

CENTRO UNIVERSITÁRIO FEEVALE

MAGNO SEVERO DE LIMA

EXTRAÇÃO DE SUBCONJUNTOS DE ONTOLOGIAS:
GENEONTOLOGY SLIM

Novo Hamburgo, novembro de 2008.

MAGNO SEVERO DE LIMA

EXTRAÇÃO DE SUBCONJUNTOS DE ONTOLOGIAS:
GENEONTOLOGY SLIM

Centro Universitário Feevale
Instituto de Ciências Exatas e Tecnológicas
Curso de Ciência da Computação
Trabalho de Conclusão de Curso

Professor Orientador: Rodrigo Rafael V. Goulart

Novo Hamburgo, novembro de 2008.

AGRADECIMENTOS

Gostaria de agradecer a todos os que, de alguma maneira, contribuíram para a realização desse trabalho de conclusão, em especial a minha esposa Cíntia e meu filho Raul pela paciência que tiveram comigo durante o desenvolvimento deste trabalho.

Aos amigos e às pessoas que convivem comigo diariamente, minha gratidão, pelo apoio emocional - nos períodos mais difíceis do trabalho.

Enfim, gostaria de agradecer ao meu orientador pela ajuda na elaboração deste trabalho.

RESUMO

A evolução da Internet e a crescente disponibilidade de informações dificultam a busca de informações co-relacionadas uma vez que os mecanismos de busca atuais não compreendem a semântica das informações. Na biologia esse problema pode ocorrer também, uma vez que em alguma parte do planeta, havendo algum biólogo realizando novos estudos e descobrindo algum novo elemento que foi utilizado para compor um novo produto, essa informação poderia ser útil para outro biólogo que pode estar realizando o mesmo estudo, porém em diferentes partes do planeta. Mas como os mesmos dados são tratados com terminologias diferentes por bases de dados, fica difícil de se encontrar essa pesquisa que muitas vezes acaba levando que a mesma se realize desde o ponto zero. A área a ser estudada trata sobre a geração de ontologias. O projeto a ser realizado visa estudar os passos necessários para o desenvolvimento de um sistema para extração de ontologias, que tem como objetivo principal gerar ontologias menores a partir de ontologias maiores, para área da biologia. Para esse fim, serão estudadas técnicas utilizadas por ferramentas já existentes no mercado. O propósito deste trabalho é efetuar testes e apresentar os resultados, e a partir destes desenvolver e documentar um protótipo de software desta aplicação.

Palavras-chave: Internet, biologia, ontologias.

ABSTRACT

The evolution of the Internet and the growing availability of information makes it difficult to search for co-related information since the current search engines do not understand the semantics of information. In biology this problem may also occur, because somewhere in the world, with some biologist conducting further studies and discovering a new element that was used to compose a new product, that information could be useful to another biologist who may be doing the the same study, but in different parts of the planet. But as the same data are treated with different terminology in databases, it is difficult to find such research that often ends up leading it to take place from the zero point. The area being studied is about the generation of ontologies. The project aims to be conducted to study the steps needed to develop a system for extraction of ontologies, which main goal is to generate ontologies minors from higher ontologies for areas of biology. To that end, will be studied by techniques used tools already on the market. The purpose of this study is to perform tests and submit the results, and from these develop and document a prototype of this software application.

Key words: Internet, biology, ontologies.

LISTA DE FIGURAS

Figura 1.1 – Ontologias em OWL_____	15.
Figura 1.2 – Código OWL_____	22.
Figura 2.1 – Componente celular_____	25.
Figura 2.2 – Processo biológico_____	26.
Figura 2.3 – Função molecular_____	27.
Figura 2.4 – Estrutura GO_____	28.
Figura 2.5 – Representação da GO_____	28.
Figura 3.1 – Ferramenta AmiGO_____	29.
Figura 3.2 – Resultado da busca feita na ferramenta AmiGO_____	30.
Figura 3.3 – Tela principal do OboEdit_____	31.

LISTA DE TABELAS

Tabela 1.1 – Metodologia para ontologias_____18.

Tabela 1.2 – Linguagens de representação de ontologias_____23.

LISTA DE ABREVIATURAS E SIGLAS

GO	Gene Ontology
SGD	Saccharomyces Genome Database
MGD	Mouse Genome Database
OWL	Web Ontology Language
DAML	Darpa Agent Markup Language
OIL	Ontology Inference Layer
XML	Extensible Markup Language
RDF	Resource Description Framework
XOL	Ontology Exchange Language
SHOE	Simple HTML Ontology Extensions
HTML	Hyper Text Markup Language
RNA	Ácido Ribonucleico

SUMÁRIO

1 INTRODUÇÃO	11
2 ONTOLOGIA.....	13
2.1 O que são ontologias	13
2.2 Vantagens do uso de ontologias.....	15
2.3 Composição de ontologias.....	16
2.4 Metodologias	17
2.5 Tipos de ontologias	19
2.6 Ferramentas para construção de ontologias.....	20
2.7 Linguagens para ontologias	21
3 GENE ONTOLOGY.....	24
3.1 O que é gene ontology.....	24
3.2 Organização da gene ontology.....	24
3.2.1 Componentes celulares	24
3.2.2 Processos biológicos	25
3.2.3 Função molecular.....	26
3.3 Estrutura da gene ontology	27
3.3.1 Árvore de representação.....	28
3.3.2 Anotação GO.....	28
3.4 Ferramentas GO	29

3.4.1 AmiGO.....	29
3.4.2 OboEdit	30
4 GO SLIM	32
4.1 O que é Go Slim.....	32
4.2 Porquê usar Go Slim	32
4.3 Produzindo um Slim utilizando o OboEdit.....	32
4.3.1 Produzindo um Slim Passo a passo	33
4.3.2 Considerações sobre o OboEdit.....	34
CONSIDERAÇÕES FINAIS.....	35
REFERÊNCIAS BIBLIOGRÁFICAS	36

1 INTRODUÇÃO

A Biologia, atualmente, perde muito tempo e esforço na busca de todas as informações disponíveis sobre cada pequena área de investigação. Fica ainda mais difícil pelas grandes variações de terminologias que podem ser de uso comum, num dado momento, que inibem a eficiência da pesquisa de ambos, computadores e pessoas. Por exemplo, buscar todos os produtos genéticos que estão relacionados a um antibiótico, duas ou mais bases de dados podem tratar uma mesma terminologia de um modo diferente tornando difícil para humanos e computadores encontrar termos equivalentes. Para auxiliar neste processo surgiu o consórcio *Gene Ontology*.

O projeto *Gene Ontology* (GO) é um esforço de colaboração para enfrentar a necessidade de coerência na descrição de produtos genéticos em diferentes bases de dados. O projeto começou com a colaboração entre organismo modelo de três bases de dados, FlyBase (*Drosophila*), *Saccharomyces Genome Database* (SGD) e do *Mouse Genome Database* (MGD), em 1998. Desde então, o GO Consórcio tem crescido para reunir bancos de dados, incluindo os maiores repositórios do mundo de vegetais, animais e genomas microbianos (www.geneontology.org GO, 2008).

De forma simples, para elaborar ontologias, definem-se categorias para as coisas que existem em um mesmo domínio. Ontologia é um "catálogo de tipos de coisas" em que se supõe existir um domínio, na perspectiva de uma pessoa que usa uma determinada linguagem (HINZ, 2006 apud SOWA, 2000). Trata-se de "uma teoria que diz respeito a tipos de entidades e, especificamente, a tipos de entidades abstratas que são aceitas em um sistema com uma linguagem" (HINZ, 2006 apud CORAZZON, 2002).

