PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Universidade Federal do Rio de Janeiro

Vivian dos Santos Silva

Uma abordagem para alinhamento de ontologias biomédicas para apoiar a anotação genômica

DISSERTAÇÃO DE MESTRADO

Instituto de Matemática

Núcleo de Computação Eletrônica

Vivian dos Santos Silva

UMA ABORDAGEM PARA ALINHAMENTO DE ONTOLOGIAS BIOMÉDICAS PARA APOIAR A ANOTAÇÃO GENÔMICA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática

Orientadores: Profa. Maria Luiza Machado Campos
Prof. João Carlos Pereira da Silva

Rio de Janeiro
2010

S586 Silva, Vivian dos Santos.
Uma abordagem para alinhamento de ontologias biomédicas para apoiar a anotação genômica / Vivian dos Santos Silva. 2010.
111 f.; il.

Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro. Instituto de Matemática. Núcleo de Computação Eletrônica, 2010.

Orientadores: Maria Luiza Machado Campos, João Carlos Pereira da Silva

1. Alinhamento de Ontologias – Teses. 2. Anotação Genômica. 3. Bioinformática. I. Campos, Maria Luiza Machado. II. Silva, João Carlos Pereira da. III. Universidade Federal do Rio de Janeiro. Instituto de Matemática. Núcleo de Computação Eletrônica. IV. Título.

CDD

Vivian dos Santos Silva

UMA ABORDAGEM PARA ALINHAMENTO DE ONTOLOGIAS BIOMÉDICAS PARA APOIAR A ANOTAÇÃO GENÔMICA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Aprovada em: Rio de Janeiro, 27 de agosto de 2010.

_____
(Profa. Maria Luiza Machado Campos, Ph.D.,PPGI/UFRJ)

_____
(Prof. João Carlos Pereira da Silva, D.Sc., PPGI/UFRJ)

_____
(Profa. Jonice de Oliveira Sampaio, D.Sc., PPGI/UFRJ)

_____
(Profa. Maria Claudia Reis Cavalcanti, D.Sc., IME)

_____
(Profa. Maria Luiza de Almeida Campos, D.Sc., UFF)

RESUMO

SILVA, Vivian dos Santos. **Uma abordagem para alinhamento de ontologias biomédicas para apoiar a anotação genômica**. Rio de Janeiro, 2010. Dissertação (Mestrado em Informática). Programa de Pós-Graduação em Informática, Universidade Federal do Rio de Janeiro, 2010.

Os recentes avanços nas pesquisas genômicas têm exigido cada vez mais recursos tecnológicos, que proveem suporte à descoberta, sequenciamento e descrição de novos organismos, assim como a disponibilização e acesso às informações resultantes dessas pesquisas. A anotação genômica, tarefa de atribuir uma descrição a cada sequência genômica descoberta, é uma atividade importante dentro do processo de sequenciamento, e conta com a utilização de ontologias para que seja mantido um vocabulário uniforme. Apesar de existirem muitas ontologias biomédicas, resultados dos esforços do consórcio *Open Biomedical Ontologies* (OBO), frequentemente apenas a Gene Ontology é utilizada no processo de anotação. A utilização de diversas ontologias, em conjunto com a GO, poderia enriquecer o vocabulário utilizado, trazendo maior riqueza de detalhes à anotação, sendo para isso necessário a identificação de equivalências entre os termos da GO e das demais ontologias. Este trabalho apresenta uma abordagem para alinhamento de ontologias biomédicas dentro do processo de anotação genômica que tem como objetivo, a partir de um termo da GO utilizado na anotação e recuperado automaticamente, identificar termos equivalentes em outras ontologias biomédicas, permitindo ao anotador escolher qual dentre eles possui a descrição mais detalhada e pode consequentemente trazer mais detalhes à anotação. Além da utilização da ferramenta FOAM - que se caracteriza por uma estratégia iterativa que considera diversas características das ontologias no cálculo de equivalências - como base da abordagem, dois pontos principais nortearam seu desenvolvimento: a escolha de um subconjunto de medidas de similaridade adequadas às características das ontologias da OBO, e a utilização de ontologias de fundamentação, que permitem a análise da natureza conceitual dos termos comparados, servindo como mais um parâmetro no cálculo de similaridades e diminuindo a possibilidade de associações entre termos derivados de categorias distintas. Realizando-se o alinhamento entre duas ontologias biomédicas, GO e INOH Event, notou-se um aumento de 14% no número de alinhamentos corretos e uma diminuição de 5% na quantidade de associações incorretas, em relação aos resultados obtidos pelo FOAM sem nenhum recurso adicional, quando a abordagem é aplicada em conjunto com um valor de corte alto para a similaridade entre os termos. A melhoria na qualidade dos alinhamentos, representada pelo maior número de alinhamentos corretos, reforça a utilidade da abordagem proposta na complementação do trabalho do anotador.

