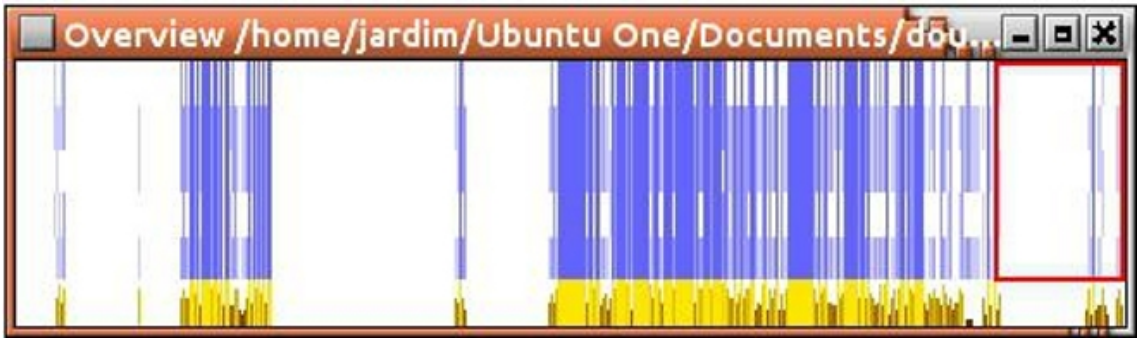



Figura 4.14: Alinhamento entre a proteína-alvo de fármaco e suas ortólogas nos 22 protozoários.

Tabela 4.5: Fármaco para *Cryptosporidium*

| **Característica** | **Descrição** |
|---|---|
| Nome genérico do fármaco | Vorinostat |
| Organismo(s) afetado(s) | Humanos e outros mamíferos |
| Indicação | Para o tratamento de manifestações cutâneas de pacientes com linfoma de células T cutâneo que têm doença progressiva, persistente ou recorrente ou em sequência de duas terapias sistémicas. |
| Mecanismo de ação | Vorinostat inibe a atividade enzimática de histonas desacetilases HDAC1, HDAC2 e HDAC3 (Classe I) e HDAC6 (Classe II) em concentrações nanomolares (IC <sub> 50 </ sub> <86 nM). Estas enzimas catalisam a remoção de grupos acetilo a partir de resíduos de lisina de proteínas, incluindo as histonas e factores de transcrição. Em algumas células de cancro, existe uma sobre-expressão de HDAC, ou um recrutamento aberrante da HDAC para os factores de transcrição, causando oncogénicos hipoacetilação de histonas nucleares nucleosomal. Hipoacetilação de histonas está associado a uma estrutura de cromatina condensada e repressão da transcrição do gene. A inibição da actividade de HDAC permite a acumulação de grupos acetilo, em que os resíduos de lisina de histona, resultando em uma estrutura de cromatina aberta e de activação da transcrição. In vitro, vorinostat provoca a acumulação de histonas acetiladas e induz a paragem do ciclo celular e / ou apoptose de algumas células transformadas. O mecanismo do efeito anti-neoplásico de vorinostat não foi completamente caracterizado. |
| Categoria do fármaco | Agente antineoplásico<br>Agente anticarcinogênico<br>Inibidores de enzimas |

*Continua na próxima página*

Tabela 4.5 – *Continuação da página anterior*

| **Característica** | **Descrição** |
|---|---|
| | Agente anti-inflamatório, não esteróide |
| Anotação funcional da proteína-alvo | Histona deacetilase 2 |
| Anotação funcional do alvo no protozoário | Histona deacetilase |
| Fenótipos do alvo no protozoário | RNA acumulação: diminuída<br>Viável<br>Recombinação meiótica: aumentada<br>Transposição de elemento transponível: aumentada<br>Eficiência de acasalamento: diminuída<br>Crescimento respiratório: ausente<br>Crescimento vegetativo: diminuição da taxa<br>Resistência a químicos: diminuída<br>Excreção de composto químico: aumentada<br>Absorção de nutrientes: diminuição da taxa |
| Vias metabólicas que o alvo no protozoário participa | Mapid:05220 (*Chronic*)<br>Mapid:05200 (*Pathways*)<br>Mapid:04330 (*Notch*)<br>Mapid:05016 (Huntingtons)<br>Mapid:04110 (*Cell*) |

#### 4.5.4.2 Gênero *Plasmodium*

Para o gênero *Plasmodium* foram encontrados 111 pares de proteína-alvo/proteína de protozoário. O par com menor similaridade foi entre XP_002260251.1 - Valina-tRNA

ligase e VARS - Valil-tRNA sintetase. O alinhamento par-a-par teve uma identidade de 41,95% (Figura 4.15).

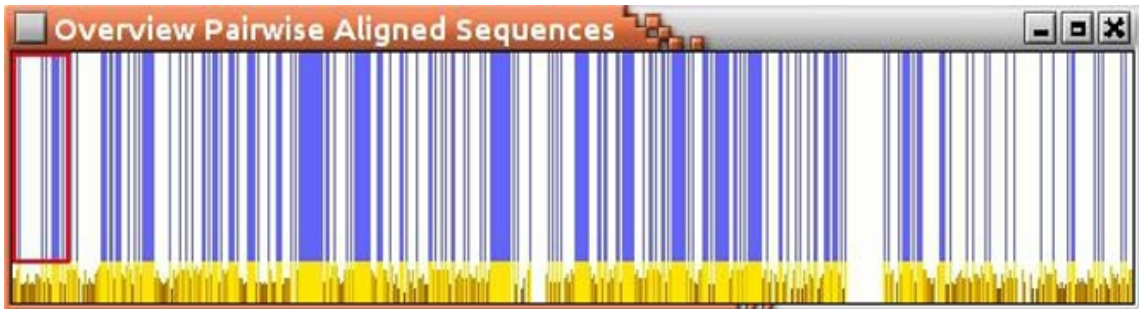



Figura 4.15: Alinhamento par-a-par entre a proteína do gênero *Plasmodium* com a proteína-alvo ortóloga do fármaco que apresentou menor similaridade.

Foram recuperadas mais 14 proteínas ortólogas entre os 22 protozoários do ProtozoaDB e um novo alinhamento foi realizado (Figura 4.16).


Figura 4.16: Alinhamento entre a proteína-alvo de fármaco e suas ortólogas nos 22 protozoários.

Para esta proteína do gênero *Plasmodium* foi identificado o fármaco **L-Valine** com as características apresentadas na Tabela 4.6.

Tabela 4.6: Fármaco para *Plasmodium*

| **Característica** | **Descrição** |
|---|---|
| Nome genérico do fármaco | L-Valine |
| Organismo(s) afetado(s) | Humanos e outros mamíferos |
| Indicação | Promove o vigor mental, coordenação muscular, e as emoções calmas. Também podem ser de utilização numa minoria de pacientes com encefalopatia hepática e, em alguns doentes com fenilcetonúria. |
| Mecanismo de ação | Aplica-se a valina, leucina e isoleucina. Este grupo de aminoácidos essenciais são identificados como os ácidos de cadeia ramificada, os BCAA. Uma vez que este arranjo de átomos de carbono não pode ser feita pelos seres humanos, estes aminoácidos são um elemento essencial na dieta. O catabolismo de todos os três compostos inicia no músculo e no rendimento de NADH e FADH2, que pode ser utilizado para a geração de ATP. O catabolismo de todos os três destes aminoácidos usa as mesmas enzimas, nos dois primeiros passos. O primeiro passo em cada caso é uma transaminação usando um único BCAA aminotransferase, com a-cetoglutarato como receptor de amina. Como resultado, os três diferentes a-ceto ácidos são produzidos e são oxidados utilizando um comum de cadeia ramificada-desidrogenase de ácido a-ceto, obtendo-se os três diferentes derivados CoA. Subsequentemente, as vias metabólicas divergem, produzindo muitos intermediários. O principal produto de valina é propionylCoA, o precursor glicogênicos de succinil-CoA. Catabolismo isoleucina termina com a produção de acetilCoA e propionylCoA; assim isoleucina é tanto glicogênicos e ketogenic. Leucina dá origem a acetilCoA e acetoacetylCoA, e é, assim, classificada como estritamente ketogenic. Há um número de doenças genéticas associadas com o catabolismo defeituoso dos AACR. O defeito mais comum é o de desidrogenase de cadeia ramificada de ácido a-ceto. Uma vez que existe apenas uma enzima desidrogenase para todos os três aminoácidos, todos os três a-ceto ácidos acumulam e são excretados na urina. A doença é conhecida como doença do xarope de bordo por causa de o odor característico da urina em indivíduos afectados. Retardo mental, nestes casos, é extensa. Infelizmente, uma vez que estes são os aminoácidos essenciais, que não pode ser fortemente limitado na dieta, em última análise, a vida dos indivíduos afectados é curta e é o desenvolvimento anormal Os principais problemas neurológicos são pobres devido à formação de mielina no SNC. |
| Categoria do fármaco | Suplemento de dieta<br>Micronutrientes<br>Aminoácidos essenciais |
| Anotação funcional da proteína-alvo | Valil-tRNA sintetase |
| Anotação funcional do alvo no protozoário | Valina-tRNA ligase |

*Continua na próxima página*

Tabela 4.6 – *Continuação da página anterior*

| **Característica** | **Descrição** |
|---|---|
| Fenótipos do alvo no protozoário | Resistência a químicos: aumentada<br>Progressão do ciclo celular: anormal<br>Crescimento vegetativo: diminuição da taxa<br>Resistência a químicos: diminuida<br>Crescimento vegetativo: diminuida<br>Aptidão competitica: aumentada<br>Inviável |
| Vias metabólicas que o alvo no protozoário participa | Mapid:00290 (*Valine*)<br>Mapid:00970 (*Aminoacyl-tRNA*) |

#### 4.5.4.3 Gênero *Theileria*

Para o gênero *Theileria* foram encontrados 34 pares de proteínas. O par com menor similaridade foi entre XP_951883.1 - CTP:fosforicolina citidililtransferase e PCYT1A - Colina-fosfato citidililtransferase A. O percentual de identidade do alinhamento par-a-par entre elas foi de 27,45% (Figura 4.17).


Figura 4.17: Alinhamento par-a-par entre a proteína do gênero *Theileria* com a proteína-alvo ortóloga do fármaco que apresentou menor similaridade.

No ProtozoaDB foram encontradas 10 proteínas ortólogas entre os 22 protozoários que foram alinhadas com o par de menor similaridade (Figura 4.18).

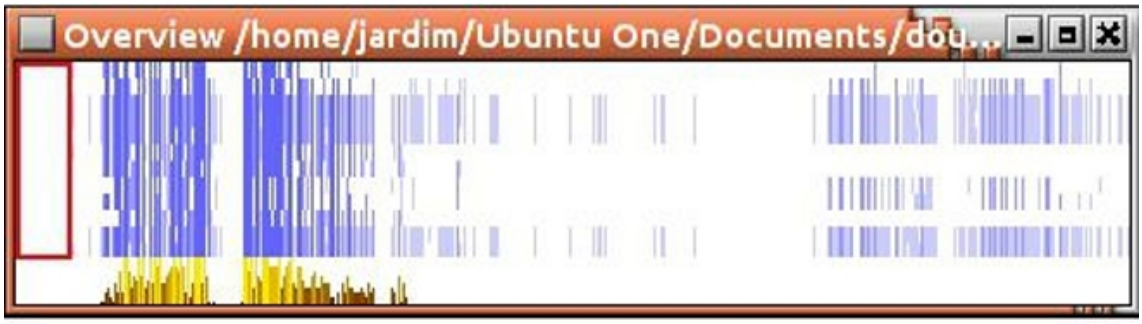



Figura 4.18: Alinhamento entre a proteína-alvo de fármaco e suas ortólogas nos 22 protozoários.

Para esta proteína do gênero *Theileria* foi identificado o fármaco **Choline** com as características apresentadas na Tabela 4.7.

Tabela 4.7: Fármaco para *Theileria*

| **Característica** | **Descrição** |
|---|---|
| Nome genérico do fármaco | Choline |
| Organismo(s) afetado(s) | Humanos e outros mamíferos |
| Indicação | Para suplementação nutricional, também para o tratamento de carência ou desequilíbrio dietético |

*Continua na próxima página*

Tabela 4.7 – *Continuação da página anterior*

| **Característica** | **Descrição** |
|---|---|
| Mecanismo de ação | A colina é uma parte principal do grupo de cabeça polar de fosfatidileolino. Fosfatidilcolina papel na manutenção da integridade da membrana celular é vital para todos os processos biológicos básicos: o fluxo de informação, comunicação intracelular e bioenergética. Ingestão inadequada de colina pode afetar negativamente todos esses processos. A colina é também uma parte importante de uma outra membrana de fosfolípidos, esfingomielina, também importante para a manutenção da estrutura e função celular. É notável e não é de surpreender que a deficiência de colina, em cultura de células causa a apoptose ou morte celular programada. Isto parece ser devido a anomalias no conteúdo da célula de fosfatidilcolina de membrana e um aumento da ceramida, um precursor, bem como um metabolito, do esf ingomielina. Acumulação de ceramida, que é causada por uma deficiência de colina, parece activar caspase, um tipo de enzima que medeia a apoptose. Betaína ou trimetilglicina é derivado a partir de colina via uma reacção de oxidação. A betaína é um dos factores que mantém baixos níveis de homocisteína pela resynthesizing L-metionina a partir da homocisteína. Níveis elevados de homocisteína é um factor de risco significativo para a aterosclerose, bem como outras perturbações cardiovasculares e neurológicas. A acetilcolina é um dos principais neurotransmissores e requer colina para a sua síntese. Os níveis de acetilcolina no cérebro adequados são acreditados para ser protectora contra certos tipos de demência, incluindo a doença de Alzheimer. |
| Categoria do fármaco | Micronutriente<br>Suplemento de dieta<br>Agente lipotrópico<br>Agente *nootropic* |
| Anotação funcional da proteína-alvo | Colina-fosfato citidililtransferase A |
| Anotação funcional do alvo no protozoário | CTP:fosforicolina citidililtransferase |
| Fenótipos do alvo no protozoário | Resistência a químicos: diminuida<br>Crescimento vegetativo: diminuição da taxa<br>Aptidão competitiva: normal<br>Resistência a químicos: aumentada<br>Viável<br>Sensibilidade ao calor: aumentada |
| Vias metabólicas que o alvo no protozoário participa | Mapid:00564 (*Glycerophospholipid*)<br>Mapid:00440 (*Phosphonate*)<br>Mapid:01100 (*Metabolic*) |

#### 4.5.4.4 Gênero *Trypanosoma*

Para o gênero *Trypanosoma* foram encontados 85 pares de proteínas. O par com menor similaridade foi entre XP_816252.1 - Gamma-glutamilcisteina sintetase e GCLC - Glutamato-cisteine ligase catalítico subunidade, que apresentou no alinhamento par-a-par uma identidade de 35,21% (Figura 4.19).


Figura 4.19: Alinhamento par-a-par entre a proteína do gênero *Trypanosoma* com a proteína-alvo ortóloga do fármaco que apresentou menor similaridade.

Mais 14 proteínas ortólogas entre os 22 protozoários foram buscadas no ProtozoaDB e realizado um novo alinhamento (Figura 4.20).

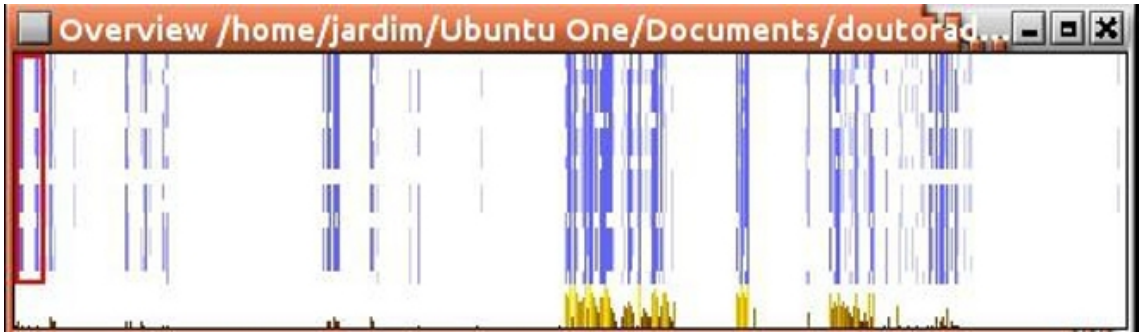



Figura 4.20: Alinhamento entre a proteína-alvo de fármaco e suas ortólogas nos 22 protozoários.