O projeto *Gene Ontology* (GO) foi desenvolvido, estruturado e controlado por 3 vocabulários (ontologias) que descrevem, gene produtos em termos dos seus processos biológicos associados, componentes celulares e funções moleculares, em uma espécie de forma independente. Há três aspectos distintos para este esforço: primeiro, o desenvolvimento

e a manutenção das ontologias, em segundo lugar, a anotação de genes produtos, o que implica em fazer associações entre ontologias, os genes e a genética, e o terceiro, o desenvolvimento de ferramentas que facilitem a criação, manutenção e uso de ontologias. O uso de termos GO, com as bases de dados colaborando facilita as consultas uniformes. Com os vocabulários controlados e estruturados de forma a poderem ser consultados em diferentes níveis: por exemplo, você pode usar GO para encontrar todos os produtos genéticos envolvidos no genoma do rato. Esta estrutura permite também anotar propriedades para atribuir aos produtos gene ou genes em diferentes níveis, em função da profundidade de conhecimento sobre essa entidade.

O foco principal deste estudo é utilizar a GO Slims que extrai um subconjunto da GO. Utilizar toda a GO inclui muitos termos na utilização de inferências. As GO Slims são criadas pelos usuários de acordo com suas necessidades, e podem ser específicas para espécies ou para determinadas áreas da ontologia. E a partir deste estudo viabilizar a criação de uma ferramenta para gerar ontologias para uma determinada área da biologia a ser definida.

2 ONTOLOGIA

Neste capítulo será apresentado definições de ontologia bem como as vantagens de uso e a sua composição, metodologias e algumas ferramentas para construção e linguagens para ontologias.

2.1 O que são ontologias

Ontologias, na filosofia têm sido usadas para descrever domínios naturais, ou seja, as coisas naturais do mundo como os tipos de existências e as relações temporais. Ontologia é um ramo da filosofia que lida com a natureza e a estrutura do ser. Esse termo foi inserido por Aristóteles em *Metafísica*. Termo também adotado pela comunidade de inteligência artificial para referir-se a conceitos e termos que podem ser usados para apresentar algumas áreas do conhecimento ou construir uma representação deste. Apesar de sua divulgação ao longo do tempo em que vem sendo utilizado, ainda não há um consenso, principalmente na comunidade da Ciência da Computação sobre a semântica do termo “ontologia”.

Uma ontologia define os termos utilizados para descrever e representar uma área de conhecimento. Ontologias são utilizadas pelas pessoas, bases de dados e aplicações que necessitam de partilhar informações de um domínio (W3C, 2004).

Ontologia é o ramo de metafísica interessada em identificar, em termos muito gerais, os tipos das coisas que existem na realidade, e como descrevê-las. Por exemplo, a observação que o mundo é feito de objetos específicos que podem ser agrupadas em classes abstratas baseadas na partilha de bens é um típico compromisso ontológico (THE SEMANTIC WEB VISION).

Segundo www.geneontology.org GO (2008), ontologias são estruturadas como grafos acíclicos dirigidos, que são similares às hierarquias, mas diferem na medida em que o prazo mais especializado (filho) pode ser associado a mais do que uma expressão menos

especializada (pai). Por exemplo, o processo biológico hexose biosynthetic tem dois pais, processo hexose metabólico e processo monossacarídeo biosynthetic. Isto é, processo biosynthetic é um tipo de processo metabólico de uma hexose que é um tipo de monossacarídeo. Quando qualquer gene hexose biosynthetic envolvidos no processo é anotado para o presente prazo, é automaticamente anotado para ambos os processo hexose metabólica e processo monossacarídeo biosynthetic.

Segundo Smith (2007), Ontologia é a ciência do que é, do real e das estruturas de objetos, propriedades, eventos, processos e relações em todos os domínios da realidade.

Segundo Matos (2008) apud Almeida (2003), historicamente o termo ontologia tem origem no grego *ontos*, ser e *logos*, palavra. É um termo introduzido na filosofia com o objetivo de distinguir o estudo do ser humano como tal, do estudo de outros seres das ciências naturais. A origem é a palavra aristotélica “categoria”, que pode ser usada para classificar e caracterizar alguma coisa.

Dentre as várias definições para ontologias a que mais é encontrada freqüentemente na literatura, e que é apresentada por Breitman (2006) proveniente de Gruber em seu artigo “*A Translation Approach to Portable Ontology Specification*”: “*Uma ontologia é uma especificação formal explícita de uma conceitualização compartilhada.*”

É importante explicar o significado das palavras contidas nesta frase. A palavra “conceitualização” traduz um modelo abstrato de algum fenômeno que identifica os conceitos relevantes para o próprio. A palavra “explícita” significa que os tipos de conceitos usados e as limitações do uso destes conceitos devem ser definidos de forma clara. A palavra “formal” refere-se que a ontologia deve ser passível de ser processada por uma máquina. E por fim a palavra “compartilhada” que reflete a noção de que ontologia captura um conhecimento consensual, quer dizer que, esse conhecimento não deve ser restrito a algumas pessoas mas sim aceito por um grupo de pessoas (BREITMAN, 2006).

Esta figura ilustra uma ontologia construída em uma linguagem OWL.

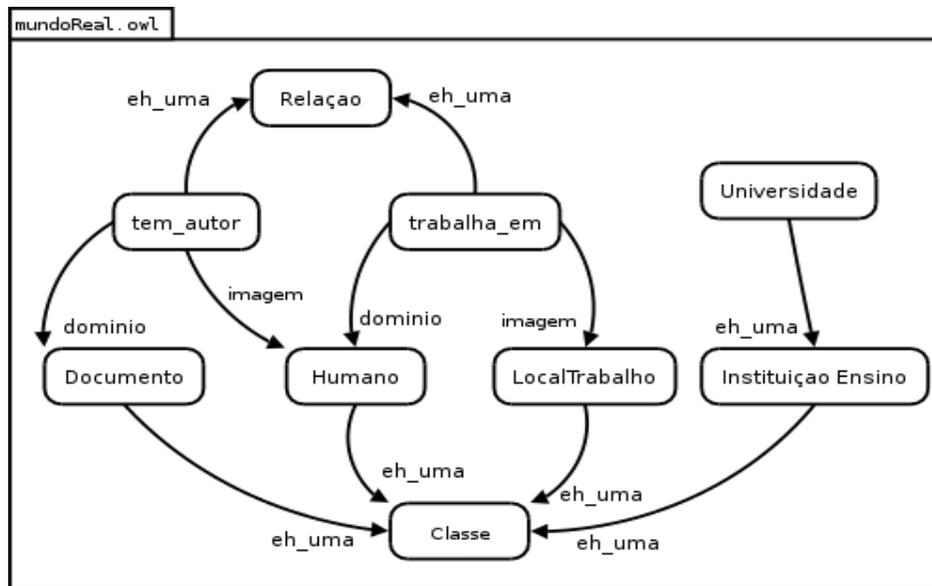


Figura 1.1: Ontologia em OWL

Fonte: <http://www.linux.ime.usp.br/~cef/mac499-04/monografias/rec/ghsilva/mundoReal.png>

2.2 Vantagens do uso de ontologias

Agora que já sabemos o que é ontologia, podemos destacar as vantagens em usá-las. Abaixo será apresentada uma lista com as principais vantagens da utilização de ontologias na computação:

- Fornecem vocabulário para representação do conhecimento.
- Vocabulário que tem por trás uma conceitualização que o mantém. Evitando assim interpretações imprecisas deste vocabulário.
- Permitem compartilhamento de conhecimento, sendo assim, caso exista uma ontologia que modele adequadamente certo domínio de conhecimento, esta pode ser compartilhada e usada por desenvolvedores de aplicações dentro desse domínio. Explicando melhor, considere que exista uma ontologia para o domínio de locadoras. Uma vez que esta ontologia está disponível, muitas locadoras podem montar seus catálogos usando o vocabulário colocado por esta ontologia sem a necessidade de fazer uma nova análise do domínio locadoras.
- Disponibiliza uma descrição exata do conhecimento. Diferente da linguagem natural, em que as palavras podem ter sua semântica muito diferente conforme seu argumento, a ontologia pode ser escrita em linguagem formal, não deixando espaço para o *gap* semântico existente na linguagem natural. Um exemplo seria

quando existe palavra de dupla interpretação como “rede”, a pessoa pode estar falando de uma rede de pescar ou uma rede para descansar. A interpretação da palavra pode ser atribuída a um conceito ou outro dependendo do estado mental da pessoa.