ABSTRACT

SILVA, Vivian dos Santos. **Uma abordagem para alinhamento de ontologias biomédicas para apoiar a anotação genômica**. Rio de Janeiro, 2010. Dissertação (Mestrado em Informática). Programa de Pós-Graduação em Informática, Universidade Federal do Rio de Janeiro, 2010.

Recent advances in genome research have required an increasingly amount of technological resources, which provide support to the discovery, sequencing and description of new organisms, as well as the availability and access to resulting information. The genome annotation, the task of assigning a description to each discovered genome sequence, is an important activity within the process of sequencing, and relies on the use of ontologies in order to maintain a uniform vocabulary. Although there are many biomedical ontologies, which are results of the Open Biomedical Ontologies (OBO) Consortium efforts, often only the Gene Ontology is used in the annotation process. The use of different ontologies, along with GO, could enrich the vocabulary used in the annotation, bringing more details to it, and, in order to make it possible, it's necessary the identification of equivalences between terms from GO and other ontologies. This work presents an approach for aligning biomedical ontologies within the genome annotation process that aims, from a term used in the GO annotation and recovered automatically, identifying equivalent terms in other biomedical ontologies, enabling the annotator to choose which among them has the more detailed description, and can therefore bring more details to the annotation. In addition to using the ontology alignment tool FOAM - which is characterized by an iterative strategy that considers several ontology features in the similarity calculation - as the approach basis, two main points have guided its development: the choice of a subset of similarity measures suited to the characteristics of OBO ontologies, and the use of foundational ontologies, allowing the analysis of the conceptual nature of each term, serving as another parameter in the similarity calculation and reducing the possibility of associations between terms derived from different categories. Performing the alignment between two biomedical ontologies, GO and INOH Event, we noted a 14% increase in the number of correct alignments and a 5% decrease in the amount of incorrect associations, compared to the results obtained by FOAM executed with no further resources, when the approach is applied in conjunction with a high similarity cutoff value. The improvement in the alignment quality, represented by the higher number of correct alignments, reinforces the usefulness of the proposed approach in complementing the annotator work.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus por estar ao meu lado em cada segundo da minha vida, e por me conceder toda a força que me ajuda a superar todas as dificuldades, e que possibilitou mais esta conquista.

Agradeço a meus pais, Ivanir e Sebastião, e a toda minha família, pelo carinho e confiança que permanecem presentes mesmo à distância.

Agradeço aos amigos do Mestrado, que estiveram ao meu lado durante estes anos: Patrícia, Eliêmia, Alan, João Vítor, Inês, Miguel e Kelli, pelo companheirismo e apoio de todos os dias.

Agradeço à Linair, por compartilhar seus conhecimentos na área e me ajudar a definir o tema e delimitar o escopo da pesquisa.

Agradeço à bióloga e amiga Nicole Scherer, pelo auxílio na classificação de termos durante o experimento e pelas palavras de incentivo nesta etapa final, e aos também biólogos da Fiocruz André Pitaluga, Kary Ocaña e Diogo Tschoeke pela ajuda na validação do experimento.