Para esta proteína do gênero *Trypanosoma* foi identificado o fármaco **L-Cysteine** com as características apresentadas na Tabela 4.8.

Tabela 4.8: Fármaco para *Trypanosoma*

| **Característica** | **Descrição** |
|---|---|
| Nome genérico do fármaco | L-Cysteine |
| Organismo(s) afetado(s) | Humanos e outros mamíferos |
| Indicação | Para a prevenção de danos no fígado e nos rins associada com a sobredosagem de acetaminofeno |
| Mecanismo de ação | Embora classificado como um não-essencial aminoácido cisteína pode ser essencial para crianças, idosos e indivíduos com doença metabólica certo ou que sofrem de síndromes de má absorção. Cisteína podem geralmente ser sintetizados pelo organismo humano sob condições fisiológicas normais, se uma quantidade suficiente de metionina está disponível. Devido à capacidade de tióis de sofrer reacções redox, cisteína tem propriedades antioxidantes. Propriedades antioxidantes da cisteína são normalmente expressos no tripeptídeo glutationa, que ocorre em seres humanos, bem como outros organismos. A disponibilidade sistémica oral de glutationa (GSH) é insignificante, de modo que deve ser biossintetizado a partir dos seus aminoácidos constituintes, cisteína, glicina e ácido glutâmico. Ácido glutâmico e glicina estão prontamente disponíveis para as dietas da maioria dos países industrializados, mas a disponibilidade de cisteína pode ser o substrato limitante. A cisteína é também uma importante fonte de sulfureto no metabolismo humano. O sulfeto de ferro-enxofre e em nitrogenase é extraído a partir de cisteína, que é convertido em alanina no processo. Num relatório de 1994 divulgado por cinco companhias de cigarro de topo, a cisteína é um dos 599 aditivos para cigarros. A sua utilização ou finalidade, no entanto, não é conhecido, assim como a maioria dos aditivos de cigarros. A sua inclusão nos cigarros poderia oferecer dois benefícios: Atuando como expectorante, uma vez que fumar aumenta a produção de muco nos pulmões, e aumentando a glutationa antioxidante benéfico (o que é diminuída em fumantes). |
| Categoria do fármaco | Suplemento nutricional<br>Suplemento de dieta<br>Micronutriente<br>Aminoácido não essencial |
| Anotação funcional da proteína-alvo | Glutamato-cisteina ligase catalítico subunidade |
| Anotação funcional do alvo no protozoário | Gamma-glutamilcisteina sintetase |
| Fenótipos do alvo no protozoário | Resistência a químicos: aumentada<br>Viabilidade: aumentada<br>Viável |

*Continua na próxima página*

Tabela 4.8 – *Continuação da página anterior*

| **Característica** | **Descrição** |
|---|---|
| | Resistência a estresse oxidativo: diminuída<br>Acúmulo de compostos químicos: diminuida<br>Índice de brotamento: diminuida<br>Tamanho da célula: aumentada<br>Aptidão competitiva: aumentada<br>Resistência a químicos: diminuida<br>Crescimento em fase exponencial: diminuição da taxa<br>Cor da colônia: anormal<br>Resistência a dessecação: diminuida<br>Auxotrofia<br>Crescimento vegetativo: diminuida<br>Resistência a estresse oxidativo: aumentada<br>Sensibilidade ao calor: aumentada<br>Acúmulo de compostos químicos: aumentada<br>Crescimento respiratório: ausente |
| Vias metabólicas que o alvo no protozoário participa | Mapid:00480 (*Glutathione*)<br>Mapid:01100 (*Metabolic*) |

## 4.6 Validação dos resultados obtidos

Com a lista dos 150 fármacos com potencial para serem reposicionados, foi realizada uma busca no PubMed que relacionasse a proteína do protozoário com fármacos para tratamento das doenças. Foram encontrados 38 artigos (Tabela 4.9).

Tabela 4.9: Quantidade de artigos encontrados no PubMed por gênero de protozoário

| Protozoário | Qtd. Artigos |
|---|---|
| *Cryptosporidium* | 5 |
| *Plasmodium* | 14 |
| *Theileria* | 1 |
| *Trypanosoma* | 18 |
| **TOTAL** | 38 |

Foi realizada uma curagem manual dos 38 resumos encontrados e selecionado um para cada gênero de protozoário, apresentados a seguir.

### 4.6.1 Gênero *Cryptosporidium*

**Termo:** O termo procurado no PubMed foi: ("Cryptosporidium parvum"AND "gpantothenate kinase"AND (drug OR target))

**Título:** Benzoylbenzimidazole-based selective inhibitors targeting Cryptosporidium parvum and Toxoplasma gondii calcium-dependent protein kinase-1 (`http://www.ncbi.nlm.nih.gov/pubmed?term=22795629`)

### 4.6.2 Gênero *Plasmodium*

**Termo:** O termo procurado no PubMed foi: ("Plasmodium vivax"AND "serine hydroxymethyltransferase"AND (drug OR target))

**Título:** Plasmodium serine hydroxymethyltransferase as a potential anti-malarial target: inhibition studies using improved methods for enzyme production and assay (`http://www.ncbi.nlm.nih.gov/pubmed?term=22691309`)

### 4.6.3 Gênero *Theileria*

**Termo:** O termo procurado no PubMed foi: ("Theileria annulata"AND "serinethreonine protein"AND (drug OR target))

**Título** Theileria-mediated constitutive expression of the casein kinase II-alpha subunit in bovine lymphoblastoid cells (`http://www.ncbi.nlm.nih.gov/pubmed?term=9211502`)

### 4.6.4 Gênero *Trypanosoma*

**Termo:** O termo procurado no PubMed foi: ("Trypanosoma cruzi"AND "gamma-glutamylcysteine synthetase"AND (drug OR target))

**Título** Drug target validation of the trypanothione pathway enzymes through metabolic modelling (`http://www.ncbi.nlm.nih.gov/pubmed?term=22394478`)

# Capítulo 5

# Discussão

## 5.1 Reposicionamento de fármaco com integração de bases de dados

Nos últimos anos foram descritos diversos trabalhos de integração de bases de dados biológicas no intuito de prover novas informações sobre alvos de fármacos e reposicionamento de fármacos. Várias técnicas foram propostas, desde o uso de *datawarehouse* (Günther *et al.*, 2008) até o uso de semântica (Qu *et al.*, 2009; Ekins *et al.*, 2011; Belleau *et al.*, 2008), passando pelas técnicas de ETL e disponibilização de aplicações web para exploração dos dados (Cockell *et al.*, 2010; von Eichborn *et al.*, 2011).

Dentre os trabalhos que utilizaram semântica para integrar dados, Qu *et al.* (2009) propuseram novos usos para fármacos já conhecidos ligando semanticamente a ação do fármaco com mecanismos de doenças, construindo, para isso, uma ontologia denominada Ontologia de Correlação Fármaco-Doença (Disease-Drug Correlation Ontology - DDCO ). Essa ontologia foi formalizada em OWL e integrou mútiplas ontologias, vocabulários controlados e esquemas de bases de dados para integrar informações de diversas bases, como DrugBank, GO e KEGG. A DDCO foi utilizada para integrar

as bases de dados e extrair informações sobre o estudo de caso do artigo. Entretanto, apesar do uso de ontologia o estudo não contemplou a publicação dos dados na nuvem do LOD.

Outro trabalho que utilizou semântica para integrar dados foi o Bio2RDF (Belleau *et al.*, 2008), que elaborou um vocabulário próprio, embora não tenha utilizado nenhuma ontologia ou vocabulário pré-existentes. O foco principal desse trabalho era disponibilizar o maior número de bases de dados possível na nuvem do LOD. Mesmo sem a utilização de ontologias, o projeto propôs a integração entre as bases de dados através de poucas ligações entre essas bases. Os resultados foram publicados na nuvem do LOD, mas poucas bases estão acessíveis para consultas SPARQL.

Neste trabalho foi estabelecida uma base de dados rica em informações relacionadas à biologia molecular, particularmente nas áreas de genômica e proteômica, com informações de fármacos e alvos de fármacos, com a finalidade de inferir novos usos para antigos fármacos (reposicionamento) para tratamento de doenças negligenciadas causadas por protozoários. Para isso, diversas bases de dados foram integradas através de semântica com o uso de um vocabulário próprio e controlado. A escolha por não se utilizar ontologia ou vocabulário pré-existente foi intencional e serviu para corroborar os resultados de outros trabalhos (Belleau *et al.*, 2008; Cockell *et al.*, 2010) que seguiram essa mesma linha e que demonstram que mesmo sem a utilização de ontologias ou reuso de vocabulários é possível integrar bases de dados. Além disso, das bases de dados utilizadas neste estudo, apenas a base de dados do ProtozoaDB será disponibilizada na nuvem do LOD, com um vocabulário mínimo, baseado no esquema do GUS. Em trabalhos futuros pode-se elaborar uma ontologia ou vocabulário mais apropriado para representar o domínio do conhecimento de genomas e proteomas de protozoários.

A abordagem de integração de bases de dados para identificar possíveis fármacos para reposicionamento obteve melhores resultados que outros trabalhos similares apresentados na literatura pesquisada, pois foi possível a identificação de uma extensa lista de fármacos já comercializados como possíveis de serem reposicionados para tratamento de doenças causadas por protozoários.

## 5.2 Módulo PostSemantic e integração através de semântica

Existem diversos métodos disponíveis na literatura para integrar bases de dados heterogêneas (Litwin *et al.*, 1990; Sheth & Larson, 1990; Kashyap & Sheth, 1996), isto é, que possuem estruturas de armazenamento diferentes. Entretanto, esses métodos necessitam do conhecimento prévio de cada estrutura para adequar a um único modelo, de forma que a integração das bases ocorra através de um esquema global. Mesmo com a utilização de descritores ou outros instrumentos terminológicos, a exemplo de ontologias, o problema de integrar bases heterogêneas ainda persiste.

Para Sheth & Larson (1990) o problema da heterogeneidade dos dados passa pelo uso de diferentes SGBD, pela diferença estrutural dos modelos de dados adotados, pela diferença sintática dos termos e pela diferença semântica dos termos. Isto evidencia a complexidade da tarefa de integração de bases de dados heterogêneas.

Além do problema da heterogeneidade, muitas vezes essas bases também estão distribuídas fisicamente em locais distintos ou até mesmo pertencem a grupos ou organizações diversos, dificultando a integração.

No processo de P&D de um novo fármaco ou de um novo uso para um fármaco há necessidade de integrar bases de dados de diversos domínios da informação (Slater

*et al.*, 2008) e, quase sempre, essas bases estão fisicamente em locais distintos e pertencem a grupos de pesquisas distintos.

Na biologia há uma predominância do uso de bases de dados relacionais (Kasprzyk & Smedley, 2006), porém os esquemas são próprios de cada grupo de pesquisa. Com o advento da internet, várias organizações passaram a disponibilizar seus dados de forma pública em páginas ou sítios. A adoção do uso de semântica na web (Berners-Lee *et al.*, 2001) permite que essas páginas ou sítios sejam acessados tanto por humanos quanto por programas de computador (agentes). Aliou-se a isso o uso dos padrões RDF e URI, disponibilizando os dados na nuvem do LOD (Berners-Lee, 2006). No gráfico mais atual do LOD (Figura 1.14) é possível observar que já existe uma grande quantidade de bases de dados biológicas disponibilizadas para consultas públicas e para integração com outras bases que seguem as diretrizes do LOD. Desta forma, diferentemente das bases relacionais, não há necessidade do conhecimento total da estrutura de cada base, mas sim do vocabulário ou ontologia utilizados pelas bases, afim de possibilitar a interoperabilidade ou até mesmo a integração das mesmas.

Entretanto, converter bases de dados já existentes, denominadas de bases legadas, para o padrão RDF, seguindo as premissas do LOD, não é uma tarefa trivial. Além disso, o volume e a heterogeneidade dos dados na biologia é outro fator desafiador (McEntire & Stevens, 2006).

Neste estudo procurou-se criar um mecanismo para tornar menos laboriosa a conversão de bases legadas para o padrão RDF. O módulo denominado PostSemantic, desenvolvido neste estudo, disponibiliza objetos de bancos de dados (Tabela 4.1) para a conversão de dados relacionais (tuplas) para o padrão RDF (triplas), sem a necessidade de construções de esquemas globais ou processos de ETL, que apesar de

alguns benefícios como captura de proveniência, são custosos e próprios para cada solução. Essa conversão se baseia em um processo controlado de mapeamento das informações que se deseja converter, ficando a critério do usuário a utilização de alguma ontologia ou vocabulário. Apesar de ter sido utilizado um vocabulário próprio para a integração das bases de dados neste estudo, recomenda-se fortemente o reuso de ontologias e vocabulários de forma a garantir uma integração melhor com as outras bases do LOD. O módulo desenvolvido ainda é capaz de gerenciar pontualmente a migração de cada tabela, com a adoção de URI específicos para cada campo, além de contemplar a conversão de chaves primárias e chaves estrangeiras.

Neste estudo, foram utilizados processos de ETL para as bases consideradas externas, ou seja, todas as bases de dados com exceção da base do ProtozoaDB. Isso em razão dessas bases ainda não estarem disponíveis na nuvem do LOD, ou em padrão RDF. O esforço deste trabalho em integrar com semântica e seguindo as diretrizes do LOD é em razão da visão de que em um futuro próximo outras bases estarão disponibilizadas na nuvem do LOD, permitindo consultas e interoperabilidade entre elas.

Já existem na literatura algumas soluções que trabalham com modelos relacionais com suporte à semântica. O primeiro projeto documentado é o D2R-Server (Bizer & Cyganiak, 2009) que converte modelos relacionais em padrão RDF através de um arquivo de configuração, escrito com uma linguagem própria (D2R-Language). Além disso, este projeto disponibiliza um servidor web com suporte aos protocolos HTTP e SPARQL, permitindo consultas à base. O núcleo desse projeto é o arquivo de configuração que é utilizado para converter consultas SPARQL em consultas SQL e para converter as tuplas retornadas em triplas de informação. Esse arquivo é elaborado através de uma linguagem própria, denominada D2R *Language*, o que dificulta o trabalho de mapeamento e conversão dos dados.

Outro projeto que utiliza o princípio de conversão de SQL é o *Triplify* (Auer *et al.*, 2009). A partir de uma consulta SQL, ou mais precisamente de uma visão, o programa é capaz de gerar triplas em diversos formatos, realizando uma busca no metadados da base de dados. Essa solução não disponibiliza uma interface para consultas SPARQL, limitando o uso de termos semânticos para consulta à base.

Outra técnica utilizada para oferecer suporte semântico às bases relacionais é criar objetos de banco de dados, tais como funções e procedimentos, que simulam uma consulta semântica (Pan & Heflin, 2003; Levshin & Markov, 2009). Esse tipo de solução também não permite consultas SPARQL, além de não gerarem informações em padrão RDF.

Por fim, Garrote & García (2011) construíram uma biblioteca de funções para permitir a construção de serviços web baseados em bases relacionais para prover triplas para Dados Ligados (*Linked Data*).

A solução adotada neste estudo converte os dados relacionais para o padrão RDF, persistindo o resultado e permitindo consultas SPARQL através da disponibilização desses dados em um servidor. A linguagem SPARQL permite consultas às bases de dados em padrão RDF, incluindo em sua sintaxe relações semânticas, tais como: relações entre termos, relações de sinonímia, entre outras. O módulo PostSemantic é precursor na disponibilização de um ferramental para o SGBD PostgreSQL com a finalidade de converter e disponibilizar dados em padrão RDF.

O módulo desenvolvido nesta tese demonstrou uma alternativa melhor que as alternativas apresentadas na literatura consultada para conversão de modelos relacionais em padrão RDF, sem a necessidade de conhecimento de linguagens para conversão e nem da execução de programas de conversão.