- O mapeamento da linguagem da ontologia é possível ser feito sem que com isso seja mudada a sua conceitualização, ou seja, a mesma conceitualização pode ser publicada em diversas línguas.
- É possível estender o uso da ontologia genérica de forma que ela se adapte a um domínio específico. Exemplo, se um indivíduo precisa de uma ontologia sobre motos para desenvolver uma aplicação é somente necessário encontrar uma ontologia sobre o domínio genérico de veículos, pode usar essa ontologia estendendo para o domínio da aplicação que neste caso é de motos.

Existem muitas outras vantagens em utilizar ontologias, essas que foram apresentadas são as principais, as outras na sua maioria são decorrentes destas. (HINZ, 2006).

2.3 Composição de ontologias

Os componentes básicos de uma ontologia são classes (organizadas em uma taxonomia), relações (representam o tipo de interação entre os conceitos de um domínio), axiomas (usados para modelar sentenças sempre verdadeiras) e instâncias (utilizadas para representar elementos específicos, ou seja, os próprios dados). (HINZ apud GRUBER, 1996).

Segundo Hinz (2006, p. 25), De forma resumida, uma ontologia é composta de: Domínio, Vocabulário, Glossário, Tesouro, Taxonomia e Rede Semântica.

O domínio define a abrangência da ontologia. Como dito em (HINZ apud RUSSEL; NORVIG, 2004, p. 245), o domínio de uma representação do conhecimento é uma parte do mundo sobre o qual deseja-se expressar algum conhecimento. Este domínio define os limites onde é definida toda a estrutura da ontologia.

Ainda segundo Hinz, vocabulário é composto de classes e instâncias. Classes apresentam conceitos dentro do domínio estimado. Conceitos que representam um conjunto abstrato de objetos.

Instâncias são eventos específicos de objetos em relação à classe avaliada. Uma instância apresenta conceitos, de uma forma individualizada, única e sólida, fazendo referência a um objeto real (documento, páginas da intranet e base de dados).

Dentro do vocabulário, as classes possuem características. Estas características são determinadas pelos seus atributos ou propriedades. Atributos que possuem valor nas instâncias. A taxonomia é composta das relações hierárquicas entre as classes. Essa relação define a estrutura onde classes pai possuem classes filhos, com atributos herdados das classes pai.

A composição das taxonomias se dá através de termos ou conceitos sobre o universo da informação registrada. Assim como o tesouro define conceitos, mas também permite outros tipos de relacionamentos entre os conceitos. Usando taxonomias ou tesouros, o usuário poderá optar por qual categoria de informação ele deseja acessar, combinando com a busca por palavra chave. Esse tipo de busca fornece resultados mais eficientes no acesso à informação. A deficiência apresentada é a não disponibilidade de uma linguagem padrão que descreva sua estrutura. Isso impede o compartilhamento dessa estrutura por diferentes sistemas, à construção de ferramentas de ajuda no desenvolvimento das estruturas e a utilização de sistemas de manipulação de regras de restrição (HINZ, 2006).

2.4 Metodologias

Encontra-se pouco desenvolvido o processo de construção de ontologias. Grande parte dos desenvolvedores usa seus próprios critérios para criar suas ontologias.

Com a falta de uma metodologia comum para o desenvolvimento, torna-se comum entre os desenvolvedores de ontologias passarem diretamente do passo de aquisição de conhecimento para o passo de implementação, o que acarreta nos seguintes problemas:

- Modelos conceituais de ontologia ficam implícitos no código da implementação.
- Dificuldades no reuso da ontologia, o *design* da ontologia e as decisões do projeto estão implícitos no código.
- Ocasiona problemas de comunicação devido às dificuldades que o *expert* no domínio da ontologia tem para entender o código da implementação. Torna-se um sério problema, pois ele tende a ser a principal fonte de informação sobre o domínio.

- Causa dificuldades no desenvolvimento de ontologias complexas, pois a passagem da aquisição de conhecimento para a implementação é abrupta.

Para evitar que estes tipos de problemas que foram citados acima ocorram devemos seguir alguma metodologia. Abaixo será apresentada uma tabela com algumas metodologias e suas definições:

Metodologia	Referência	Fases
Enterprise Ontology	(USCHOLD;KING,1995)	<ul style="list-style-type: none"> • Identificar a proposta da ontologia. • Construir a ontologia capturando, codificando e integrando conhecimento apropriado a partir de ontologias existentes. • Avalia a ontologia. • Documenta a ontologia.
TOVE (Toronto Virtual Enterprise)	(GRUNINGER;FOX,1995)	<ul style="list-style-type: none"> • Capturar cenários de motivação. • Formular questões de competência informal. • Terminologia Formal. • Formular questões em FOL. • Especificar axiomas. • Avaliar a ontologia.
Methontology	(GÓMEZ PÉREZ, 1998)	<ul style="list-style-type: none"> • Especificar o requisito. • Conceitualizar o domínio do conhecimento. • Formalizar o modelo conceitual em uma linguagem formal. • Implementar um modelo formal. • Fazer a manutenção de ontologias

		implementadas.
On-To-Knowledge	(STAAB et al.,2001)	<ul style="list-style-type: none"> • Requisitos da ontologia são capturados e especificados. • Questões de competência são identificadas. • Ontologias potencialmente reusadas são estudadas e uma versão “draft” da ontologia é construída.

Tabela 1.1: Metodologias para ontologias

Fonte: <http://ppginf.ucpel.tche.br/TI-arquivos/2006/VerlaniHinz/PPGINF-UCPel-TI-2006-2-10.pdf>

As metodologias apresentadas acima possuem características e abordagens diversas.

Não podemos fazer a união destas propostas apresentadas em uma única metodologia. Para verificar a utilidade das metodologias e compará-las, é necessário avaliar a ontologia resultante da aplicação de cada metodologia.

Segundo Hinz (2006) aput Gómez-Pérez (1999), “Apesar dos princípios de construção propostos, ainda não existem ferramentas de metodologia que guiem uma elaboração completamente padronizada, como em outros tipos de software. Em consequência disso, não há métodos sedimentados de validação, verificação, desenvolvimento e mesmo documentação de ontologias, que muito facilitaria seu reuso e aplicação. Engenharia e metodologias têm sido estudadas e propostas, visando preencher este vácuo”.

2.5 Tipos de ontologias

Segundo Becker (2005) aput Muñoz (2002), há vários critérios adotados para classificar as ontologias existentes. De acordo com Becker (2005) aput Guérios (2005), as ontologias podem ser classificadas quanto ao tipo e quanto à profundidade.

Quanto ao tipo, Becker (2005) aput Guarino (1998) classifica as ontologias nas seguintes categorias:

- **Ontologias Genéricas:** não dependem de um problema ou domínio em particular. Descrevem conceitos gerais, tais como espaço, tempo, matéria e objeto, ou seja, teorias básicas do mundo, de caráter abstrato e que sejam aplicáveis a qualquer

domínio (que constituam conhecimento de senso comum). São construídas com o intuito de serem especializadas na definição de conceitos em uma ontologia de domínio;

- **Ontologias de Domínio:** aplicam-se a um domínio do conhecimento, descrevendo conceituações particulares deste, por exemplo, de Medicina ou Direito. Constituem o tipo mais comumente desenvolvido;
- **Ontologias de Tarefas:** descrevem o vocabulário relacionado a uma atividade ou tarefa genérica, independentemente do domínio desta, tal como diagnose ou vendas;
- **Ontologias de Aplicação:** tal como as Ontologias de Tarefas, elas descrevem conceitos relacionados à execução de uma atividade ou tarefa, contudo, tais conceitos são dependentes do domínio e da tarefa representados;
- **Ontologias de Representação:** explicam as conceituações que fundamentam os formalismos de representação de conhecimento;

2.6 Ferramentas para construção de ontologias

Neste item será explicado o funcionamento de algumas ferramentas utilizadas na construção e edição de ontologias, não será um estudo completo, pois seria impossível apresentar todas as ferramentas disponíveis atualmente.