Agradeço aos meus orientadores, professores Maria Luiza Machado Campos e João Carlos Pereira da Silva, pelo apoio na realização deste trabalho e pela liberdade e confiança que me depositaram na condução desta pesquisa.

Agradeço às professoras Jonice de Oliveira Sampaio e Maria Luiza de Almeida Campos, por aceitarem o convite para fazer parte da banca, e em especial à professora Maria Cláudia Reis Cavalcanti, por participar deste trabalho não só como membro da banca, mas também como colaboradora, tendo sido de fundamental importância no início desta pesquisa, ajudando a definir vários passos importantes na elaboração da abordagem.

A estes e a todos os outros que de alguma forma contribuíram para a conclusão deste trabalho, meu sincero muito obrigada!

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Com a evolução das pesquisas genômicas e descobertas frequentes de novos organismos a serem estudados, cresce a necessidade de *softwares* que permitam sua descrição detalhada, além de armazenar e gerenciar o grande número de informações geradas. Nesse contexto, cresce também a importância da Bioinformática, explorando tecnologias que, dentre outras atividades, permitem aos pesquisadores a manipulação de grandes bancos de dados genômicos e o compartilhamento de informações, como, por exemplo, sequências genéticas, com outros laboratórios e centros de pesquisa.

O processo de anotação genômica "compreende atividades que registram o recebimento e pré-análise de sequências e, em seguida, a análise, anotação e consolidação da anotação realizada sobre a sequência" (BELLOZE, 2007). As sequências e suas respectivas anotações passam a compor grandes bases de dados genômicas, que podem beneficiar pesquisas em diversas áreas como Saúde e Agricultura, entre outras.

Uma anotação é o registro do significado biológico de cada sequência descoberta (BELLOZE, 2007). Além de conter informações como observações feitas pelo biólogo e referências a publicações, entre outras, as anotações sobre os dados genômicos podem ser feitas com base em uma ontologia. Ontologias são formalmente descritas como "um conjunto de primitivas representacionais utilizadas para modelar um domínio de conhecimento ou discurso" (GRUBER, 2008). Estas primitivas consistem de classes, atributos e relacionamentos que compõem um vocabulário do domínio, permitindo que uma linguagem padronizada seja utilizada por todos os membros de uma comunidade.

A ontologia mais utilizada para a tarefa de anotação genômica é a Gene Ontology, resultado de um consórcio criado para desenvolver e utilizar ontologias que deem suporte a anotações de genes e seus produtos em uma ampla variedade de organismos de forma biologicamente significativa. A GO fornece uma linguagem sistemática que possibilita descrições consistentes dentro de três domínios biológicos chave: componente celular, processo biológico e função molecular (THE GENE ONTOLOGY CONSORTIUM, 2008).

1.1 Caracterização do problema

Apesar da grande utilidade da Gene Ontology, a rápida evolução das pesquisas biológicas começa a demandar a utilização de outras ontologias. O projeto OBO (*Open Biomedical Ontologies*), do qual a GO também faz parte, trabalha no desenvolvimento de diversas ontologias que têm como objetivo servir como referência dentro do domínio biomédico (OBO FOUNDRY, 2009). Além das ontologias de propósito geral, que podem ser utilizadas na descrição de qualquer organismo, outras são também voltadas para uma espécie ou assunto em particular, como anfíbios ou doenças humanas. A utilização em conjunto da Gene Ontology com as demais ontologias da OBO pode enriquecer a anotação genômica, gerando descrições mais completas e consistentes.

Apesar disso, existe uma grande dificuldade na utilização de outras ontologias durante o processo de anotação. A preferência pela GO se deve ao fato de a parte automatizada da anotação ser feita por similaridade, ou seja, quando uma sequência é anotada, sequências similares de outros organismos previamente anotados são buscadas em bases de dados genômicos. Como essas bases normalmente já foram anotadas utilizando-se a GO, os termos utilizados são transferidos para a anotação da nova sequência automaticamente.