## 5.3 Dados em RDF e Nuvem de dados

A nuvem de dados (Figura 4.1) criada neste estudo foi fruto da integração entre as bases de dados utilizadas e do vocabulário utilizado. Tal vocabulário foi, intencionalmente, elaborado especificamente para este estudo com a finalidade de comprovar que, mesmo sem o reuso de vocabulários e ontologias, é possível integrar bases de dados com o mínimo de acoplamento possível e conseguir gerar novas informações para inferências científicas, assim como também proposto no projeto Bio2RDF (Belleau *et al.*, 2008).

Apesar do conjunto de dados ter sido disponibilizado para consultas públicas, apenas a base de dados em padrão RDF do ProtozoaDB será integrada à nuvem do LOD, em virtude de pertencer ao grupo deste estudo.

A nuvem de dados criada neste estudo é similar às criadas em outros trabalhos descritos na literatura pesquisada e serviu para demonstrar o potencial de exploração de nuvem de dados integradas por semântica, através do uso do padrão RDF e descritores URI, tal como a nuvem do LOD.

## 5.4 Tradução de termos computacionalmente complexos

Os termos utilizados pelos pesquisadores nem sempre são computacionalmente simples, assim como, por exemplo, o termo **proteína**. Termos dessa natureza normalmente são expressos, em RDF, como sujeito ou predicado, sendo simples de serem processados por consultas SPARQL. Contudo, há termos utilizados pelos pesquisadores que possuem intrinsecamente uma complexidade computacional. Os três termos criados para este estudo (Tabela 4.3) são exemplos de termos computacionalmente

complexos, ou seja, que precisam antes serem traduzidos por uma consulta SPARQL para depois serem processados.

A solução formulada neste estudo traduz os termos mapeados no momento da realização da consulta à base, sendo o resultado persistido na mesma. Essa solução permite que o usuário utilize uma linguagem mais próxima de seu domínio (jargão) e se abstenha de termos mais computacionais. Além disso, a solução também permite que as ligações entre as bases de dados se realizem no momento da consulta, através da renomeação de um termo previamente mapeado. Isso facilita a utilização de vocabulários próprios e permite a normalização de termos diferentes que possuam mesmo significado. Embora esse processamento aumente o tempo de resposta de uma consulta, os benefícios de se utilizar termos mais conhecidos pelos usuários compensa esse aumento.

A abordagem utilizada neste estudo para tradução de termos em tempo de execução é original e não foi encontrada técnica similar na literatura pesquisada.

## 5.5 Informações quantitativas

A Tabela 4.4 exibe os dados brutos de cada base de dados convertida para o padrão RDF e dados sobre as relações entre esses dados. Nos 22 protozoários do estudo foram encontrados 386 proteínas ortólogas às proteínas de alvos para fármacos já comercializados. No intuito de prover uma quantidade maior de informações para a relação final de fármacos com potencial para reposicionamento, foram considerados apenas os fármacos cujos alvos fossem ortólogos às proteínas de protozoários, que tivessem fenótipos associados e que pertencessem a pelo menos uma via metabólica. Sendo assim, foram encontrados 150 fármacos com essas características (Apêndice D). Essas informações tornam possível uma análise mais detalhada dos mecanismos

de ação, vias metabólicas das proteínas-alvo, características das proteínas-alvo (fenótipos) e demais dados genômicos e proteômicos de interesse para o estudo de reposicionamento de fármacos. Não foi encontrado na literatura nenhum resultado próximo ao encontrado por este estudo. Os trabalhos relacionados avaliam estudos de caso e sugerem a aplicação da metodologia para encontrar novos resultados. Com a divulgação deste trabalho será possível, partir da lista de fármacos e alvos identificados, um estudo mais detalhado de cada um, com a finalidade de reposicioná-los para tratamento de doenças causadas por protozoários.

Não foi encontrado resultado similar na literatura pesquisada.

## 5.6 Informações qualitativas

Com a finalidade de explorar a nuvem de dados para obter além da lista de 150 fármacos, outras informações relevantes para o processo de análise de quais fármacos são realmente interessantes para reposicionamento, diversas consultas foram realizadas com o montante geral de fármacos com alvos com ortólogos em protozoários (394).

A Figura 4.2 utiliza a informação da base de dados do DrugBank para organizar os 394 fármacos de acordo com os seus organismos afetados. Apesar de grande parte dos fármacos não terem o registro dessa informação, pode-se verificar que existem fármacos que afetam vírus, bactéria, levedura, fungo, além do humano. O fato de terem sido encontradas proteínas de protozoários ortólogas aos alvos originais desses fármacos, com tal diversidade de organismos, ocorre, provavelmente, em razão da conservação de algumas proteínas-alvo entre esses organismos (Ciccarelli *et al.*, 2006).

Em relação à categoria dos fármacos (Figura 4.3), classificação dada também pelo DrugBank, é importante frisar que um fármaco pode estar associado a mais de uma categoria em razão de suas proteínas-alvo. A categoria mais representativa nos 394 fármacos foi a de Suplementos de Dieta e micronutrientes. Essas categorias de fármacos agem normalmente em alvos ligados ao metabolismo de carboidratos e à produção de energia (ATP), vias que são conservadas entre os organismos conhecidos. Isso, em tese, não caracteriza essas proteínas como bons alvos de fármacos, uma vez que qualquer fármaco que altere essas vias nos protozoários poderia afetar também nos humanos.

Dentre as 22 espécies de protozoários do estudo em apenas nove espécies foram encontrados ortólogos de alvos de fármacos (Figura 4.4). Em *Trypanosoma cruzi* foi encontrado o maior número de ortólogos, devido, talvez, ao grande número de parálogos que esse organismo possui (El-Sayed *et al.*, 2005).

Para uma compreensão melhor sobre o significado de cada fenótipo, se faz necessário recorrer ao SGD para avaliar o experimento realizado. No Anexo A encontra-se a tabela de todos os fenótipos da base de *S. cereviasae*, organizados de acordo com a hierarquia. Desta forma, a Figura 4.5 mostra que dentre os alvos dos 394 fármacos, o fenótipo mais encontrado foi o de Resistência diminuída a produtos químicos. Este fenótipo está relacionado à classe de fenótipos de Resistência ao estresse e é necessário observar o experimento realizado para saber a qual substância química o gene é afetado. Esta análise será realizada em trabalhos futuros.

Para fins desta discussão, o conceito de fenótipo do gene foi estendido para fenótipo da proteína codificada pelo gene, denominada simplesmente de fenótipo da proteína. A Figura 4.6 apresenta o número de fenótipos por espécie de protozoário e o número de proteínas associadas a algum fenótipo. Novamente a espécie *T. cruzi* aparece

como a que mais possui proteínas associadas a fenótipos, provavelmente em razão do grande número de parálogos desse organismo (El-Sayed *et al.*, 2005).

A Figura 4.7 mostra a informação de fenótipos associados às proteínas de protozoários ortólogas aos alvos dos 394 fármacos. Pode-se notar que os possíveis alvos de fármacos em protozoários estão associados a vários fenótipos.

Este estudo contemplou 393 mapas de vias metabólicas, dos quais 255 (65%) possuem pelo menos um representante dos 22 protozoários do ProtozoaDB. Observa-se pela Figura 4.8 que a maior parte das proteínas de protozoários encontra-se no mapa denominado Metabolismo (map01100).

A Figura 4.9 mostra o percentual de mapas de vias metabólicas em que pelo menos uma proteína do protozoário está presente. Com as informações geradas por este estudo é possível remontar algumas vias metabólicas de protozoários. *T. cruzi* foi o organismo com maior número de mapas identificados.

Foram encontrados 216 fármacos que possuem alvos com ortólogos em protozoários e cuja as proteínas dos protozoários foram associadas a pelo menos um fenótipo (Figura 4.10). Observa-se que para esses, o fenótipo mais representativo foi o de Viabilidade, que possui relação direta com a essencialidade do gene e o segundo mais representativo foi o fenótipo de Resistência a produtos químicos diminuída. Não foi encontrado resultado similar na literatura pesquisada.

Dos 394 fármacos inicialmente identificados foram encontrados 150 fármacos que possuem alvos ortólogos a proteínas de protozoários e essas proteínas possuem fenótipo associado e pertencem a pelo menos uma via metabólica descrita no Kegg (Figura 4.11). A Figura 4.12 mostra os alvos identificados nos protozoários que são ortólogos

aos alvos dos 150 fármacos. Os dois alvos mais representativos estão ligados diretamente à produção de energia, corroborando os demais resultados que apontaram para vias de metabolismo e categorias de fármacos. Embora o estudo tenha inciado com 22 protozoários, esses 150 fármacos apontaram para apenas nove espécies, dentro de 4 gêneros: *Cryptosporidium*, *Plasmodium*, *Theileria* e *Trypanosoma*.

A lista dos 150 fármacos é um ponto inicial para pesquisas futuras e não foi encontrado resultado similar na literatura pesquisada. Neste estudo foram feitas análises para os quatro gêneros de protozoários identificados nesses 150 fármacos.

## 5.7 Possíveis alvos para o gênero *Cryptosporidium*

A doença causada pelo protozoário do gênero *Cryptosporidium* é denominada criptosporidiose e é caracterizada como uma doença que causa diarreia aquosa e não sanguinolenta. Pacientes imunocompetentes apresentam cura espontânea após uma ou duas semanas, enquanto que em pacientes imunocomprometidos pode levar a óbito (O'Donoghue, 1995).

A proteína-alvo no protozoário identificada nesse estudo é a Histona deacetilase (HDAC). Essa proteína possui uma atividade oposta às histonas acetiltransferases, diminuindo os níveis de acetilação e, em geral, estão associadas a expressão gênica. Quanto maior for o nível de acetilação, maior a atividade transcricional. Por outro lado, níveis baixos de acetilação estão associados com repressão gênica (de Ruijter *et al.*, 2003).

O alinhamento entre a proteína-alvo do fármaco, uma histona deacetilase 2 e a proteína-alvo do protozoário (Figura 4.13) demonstra vários domínios conservados entre essas proteínas ortólogas. Após o alinhamento com outras Histonas deacetilase

de protozoários, percebe-se que, apesar do alto grau de similaridade, existem domínios conservados apenas em histonas de protozoários, conforme destacado na Figura 4.14.

Estudos recentes têm avaliado o papel das HDAC em *Cryptosporidium* (Farmar *et al.*, 2009) e associado esse alvo a tratamento contra a malária (Andrews *et al.*, 2009).

Dados gerados neste estudo mostram que o fármaco detectado para reposicionamento, o Vorinostat, pertence a quatro categorias, sendo uma delas a de inibidor de enzima. Os alvos desse fármaco, em sua maioria, são HDAC de humanos (HDAC1, HDAC2, HDAC3, HDAC6 e HDAC8), tendo sido o HDAC2 o ortólogo a proteína do protozoário. Esse fármaco inibe a atividade enzimática das HDAC, causando hipoacetilação e consequente repressão transcricional por induzir uma estrutura condensada da cromatina.

Dentre os fenótipos encontrados para essa proteína destacam-se dois: RNA acumulação diminuída e viável. O primeiro em razão da repressão transcricional e o segundo por estar associado a processos de apoptose da célula, conforme descritos nos mapas de vias metabólicas encontrados no estudo, principalmente o mapa 05220 . Estudos têm comprovado que inibidores de HDAC induzem a apoptose da célula e sugerem o uso para tratamento em diversos cânceres (Ganslmayer *et al.*, 2012; Dasmahapatra *et al.*, 2012).

## 5.8 Possíveis alvos para o gênero *Plasmodium*

O gênero *Plasmodium* é responsável pela malária, doença que acomete milhões de pessoas em todo o mundo. Essa doença é caracterizada pelo rompimento dos

glóbulos vermelhos em razão da reprodução exacerbada do protozoário dentro dessas células.

O alvo em protozoário identificado por esse estudo é a Valina-tRNA ligase (VARS), também denominada Valil-tRNA sintetase. VARS é uma enzima que catalisa a ligação de um aminoácido com a molécula de tRNA e pertence a classe 1A da aminoacil-tRNA ligase. As aminoacil-tRNA ligase são responsáveis pelo processo de aminoacilação, isto é, produção de moléculas de tRNA com a extremidade 3' com uma ligação covalente a um aminoácido.

O alinhamento par-a-par mostrou diversas regiões conservadas entre a proteína-alvo do fármaco e a proteína-alvo do protozoário (Figura 4.15). No entanto, quando utilizadas outras proteínas homólogas à proteína do protozoário (Figura 4.16), não foi possível estabelecer domínios só de protozoários.

O fármaco identificado neste estudo foi L-Valine. Este fármaco age no alvo l-valine, um aminoácido essencial que participa da reação catalítica da enzima VARS. As categorias do fármaco ainda incluem Suplementos de Dieta e Micronutrientes. Em tese, tais fármacos são utilizados para hiperexpressar seus alvos. Neste caso, garantir a presença do aminoácido valina para o processo catalítico da enzima. Isto pode ser comprovado pela indicação do fármaco e pelo mecanismo de ação.

Dos fenótipos identificados, dois se relacionam a inibidores desta enzima: Progressão do ciclo celular anormal e inviabilidade. Um fenótipo está relacionado a potencialização do aminoácido Valina: Aptidão competitiva aumentada (*fitness increased*).

Apesar de não poder ser indicado para o tratamento da malária, a informação deste estudo pode ser vista como um ponto de partida para análises mais detalhadas sobre a inibição da enzima VARS para o tratamento da malária.

## 5.9 Possíveis alvos para o gênero *Theileria*

Os protozoários do gênero *Theileria* deste estudo incluem o *T. parva* e o *T. annulata*. Ambos causam doenças que acometem gado bovino, sendo denominadas de febre da costa oeste e theileriose, respectivamente. São caracterizadas por febre alta e linfonodos próxima à picada do carrapato, vetor desses protozoários.

A proteína-alvo do protozoário identificada neste estudo foi a Fosforicolina citidililtransferase, também denominada de colina-fosfato citidililtransferase. Pertence a família das transferases e é uma enzima que participa do metabolismo de aminofosfato. Essa enzima é importante para o processo de manutenção da integridade das membranas celulares. A falta de colina é associada a processos de apoptose.

O alinhamento par-a-par (Figura 4.17) mostrou poucos domínios conservados entre as proteínas ortólogas, tendo apenas 27,45% de identidade. Quando observado o alinhamento com outras proteínas ortólogas de protozoários, verifica-se diversas regiões conservadas apenas em protozoários (Figura 4.18). Isso pode sugerir regiões específicas e bons marcadores. No entanto, estudos mais detalhados das estruturas dessas proteínas precisam ser realizados para resultados mais conclusivos.

O fármaco identificado no estudo é o Choline que pertence à categoria de Agentes lipotrópicos, entre outras. Seu mecanismo de ação é basicamente fornecer colina para o organismo, de forma a garantir a manutenção da integridade das membranas celulares.

Se por um lado esse fármaco potencializa seu alvo ao fornecer colina para o processo da enzima, por outro lado, possíveis inibidores dessa enzima são considerados bons alvos para fármacos (de Freitas-Junior *et al.*, 2012).

Dos fenótipos identificados dois destacam-se pela relação com exposto acima: viável e sensibilidade ao calor aumentada. Ambos com relação direta à manutenção da membrana celular.

## 5.10 Possíveis alvos para o gênero *Trypanosoma*

Neste estudo foram avaliadas duas espécies desse gênero: *T. cruzi* e *T. brucei*. A primeira causa a doença de Chagas e a segunda a doença do sono.

A proteína-alvo identificada foi a Gamma-glutamilcisteina sintetase que é uma enzima que participa da biossíntese da glutationa, com importância antioxidante, prevenindo danos causados por radicais livres e peróxidos (Pompella *et al.*, 2003).

O alinhamento par-a-par (Figura 4.19) mostrou regiões conservadas entre as proteínas-alvo. Porém, o alinhamento com outras proteínas ortólogas de protozoários mostrou que existem regiões conservadas somente para os protozoários (Figura 4.20), demonstrando a divergência entre essas proteínas-alvo.

O fármaco identificado foi o L-Cysteine que pertence as categorias de suplemento nutricional, suplemento de dieta, micronutriente e aminoácido não essencial. Embora esse tipo de fármaco potencialize o alvo, estudo recente demonstrou que a gamma-glutamilcisteina sintetase pode ser um alvo promissor para fármacos contra *T. cruzi* (Olin-Sandoval *et al.*, 2012).