Como se trata de um trabalho custoso, qualquer ajuda na construção de ontologias pode representar ganhos significativos. Alguns exemplos de ferramentas para construção de ontologias são apresentados a seguir:

OilEd – é um simples editor, funciona como o “NotePad” dos editores de ontologias. Ele oferece suporte a ontologias desenvolvidas nas linguagens DAML+OIL e OWL. Essa versão não apresenta suporte para nenhuma metodologia de desenvolvimento de ontologias. O OilEd não tem suporte para a integração, versionamento e argumentação entre ontologias. Ele simplesmente permite ao usuário escrever ontologias e demonstra como usar o verificador FACT para comparação das mesmas.

OntoEdit – é um editor gráfico para edição de ontologias, permite a inspeção, navegação, codificação e alteração de ontologias. Este editor foi desenvolvido dentro do projeto On-To-Knowledge. Ele também implementa um processo específico para construção

de ontologias. Seu funcionamento ocorre da seguinte maneira: na primeira fase são coletados requisitos que descrevem as atividades em que a ontologia dará suporte. De um modo geral, essa tarefa será realizada pelos especialistas do domínio acompanhados pelos especialistas da modelagem. Essa fase deverá gerar subsídios que conduzirão o engenheiro de ontologia na decisão sobre os conceitos relevantes e a estrutura hierárquica na ontologia. Existe também a fase de refinamento que uma ontologia madura é produzida e orientada a aplicação de acordo com as especificações inseridas na fase anterior. No OntoEdit as ontologias são armazenadas em diferentes bancos relacionais e podem ser desenvolvidas em XML, FLogic, RDF(S) e DAML+OIL. (BREITMAN, 2006).

Ontolíngua – é um conjunto de serviços que permite o desenvolvimento de ontologias compartilhadas entre grupos. Também permite acesso a uma biblioteca de ontologias, tradutores para linguagem e um editor para criar e navegar pela ontologia (HINZ, 2006).

Protege – ambiente interativo para projeto de ontologias com código aberto, que proporciona uma interface gráfica para edição de ontologias e uma arquitetura para criação de ferramentas baseadas em conhecimento. A arquitetura é modulada e permite a inserção de novos recursos (HINZ, 2006).

Protege é um ambiente para criação e edição de ontologias, ferramenta de software livre, desenvolvida em java, tem sua arquitetura aberta, permitindo que os usuários possam desenvolver novos plugins capazes de aumentar o conjunto de funcionalidades da ferramenta. O Protege também utiliza um mecanismo de inferência para a verificação de ontologias e a classificação automática. É uma das ferramentas mais utilizadas no desenvolvimento de ontologias em OWL junto com OilEd (BREITMAN, 2006).

Chimaera – é um sistema de software que ajuda o usuário na criação e manutenção de ontologias. Suas principais funções são: combinação de múltiplas ontologias e diagnóstico de ontologias individuais ou múltiplas. Ajuda o usuário na tarefa de carregamento de bases de conhecimento em diversos formatos, reorganização taxionômica, resolução de conflitos com nomes, edição de termos e busca de ontologias. A intenção original desta ferramenta era ser aproveitada para combinar fragmentos de bases de conhecimento heterogêneas.

2.7 Linguagens para ontologias

Muitas linguagens foram criadas para representação de ontologias. Exemplos como: XOL, SHOE, DAML, RDF, RDF(S), OIL e OWL.

Dentre essas linguagens apresentadas, a OWL é a mais promissora e sofisticada do que suas linguagens bases, porque seus conceitos podem ser especificados por combinações lógicas como interseção, junção, ou complemento de outros conceitos. O OWL pode indicar que uma propriedade é transitiva, funcional, transitiva ou inversa, em relação a uma outra propriedade, instâncias, pertencem à quais conceitos, que conceitos e propriedades podem ser utilizadas entre indivíduos.

O OWL foi desenvolvida a partir das linguagens DAML e OIL e criada em cima da arquitetura XML e RDF. Foi proposta como padrão pelo W3C, agregando diversos pontos positivos das linguagens anteriores.

Na figura que vem a seguir está mostrando um pequeno código desenvolvido na linguagem OWL.

```

</owl:Ontology>
  <owl:Class rdf:ID="CoberturaDeTomate">
    <rdf:subClassOf>
      <owl:Restriction>
        <owl:onProperty>
          <owl:FunctionalProperty rdf:ID="temTempero"/>
        </owl:onProperty>
        <owl:someValuesFrom>
          <owl:Class rdf:ID="Suave"/>
        </owl:someValuesFrom>
      </owl:Restriction>
    </rdf:subClassOf>
  </owl:Class>

```

Figura 1.2: Código OWL
Fonte: <http://protege.stanford.edu>.

Nesta figura é apresentado um trecho do código em OWL de uma ontologia de pizzas retirado do site do PROTEGÉ.

Esta tabela apresentada um comparativo entre as principais linguagens de representação de ontologias.

Linguagens	Objetivo	Vantagens
RDF	Tornar possível a especificação de regras semânticas para dados baseados em XML de maneira padronizável e interoperável. Definir	RDFs define o significado da estrutura usando os tags estabelecidos em seu modelo. Provê tags para definir classes,

	um mecanismo independente de domínio. Definir um mecanismo para descrever recursos.	subclasses e propriedades. Permite restrições de domínio e de valores (range) e, propriedades.
OIL	Representação de uma semântica de informação acessível para a máquina na web.	Consegue promover pesquisas em torno do conceito semântico contido das informações das páginas web e não apenas através das palavras-chave.
DAML-OIL	Foi construída para ser uma linguagem de marcação semântica para recursos web (facilitar a representação dos conceitos de web semântica, descrevendo a estrutura de um domínio, ser compatível com padrões web existentes, fácil compreensão e uso, ter uma semântica formal associada, possuir um relativo poder de expressão)	Restrições de cardinalidade, tipos de dados, classes definidas e numerações, equivalência, extensibilidade, semântica formal, herança, inferência.
OWL	É um documento XML, RDF, RDFS Originado de vários projetos: DARPA (DAML) + EU Project (OIL)	Subdividido em 3 sub-linguagens: lite, DL e FULL
CYC	Proposta de uma ontologia geral com possibilidade de tratamento de contextos específicos (microteorias). Manipulação de grandes quantidades de conceitos.	Representação e tratamento de grandes ontologias suporte para inferência.

Tabela 1.2: Linguagem de representação de ontologias

3 GENE ONTOLOGY

Neste capítulo será explicado o funcionamento do consórcio Gene Ontology, assim como sua organização, estrutura e algumas ferramentas.

3.1 O que é gene ontology

É um conjunto de vocabulário controlado que definem conceitos e classificam a relação entre eles. Gene Ontology (GO) - Fornece vocabulário controlado para descrever o conhecimento biológico de genes e produtos genéticos.

O projeto da GO (Gene Ontology) é um consórcio entre vários centros de pesquisa na área de genoma. Essa ontologia refere-se a genes e o objetivo desse projeto é fornecer um conjunto de vocabulários estruturados para um domínio biológico específico que pode ser usado para descrever produtos de genes em um organismo. (www.geneontology.org GO, 2008).

Segundo www.lge.ibi.unicamp.br, Ontologia é um vocabulário de produtos de genes dinâmico controlado com potencial de ser aplicado a todos os organismos.

3.2 Organização da gene ontology

Os três princípios de organização da GO são componentes celulares, processos biológicos e função molecular. Um gene produto pode ser associado e localizado em um ou mais componentes celulares e é ativo em um ou mais processos biológicos, durante o qual desempenha uma ou mais função molecular (www.geneontology.org GO 2008).

3.2.1 Componentes celulares

Um componente celular é um elemento de uma célula, mas com a ressalva de que é parte de alguns objetos maiores, esta pode ser uma estrutura anatômica (por exemplo, retículo endoplasmático ou núcleo áspero) de um gene ou grupo de produtos (por exemplo) ribossomo, (proteassoma ou uma proteína dimer).