Essa atribuição automática de termos da GO pode ser alterada na etapa manual da anotação, onde novos termos podem ser adicionados. Novamente, as facilidades oferecidas pela GO, que possui um amplo conjunto de ferramentas que permitem busca e navegação em sua hierarquia, fazem com que esta seja a opção mais adotada, em detrimento de outras ontologias.

Embora a GO capture boa parte dos aspectos que os biólogos desejam descrever sobre as sequências (BODENREIDER, STEVENS, 2006), com a ampla gama de ontologias oferecidas pela OBO, muitas vezes pode haver uma ontologia mais específica, voltada especialmente para a descrição do tipo de organismo que está sendo sequenciado, e cujos termos podem apresentar uma definição mais detalhada, o que pode influenciar diretamente na qualidade da anotação gerada. Porém, o anotador precisaria, em muitos casos, percorrer a hierarquia desta ontologia manualmente para encontrar o termo desejado, o qual em alguns casos, apesar de relativo ao mesmo conceito, pode ter um nome diferente daquele adotado pela GO. Como gera um esforço adicional, a utilização de outras ontologias biomédicas normalmente é desconsiderada, mesmo considerando que possam trazer benefícios ao processo de anotação.

1.2 Motivação

Para que seja possível a utilização de mais de uma ontologia durante a anotação, é necessário identificar equivalências entre seus termos. Se, durante a anotação de uma determinada sequência, é utilizado o termo *x* da GO, o anotador precisa saber quais termos nas demais ontologias possuem o mesmo significado de *x*. Deste modo, torna-se possível comparar suas definições e verificar quais possuem maior riqueza de detalhe, podendo, deste modo, enriquecer também a anotação.

A identificação de equivalências entre os termos de duas ou mais ontologias é possível através do alinhamento de ontologias. Ehrig (2007) define a tarefa de alinhar duas ontologias como "para cada entidade (conceito, relação ou instância) na primeira ontologia, tentar encontrar uma entidade correspondente, com o mesmo significado pretendido, na segunda ontologia". Embora essa definição implique em uma relação de igualdade entre as entidades alinhadas, interpretações mais genéricas também são possíveis. Por exemplo, Kalfoglou e Schorlemmer (2003) chamam de alinhamento de ontologias "a tarefa de estabelecer uma coleção de relações binárias entre os vocabulários de duas ontologias". Uma relação binária pode abranger tanto relações hierárquicas (entre superclasses e subclasses), como qualquer outra relação definida na ontologia. Deste modo, podemos utilizar o alinhamento para encontrar duas entidades similares, mas também uma entidade que é mais geral que outra, ou parte de outra, e assim por diante.

Como o anotador já dispõe do termo da GO, encontrado automaticamente na primeira etapa do processo de anotação, o alinhamento seria responsável por buscar este termo em outras ontologias, deixando a cargo do usuário apenas a tarefa de validação dos resultados encontrados. Desta forma, novos termos são disponibilizados sem que suas ontologias tenham que ser pesquisadas manualmente, podendo-se encontrar conceitos equivalentes àqueles expressos na GO mesmo que eles tenham recebido nomes diferentes.

1.3 Objetivo

O objetivo deste trabalho é elaborar uma abordagem para alinhamento de ontologias voltada para o domínio de Bioinformática, realizando uma análise das medidas de similaridade mais adequadas às características das ontologias da OBO, e contando também com o auxílio de

ontologias de fundamentação (*Foundational Ontologies*), que permitem a distinção de conceitos de acordo com sua natureza no nível ontológico (GUARINO, 1994).

Segundo Guizzardi (2005), uma ontologia de fundamentação é "um sistema formal, independente de domínio, de categorias e suas ligações que podem ser usadas para construir modelos de domínio específico". Essas categorias deixam clara a distinção entre uma *propriedade* de uma coisa, o *tipo* ao qual a coisa pertence, o *papel* desempenhado pela coisa, e assim por diante. Assim, além de características sintáticas e estruturais, podemos considerar também a natureza do conceito, fornecendo mais parâmetros para o cálculo da similaridade e descartando associações entre entidades de naturezas distintas.