Para o alvo identificado, a gamma-glutamilcisteina sintetase, foram encontrados diversos fenótipos, incluindo o fenótipo viável e resistência a estresse oxidativo aumentada.

## 5.11 Validação dos resultados

De posse da lista dos 150 fármacos e com os potenciais alvos em protozoários identificados foi realizada uma busca no PubMed com a finalidade de verificar a relação entre esses potenciais alvos e estudos para fármacos em protozoários. A Tabela 4.9 mostra, por gênero de protozoário, a quantidade de artigos encontrados, para a busca que levou em consideração apenas o primeiro resultado encontrado. De certo que nem todo artigo relacionado com os termos buscados tiveram relevância para este estudo. Isso deve-se ao fato de que a busca no PubMed é realizada por sintaxe e que, por vezes, esses termos podem estar em um mesmo artigo para relacionar-se com outro assunto diferente do foco deste estudo.

Entretanto, a leitura realizada nos 38 resumos revelou que 60% dos artigos possuem relevância com o assunto deste estudo, tendo sido incluídos no Apêndice E.

Para o gênero *Cryptosporidium* foram encontrados 5 artigos, tendo 4 (80%) artigos relevância com este estudo. O artigo destacado descreve a relação entre as proteínas kinase e alvos para fármacos em protozoários. Os autores realizaram experimentos onde foram sintetizados uma série de inibidores para proteína kinase dependente de cálcio (CDPK1) à base de benzoilbenzimidazole, obtendo resultados que indicam que esses alvos podem ser bons alvos para fármacos anticriptospiridose (Zhang *et al.*, 2012).

Foram recuperados 14 artigos para o gênero *Plasmodium*, sendo que 11 (79%) possuem relação com este estudo. O termo procurado para o gênero *Plasmodium* incluiu o alvo serina hidroxymetiltransferase que foi encontrado neste estudo entre os alvos de protozoários homólogos aos alvos dos 150 fármacos. Sopitthummakhun *et al.* (2012) utilizaram técnicas de biologia molecular para testar inibidores da serina

hidroxymetiltransferase de *Plasmodium* com a finalidade de prover informações úteis para fabricação de fármacos antimalariais.

Para o gênero *Theileria* foi encontrado apenas um artigo relacionado com os termos buscados. Esse artigo descreve a relação entre a proteína kinase serina/treonina e o fármaco buparvaquona (Shayan & Ahmed, 1997).

Foram encontrados 18 artigos para o gênero *Trypanosoma*, dos quais 7 (39%) possuem relação com este estudo. No artigo selecionado, os autores discutem sobre os níveis de inibição de alvos em *T. cruzi*, entre eles a gamma-glutamilcisteina sintetase, de forma a afetar o metabolismo de tripanotiona. Com isso, inferem que esse alvo pode ser um bom alvo para fármacos (Olin-Sandoval *et al.*, 2012).

O objetivo principal deste estudo foi a identificação de fármacos já comercializados que possuíssem potencial para serem reposicionados para tratamento de doenças causadas por protozoários. Após a integração semântica das bases de dados do estudo, foram encontrados inicialmente 394 fármacos com esse potencial. Entretanto, com a finalidade de prover mais informações acerca dos alvos desses fármacos, dois filtros foram realizados: o cruzamento das informações dos alvos com os fenótipos e a participação desses alvos em alguma via metabólica. Com isso, 150 fármacos foram identificados como potenciais para o reposicionamento, tendo as características de terem alvos ortólogos às proteínas de protozoários, com fenótipos associados e participantes de uma ou mais vias metabólicas.

Cabe salientar que nas análises feitas neste estudo em relação aos gêneros de protozoários, apenas para o gênero *Theileria* foi identificado um fármaco que age diretamente em um alvo importante no protozoário, cabendo estudos mais aprofundados. Nos três demais gêneros, a metodologia empregada conseguiu identificar alvos im-

portantes, já descritos na literatura. Isso se deu em função da base de fármacos e alvos escolhida, o Drugbank, categorizar aminoácidos essenciais e suplementos de dieta como sendo fármacos. Um filtro para retirar tais categorias poderia solucionar esse problema. Entretanto, o objetivo deste estudo foi de demonstrar a capacidade da metodologia em garimpar a informação de possíveis fármacos para reposicionamento utilizando as informações originais das bases de dados, uma vez que, a partir da disponibilização dessas bases na nuvem do LOD, consultas poderão ser realizadas diretamente nas bases originais.

# Capítulo 6

# Conclusão

1. Neste trabalho foi proposta a identificação de novos usos para fármacos (já comercializados) visando o tratamento de doenças causadas por protozoários, utilizando o conceito biológico de homologia, assim, a integração de bases de dados biológicas permitiu que fossem realizadas inferências de ortologias entre proteínas de humanos e protozoários sem a necessidade da execução de *software* de bioinformática para inferência de ortologia.

2. Nessa mesma linha, foram realizadas inferências de ortologias entre proteínas de *S. cerevisae* e protozoários para mapear fenótipos que poderiam ser extrapolados em protozoários. Essa abordagem se mostrou útil contribuindo na identificação de fenótipos associados a alvos de fármacos.

3. A integração de bases de dados biológicas para P&D de novos fármacos, quer seja pelo método *de novo*, quer seja por reposicionamento, é uma realidade proposta pela comunidade científica, assim como o uso de semântica para integração de bases de dados. Este estudo corroborou outros trabalhos da literatura que demonstram que a integração através de semântica possui características próprias que enriquecem as ligações entre as bases de dados.

4. A utilização das premissas do LOD para conversão e integração das bases de

dados permitiu o enriquecimento das informações na nuvem do LOD, com o acréscimo de dados de genoma e proteoma de 22 protozoários patogênicos.

5. Os 150 fármacos listados, originalmente identificados neste estudo, possuem potencial para serem reposicionados para tratamento de doenças causadas por protozoários. Algumas das proteínas-alvo desses fármacos já foram associadas a bons alvos pela literatura recente, entretanto outras ainda não foram relacionadas e constituem um ponto de partida para estudos mais aprofundados de alvos e reposicionamento de fármacos.

# Referências Bibliográficas


Andrews KT, Tran TN, Wheatley NC, Fairlie DP. Targeting histone deacetylase inhibitors for anti-malarial therapy. *Current topics in medicinal chemistry*. 2009, 9(3):292–308.

Aronson JK. Old drugs – new uses. *British Journal of Clinical Pharmacology*. 2007, page 3.

Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews*. 2004, 3:673–683.

Ashburner M, Dwight, Ringwald M. Gene Ontology : tool for the. *Nature America Inc*. 2000, 25(may):25–29.

Auer S, Dietzold S, Lehmann J, Hellmann S, Aumueller D. Triplify - Light-Weight Linked Data Publication from Relational Databases. In *International World Wide Web Conference Com- mittee (IW3C2*. 2009.

Baldauf SL. The Deep Roots of Eukaryotes. *Science*. 2003, 300(5626):1703–1706.

Barbosa CP, Biancardi C, Silvestre LJ. Integração de Dados Heterogêneos em Ambiente Web. Technical report. 2001.

Bauer F, Kaltenböck M. *Linked Open Data : The Essentials*. edition mono/monochrom, Vienna, Austria. 2012.

Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup

to build bioinformatics knowledge systems. *Journal of biomedical informatics*. 2008, 41(5):706–16.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucl. Acids Res.*. 2011, 39(suppl 1):D32–D37.

Berners-Lee T. What the Semantic Web can represent. 1998. [*online*]. Disponível em `http://www.w3.org/DesignIssues/RDFnot.html`. Acessado em novembro de 2012.

Berners-Lee T. Linked Data. 2006. [*online*]. Disponível em `http://www.w3.org/DesignIssues/LinkedData.html`. Acessado em novembro de 2012.

Berners-Lee T, Connolly D, Swick RR. Web Architecture: Describing and Exchanging Data. 1999. [*online*]. Disponível em `http://www.w3.org/1999/04/WebData`. Acessado em novembro de 2012.

Berners-Lee T, Hendler J, Lassila O. The Semantic Web. *Scientific American*. 2001.

Bizer C, Cyganiak R. Publishing Databases on the Semantic Web. Technical report, Freie Universitat Berlin. 2009.

Bizer C, Heath T, Ayers D, Raimond Y. Interlinking Open Data on the Web. In *ESWC*. 2007.

Cançado JR. Long term evaluation of etiological treatment of Chagas disease with benznidazole. *Revista do Instituto de Medicina Tropical de São Paulo*. 2002, 44:29–37.

Chen B, Ding Y, Wang H, Wild DJ, Dong X, Sun Y, *et al*. Chem2Bio2RDF: A Linked Open Data Portal for Systems Chemical Biology. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 2010, pages 232–239.

Chen F, Mackey AJ, Jr CJS, Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*. 2006, 34(suppl 1:D363–368.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, *et al*. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res*. 2012, 40 (Database):D700–5.

Chirac P, Torreele E. Global framework on essential health R&D. *Lancet*. 2006, 367(9522):1560–1561.

Chong CR, Sullivan Jr DJ. New uses for old drugs. *Nature*. 2007, 448(August):645–646.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*. 2006, 311(5765):1283–7.

Clive S, Gardiner J, Leonard RC. Miltefosine as a topical treatment for cutaneous metastases in breast carcinoma. *Cancer Chemother Pharmacol*. 1999, 44 (Suppl):29–30.

Cockell SJ, Weile J, Lord P, Wipat C, Andriychenko D, Pocock M, *et al*. An integrated dataset for in silico drug discovery. *Journal of integrative bioinformatics*. 2010, 7(3):1–13.

Coura JR. *Síntese das doenças infecciosas e parasitárias*. Editora Guanabara Koogan. 2008.

Cuadra A, Cutanda MM, Fuentes-Lorenzo D, Sanchez L. A semantic web-based integration framework. In *Next Generation Web Services Practices (NWeSP)*. 2011.

Dasmahapatra G, Patel H, Nguyen TK, Attkisson E, Grant S. PLK1 inhibitors synergistically potentiate HDAC inhibitor lethality in IM -sensitive or -resistant BCR/ABL+ leukemia cells in vitro and in vivo. *Clin Cancer Res.*. 2012.

Davidson SB, Crabtree J, Brunk B, Schug J, Tannen V, Overton C, *et al*. K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources. 2000.

Dávila AMR, Mendes PN, Wagner G, Tschoeke DA, Cuadrat RRC, Liberman F, *et al*. ProtozoaDB: dynamic visualization and exploration of protozoan genomes. *Nucl. Acids Res*. 2008, 36(Database issue):D547–52.

de Freitas-Junior PRG, Catta-Preta CMC, Andrade IDS, Cavalcanti DP, de Souza W, Einicker-Lamas M, *et al*. Effects of miltefosine on the proliferation, ultrastructure, and phospholipid composition of Angomonas deanei, a trypanosomatid protozoan that harbors a symbiotic bacterium. *FEMS microbiology letters*. 2012, 333(2):129–37.

de Ruijter AJM, van Gennip AH, Caron HN, Kemp S, van Kuilenburg ABP. Histone deacetylases (HDACs): characterization of the classical HDAC family. *The Biochemical journal*. 2003, 370(Pt 3):737–49.

DNDi. Drugs for Neglected Diseases initiative. 2010. [*online*]. Disponível em `http://www.dndi.org.br`. Acessado em novembro de 2012.

Ekins S, Williams AJ, Krasowski MD, Freundlich JS. In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discovery Today*. 2011, 00:13.

El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, *et al*. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science (New York, N.Y.)*. 2005, 309(5733):409–15.

Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res*. 2002, 30(7):1575–1584.

Farmar JG, Nika H, Che Fy, Weiss L, Angeletti RH. IVICAT for the Masses : an Improved Technique for Permethylation of Peptides Analysis of Histone Methylation via Quaternization. *Journal of biomolecular techniques*. 2009, 10:285–292.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, *et al*. Ensembl 2011. *Nucleic Acids Res*. 2011, 39(Databas:D800–D806.

Ganslmayer M, KONTUREK P, HEROLD C, NEURATH MF, ZOPF S. Antitumoral efficacy of four histone deacetylase inhibitors in hepatoma in vitro and in vivo. *Anticancer Research*. 2012.

Garrote A, García MNM. RESTful writable APIs for the web of Linked Data using relational storage solutions. In *LDOW*. 2011.

Goto S, Bono H, Ogata H, Fujibuchi W, Nishioka T, Sato K, *et al*. Organizing and computing metabolic pathway data in terms of binary relations. In *Pac Symp Biocomput.*. 1997, pages 175–86.

Gruber TR. A translation approach to portable ontology specifications. *KNOWLEDGE ACQUISITION*. 1993, 5:199–220.

Guido RVC, Andricopulo AD, Oliva G. Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. *Estudos Avançados*. 2010, 24:81–98.

Guido RVC, Oliva G, Andricopulo AD. Virtual screening and its integration with modern drug design technologies. *Curr Med Chem*. 2008, 15(1):37–46.

Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, *et al*. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic acids research*. 2008, 36(Database issue):D919–22.

Halevy AY. Data Integration: A status report. In *10th Conference on Database Systems for Business, Technology and Web*. 2003.

Hausenblas M, Karnstedt M. Understanding Linked Open Data as a Web-Scale Database. *2010 Second International Conference on Advances in Databases, Knowledge, and Data Applications*. 2010, pages 56–61.

Hors AJL, Nally M, Speicher SK. Using read write Linked Data for Application Integration âĂŞ Towards a Linked Data Basic Profile. In *LDOW 2012*. 2012.

Inmon WH. *Como construir o data warehouse*. Editora Campus, RJ. 1997.

Ito T, Ando H, Suzuki T, Ogura T, Hotta K, Imamura Y, *et al*. Identification of a Primary Target of Thalidomide Teratogenicity. *Science*. 2010, 327(5971):1345–1350.

Kashyap V, Sheth A. Semantic and Schematic Similarities between Database Objects : A Context-based approach. *The International Journal on Very Large Data Bases*. 1996, 5(4):276–304.

Kasprzyk A, Smedley D. 2006. *Database Management*, chapter 16, pages 389–401. CRC Press.

Kim W, Seo J. Classifying schematic and data heterogeinity in multidatabase systems. *IEEE Computer*. 1991, 24(12):12–18.

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, *et al*. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*. 2011, 39 (Database):D1035–41.

Kohara Y, Akiyama K, Isono K. The physical map of the whole E. coli chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell*. 1987, 50:495–508.

Koonin EV. Orthologs, Paralogs, and Evolutionary Genomics. *Annu. Rev. Genet.*. 2005, 39:309–338.

Kotpal RL. *Modern Text Book of Zoology Invertebrates*. Editora Rakesh Kumar Rastogi. 2010.

Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*. 2010, 6:343.

León D, Markel S. 2006. *In Silico Technologies in Drug Target Identification adn Validation*, chapter Introducti, pages 1–9. CRC Press.

Levshin DV, Markov AS. Algorithms for Integrating PostgreSQL with the Semantic Web. *Programming and Computer Software*. 2009, 35:136–144.

Li L, Stoeckert CJJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*. 2003, 13(9):2178–2189.

Litwin W, Mark L, Roussopoulos N. Interoperability of multiple autonomous databases. *ACM Comput. Surv.*. 1990, 22(3):267–293.

Liu Y, Liu X, Yang L. Analysis and design of heterogeneous bioinformatics database integration system based on middleware. In *Information Management and Engineering (ICIME)*. 2010.

Lombardino JG, Lowe JA. A guide to drug discovery: The role of the medicinal chemist in drug discovery - then and now. *Nature Reviews Drug Discovery*. 2004, 3:853–862.

Matthews SJ, McCoy C. Thalidomide: A Review of Approved and Investigational Uses. *Clin Ther*. 2003, 25(2):342–395.

McEntire RA, Stevens R. Ontologies. In Léon D, Markel S, editors, *In silico Technologies in Drug Target Identification and Validation*, chapter 20, pages 451–480. CRC Press. 2006.

Moran M, Guzman J, Ropars AL, McDonald A, Jameson N, Omune B, *et al*. Neglected

Disease Research and Development: How Much Are We Really Spending? *PLoS Med*. 2009, 6(2):e1000030.