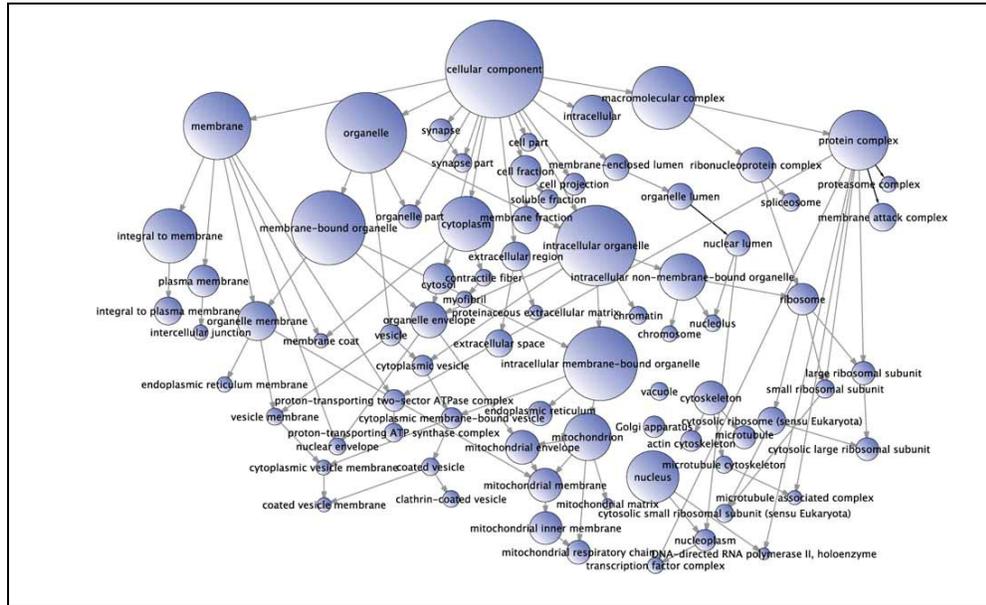


Figura 2.1: Componente celular

Fonte: <http://www.biomedcentral.com/1471-2164/9/96>

3.2.2 Processos biológicos

Um processo biológico é uma série de eventos realizados por um ou mais conjuntos ordenados de funções moleculares (apresentados a seguir). Exemplos de processo biológicos, no sentido amplo, são *cellular physiological process* ou *signal transduction*. Exemplos mais específicos são *pyrimidine metabolic process* ou *alpha-glucoside transport*. Pode ser difícil distinguir entre um processo biológico e uma função molecular, mas a regra geral é que um processo deve ter mais do que uma etapa distinta (www.geneontology.org GO 2008).

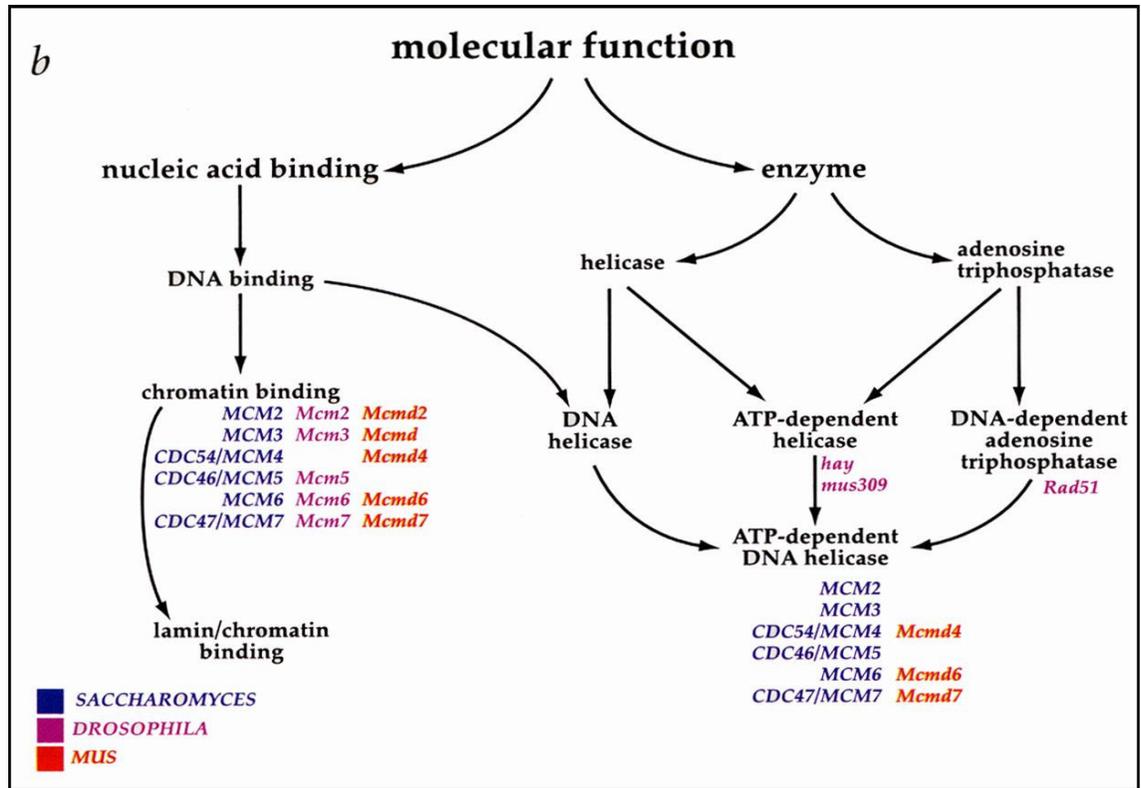


Figura 2.2: Processo biológico

Fonte: http://gchelpdesk.ualberta.ca/news/15apr04/AmiGO/amigo_images/tree.png

3.2.3 Função molecular

Uma função molecular descreve as atividades que ocorrem em nível molecular. Exemplos de termos funcionais gerais são catalytic activity, transporter activity, e binding. Exemplos de termos mais restritos *adenylate cyclase activity* ou *Toll receptor binding*. O nome do produto de um gene pode se confundido com a sua função molecular, e por esse motivo que no GO muitas funções moleculares são adicionados com a palavra “activity”.



Figuar 2.3: Função molecular
 Fonte: <http://proteinfuction.net/index.php/GO>

Este é um exemplo de uma classificação genética de um termo GO.

Nome do Gene: **BRCA1**

Função Molecular: **protein binding**

Processo Biológico: **DNA Replication and Chromosome Cycle**

Componente Celular: **nucleus**

3.3 Estrutura da gene ontology

A figura quem vem assegurar mostra como é a estrutura do consórcio GO, e como é definida a relação entre os conceitos. E como se relacionam os termos: “cell” “nucleus” “membrane”.

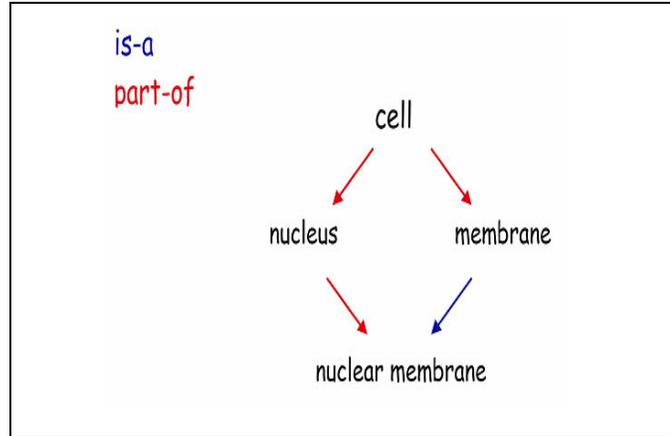


Figura 2.4: Estrutura GO

Fonte: www.geneontology.org GO 2008

3.3.1 Árvore de representação



Figura 2.5: Representação da GO

Fonte: www.geneontology.org GO 2008

3.3.2 Anotação GO

Anotação é o processo de atribuição GO termos de gene produtos. A anotação de dados no banco de dados GO é feito por membros do Consórcio GO, e do Consórcio é

incentivar ativamente novos grupos a contribuem para iniciar a declaração. Esta anotação pode ser manual ou eletrônica.