Espera-se que assim sejam obtidos alinhamentos mais confiáveis, isto é, que seja retornado um número maior de associações corretas, automatizando ainda mais o processo e diminuindo o esforço necessário para a validação dos resultados obtidos. Assim, espera-se permitir ao anotador identificar rapidamente os termos de interesse, enriquecendo o vocabulário a ser utilizado e, consequentemente, a anotação gerada.

1.4 Organização do trabalho

Este trabalho está estruturado da seguinte forma: o Capítulo 2 introduz o processo de anotação genômica, bem como os esforços de padronização representados pela utilização de ontologias nesta tarefa. São levantadas também as principais característica das ontologias desenvolvidas e utilizadas no domínio biomédico.

O Capítulo 3 detalha a tarefa de alinhamento de ontologias, definindo-a formalmente, expondo algumas de suas aplicações e listando as principais técnicas utilizadas, em especial a técnica baseada em ontologias de topo, que será utilizada na abordagem proposta. Conceitos relacionados, como integração, *merge* e outros, que normalmente causam confusão devido à variedade de definições, são também apresentados. É realizado também um estudo sobre as medidas de similaridade usadas na identificação de termos equivalentes.

O Capítulo 4 apresenta um estudo sobre algumas das ferramentas de alinhamento de ontologias mais recentes, destacando seus pontos fortes e desvantagens. É realizada também uma análise comparativa das abordagens adotadas por cada uma dessas ferramentas, além de uma

avaliação do ponto de vista prático, com o intuito de definir qual apresenta maior potencial de reutilização.

O Capítulo 5 apresenta a abordagem proposta, detalhando sua estrutura, e aprofundando seu passo principal: o alinhamento de ontologias da OBO. São destacados os dois diferenciais deste passo: a escolha das medidas de similaridade adequadas às características das ontologias da OBO e a forma como ontologias de fundamentação apoiam o processo de alinhamento. Alguns trabalhos relacionados nesta área são também apresentados.

O Capítulo 6 descreve o experimento realizado com o objetivo de validar a abordagem proposta, analisando-se os resultados obtidos. Por fim, o Capítulo 7 traça as conclusões do trabalho, listando suas principais contribuições e indicando direções para possíveis trabalhos futuros.

2 ONTOLOGIAS BIOMÉDICAS NA ANOTAÇÃO GENÔMICA

Nos últimos anos, as pesquisas genômicas têm avançado rapidamente. Com o número crescente de genomas sendo sequenciados, começa a ficar cada vez mais evidente a explosão de informação genômica com as quais os biólogos precisam lidar. Para manipular esta grande quantidade de dados, o apoio computacional é fundamental, tanto para processar quanto para armazenar e permitir o acesso às informações por parte de diversos grupos de pesquisa. Deste modo, a Bioinformática, área multidisciplinar que viabiliza a análise em larga escala dos resultados de pesquisas de sequenciamento genômico, evolui também rapidamente.

Segundo Lacroix e Critchlow (2003), "Bioinformática pode se referir a praticamente qualquer esforço de colaboração entre biólogos ou geneticistas e cientistas da computação e deste modo cobre uma ampla variedade de domínios tradicionais da ciência da computação, incluindo modelagem, recuperação, mineração, integração, gerência e limpeza de dados, *data warehousing*, ontologias, simulação, computação paralela, tecnologia baseadas em agentes, computação em *grid* e visualização." O apoio computacional às pesquisas genômicas é indispensável por dois motivos: primeiro, porque muitos problemas na Bioinformática requerem que a mesma tarefa seja realizada diversas vezes, por exemplo, comparar uma nova sequência genômica com outras sequências de mesmo tipo armazenadas em bancos de dados a fim de descobrir similaridades. Segundo, porque *softwares* específicos podem resolver problemas como inferir o que cada sequência de nucleotídeos ou aminoácidos representa e interpretar biologicamente esses dados (BELLOZE, 2007).