Morel CM, Serruya SJ, Penna GO, Guimarães R. Co-authorship Network Analysis: A Powerful Tool for Strategic Planning of Research, Development and Capacity Building Programs on Neglected Diseases. *PLoS Negl Trop Dis*. 2009, 3(8):e501.

Musen Ma, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, *et al*. The National Center for Biomedical Ontology. *Journal of the American Medical Informatics Association : JAMIA*. 2012, 19(2):190–5.

Neumann E. A life science Semantic Web: are we there yet? *Science's STKE : signal transduction knowledge environment*. 2005, 283:pe22.

Neumann EK, Miller E, Wilbanks J. What the semantic web could do for the life sciences. *Drug Discovery Today: BIOSILICO*. 2004, 2(6):228–236.

O'Donoghue PJ. Cryptosporidium and cryptosporidiosis in man and animals. *International journal for parasitology*. 1995, 25(2):139–95.

Olin-Sandoval V, González-Chávez Z, Berzunza-Cruz M, Martínez I, Jasso-Chávez R, Becker I, *et al*. Drug target validation of the trypanothione pathway enzymes through metabolic modelling. *FEBS J.*. 2012, 279(10):1811–33.

Pan Z, Heflin J. DLDB : Extending Relational Databases to Support Semantic Web Queries. In *PSSS*. 2003.

Pasquier C. Biological data integration using Semantic Web technologies. *Biochimie*. 2008, 90(4):584–594.

Pompella A, Visvikis A, Paolicchi A, Tata VD, Casini AF. The changing faces of glutathione, a cellular protagonist. *Biochemical Pharmacology*. 2003, 66(8):1499–1503.

Qu Xa, Gudivada RC, Jegga AG, Neumann EK, Aronow BJ. Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. *BMC bioinformatics*. 2009, 10 Suppl 5:S4.

Remm M, Storm CE, Sonnhammer ELL. Automatic clustering of otrhologs and inparalogs from pairwise species comparisons. *J. Mol. Biol.*. 2001, 314:1041–1052.

Richon AB. A Short History of Bioinformatics. 2012. [*online*]. Disponível em `http://www.netsci.org/Science/Bioinform/feature06.html`. Acessado em 2012.

Ruppert EE, Barnes RD. *Zoologia dos Invertebrados*. 1996, 6a. edition.

Sachs JD. Macroeconomics and Health: Investing in Health for Economic Development. Technical report, Organização Mundial da Saúde. 2001.

Shayan P, Ahmed J. Theileria-mediated constitutive expression of the casein kinase II-alpha subunit in bovine lymphoblastoid cells. *Parasitol Res.*. 1997, 83(6):526–32.

Sheskin J. Thalidomide in the treatment of lepra reactions. *Clin Pharmacol Ther*. 1965, 6:303–306.

Sheth A, Larson J. Federated Database Systems for Managing Distributed , Heterogeneous , and Autonomous Databases. *ACM Computing Surveys*. 1990, 22(3).

Slater T, Bouton C, Huang ES. Beyond data integration. *Drug Discovery Today*. 2008, 13:584–589.

Sopitthummakhun K, Thongpanchang C, Vilaivan T, Yuthavong Y, Chaiyen P, Leartsakulpanich U. Plasmodium serine hydroxymethyltransferase as a potential antimalarial target: inhibition studies using improved methods for enzyme production and assay. *Malaria journal*. 2012, 11:194.

Storm CV, Sonnhammer EL. Automated ortholog inference from phylogenetic trees and calculation of ortholog reliability. *Bioinformatics*. 2002, 18:92–99.

Sundar S, Rosenkaimer F, Makharia MK, Goyal AK, Mandal AK, Voss A, *et al*. Trial of oral miltefosine for visceral leishmaniasis. *Lancet*. 1998, 352(9143):1821–3.

Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett.*. 1999, 174(2):247–50.

Terrett NK, Bell AS, Brown D, Ellis P. Sildenafil (Viagra), a potent and selective inhibitor of type 5 CGMP Phosphodiesterase with utility for the treatment of male erectile dysfunction. 1996, 6(15):1819–1824.

The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*. 2012, 40(D1):D71–D75.

Uschold M, Gruninger M. Ontologies: Principles, methods and applications. *Knowledge Engineering Review*. 1996, 11:93–136.

von Eichborn J, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R. PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic acids research*. 2011, 39(Database issue):D1060–6.

Waterhouse AM, Procter JB, Martin DMa, Clamp M, Barton GJ. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)*. 2009, 25(9):1189–91.

WHO. Fact Sheet. 2012a. [*online*]. Disponível em `http://www.who.int/mediacentre/factsheets/fs259/en/index.html`. Acessado em novembro de 2012.

WHO. Fact Sheet. 2012b. [*online*]. Disponível em `http://www.who.int/mediacentre/factsheets/fs094/en/index.html`. Acessado em novembro de 2012.

WHO. Fact Sheet. 2012c. [*online*]. Disponível em `http://www.who.int/mediacentre/factsheets/fs340/en/index.html`. Acessado em novembro de 2012.

Wiederhold G. Interoperation, Mediation, and Ontologies. In *International Symposium on Fifth Generation Computer Systems*. 1994.

Zhang Z, Ojo KK, Johnson SM, Larson ET, He P, Geiger Ja, *et al*. Benzoylbenzimidazole-based selective inhibitors targeting Cryptosporidium parvum and Toxoplasma gondii calcium-dependent protein kinase-1. *Bioorganic & medicinal chemistry letters*. 2012, 22(16):5264–7.

# Apêndice A

# Consulta SQL na base do ProtozoaDB

Consulta construída para recuperar os dados dos proteomas dos 22 protozoários na base de dados do ProtozoaDB.

```
SELECT
  aasequence.aa_sequence_id,
  aasequence.description,
  aasequence.source_id,
  taxonname."name",
  aafeature.na_feature_id,
  genefeature.source_id,
  genefeature.gene,
  dbref.primary_identifier
FROM
  dots.aasequence,
  sres.taxonname,
  dots.aafeature,
  dots.genefeature,
  dots.dbrefnafeature,
  sres.dbref
```

```
WHERE
  aasequence.taxon_id = taxonname.taxon_id AND
  aasequence.aa_sequence_id = aafeature.aa_sequence_id AND
  aafeature.na_feature_id = genefeature.na_feature_id AND
  aafeature.na_feature_id = dbrefnafeature.na_feature_id AND
  dbref.db_ref_id = dbrefnafeature.db_ref_id AND
  dbref.external_database_release_id = 1410 AND
  aasequence.external_database_release_id IN (2,1421)
  LIMIT 10;
```

# Apêndice B

# Programas construídos

## Programas de conversão

Abaixo estão os códigos-fontes dos programas utilizados para conversão de arquivos das bases de dados do projeto:

### Conversão da base de dados do SIDER

```
#!/usr/lib/env ruby
=begin
    Este programa abre o arquivo medra do SIDER e carrega para
    a base de dados
=end
require 'logger'
require 'pg'
log = Logger.new('log_sider.txt')
log.level = Logger::INFO
log.info('INICIO')
if (ARGV[0].nil?)
    log.error('Use: ruby parser.rb <fileIn>')
```

```
    exit 1
end
log.info('Abrindo conexao com o BD')
conn = nil
.
.
.
conn.exec('commit')
conn.close()
log.info("Registros inseridos: #{inseridas}")
log.info('FIM')
exit 0
```

O código-fonte completo pode ser encontrado no material suplementar, disponibilizado em mídia digital (cdrom) ou no endereço `http://wiki.biowebdb.org/index.php/Drug_Repositioning#Base_de_dados_SIDER`

### Conversão da base de dados do OrthoMCLDB

```
#!/usr/lib/env ruby
=begin
    Este programa abre o arquivo do Orthomcl e carrega para
    a base de dados
=end
require 'logger'
require 'pg'
.
.
.
```

```
begin
    # Insere os registros
    res = conn.exec("INSERT INTO dots.tese_orthomcl(orthologous, organism, accessio

    log.debug('Registro inserido')
.
.
.
conn.exec('commit')
conn.close()
log.info("Registros inseridos: #{inseridas}")
log.info('FIM')
exit 0
```

O código-fonte completo pode ser encontrado no material suplementar, disponibilizado em mídia digital (cdrom) ou no endereço `http://wiki.biowebdb.org/index.php/Drug_Repositioning#Base_de_dados_OrthoMCLDB`

### Servico WEB para buscar dados do KEGG

```
#!/usr/bin/env ruby
require 'logger'
require 'rubygems'
require 'csv'
require '~/work/biowebdb/svn.biowebdb.org/biowebdb-ruby-lib/services/kegg.rb'
# Criando Log
logger = Logger.new('log.log')
logger.info("Inicio")
restart = true
```

```
.
.
.
# Converte GI para Kegg ID
keggid = @kegg.getKeggIdByGI(line[1])
    if keggid != ""
        log += "KeggID:#{keggid}|"
        # Busca grupo de ortologos para o Kegg ID
        ko = @kegg.getKoByKeggId(keggid)
        log += "KO:#{ko}|"
        aFile.syswrite "#{line[0]};#{line[1]};#{keggid};#{ko}\n"
    else
        aFile.syswrite "#{line[0]};#{line[1]};;\n"
    end
.
.
.
aFile.close
logger.info("Fim!")
```

O código-fonte completo pode ser encontrado no material suplementar, disponibilizado em mídia digital (cdrom) ou no endereço `http://wiki.biowebdb.org/index.php/Drug_Repositioning#Servi.C3.A7o_WEB_do_KEGG`

### Conversão da Base de Dados do KEGG

```
#!/usr/lib/env ruby
=begin
    Este programa abre o arquivo do Kegg e carrega para
```

```
    a base de dados
=end
require 'logger'
require 'pg'
require 'csv'
log = Logger.new('log_kegg.txt')
log.level = Logger::INFO
log.info('INICIO')
.
.
.
res = conn.exec("INSERT INTO dots.tese_kegg(keggid, ko, genbankidprotein) VALUES ('
log.debug('Registro inserido')
inseridas+=1
.
.
.
conn.exec('commit')
conn.close()
log.info("Registros inseridos: #{inseridas}")
log.info('FIM')
exit 0
```

O código-fonte completo pode ser encontrado no material suplementar, disponibilizado em mídia digital (cdrom) ou no endereço `http://wiki.biowebdb.org/index.php/Drug_Repositioning#Base_de_dados_KEGG`

### Conversão da Base de Dados do KEGG - Vias Metabolicas

```
#!/usr/lib/env ruby
=begin
    Este programa abre o arquivo de Mapas e KO do KEGG e carrega
    para a base de dados
=end
require 'logger'
require 'pg'
require 'csv'
log = Logger.new('log_kegg.txt')
log.level = Logger::DEBUG
log.info('INICIO')
.
.
.
res = conn.exec("INSERT INTO dots.tese_map_ko(ko, mapid) VALUES ('#{ko}', '#{map}')
log.debug('Registro inserido')
inseridas+=1
.
.
.
conn.exec('commit')
conn.close()
log.info("Registros inseridos: #{inseridas}")
log.info('FIM')
exit 0
```

O código-fonte completo pode ser encontrado no material suplementar, disponibilizado em mídia digital (cdrom) ou no endereço http://wiki.biowebdb.org/index.php/Drug_Repositioning#Base_de_dados_KEGG_Pathway

### Conversão da Base de Dados do Ensembl

```
#!/usr/lib/env ruby
=begin
    Este programa abre o arquivo do Ensembl GTF e carrega para
    a base de dados
=end
require 'logger'
require 'pg'
require 'bio'
log = Logger.new('log_ensembl.txt')
log.level = Logger::INFO
log.info('INICIO')
if (ARGV[0].nil?)
    log.error('Use: ruby parser.rb <fileIn>')
    exit 1
end
log.info('Abrindo conexao com o BD')
conn = nil
.
.
.
res = conn.exec("INSERT INTO dots.tese_ensembl_gtf(accession_number, gene_name) VAL
log.debug('Registro inserido')
```

```
inseridas+=1
.
.
.
conn.exec('commit')
conn.close()
log.info("Registros inseridos: #{inseridas}")
log.info('FIM')
exit 0
```

O código-fonte completo pode ser encontrado no material suplementar, disponibilizado em mídia digital (cdrom) ou no endereço `http://wiki.biowebdb.org/index.php/Drug_Repositioning#Base_de_dados_Ensembl`

### Conversão da Base de Dados do SGD

```
#!/usr/lib/env ruby
=begin
    Este programa abre o arquivo do SGD e carrega para
    a base de dados
=end
require 'logger'
require 'pg'
require 'csv'
log = Logger.new('log_kegg.txt')
log.level = Logger::INFO
log.info('INICIO')
if (ARGV[0].nil?)
    log.error('Use: ruby parser.rb <fileIn>')
```

```
    exit 1
end
log.info('Abrindo conexao com o BD')
conn = nil
.
.
.
res = conn.exec("INSERT INTO dots.tese_phenotype(accession_number, phenotype) VALUE
log.debug('Registro inserido')
inseridas+=1
.
.
.
conn.exec('commit')
conn.close()
log.info("Registros inseridos: #{inseridas}")
log.info('FIM')
exit 0
```

O código-fonte completo pode ser encontrado no material suplementar, disponibilizado em mídia digital (cdrom) ou no endereço http://wiki.biowebdb.org/index.php/Drug_Repositioning#Base_de_dados_SGD

# Funçoes do PostSemantic

Os códigos-fontes completos das funções do PostSemantic podem ser encontrados no material suplementar, disponibilizado em mídia digital (cdrom) ou no endereço http://wiki.biowebdb.org/index.php/Drug_Repositioning#Fun.C3.A7.C3.B5es_do_

PostSemantic

### Função Insert Mapping Classes

```
/* Apagando os Dados de Mapeamento anterior */
truncate postsemantic.pg_mapping_classes;
truncate postsemantic.pg_mapping_properties;

/* Verificando Schema.Tabela */
SELECT count(1)
INTO achou
FROM pg_class c
INNER JOIN pg_namespace n on (c.relnamespace = n.oid)
WHERE c.relname = v_table
AND      n.nspname = v_schema
AND      n.nspname NOT IN ('pg_catalog', 'pg_toast');
if (achou = 0) then
return 'Table/Schema not found!';
end if;
.
.
.
uriprop := (v_uri)::varchar || '/' || propname;
            INSERT INTO postsemantic.pg_mapping_properties(relnamespace, attrelid,
            VALUES (oidschema, oidtable, propname, proptype, (uriprop)::uri, ispk);
.
.
.
```

```
return 'ok';
```

### Função Triplification

```
/* Criando table triplify
CREATE TEMPORARY TABLE triplify(s varchar, p varchar, o varchar);
*/
/* Limpando a tabela de triplificacao */
truncate public.nodes;
truncate public.triples;
.
.
.
IF pOffset > total THEN
    EXIT;
END IF;
PERFORM postsemantic.tuple2triple(table_name, pLimit, pOffset);
pOffset = pOffset + pLimit + 1;
.
.
.
END;
```

### Função Tuple2Triple

```
warn(PL::args_type)
warn("Rodando em #{args[0]}")
warn("Buscando nome da coluna PK")
command = "SELECT attname
```

```
            FROM postsemantic.pg_mapping_properties
            WHERE indisprimary = TRUE
            AND   attrelid = (
                  SELECT distinct c.relfilenode
                  FROM pg_class c, pg_attribute a, pg_type t
                  WHERE c.relname = '#{args[0]}'
                  AND a.attnum > 0
                  AND a.attrelid = c.oid
                  AND a.atttypid = t.oid);"

column_key = ""
.
.
.
# Insere object
command = "SELECT postsemantic.insert_node('#{o}', #{type_node}) as value;"
object_node = ""
res = PL.exec(command)
object_node = res[0]['value']
# Insere tripla
command = "INSERT INTO public.triples(s, p, o)
            VALUES (#{subject},#{property},#{object_node});"
res = PL.exec(command)
.
.
.
return "Ok"
```

### Função Hash

```
#!/usr/bin/env ruby
require 'digest/md5'
class Hash
    def hash(vhash_node, vtype_node)
        toHash = vhash_node.to_s
        type = vtype_node
.
.
.
return vhash.unpack('q')
```