3.4 Ferramentas GO

Atualmente no mercado existem muitas ferramentas para criação e edição de ontologias, mas para conseguir estudar bem o funcionamento delas vou selecionar dentre elas duas que fazem parte do consórcio Gene Ontology e que estão relacionados com meu estudo. As ferramentas que escolhi para fazer o estudo são essas: AmiGO e OboEdit.

3.4.1 AmiGO

A ferramenta AmiGO está disponível no site da Gene Ontology para que os usuários possam pesquisar produtos genéticos e visualizar todos os termos com os quais estão relacionados, ele também permite que você procure por um termo “RNA” e vê todos os gene produtos anotados a ele, ou procurar por um outro gene produto e vê todas suas associações. Você também pode utilizar o browse de ontologias para ver os relacionamentos entre os termos assim como o número de gene produtos (www.geneontology.org GO, 2008).

Esta figura mostra a tela principal da ferramenta AmiGO.

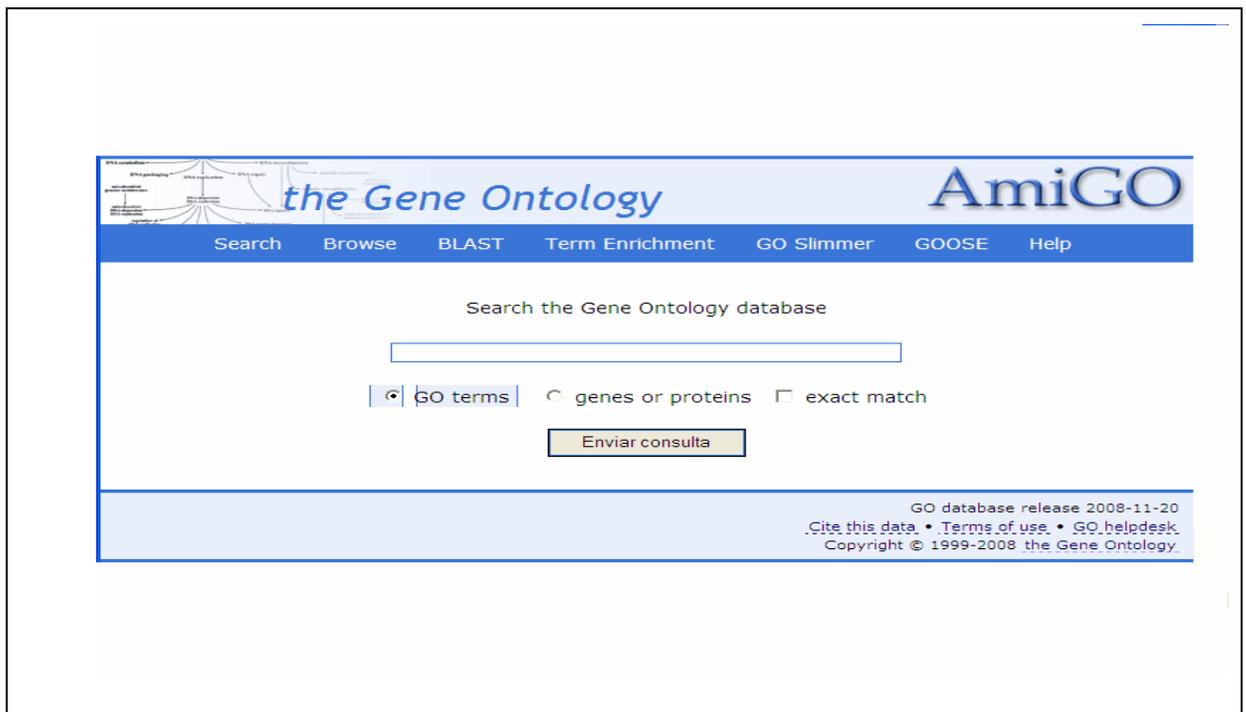


Figura 3.1: Ferramenta AmiGO
Fonte: www.geneontology.org

Com a ferramenta AmiGO você insere um termo GO e ele mostra todos os produtos de gene anotados para este, ou procura por produto de gene e mostra suas associações. Na figura a seguir foi colocado para busca o termo “RNA” e o resultado para esta busca foi de 932 associações, você pode navegar através dos resultados, é só selecionar o termo desejado. Neste caso não foi colocado nenhum tipo de filtro, mas ele permite que você insira filtros para diminuir o número de resultados, como exemplo pode-se colocar um filtro para que o resultado mostrado seja só de termos relacionados com o componente celular ou um outro filtro em que mostre somente as associações onde aparece no nome o termo “RNA” e não nas definições. Se alguém tiver interesse em explorar essa ferramenta ela está disponível no site da GO em: <http://www.godatabase.org/cgi-bin/go.cgi>.

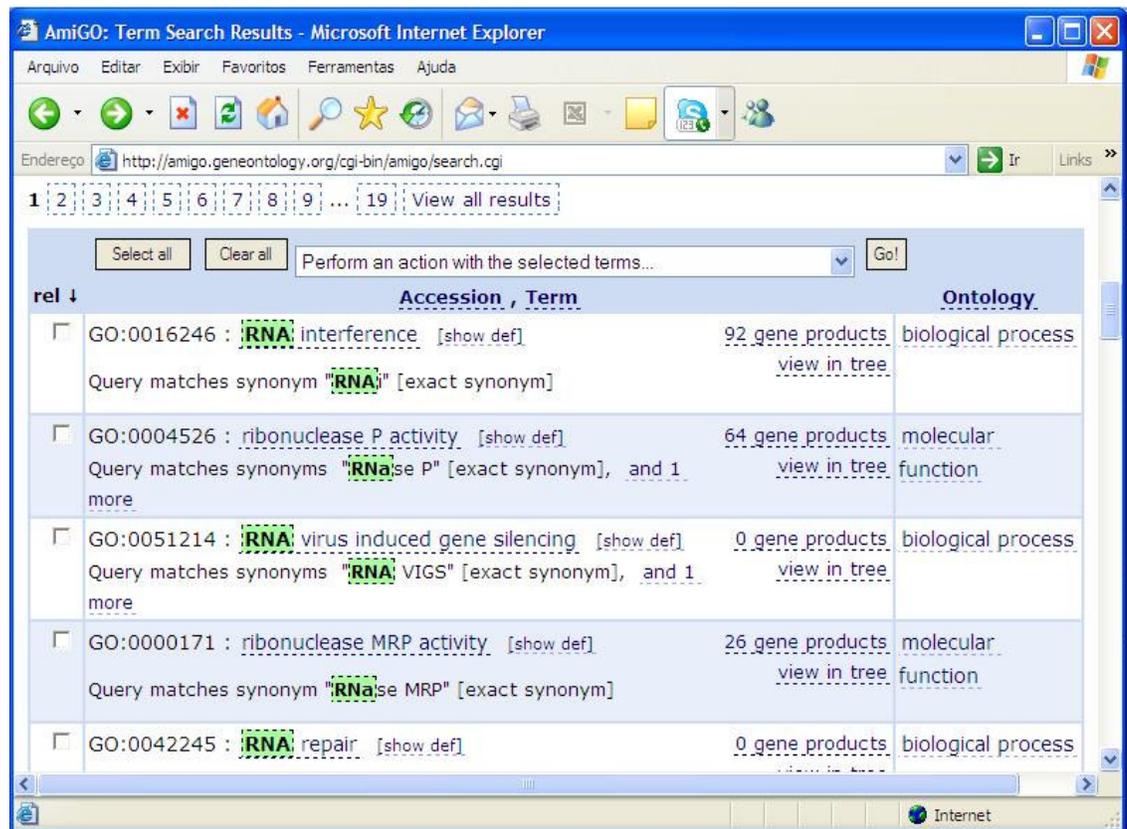


Figura 3.2: Resultado da busca feita na ferramenta AmiGO

3.4.2 OboEdit

É uma ferramenta gráfica que foi desenvolvida e é mantida pelo consórcio da Gene Ontology. Este editor foi implementado em java, e serve para editar e visualizar ontologias. O OboEdit é particularmente muito valioso para edição e visualização de ontologias na área da

biomedicina. Fornece uma interface amigável para os biólogos, e é excelente para a rápida geração de grandes ontologias, incidindo sobre as relações entre classes relativamente simples. OboEdit por tratar-se de uma ferramenta open source pode ser baixado gratuitamente no site da Go, em: www.geneontology.org.