A interpretação dos dados gerados é uma das fases mais importantes deste tipo de pesquisa. A anotação genômica consiste na descrição das sequências descobertas, atribuindo-lhes características biológicas. Deste modo, uma anotação é o registro do significado biológico de cada gene identificado. A composição de cada gene sequenciado, em termos de domínio de proteínas, após ser computada pode ser disponibilizada em bases de dados privadas ou públicas. Essas características podem ser imaginadas como a "sintaxe" do genoma, dado que sua definição é análoga à tarefa de interpretar uma linguagem natural. A "semântica" da anotação genômica pode ser imaginada como a determinação da função dos genes, pois muito pouco pode ser dito a esse respeito simplesmente através da composição da sequencia (FRISHMAN, VALENCIA,

2008). Associar anotações a fragmentos de genes, portanto, proporciona o contexto genômico para interpretar os dados obtidos.

A tarefa de anotação inicia-se com a extração e fragmentação do DNA do organismo estudado. De acordo com suas interpretações na bancada de experimentos biológicos, os pesquisadores consultam fontes de dados externas e executam programas de análise de sequências. Várias ferramentas automatizam parte do processo de anotação. Esse processo envolve filtragem, transformação e manipulação computacional de dados, porém frequentemente requer também esforço humano para correção e curagem. Deste modo, podemos identificar dois tipos de anotação: *automática*, que é gerada por programas de análise ou importadas por fontes de dados públicas, e *manual*, que é criada diretamente pelo pesquisador.

O processo de anotação normalmente é dividido em três etapas. Na primeira etapa são importados dados referentes às sequências das fontes de dados públicas, com o auxílio de ferramentas de Bioinformática. Essa etapa corresponde à anotação automática, onde as sequências obtidas passam por diversos programas que ajudam os anotadores a identificá-las. Na segunda etapa, especialistas observam os dados obtidos na primeira etapa e identificam as sequências de acordo com critérios pré-definidos. Após a identificação dos genes, na terceira etapa é feita a anotação manual, realizada pelo próprio pesquisador (LEMOS, 2005). A Tabela 1 mostra as atividades envolvidas na anotação e algumas ferramentas utilizadas para auxiliar cada uma delas. Essas tarefas não serão detalhadas aqui, Wagner (2006) apresenta um trabalho onde cada uma delas, bem como as ferramentas que as apoiam, são descritas mais detalhadamente.

Tabela 1. Atividades do processo de anotação genômica e principais ferramentas utilizadas
Fonte: Wagner (2006)

Etapas da Anotação	***Ferramentas***	***Referência***
Avaliação de qualidade	*Phred*	EWING, GREEN, 1998
Limpeza de vetores	*Crossmatch*	EWING *et al*., 1998
Montagem ou agrupamento de sequências	*Phrap* *CAP3*	EWING *et al*., 1998 HUANG, MADAN, 1999
Predição *in silico* de genes	*GLIMMER* *GeneMark*	DELCHER *et al*., 1999 BORODOVSKY, MCININCH, 1993
Análise de similaridade	*BLAST* *FASTA* *Interpro*	ALTSCHUL *et al*., 1997 PEARSON, LIPMAN, 1988 MULDER *et al*., 2005
Busca de homologias distantes	*HMMER* *SAM*	EDDY, 2003 KARPLUS *et al.,* 1998
Alinhamento de sequências	*Clustalw*	THOMPSON *et al*., 1994

	T-Coffee	NOTREDAME *et al*., 2000
Análises filogenéticas	*Phylip* *MEGA*	FELSENSTEIN, 2005 KUMAR *et al*., 2004
Busca de genes ortólogos/parálogos	*OrthoMCL*	LI *et al*., 2003
Vias metabólicas	*KEGG*	KANEHISA, GOTO, 2000
Códons usuais (*Codon Usage*)	*Cusp*	RICE *et al*., 2000
Conteúdo G+C	*Geecee*	RICE *et al*., 2000

Como as anotações poderão ser acessadas por diversos centros de pesquisa de qualquer lugar do mundo, é importante que a linguagem utilizada seja de entendimento comum, baseada em um vocabulário padrão, que seja compartilhado por todos dentro deste domínio. Assim, às etapas do processo de anotação citadas anteriormente, soma-se também a anotação baseada em ontologia. Ontologias fornecem um vocabulário uniforme, permitindo que os genes sejam sempre identificados através dos mesmos termos pertencentes ao domínio da biologia molecular, independente do grupo que realizou a anotação. As ontologias utilizadas no processo de anotação genômica são descritas a seguir.