### Função Insert Node

```
SELECT value_node || '||'
INTO node;
SELECT postsemantic.hash(node ,type_object::varchar)
INTO pk;
.
.
.
/* Not Found, creating new */
INSERT INTO public.nodes (hash, lex, lang, datatype, "type")
VALUES(pk::bigint, value_node, '', '', type_object);
.
.
.
RETURN pk::bigint;
```

## Programa para validação dos resultados

Com a finalidade de validar os resultados da tese, foi escrito um programa para buscar artigos no PubMed. O código-fonte completo desse programa pode ser obtido no material suplementar ou no endereço `http://wiki.biowebdb.org/index.php/Drug_Repositioning#Programa_para_recuperar_artigos`.

```
#!/usr/bin/env ruby
require "rubygems"
require "bio"
require "rinruby"
abstracts = Array.new()
Bio::NCBI.default_email = "rodrigo_jardim@fiocruz.br"
.
.
.
texto = "(\"#{taxon}\" AND #{annot} AND (drug or target))"
# Buscando os PMID's dos textos
if (ant !=texto)
.
.
.
end
```

# Apêndice C

# Consultas SPARQL

Todas as consultas SPARQL podem ser encontradas no material suplementar, disponibilizado em mídia digital (cdrom) ou no endereço `http://wiki.biowebdb.org/index.php/Drug_Repositioning#Queries_SPARQL`

## Prefixos

```
PREFIX nci: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.
    owl#>
PREFIX knoesis: <http://knoesis.org/tcruzipse/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX drug:<http://www4.wiwiss.fu-berlin.de/drugbank/resource
    /drugbank/>
PREFIX protozoadb:<http://biowebdb.org/protozoadb/>
PREFIX sideeffects:<http://sideeffects.embl.de/>
PREFIX ensembl:<http://www.ensembl.org/>
PREFIX kegg:<http://www.genome.jp/kegg/>
PREFIX orthomcl:<http://www.orthomcl.org/>
```

```
PREFIX sce:<http://www.yeastgenome.org/>
```

# Consultas para criação de termos computacionalmente complexos

## Criação do termo reuseTo

```
CONSTRUCT {
    ?farmacoId <http://biowebdb.org/protozoadb/reuseTo> ?
        protozoa
}
WHERE{
    ?protozoa <http://biowebdb.org/protozoadb/accession_number
        > ?an .
    ?an <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <
        http://www.orthomcl.org/> .
    ?protozoa <http://www.w3.org/1999/02/22-rdf-syntax-ns#type
        > <http://biowebdb.org/protozoadb/> .
    ?protozoa <http://biowebdb.org/protozoadb/gi> ?gi .
    ?an <http://www.orthomcl.org/gi> ?gi .
    ?grupo <http://biowebdb.org/protozoadb/accession_number> ?
        an .
    ?grupo <http://www.orthomcl.org/orthologous_group> ?
        string_grupo .
    ?grupo <http://www.ensembl.org/accession_number> ?
        codigoensembl .
    ?ncbiAccession <http://www.ensembl.org/accession_number> ?
        codigoensembl .
```

```
    ?ncbiAccession <http://www.ncbi.nlm.nih.gov/
        accession_number> ?accession_number_string .
    ?alvo <http://www4.wiwiss.fu-berlin.de/drugbank/resource/
        drugbank/geneName> ?accession_number_string .
    ?farmacoId <http://www4.wiwiss.fu-berlin.de/drugbank/
        resource/drugbank/target> ?alvo .
    ?farmacoId <http://www4.wiwiss.fu-berlin.de/drugbank/
        resource/drugbank/genericName> ?farmaco .
    ?protozoa <http://biowebdb.org/protozoadb/taxonname> ?
        taxon
}
```

## Criação do termo phenotype

```
CONSTRUCT {
    ?protozoa <http://biowebdb.org/protozoadb/phenotype> ?
        phenotype
}
WHERE {
    ?protozoa <http://biowebdb.org/protozoadb/gi> ?gi .
    ?ko <http://www.genome.jp/kegg/gi> ?gi .
    ?ko <http://www.yeastgenome.org/accession_number> ?yeast .
    ?yeast <http://www.yeastgenome.org/phenotype> ?phenotype
}
```

## Criação do termo pathway

```
CONSTRUCT {
    ?protozoa <http://biowebdb.org/protozoadb/pathway> ?mapid
```

```
}
WHERE {
    ?protozoa <http://biowebdb.org/protozoadb/gi> ?gi .
    ?ko <http://www.genome.jp/kegg/gi> ?gi .
    ?ko <http://www.genome.jp/kegg/mapid> ?mapid
}
```

# Consultas dos totais de dados na nuvem

## Total geral

```
SELECT count(?s)
WHERE{
    ?s ?p ?o
}
```

## Por base de dados

```
SELECT count(?s)
WHERE{
    ?s ?p ?o .
       FILTER regex(str(?s), 'protozoadb', 'i')
}
```

Obs: Altere o nome protozoadb para as outras bases de dados (kegg, drugbank, sider, orthomcl, yeast)

# Consultas para dados quantitativos

## Quantidade de classes

```
SELECT DISTINCT *
WHERE{
    ?s a rdfs:Class
}
```

### Total de ligações (*links*)

```
SELECT count(?p)
WHERE{
    ?s ?p ?o
}
```

### Quantidade de fármacos

```
SELECT (count(DISTINCT ?o) as ?count)
WHERE{
    ?s drug:genericName ?o
}
```

### Quantidade de alvos de fármacos

```
  SELECT (count(DISTINCT ?o) as ?count)
WHERE{
    ?s drug:target ?o
}
```

### Quantidade de efeitos colaterais

```
SELECT (count(DISTINCT ?o) as ?count)
WHERE{
    ?s sideeffects:side_effect ?o
```

```
}
```

### Quantidade de grupos de ortólogos pelo OrthoMCLDB

```
SELECT (count(DISTINCT ?o) as ?count)
WHERE{
    ?s orthomcl:orthologous_group ?o FILTER isLiteral(?o)
}
```

### Quantidade de mapas de vias metabólicas

```
SELECT (count(DISTINCT ?o) as ?count)
WHERE{
    ?s kegg:mapid ?o FILTER isLiteral(?o)
}
```

## Consultas para dados qualitativos

### Fármacos

#### Com alvos com ortólogos em protozoários

```
SELECT (count(DISTINCT ?farmaco) as ?farm) (count(DISTINCT ?
   protozoa) as ?proto)
WHERE{
    ?farmaco protozoadb:reuseTo ?protozoa
}
```

### Por organismo afetado

```
SELECT ?organismo (count(DISTINCT ?farmaco) )
WHERE{
```

```
    ?farmaco protozoadb:reuseTo ?protozoa .
    OPTIONAL {?farmaco drug:affectedOrganism ?organismo}
} group by ?organismo
```

### Por categoria

```
SELECT ?categoria (count(DISTINCT ?farmaco) )
WHERE{
    ?farmaco protozoadb:reuseTo ?protozoa .
    OPTIONAL { ?farmaco drug:drugCategory ?categoria } .
} group by ?categoria
```

### Por organismo

```
SELECT ?organismo (count(DISTINCT ?farmaco) )
WHERE{
    ?farmaco protozoadb:reuseTo ?protozoa .
    OPTIONAL { ?protozoa protozoadb:taxonname ?organismo }
       FILTER regex(?organismo, 'cruzi' ,'i').
} group by ?organismo
```

Obs: Substituir pelo organismo desejado (Ex: brucei, parva, plasmodium, etc.)

## Fenótipos

### Total de fenótipos

```
SELECT (count(DISTINCT ?fenotipo) as ?count)
WHERE{
    ?protozoa protozoadb:phenotype ?fenotipo FILTER regex(str
       (?protozoa), 'accession', 'i')
```

```
}
```

### Total de proteínas por fenótipo

```
SELECT  ?fenotipo (count(DISTINCT ?protozoa) as ?count)
WHERE{
    ?protozoa protozoadb:phenotype ?fenotipo FILTER regex(str
       (?protozoa), 'accession', 'i')
} GROUP BY ?fenotipo
```

### Total de proteínas por fenótipo / organismo

```
SELECT  ?taxon (count(DISTINCT ?protozoa) as ?countProtozoa) (
   count(DISTINCT ?fenotipo) as ?countFenotipo)
WHERE{
    ?protozoa protozoadb:phenotype ?fenotipo FILTER regex(str
       (?protozoa), 'accession', 'i') .
    ?protozoa protozoadb:taxonname ?taxon FILTER regex(?taxon,
        'bovis', 'i') .
} GROUP BY ?taxon
```

### Total de fármacos por fenótipo

```
SELECT ?fenotipo (count(DISTINCT ?farmaco) as ?count)
WHERE{
    ?farmaco protozoadb:reuseTo ?protozoa .
    ?protozoa protozoadb:phenotype ?fenotipo .
} GROUP BY ?fenotipo
```

# Mapas de Vias Metabólicas

## Proteínas por mapas

```
SELECT ?mapa (count (DISTINCT ?protozoa))
WHERE{
    ?protozoa protozoadb:pathway ?via FILTER regex(str(?
       protozoa), 'accession', 'i') .
    ?via kegg:description ?mapa
} GROUP BY ?mapa
```

## Consulta de Detalhes para os fármacos discutidos

```
SELECT DISTINCT *
WHERE{
    ?farmaco protozoadb:reuseTo <http://biowebdb.org/
       protozoadb/accession_number/XP_002260251.1> .
    ?farmaco drug:target ?target .
    ?target drug:geneName ?geneName FILTER (?geneName = "VARS
       ") .
    OPTIONAL {?farmaco drug:genericName ?name .}
    OPTIONAL {?farmaco drug:drugCategory ?category .}
    OPTIONAL {?farmaco drug:indication ?indication .}
    OPTIONAL {?farmaco drug:mechanismOfAction ?acao .}
    OPTIONAL {?farmaco drug:affectedOrganism ?organismo .}
    OPTIONAL {?farmaco drug:proteinBinding ?proteinBinding .}
    OPTIONAL {?target drug:name ?function .}
    OPTIONAL {?conceptid sideeffects:genericName ?name . ?
       conceptid sideeffects:side_effect ?sider} .
```

```
    <http://biowebdb.org/protozoadb/accession_number/
        XP_002260251.1> protozoadb:annotation ?alvo_protozoa .
    OPTIONAL {<http://biowebdb.org/protozoadb/accession_number
        /XP_002260251.1> protozoadb:phenotype ?fenotipo .
    <http://biowebdb.org/protozoadb/accession_number/
        XP_002260251.1> protozoadb:pathway ?mapa .
    ?mapa kegg:description ?via
}
```

# Apêndice D

# Lista dos Fármacos encontrados com potencialidade de reposicionamento

## Lista dos 394 fármacos

Tabela D.1: Lista dos 394 fármacos cujos alvos possuem ortólogos à proteínas de protozoários

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| (2e,3s)-3-Hydroxy-5'-[(4-Hydroxypiperidin-1-Yl)Sulfonyl]-3-Methyl-1,3-Dihydro-2,3'-Biindol-2'(1'h)-One | Flavopiridol |
| (3-Carboxy-2-(R)-Hydroxy-Propyl)-Trimethyl-Ammonium | Flupenthixol |
| (4-sulfamoyl-phenyl)-thiocarbamic acid O-(2-thiophen-3-yl-ethyl) ester | Formic Acid |
| (4s-Trans)-4-(Ethylamino)-5,6-Dihydro-6-Methyl-4h-Thieno(2,3-B)Thiopyran-2-Sulfonamide-7,7-Dioxide | Fructose |
| (4s-Trans)-4-(Methylamino)-5,6-Dihydro-6-Methyl-4h-Thieno(2,3-B)Thiopyran-2-Sulfonamide-7,7-Dioxide | Fumarate |
| (5-Chloropyrazolo[1,5-a]Pyrimidin-7-Yl)-(4-Methanesulfonylphenyl)Amine | Gamma-Glutamyl[S-(2-Iodobenzyl)Cysteinyl]Glycine |
| (Aminooxy)Acetic Acid | Geldanamycin |

*Continua na próxima página*

Tabela D.1 – *Continuação da página anterior*

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| (R)-Mesopram | Ginkgo biloba |
| (R)-N-(3-Indol-1-Yl-2-Methyl-Propyl)-4-Sulfamoyl-Benzamide | Glibenclamide |
| (R)-Rolipram | Gliclazide |
| (S)-N-(3-Indol-1-Yl-2-Methyl-Propyl)-4-Sulfamoyl-Benzamide | Glutathione |
| (S)-Rolipram | Glycerol |
| [1-(4-Fluorobenzyl)Cyclobutyl]Methyl (1s)-1-[Oxo(1h-Pyrazol-5-Ylamino)Acetyl]Pentylcarbamate | Glycine |
| [2-Amino-6-(2,6-Difluoro-Benzoyl)-Imidazo[1,2-a]Pyridin-3-Yl]-Phenyl-Methanone | Guanosine-5'-Diphosphate |
| [3-(4-Bromo-2-Fluoro-Benzyl)-7-Chloro-2,4-Dioxo-3,4-Dihydro-2h-Quinazolin-1-Yl]-Acetic Acid | Guanosine-5'-Triphosphate |
| [4-(2-Amino-4-Methyl-Thiazol-5-Yl)-Pyrimidin-2-Yl]-(3-Nitro-Phenyl)-Amine | Heme C |
| 1-(2-Chlorophenyl)-3,5-Dimethyl-1h-Pyrazole-4-Carboxylic Acid Ethyl Ester | Hexane |
| 1-[(2-Amino-6,9-Dihydro-1h-Purin-6-Yl)Oxy]-3-Methyl-2-Butanol | Hydrochlorothiazide |
| 1-[2-(3-Biphenyl)-4-Methylvaleryl)]Amino-3-(2-Pyridylsulfonyl)Amino-2-Propanone | Hydroflumethiazide |
| 1-[N[(Phenylmethoxy)Carbonyl]-L-Leucyl-4-[[N/N-[(Phenylmethoxy)Carbonyl]-/Nl-Leucyl]Amino]-3-Pyrrolidinone/N | Hydroxyalanine |
| 1-Amino-6-Cyclohex-3-Enylmethyloxypurine | Hydroxydimethylarsine Oxide |
| 1-Hydroxy-2-Amino-3-Cyclohexylpropane | Hymenialdisine |
| 1-Hydroxy-3-Methylbutane | I-5 |
| 1-Methoxy-2-(2-Methoxyethoxy)Ethane | Icosapent |
| 1-Methyl-2-Oxy-5,5-Dimethyl Pyrrolidine | IDD552 |
| 1-Methyl-3-Oxo-1,3-Dihydro-Benzo[C]Isothiazole-5-Sulfonic Acid Amide | Iloprost |
| 1-N-(4-SULFAMOYLPHENYL-ETHYL)-2,4,6-TRIMETHYLPYRIDINIUM | Imatinib |
| 1,2,4-Triazole | Imidazole |
| 1,5-Bis(N-Benzyloxycarbonyl-L-Leucinyl)Carbohydrazide | Indirubin-3'-Monoxime |