A figura abaixo ilustra a tela principal do OboEdit.

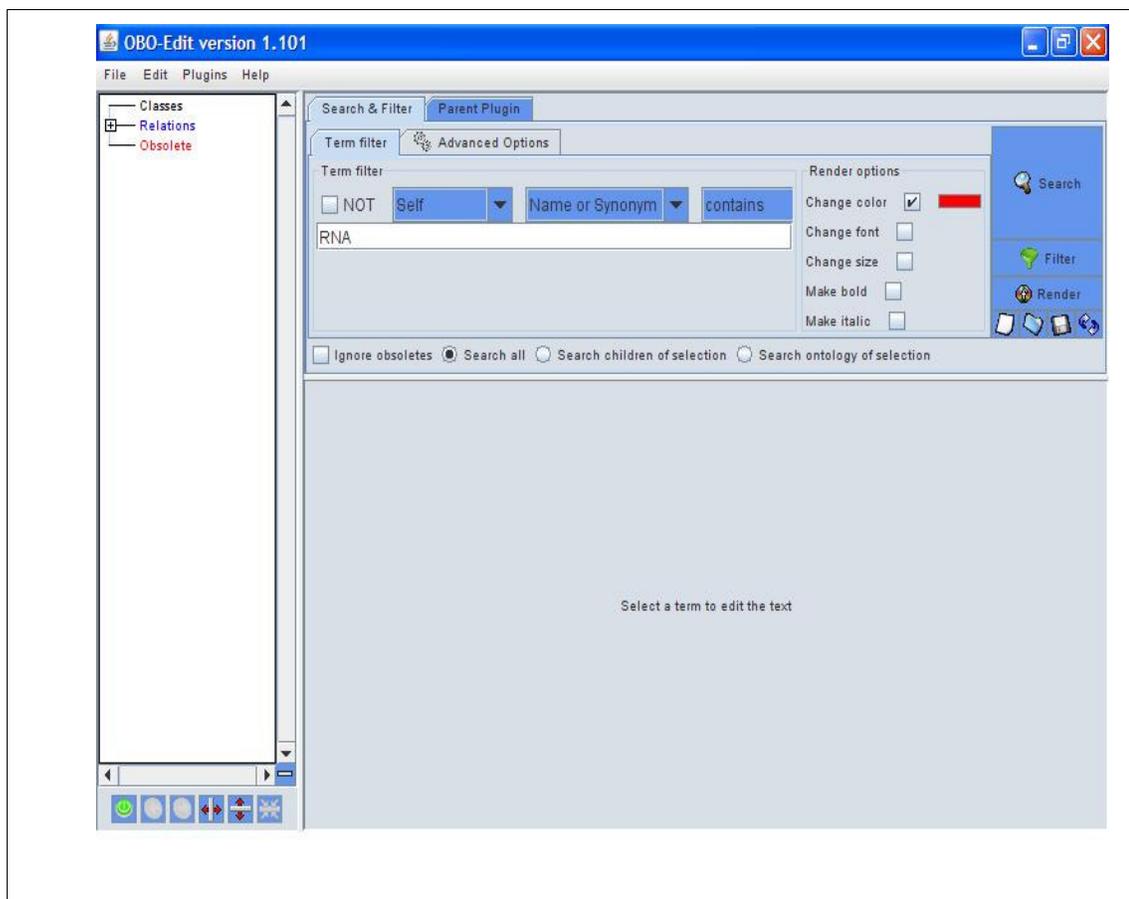


Figura 3.3: Tela principal do OboEdit

4 GO SLIM

Neste capítulo será explicado como funciona o Go Slim e porquê é mais fácil utilizá-lo do que usar toda a GO e também vou demonstrar a produção de um Slim utilizando a ferramenta OboEdit.

4.1 O que é Go Slim

É um subconjunto da GO, particularmente são úteis para dar um resumo dos resultados da anotação GO de um genoma. GO Slims são criados pelos usuários de acordo com suas necessidades, e podem ser específicas ou para uma área particular da ontologia.

4.2 Porquê usar Go Slim

Go Slim é usado simplesmente pelo motivo de que usar toda a GO não é satisfatório pelo grande volume de termos a ela associados. Os Slims são gerados a partir de uma parte da Go ou também de outros Slims. Você pode pegar as anotações de uma GO plantas e a partir dessa gerar um novo Slim específico de uma determinada espécie de planta.

4.3 Produzindo um Slim utilizando o OboEdit

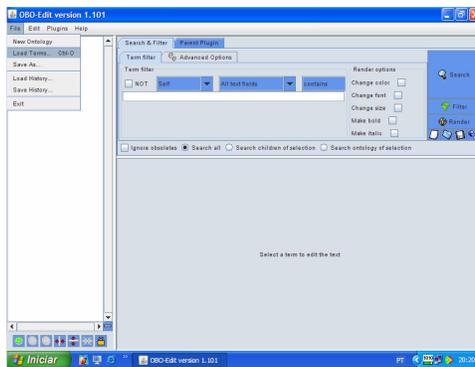
Pode ser utilizado um dos exemplos de Slims baixados do site da GO.

Após abrir o OboEdit selecione no menu file a opção: Load Terms, em seguida aparecerá uma janela onde deverá ser selecionado o tipo de adaptador, selecione OBO Flat File Adapter, após utilize a opção browse para selecionar o slim que efetuará a busca, clique em OK e o arquivo será carregado. Na opção Term filter insira um termo para efetuar a busca em seguida clique em Search, caso seja encontrado algum termo relacionado com este aparecerá na árvore à esquerda, caso queira realçar os termos você pode marcar na opção Render options o item Change color com a cor desejada para os termos. Em seguida selecione os

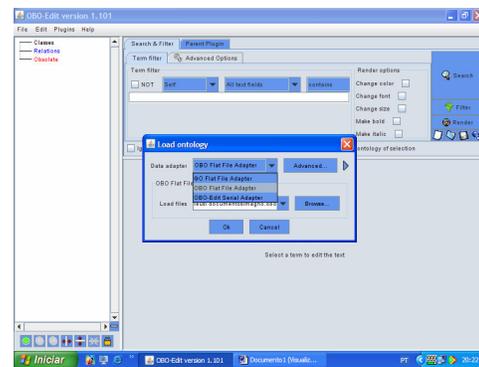
termos que você quer nos resultados encontrados e clique em filter. Este filtro pode ser salvo para que possa utilizá-lo em outros slims. O filtro pode ser salvo no disquete que aparece logo abaixo da opção Render em Save a filter to disk e nesse mesmo item você pode carregar os filtros na opção ao lado Load a filter from disk ou criar novo filtro em Create new filter.

Após aplicar o filtro pode-se salvar esse slim junto com o filtro criado na opção file → Save as. Para testar o slim criado abra ele no OboEdit.

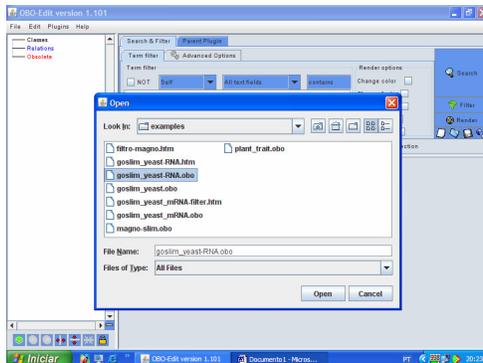
4.3.1 Produzindo um Slim Passo a passo



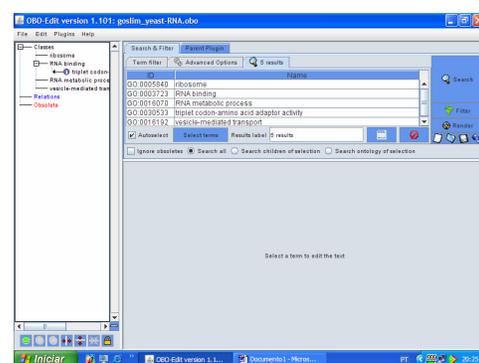
1ª Tela – Busca arquivo para análise.