2.1 Ontologias biomédicas utilizadas na anotação genômica

Por serem desenvolvidas a princípio de forma isolada, restrita a um grupo de estudos, as pesquisas genômicas podem levar à adoção de um vocabulário próprio, fazendo com que cada centro de pesquisa empregue termos diferentes em suas anotações. Isso se torna um problema quando estes bancos particulares se tornam públicos, e informações de diversas fontes precisam ser integradas e utilizadas em conjunto. Para evitar esse tipo de problema, os pesquisadores adotam a anotação baseada em ontologia, garantindo um padrão terminológico tanto dentro do mesmo banco como entre bases de dados distintas. Muitas ferramentas que auxiliam na criação de anotações dão suporte à anotação baseada em ontologia, como o *Artemis* (RUTHERFORD *et al.*, 2000), o DAS (DOWELL *et al.*, 2001) e o GARSA (DÁVILA *et al.*, 2005).

O consórcio *Open Biomedical Ontologies* (OBO) vem trabalhando no desenvolvimento de ontologias biológicas que permitem a padronização das anotações resultantes de pesquisas de sequenciamento genômico. Esta iniciativa tem como objetivo estabelecer princípios bem definidos para o desenvolvimento de novas ontologias, para evitar que a multiplicação desordenada destes vocabulários controlados comprometa a integração de informações. Os

resultados obtidos até agora incluem uma família extensível de ontologias projetadas para serem interoperáveis e bem formadas logicamente, incorporando representações precisas da realidade biológica (SMITH *et al.*, 2007).

A OBO é um esforço colaborativo, no qual estão envolvidos diversos grupos, como *National Center form Biomedical Ontologies* (NCBO, 2007), *Berkeley Bioinformatics and Open-source Projects* (BBOP, 2007), *Ontology Research Group* (ORG, 2007), entre outros. Existem atualmente mais de 60 ontologias desenvolvidas pela OBO. A principal e também mais utilizada é a *Gene Ontology*, outras ontologias de destaque incluem a *Foundational Model of Anatomy* (FMA), *Common Anatomy Reference Ontology* (CARO) e a *Ontology for Biomedical Investigations* (OBI). Pesquisas nas áreas de neurofisiologia e neuroanatomia são exemplos concretos de aplicações da metodologia da OBO na representação formal de grandes domínios (SMITH *et al.*, 2007).

A GO é com certeza o esforço de padronização mais notável para a anotação de genes. Seu objetivo é fornecer um vocabulário controlado dinâmico para descrever o papel de genes e produto de genes. A GO é composta por três ontologias que descrevem produtos de genes de acordo com suas associações: **componente celular** (componente de uma célula, pode ser também parte de algum objeto maior, como uma estrutura anatômica), **processo biológico** (série reconhecida de eventos ou funções moleculares) e **função molecular** (atividades que ocorrem no nível molecular). A divisão em três estruturas facilita não só o desenvolvimento e a manutenção das próprias ontologias, como também da anotação de produtos de genes, que fazem associações entre as ontologias e os genes e produtos de genes na colaboração entre bancos de dados (BELLOZE, 2007). A GO está em constante crescimento, assim como a quantidade de produtos de genes anotados a partir dela, que atualmente já passa de um milhão.