*Continua na próxima página*

Tabela D.1 – *Continuação da página anterior*

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| 17-Dmag | Inhibitor Idd 384 |
| 2-[Trans-(4-Aminocyclohexyl)Amino]-6-(Benzyl-Amino)-9-Cyclopentylpurine | Insulin Glargine recombinant |
| 2-Amino-6-Chloropyrazine | Insulin Lyspro recombinant |
| 2-Chlorodideoxyadenosine | Insulin recombinant |
| 2-Cyclopropylmethylenepropanal | Insulin, porcine |
| 2-Methyl-2,4-Pentanediol | Irinotecan |
| 2-Methyl-3-(2-Aminothiazolo)Propanal | Isopropyl Alcohol |
| 2-Sulfhydryl-Ethanol | Itraconazole |
| 2,6-Difluorobenzenesulfonamide | L-Alanine |
| 2'-Monophosphoadenosine 5'-Diphosphoribose | L-Arginine |
| 2',3'-Dideoxythymidine-5'-Monophosphate | L-Asparagine |
| 3-[3-(2,3-Dihydroxy-Propylamino)-Phenyl]-4-(5-Fluoro-1-Methyl-1h-Indol-3-Yl)-Pyrrole-2,5-Dione | L-Aspartic Acid |
| 3-[4-(2,4-Dimethyl-Thiazol-5-Yl)-Pyrimidin-2-Ylamino]-Phenol | L-Carnitine |
| 3-Amino-5-Phenylpentane | L-Cysteine |
| 3-Mercuri-4-Aminobenzenesulfonamide | L-Glutamic Acid |
| 3-Nitro-4-(2-Oxo-Pyrrolidin-1-Yl)-Benzenesulfonamide | L-Glutamine |
| 3-Phenyl-1,2-Propandiol | L-Isoleucine |
| 3-Pyridin-4-Yl-2,4-Dihydro-Indeno[1,2-.C.]Pyrazole | L-Ornithine |
| 3-Sulfinoalanine | L-Phenylalanine |
| 3-Thiaoctanoyl-Coenzyme A | L-Proline |
| 3,5-Difluorobenzenesulfonamide | L-Serine |
| 3,5-Dimethyl-1-(3-Nitrophenyl)-1h-Pyrazole-4-Carboxylic Acid Ethyl Ester | L-Tryptophan |
| 3'-Azido-3'-Deoxythymidine-5'-Monophosphate | L-Tyrosine |
| 4-((3r,4s,5r)-4-Amino-3,5-Dihydroxy-Hex-1-Ynyl)-5-Fluoro-3-[1-(3-Methoxy-1h-Pyrrol-2-Yl)-Meth-(Z)-Ylidene]-1,3-Dihydro-Indol-2-One | L-Valine |
| 4-(1,3-Benzodioxol-5-Yl)-5-(5-Ethyl-2,4-Dihydroxyphenyl)-2h-Pyrazole-3-Carboxylic Acid | Lactose |
| 4-(1h-Imidazol-4-Yl)-3-(5-Ethyl-2,4-Dihydroxy-Phenyl)-1h-Pyrazole | Lipoic Acid |

*Continua na próxima página*

Tabela D.1 – *Continuação da página anterior*

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| 4-(2,4-Dimethyl-Thiazol-5-Yl)-Pyrimidin-2-Yl]-(4-Trifluoromethyl-Phenyl)-Amine | Lithium |
| 4-(2,4-Dimethyl-Thiazol-5-Yl)-Pyrimidin-2-Ylamine | Lucanthone |
| 4-(2,5-Dichloro-Thiophen-3-Yl)-Pyrimidin-2-Ylamine | Lysine Nz-Carboxylic Acid |
| 4-(5-Bromo-2-Oxo-2h-Indol-3-Ylazo)-Benzenesulfonamide | Malate Ion |
| 4-(Aminosulfonyl)-N-[(2,3,4-Trifluorophenyl)Methyl]-Benzamide | Maleic Acid |
| 4-(Aminosulfonyl)-N-[(2,4-Difluorophenyl)Methyl]-Benzamide | Menadione |
| 4-(Aminosulfonyl)-N-[(2,4,6-Trifluorophenyl)Methyl]-Benzamide | Mercuribenzoic Acid |
| 4-(Aminosulfonyl)-N-[(2,5-Difluorophenyl)Methyl]-Benzamide | Methyclothiazide |
| 4-(Aminosulfonyl)-N-[(3,4,5-Trifluorophenyl)Methyl]-Benzamide | Methyl Mercury Ion |
| 4-(Aminosulfonyl)-N-[(4-Fluorophenyl)Methyl]-Benzamide | Milrinone |
| 4-(Hydroxymercury)Benzoic Acid | Mimosine |
| 4-[(3-BROMO-4-O-SULFAMOYLBENZYL)(4-CYANOPHENYL)AMINO]-4H-[1,2,4]-TRIAZOLE | Mitiglinide |
| 4-[(4-Imidazo[1,2-a]Pyridin-3-Ylpyrimidin-2-Yl)Amino]Benzenesulfonamide | Mitoxantrone |
| 4-[(4-O-SULFAMOYLBENZYL)(4-CYANOPHENYL)AMINO]-4H-[1,2,4]-TRIAZOLE | Morpholine-4-Carboxylic Acid (1-(3-Benzenesulfonyl-1-Phenethylallylcarbamoyl)-3-Methylbutyl)-Amide |
| 4-[(6-Amino-4-Pyrimidinyl)Amino]Benzenesulfonamide | Morpholine-4-Carboxylic Acid [1-(2-Benzylsulfanyl-1-Formyl-Ethylcarbamoyl)-2-Phenyl-Ethyl]-Amide |
| 4-[3-(Cyclopentyloxy)-4-Methoxyphenyl]-2-Pyrrolidinone | Morpholine-4-Carboxylic Acid [1s-(2-Benzyloxy-1r-Cyano-Ethylcarbamoyl)-3-Methyl-Butyl]Amide |
| 4-[3-Hydroxyanilino]-6,7-Dimethoxyquinazoline | Myo-Inositol |
| 4-[4-(4-Methyl-2-Methylamino-Thiazol-5-Yl)-Pyrimidin-2-Ylamino]-Phenol | N-(2-Flouro-Benzyl)-4-Sulfamoyl-Benzamide |

*Continua na próxima página*

Tabela D.1 – *Continuação da página anterior*

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| 4-[5-(Trans-4-Aminocyclohexylamino)-3-Isopropylpyrazolo[1,5-a]Pyrimidin-7-Ylamino]-N,N-Dimethylbenzenesulfonamide | N-(2-Thienylmethyl)-2,5-Thiophenedisulfonamide |
| 4-Flourobenzenesulfonamide | N-(2,3,4,5,6-Pentaflouro-Benzyl)-4-Sulfamoyl-Benzamide |
| 4-Hydroxy-1,2,5-Oxadiazole-3-Carboxylic Acid | N-(2,6-Diflouro-Benzyl)-4-Sulfamoyl-Benzamide |
| 4-Methylimidazole | N-(4-Methoxybenzyl)-N'-(5-Nitro-1,3-Thiazol-2-Yl)Urea |
| 4-Methylpiperazin-1-Yl Carbonyl Group | N-(5-Cyclopropyl-1h-Pyrazol-3-Yl)Benzamide |
| 4-Morpholin-4-Yl-Piperidine-1-Carboxylic Acid [1-(3-Benzenesulfonyl-1-Propyl-Allylcarbamoyl)-2-Phenylethyl]-Amide | N-[2-(1h-Indol-5-Yl)-Butyl]-4-Sulfamoyl-Benzamide |
| 4-Sulfonamide-[1-(4-Aminobutane)]Benzamide | N-[4-(2-Methylimidazo[1,2-a]Pyridin-3-Yl)-2-Pyrimidinyl]Acetamide |
| 4-Sulfonamide-[4-(Thiomethylaminobutane)]Benzamide | N-[4-(2,4-Dimethyl-1,3-Thiazol-5-Yl)Pyrimidin-2-Yl]-N'-Hydroxyimidoformamide |
| 4'-Deoxy-4'-Acetylyamino-Pyridoxal-5'-Phosphate | N-[4-(AMINOSULFONYL)BENZYL]-5-(5-CHLORO-2,4-DIHYDROXYPHENYL)-1H-PYRAZOLE-4-CARBOXAMIDE |
| 5-{[(2-Amino-9h-Purin-6-Yl)Oxy]Methyl}-2-Pyrrolidinone | N-Acetylalanine |
| 6-O-Cyclohexylmethyl Guanine | N-Benzyl-4-Sulfamoyl-Benzamide |
| 6-Oxo-8,9,10,11-Tetrahydro-7h-Cyclohepta[C][1]Benzopyran-3-O-Sulfamate | N-Carbamoyl-Alanine |
| 7-Hydroxy-2-Oxo-Chromene-3-Carboxylic Acid Ethyl Ester | N-Ethyl-5'-Carboxamido Adenosine |
| 8-(2-Chloro-3,4,5-Trimethoxy-Benzyl)-2-Fluoro-9-Pent-4-Ylnyl-9h-Purin-6-Ylamine | N-Methyl-N-(Methylbenzyl)Formamide |
| 8-(2-Chloro-3,4,5-Trimethoxy-Benzyl)-9-Pent-4-Ylnyl-9h-Purin-6-Ylamine | N-Octane |
| 8-(2,5-Dimethoxy-Benzyl)-2-Fluoro-9-Pent-9h-Purin-6-Ylamine | N-Pyridoxyl-Glycine-5-Monophosphate |
| 8-(2,5-Dimethoxy-Benzyl)-2-Fluoro-9h-Purin-6-Ylamine | N-Trimethyllysine |
| 8-Benzo[1,3]Dioxol-,5-Ylmethyl-9-Butyl-2-Fluoro-9h-Purin-6-Ylamine | N'-(Pyrrolidino[2,1-B]Isoindolin-4-On-8-Yl)-N-(Pyridin-2-Yl)Urea |
| 8-Bromo-Adenosine-5'-Monophosphate | N'-Pyridoxyl-Lysine-5'-Monophosphate |

*Continua na próxima página*

Tabela D.1 – *Continuação da página anterior*

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| 9-Butyl-8-(2-Chloro-3,4,5-Trimethoxy-Benzyl)-9h-Purin-6-Ylamine | N2-({[(4-Bromophenyl)Methyl]Oxy}Carbonyl)-N1-[(1s)-1-Formylpentyl]-L-Leucinamide |
| 9-Butyl-8-(2,5-Dimethoxy-Benzyl)-2-Fluoro-9h-Purin-6-Ylamine | N2-[(Benzyloxy)Carbonyl]-N1-[(3s)-1-Cyanopyrrolidin-3-Yl]-L-Leucinamide |
| 9-Butyl-8-(2,5-Dimethoxy-Benzyl)-9h-Purin-6-Ylamine | NADH |
| 9-Butyl-8-(3-Methoxybenzyl)-9h-Purin-6-Amine | Nateglinide |
| 9-Butyl-8-(3,4,5-Trimethoxybenzyl)-9h-Purin-6-Amine | Nicardipine |
| 9-Butyl-8-(4-Methoxybenzyl)-9h-Purin-6-Amine | Nicotinamide Mononucleotide |
| Acenocoumarol | Nicotinamide-Adenine-Dinucleotide |
| Acetate Ion | Nitrendipine |
| Acetazolamide | Norgestimate |
| Acetic Acid | Octanoyl-Coenzyme A |
| Adenosine monophosphate | Oleic Acid |
| Adenosine Monotungstate | Olomoucine |
| Adenosine triphosphate | OLOMOUCINE II |
| Adenosine-5'-Diphosphate | Oxalate Ion |
| Ado-P-Ch2-P-Ps-Ado | Oxamic Acid |
| Al-6619, [2h-Thieno[3,2-E]-1,2-Thiazine-6-Sulfonamide,2-(3-Hydroxyphenyl)-3-(4-Morpholinyl)-, 1,1-Dioxide] | P1-(5'-Adenosyl)P5-(5'-(3'azido-3'-Deoxythymidyl))Pentaphosphate |
| Al-6629, [2h-Thieno[3,2-E]-1,2-Thiazine-6-Sulfonamide,2-(3-Methoxyphenyl)-3-(4-Morpholinyl)-, 1,1-Dioxide] | Paclitaxel |
| AL4623 | Papaverine |
| AL5300 | Pentoxifylline |
| AL5424 | Perhexiline |
| AL5927 | Phenindione |
| AL6528 | Phenprocoumon |
| Al7089a | Phentolamine |
| AL7099A | PHENYLAMINOIMIDAZO(1,2-ALPHA)PYRIDINE |
| AL7182 | Phosphatidylserine |
| Alfentanil | Phosphoaminophosphonic Acid-Adenylate Ester |
| Aliskiren | Phosphonoserine |
| Alpha-D-Glucose-6-Phosphate | Phosphonothreonine |

*Continua na próxima página*

Tabela D.1 – *Continuação da página anterior*

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| Alpha-Ketomalonic Acid | Phosphoric Acid Mono-[3-Amino-5-(5-Methyl-2,4-Dioxo-3,4-Dihydro-2h-Pyrimidin-1-Yl)-Tetrahydro-Furan-2-Ylmethyl] Ester |
| Alpha-Linolenic Acid | Phosphoric Acid Mono-[3-Fluoro-5-(5-Methyl-2,4-Dioxo-3,4-Dihydro-2h-Pyrimidin-1-Yl)-Tetrahyro-Furan-2-Ylmethyl] Ester |
| Alpha,Beta-Methyleneadenosine-5'-Triphosphate | Phosphothiophosphoric Acid-Adenylate Ester |
| Alrestatin | Pranlukast |
| Alsterpaullone | Pravastatin |
| Aminodi(Ethyloxy)Ethylaminocarbonylbenzenesulfonamide | Prodigiosenecid |
| Amrinone | Propafenone |
| Antihemophilic Factor | Purvalanol |
| Aplyronine A | Putrescine |
| Arsenic trioxide | Pyridoxal Phosphate |
| Atazanavir | Quinacrine |
| Atorvastatin | Quinethazone |
| Auranofin | Quinidine |
| Becaplermin | Radicicol |
| Bendroflumethiazide | Ranolazine |
| Benzofuran-2-Carboxylic Acid {(S)-3-Methyl-1-[3-Oxo-1-(Pyridin-2-Ylsulfonyl)Azepan-4-Ylcarbamoyl]Butyl}Amide | Rapamycin Immunosuppressant Drug |
| Benzoic Acid | Reidispongiolide A |
| Benzthiazide | Reidispongiolide C |
| Beta-D-Glucose | Remikiren |
| Beta-Mercaptoethanol | Repaglinide |
| Bezafibrate | Rifabutin |
| Bis(Adenosine)-5'-Pentaphosphate | Rifampin |
| Bortezomib | Rosuvastatin |
| Bosentan | Roxithromycin |
| Brinzolamide | S-(N-Hydroxy-N-Iodophenylcarbamoyl)Glutathione |
| Cacodylate Ion | S-2-(Boronoethyl)-L-Cysteine |
| Caffeine | S-Adenosyl-L-Homocysteine |
| Camptothecin | S-Adenosylmethionine |
| Carvedilol | S-Benzyl-Glutathione |
| Cefazolin | S-Hexylglutathione |
| Chlorothiazide | S-Hydroxycysteine |

*Continua na próxima página*

Tabela D.1 – *Continuação da página anterior*

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| Choline | S-P-Nitrobenzyloxycarbonylglutathione |
| Cilostazol | S,S-(2-Hydroxyethyl)Thiocysteine |
| Cis-4-Cyano-4-[3-(Cyclopentyloxy)-4-Methoxyphenyl]Cyclohexanecarboxylic Acid | Saquinavir |
| Citric Acid | Sodium stibogluconate |
| Coenzyme A | Sorbinil |
| Coenzyme a Persulfide | Spermine |
| Colamine Phosphoric Acid | Sphingosine |
| Colchicine | Sphinxolide B |
| Cordycepin Triphosphate | Staurosporine |
| Cyanamide | SU9516 |
| Cyclosporine | Succinic acid |
| Cyclothiazide | Sulfamic Acid 2,3-O-(1-Methylethylidene)-4,5-O-Sulfonyl-Beta-Fructopyranose Ester |
| Cysteine Sulfenic Acid | Sulfasalazine |
| Cytarabine | Sulfinpyrazone |
| Cytidine | Sulindac |
| Cytidine-5'-Monophosphate | Suramin |
| Cytidine-5'-Triphosphate | Tadalafil |
| D-tartaric acid | Tamoxifen |
| Dansylamide | Tartronate |
| Daunorubicin | Tert-Butyl(1s)-1-Cyclohexyl-2-Oxoethylcarbamate |
| Deamido-Nad+ | Tetraethylene Glycol |
| Diazoxide | Tetrahydrofolic acid |
| Dicumarol | Thenoyltrifluoroacetone |
| Dimethyl sulfoxide | Theophylline |
| Dimethylformamide | Thiamin Diphosphate |
| Diminazene | Thymidine-5'-Phosphate |
| Dipyridamole | Thymidine-5'-Triphosphate |
| Dithioerythritol | Tolrestat |
| Docetaxel | Topiramate |
| Dornase Alfa | Topotecan |
| Dorzolamide | TRIAZOLOPYRIMIDINE |
| Double Oxidized Cysteine | Trichlormethiazide |
| Dutasteride | Triglu-5-Formyl-Tetrahydrofolate |
| Dyphylline | Triphospate |
| Enalkiren | Tris(Hydroxymethyl)Aminomethane |