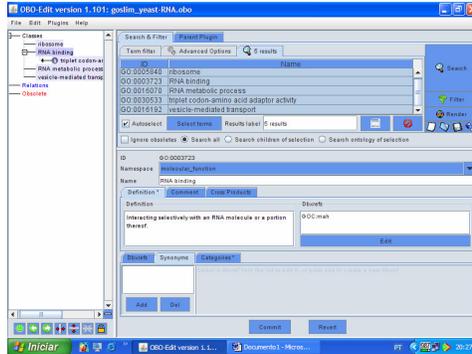
2ª Tela – Selecionar tipo do adaptador.



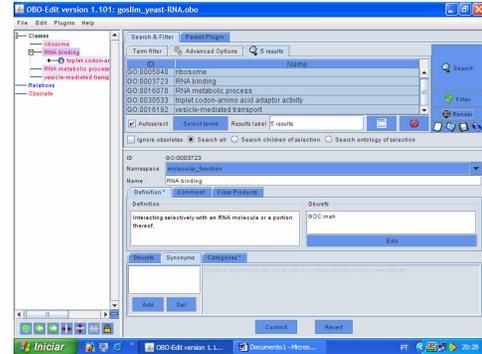
3ª Tela – Seleção do arquivo na pasta exemplos do OboEdit.



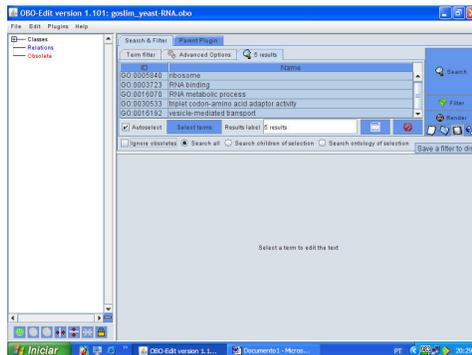
4ª Tela – Inserir o termo para pesquisa, clicar em Search (neste exemplo foi inserido o termo “RNA”), que obteve 5 associações).



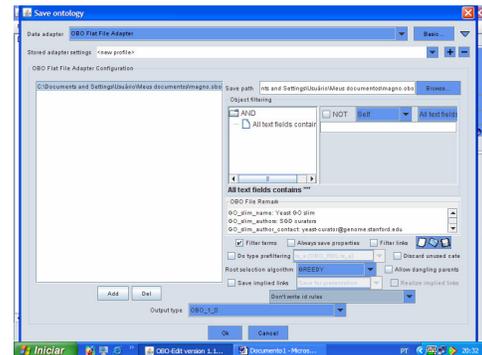
5ª Tela – Selecionar os termos que deseja para o filtro, ao selecionar você notará que eles ficam com fundo azul.



6ª Tela – Caso queira mudar a cor dos itens selecionados, você deve clicar na aba Term filter e marcar a opção Change color e ao lado escolher a cor conforme desejado, neste exemplo foi selecionada a cor vermelha.



7ª Tela – Clicar em Filter para criá-lo e depois conforme figura selecionar a opção para salvar este filtro. Caso queira salvar o Slim, clique no menu File e depois Save As, que abrirá a tela seguinte.



8ª Tela – Nesta tela você salva o Slim com a opção de salvar junto o filtro. Em Save path é inserido o caminho e o nome do Slim. Para salvar o filtro junto você deve carregar o filtro na opção que aparece ao lado da caixa de seleção Filter links, na opção salvar pode ser criado novos filtros.

4.3.2 Considerações sobre o OboEdit

Nos testes que realizei com o OboEdit pude observar que é uma ferramenta muito eficiente para gerar ontologias, através do help você consegue gerar ontologias, seguindo as instruções passo a passo, claro que você tem que saber pelo menos um pouco de inglês para conseguir utilizar a ferramenta. Após criar as ontologias você pode testá-las abrindo no OboEdit. Depois de você entender o funcionamento da ferramenta ela se torna fácil e muito útil para a criação e manutenção de ontologias. Uma vantagem que considero importante é que o OboEdit está disponível tanto em plataforma Windows como também em Linux.

CONSIDERAÇÕES FINAIS

No decorrer deste trabalho pude observar o quanto às ontologias são úteis para o desenvolvimento do conhecimento, seja ela aplicada na área da biologia ou em outras áreas, com a evolução da informática e da Internet ficou mais fácil desenvolver esse projeto uma vez que se consegue atingir cada vez mais grupos interessados a contribuir com a evolução das ontologias. As ontologias foram criadas para relacionar um termo ou produto com outros, um exemplo que se pode citar é uma ontologia de vinhos, onde podemos criar uma ontologia especificando a relação entre os tipos de vinhos com as uvas utilizadas na produção desses vinhos, também podem ser criadas ontologias relacionando os vinhos com algum alimento por exemplo podemos dizer que vinho branco pode ser servido com peixe, etc. De certa forma a ontologia nos ajuda a não ter que criar sempre do zero algum projeto uma vez que alguém já tenha criado algo que podemos utilizar e a partir desse projeto gerar um novo, seria como se sempre tivéssemos que reinventar a roda. Na biologia não é diferente podemos utilizar ontologias já criadas e a partir dessas gerar novas e sempre deixar a disposição de outras pessoas que queiram utilizá-las como fonte de estudo para criar outras ontologias. O projeto Gene Ontology (GO) é um grande repositório de ontologias, para quem deseja se aprofundar mais sobre o assunto ontologia na biologia, sugiro que consulte o site da GO em: www.geneontology.org.

REFERÊNCIAS BIBLIOGRÁFICAS

BREITMAN, Karin Koogan. **Web Semântica a Internet do Futuro**. Editora LTC, Rio de Janeiro, 2005.

HINZ, Verlaní Timm. **Proposta de Criação de uma Ontologia de Ontologias**. Disponível em: <http://ppginf.ucpel.tche.br/TI-arquivos/2006/VerlaníHinz/PPGINF-UCPel-TI-2006-2-10.pdf>. Acessado em: 05/09/2008.

Gene Ontology. **The Gene Ontology**. Disponível em: <http://www.geneontology.org>. Acessado em: 18/08/2008.

SMITH, Barry. **Ontology and Information Systems**. [s.l.: s.n.], 2003. Disponível em: < <http://ontology.buffalo.edu/smith/articles/ontologies.htm> >. Acesso em: 20/10/2008.

Gene Ontology. **What is Ontology**. Disponível em: http://www.geneontology.org/teaching_resources/presentations/2007-11_lungmeeting_dhill.ppt. Acessado em: 03/11/2008

W3C. **OWL Web Ontology Language Use Cases and Requirements**. Disponível em: <http://www.w3.org/TR/webont-req/#onto-def>. Acessado em: 05/09/2008.

The Semantic Web Vision. Disponível em: <http://mitpress.mit.edu/books/chapters/0262012103chap1.pdf>. Acessado em: 04/09/2008.

LGE. **Bancos de Dados de Genomas**. Disponível em: http://www.lge.ibi.unicamp.br/pagina/estrutura/montagem_script/exibe_ferramenta_pagina.php. Acessado em: 11/11/2008.

BECKER, Júnior. **Ontologia Lingüística para apoio a Ferramentas de Recuperação de Informações e de Text Mining**. Disponível em: <http://ead.feevale.br/tc>. Acessado em: 13/11/2008.

MATOS, Ely Edison da Silva. **Um Framework baseado em Ontologias com Serviços Web para Modelagem Conceitual em Biologia Sistêmica**. Disponível em: http://www.ufjf.br/mmc/files/2008/09/dissertacao_ely_matos.pdf. Acessado em: 14/11/2008.