*Continua na próxima página*

Tabela D.1 – *Continuação da página anterior*

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| Enprofylline | Tryptophanyl-5'amp |
| Epothilone B | Urea |
| Epothilone D | Uridine 5'-Triphosphate |
| Estramustine | Vinblastine |
| Ethinamate | Vincristine |
| Ethylene Glycol | Vindesine |
| Fidarestat | Vinorelbine |
| Fidarestat(Stereoisomer) | Vitamin A |
| Filaminast | Vitamin E |
| Finasteride | Vorinostat |
| Flavin-Adenine Dinucleotide | Warfarin |
| | |

# Lista dos 150 fármacos

Tabela D.2: Lista dos 150 fármacos cujos alvos possuem ortólogos à proteínas de protozoários, estão associados a fenótipos e estão presentes a pelo menos uma via metabólica

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| "(3-Carboxy-2-(R)-Hydroxy-Propyl)-Trimethyl-Ammonium" | "L-Arginine" |
| "1-Methoxy-2-(2-Methoxyethoxy)Ethane" | "L-Asparagine" |
| "17-Dmag" | "L-Aspartic Acid" |
| "2-Chlorodideoxyadenosine" | "L-Carnitine" |
| "2-Methyl-2,4-Pentanediol" | "L-Cysteine" |
| "2-Sulfhydryl-Ethanol" | "L-Glutamic Acid" |
| "2'-Monophosphoadenosine 5'-Diphosphoribose" | "L-Glutamine" |
| "2',3'-Dideoxythymidine-5'-Monophosphate" | "L-Isoleucine" |
| "3-Sulfinoalanine" | "L-Ornithine" |
| "3'-Azido-3'-Deoxythymidine-5'-Monophosphate" | "L-Phenylalanine" |

*Continua na próxima página*

Tabela D.2 – *Continuação da página anterior*

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| "4-(1,3-Benzodioxol-5-Yl)-5-(5-Ethyl-2,4-Dihydroxyphenyl)-2h-Pyrazole-3-Carboxylic Acid" | "L-Proline" |
| "4-(1h-Imidazol-4-Yl)-3-(5-Ethyl-2,4-Dihydroxy-Phenyl)-1h-Pyrazole" | "L-Tryptophan" |
| "4'-Deoxy-4'-Acetylyamino-Pyridoxal-5'-Phosphate" | "L-Valine" |
| "8-(2-Chloro-3,4,5-Trimethoxy-Benzyl)-2-Fluoro-9-Pent-4-Ylnyl-9h-Purin-6-Ylamine" | "Lipoic Acid" |
| "8-(2-Chloro-3,4,5-Trimethoxy-Benzyl)-9-Pent-4-Ylnyl-9h-Purin-6-Ylamine" | "Malate Ion" |
| "8-(2,5-Dimethoxy-Benzyl)-2-Fluoro-9-Pent-9h-Purin-6-Ylamine" | "Maleic Acid" |
| "8-(2,5-Dimethoxy-Benzyl)-2-Fluoro-9h-Purin-6-Ylamine" | "Mimosine" |
| "8-Benzo[1,3]Dioxol-,5-Ylmethyl-9-Butyl-2-Fluoro-9h-Purin-6-Ylamine" | "Myo-Inositol" |
| "9-Butyl-8-(2-Chloro-3,4,5-Trimethoxy-Benzyl)-9h-Purin-6-Ylamine" | "N-[4-(AMINOSULFONYL)BENZYL]-5-(5-CHLORO-2,4-DIHYDROXYPHENYL)-1H-PYRAZOLE-4-CARBOXAMIDE" |
| "9-Butyl-8-(2,5-Dimethoxy-Benzyl)-2-Fluoro-9h-Purin-6-Ylamine" | "N-Ethyl-5'-Carboxamido Adenosine" |
| "9-Butyl-8-(2,5-Dimethoxy-Benzyl)-9h-Purin-6-Ylamine" | "N-Pyridoxyl-Glycine-5-Monophosphate" |
| "9-Butyl-8-(3-Methoxybenzyl)-9h-Purin-6-Amine" | "N-Trimethyllysine" |
| "9-Butyl-8-(3,4,5-Trimethoxybenzyl)-9h-Purin-6-Amine" | "NADH" |
| "9-Butyl-8-(4-Methoxybenzyl)-9h-Purin-6-Amine" | "Nicardipine" |
| "Acetate Ion" | "Nicotinamide Mononucleotide" |
| "Acetic Acid" | "Nicotinamide-Adenine-Dinucleotide" |
| "Adenosine monophosphate" | "Oxalate Ion" |
| "Adenosine Monotungstate" | "P1-(5'-Adenosyl)P5-(5'-(3'azido-3'-Deoxythymidyl))Pentaphosphate" |
| "Adenosine triphosphate" | "Paclitaxel" |
| "Adenosine-5'-Diphosphate" | "Phosphoaminophosphonic Acid-Adenylate Ester" |

*Continua na próxima página*

Tabela D.2 – *Continuação da página anterior*

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| "Ado-P-Ch2-P-Ps-Ado" | "Phosphonoserine" |
| "Alfentanil" | "Phosphonothreonine" |
| "Alpha-Ketomalonic Acid" | "Phosphoric Acid Mono-[3-Amino-5-(5-Methyl-2,4-Dioxo-3,4-Dihydro-2h-Pyrimidin-1-Yl)-Tetrahydro-Furan-2-Ylmethyl] Ester" |
| "Alpha,Beta-Methyleneadenosine-5'-Triphosphate" | "Phosphoric Acid Mono-[3-Fluoro-5-(5-Methyl-2,4-Dioxo-3,4-Dihydro-2h-Pyrimidin-1-Yl)-Tetrahyro-Furan-2-Ylmethyl] Ester" |
| "Antihemophilic Factor" | "Phosphothiophosphoric Acid-Adenylate Ester" |
| "Aplyronine A" | "Propafenone" |
| "Arsenic trioxide" | "Putrescine" |
| "Atazanavir" | "Pyridoxal Phosphate" |
| "Atorvastatin" | "Quinacrine" |
| "Beta-D-Glucose" | "Quinidine" |
| "Bis(Adenosine)-5'-Pentaphosphate" | "Radicicol" |
| "Bortezomib" | "Ranolazine" |
| "Bosentan" | "Reidispongiolide A" |
| "Carvedilol" | "Reidispongiolide C" |
| "Choline" | "Rifabutin" |
| "Coenzyme A" | "Rifampin" |
| "Colchicine" | "Roxithromycin" |
| "Cordycepin Triphosphate" | "S-(N-Hydroxy-N-Iodophenylcarbamoyl)Glutathione" |
| "Cyclosporine" | "S-2-(Boronoethyl)-L-Cysteine" |
| "Cysteine Sulfenic Acid" | "S-Adenosylmethionine" |
| "D-tartaric acid" | "S-Benzyl-Glutathione" |
| "Daunorubicin" | "S-Hexylglutathione" |
| "Deamido-Nad+" | "S-Hydroxycysteine" |
| "Dipyridamole" | "S-P-Nitrobenzyloxycarbonylglutathione" |
| "Docetaxel" | "Saquinavir" |
| "Dornase Alfa" | "Spermine" |
| "Epothilone B" | "Sphinxolide B" |
| "Epothilone D" | "Succinic acid" |
| "Estramustine" | "Sulfasalazine" |
| "Flupenthixol" | "Tamoxifen" |
| "Fructose" | "Tartronate" |
| "Fumarate" | "Tetraethylene Glycol" |
| "Geldanamycin" | "Tetrahydrofolic acid" |
| "Ginkgo biloba" | "Thenoyltrifluoroacetone" |

*Continua na próxima página*

Tabela D.2 – *Continuação da página anterior*

| **Nome do Fármaco** | **Nome do Fármaco** |
|---|---|
| "Glibenclamide" | "Thiamin Diphosphate" |
| "Glutathione" | "Thymidine-5'-Phosphate" |
| "Glycine" | "Triglu-5-Formyl-Tetrahydrofolate" |
| "Guanosine-5'-Diphosphate" | "Triphospate" |
| "Heme C" | "Tris(Hydroxymethyl)Aminomethane" |
| "Hydroxyalanine" | "Tryptophanyl-5'amp" |
| "Icosapent" | "Vinblastine" |
| "Imatinib" | "Vincristine" |
| "Imidazole" | "Vindesine" |
| "Isopropyl Alcohol" | "Vinorelbine" |
| "Itraconazole" | "Vorinostat" |
| | |

# Apêndice E

# Artigos para validação dos resultados

http://www.ncbi.nlm.nih.gov/pubmed?term=14736154

http://www.ncbi.nlm.nih.gov/pubmed?term=14736154

http://www.ncbi.nlm.nih.gov/pubmed?term=19428657

http://www.ncbi.nlm.nih.gov/pubmed?term=19136004

http://www.ncbi.nlm.nih.gov/pubmed?term=22795629

http://www.ncbi.nlm.nih.gov/pubmed?term=18675305

http://www.ncbi.nlm.nih.gov/pubmed?term=22384084

http://www.ncbi.nlm.nih.gov/pubmed?term=22404785

http://www.ncbi.nlm.nih.gov/pubmed?term=20104620

http://www.ncbi.nlm.nih.gov/pubmed?term=22546550

http://www.ncbi.nlm.nih.gov/pubmed?term=22639416

http://www.ncbi.nlm.nih.gov/pubmed?term=22626931

http://www.ncbi.nlm.nih.gov/pubmed?term=22151036

http://www.ncbi.nlm.nih.gov/pubmed?term=22546550

http://www.ncbi.nlm.nih.gov/pubmed?term=22508312

http://www.ncbi.nlm.nih.gov/pubmed?term=18456347

http://www.ncbi.nlm.nih.gov/pubmed?term=22079692

http://www.ncbi.nlm.nih.gov/pubmed?term=22691309

http://www.ncbi.nlm.nih.gov/pubmed?term=21629662

http://www.ncbi.nlm.nih.gov/pubmed?term=11697726

http://www.ncbi.nlm.nih.gov/pubmed?term=9211502

http://www.ncbi.nlm.nih.gov/pubmed?term=22363749

http://www.ncbi.nlm.nih.gov/pubmed?term=11286798

http://www.ncbi.nlm.nih.gov/pubmed?term=22363749

http://www.ncbi.nlm.nih.gov/pubmed?term=11286798

http://www.ncbi.nlm.nih.gov/pubmed?term=23230440

http://www.ncbi.nlm.nih.gov/pubmed?term=16332290

http://www.ncbi.nlm.nih.gov/pubmed?term=22508306

http://www.ncbi.nlm.nih.gov/pubmed?term=22508306

http://www.ncbi.nlm.nih.gov/pubmed?term=18657458

http://www.ncbi.nlm.nih.gov/pubmed?term=7007881

http://www.ncbi.nlm.nih.gov/pubmed?term=22394478

http://www.ncbi.nlm.nih.gov/pubmed?term=21138769

http://www.ncbi.nlm.nih.gov/pubmed?term=21629693

http://www.ncbi.nlm.nih.gov/pubmed?term=19900395

http://www.ncbi.nlm.nih.gov/pubmed?term=18435557

http://www.ncbi.nlm.nih.gov/pubmed?term=6370265

http://www.ncbi.nlm.nih.gov/pubmed?term=23230440

http://www.ncbi.nlm.nih.gov/pubmed?term=20735352

http://www.ncbi.nlm.nih.gov/pubmed?term=21732174

http://www.ncbi.nlm.nih.gov/pubmed?term=18066434

# Anexo A

# Lista de fenótipos do SGD

Relação dos fenótipos do Saccharomyces Genome Database, que pode ser acessada no endereço `http://www.yeastgenome.org/cache/PhenotypeTree.html`.

```
observable
    cellular processes
        cell death
            apoptosis
            necrotic cell death
        chromosome/plasmid maintenance
            chromosome segregation
            mitotic recombination
            mutation frequency
            silencing
            telomere length
            transposable element transposition
        intracellular transport
            autophagy
                mitophagy
                pexophagy
```

endocytosis

organelle distribution

endoplasmic reticulum distribution

Golgi distribution

mitochondrial distribution

nuclear position

peroxisomal distribution

vacuolar distribution

vesicle distribution

protein transport

mitochondrial transport

nuclear transport

nuclear export

nuclear import

peroxisomal transport

protein secretion

vacuolar transport

RNA localization

small molecule transport

mitotic cell cycle

cell cycle progression

cell cycle progression in G1 phase

cell cycle passage through START

cell cycle progression in G2 phase

cell cycle progression in M phase

cell cycle passage through the metaphase-anaphase transition

cell cycle progression in anaphase

cell cycle progression in early anaphase

cell cycle progression in late anaphase

cell cycle progression in mid anaphase

cell cycle progression in metaphase

cell cycle progression in telophase

cell cycle progression in S phase

cell cycle progression through the G1/S phase transition

cell cycle progression through the G2/M phase transition

cytokinesis

entry into G0 (stationary phase)

exit from G0 (stationary phase)

mitotic recombination

prion state

prion formation

prion inheritance

prion loss

stress resistance

dessication resistance

freeze-thaw resistance

hydrostatic pressure resistance

radiation resistance

ionizing radiation resistance

gamma ray resistance

X ray resistance

UV resistance

resistance to chemicals

acid pH resistance

alkaline pH resistance

ionic stress resistance

```
            metal resistance
            osmotic stress resistance
                hyperosmotic stress resistance
                hypoosmotic stress resistance
            oxidative stress resistance
        resistance to enzymatic treatment
        starvation resistance
        temperature sensitive growth
            cold sensitivity
            heat sensitivity
        thermotolerance
        toxin resistance
            killer toxin resistance
development
    budding
        bud direction
        budding index
        budding pattern
            axial budding pattern
            bipolar budding pattern
    filamentous growth
        invasive growth
        pseudohyphal growth
    lifespan
        chronological lifespan
        replicative lifespan
    sexual cycle
        mating response
```

```
            cell fusion
            mating efficiency
            nuclear fusion during mating
            pheromone production
            pheromone sensitivity
                pheromone-induced cell cycle arrest
                recovery from pheromone-induced cell cycle arrest
            shmoo formation
        mating type switching
        meiosis
            meiotic recombination
        sporulation
            spore germination
            spore wall formation
            sporulation efficiency
        sterile
essentiality
    inviable
    viable
fitness
    competitive fitness
    haploinsufficient
    viability
interaction with host/environment
    adhesion
    virulence
metabolism and growth
    anaerobic metabolism
```

```
    anaerobic growth
chemical compound accumulation
chemical compound excretion
nutrient utilization
    auxotrophy
    nutrient uptake
    utilization of carbon source
        fermentative metabolism
            fermentative growth
        respiratory metabolism
            mitochondrial genome maintenance
                mitochondrial rho- mutation frequency
            oxidative phosphorylation
            petite
            petite-negative
            respiratory growth
    utilization of iron source
    utilization of nitrogen source
    utilization of phosphorus source
    utilization of sulfur source
protein activity
protein/peptide accumulation
protein/peptide distribution
protein/peptide modification
redox state
RNA accumulation
RNA modification
vegetative growth
```

growth in exponential phase

growth in post-diauxic phase

survival rate in stationary phase

morphology

cellular morphology

bud morphology

bud neck morphology

cell shape

cell size

critical cell size at G2/M (cryptic G2/M cell size checkpoint)

critical cell size at START (G1 cell-size checkpoint)

mating projection morphology

subcellular morphology

cell wall morphology

chitin deposition

septum formation

cytoskeleton morphology

actin cytoskeleton morphology

spindle morphology

position of spindle pole body

endomembrane system morphology

endoplasmic reticulum morphology

Golgi morphology

nuclear morphology

nucleolar morphology

size of nucleus

peroxisomal morphology

vacuolar morphology

```
        lipid particle morphology
        mitochondrial morphology
        plasma membrane morphology
culture appearance
    biofilm formation
    colony appearance
        colony color
        colony sectoring
        colony shape
        colony size
    liquid culture appearance
        flocculation
```