A preferência pela GO na anotação genômica se deve também ao desenvolvimento da base de dados genômica Gene Ontology Annotation (GOA). O projeto GOA oferece anotações de alta qualidade baseadas nos termos da GO, feitas de forma automática e manual (BARREL *et al.*, 2009). Com mais de 32 milhões de anotações, essa base de dados é um dos recursos mais utilizados na anotação automática ou semi-automática de novos organismos, onde sequências similares àquelas descobertas são pesquisadas para que sua anotação e, consequentemente, o termo da GO utilizado, sejam copiados para a descrição de gene recém-descoberto. A ampla

utilização desse recurso como ponto de partida para novas anotações consolida a GO como principal (e frequentemente única) ontologia utilizada no processo de anotação genômica.

O consórcio OBO é uma iniciativa importante em direção à padronização, pois um vocabulário consistente é fundamental para que seja possível realizar consulta em diversas bases de dados. Entretanto, segundo Chung e Wooley (2003), dada a diversidade de domínios de conhecimento e especialização de disciplinas científicas, é pouco provável que no futuro uma ontologia global comum, cobrindo disciplinas biológicas amplas, seja desenvolvida. Ao contrário, na pesquisa biomédica, haverá múltiplas ontologias para genomas, expressão de genes, proteomas, etc. Interoperabilidade semântica é uma área de pesquisa ativa dentro da ciência da computação. A integração de informações de múltiplas disciplinas e sub-disciplinas biológicas dependerá da colaboração de especialistas do domínio e profissionais de TI para desenvolver algoritmos e abordagens flexíveis para preencher a lacuna entre múltiplas ontologias biológicas (CHUNG, WOOLEY, 2003).

2.2 Características das ontologias biomédicas

Além da anotação genômica, existe uma grande diversidade de aplicações que fazem uso de ontologias biomédicas, como busca e consulta em bases de dados biológicos heterogêneas, processamento de linguagem natural e inferência de dados, entre outras. Estas aplicações refletem um pouco das diferenças existentes entre as ontologias desta área, em termos de conteúdo e estrutura, dividindo-as em grupos que visam a atender a necessidades biomédicas específicas (RUBIN, SHAH, NOY, 2007).

O tipo mais comum de artefato utilizado são t*erminologias*, ou *vocabulários controlados* (VCs). Rubin, Shah e Noy (2007) introduzem o termo “artefato ontológico” referindo-se à grande variedade de recursos existentes no domínio biológico comumente chamados de ontologias, mas que não são totalmente aderentes à sua definição formal. Um VC seria um destes artefatos, pois correspondem simplesmente a uma lista de conceitos e descrições textuais de seus significados, além de uma lista de termos léxicos correspondentes a cada conceito, os quais geralmente são organizados em uma hierarquia. A GO é um exemplo de VC, sendo atualmente o recurso mais utilizado em pesquisas biomédicas.

Modelos de informação (ou modelos de dados) são outro tipo de artefato ontológico, que consistem de uma estrutura que organiza informação de um determinado domínio de interesse e descreve como diferentes partes desta informação se relacionam umas com as outras. O *Microarray Gene Expression Object Model* (MAGE-OM[1]) é um exemplo de modelo de informação, largamente utilizado na troca de dados relativos a experimentos envolvendo expressão de genes por *microarray* entre sistemas distintos.

Por fim, *ontologias* em seu sentido formal (um conjunto de definições de conceitos, seus atributos e relações entre eles expressos em forma de axiomas baseados em uma lógica bem definida) são também utilizadas, a exemplo da *Foundational Model of Anatomy* (FMA[2]), que descreve as classes e relacionamentos necessários para modelar a estrutura do corpo humano de uma forma compreensível por humanos e interpretável por sistemas computacionais.

A maioria das ontologias da OBO, especialmente aquelas que interessam à anotação genômica, se encaixam na categoria de vocabulários controlados, priorizando a organização de hierarquias e formalização de terminologias. Algumas das suas características principais são: desenvolvimento baseado em objetivos claros (a GO, por exemplo, tem o objetivo específico de promover anotação consistente para produtos de genes), escopo limitado, estrutura simples (embora a linguagem introduzida pela OBO tenha evoluído em termos de expressividade, a estrutura baseada em um grafo direcionado acíclico, adotada pela GO, ainda é a mais utilizada), representação comum (em formato OBO ou OWL), identificadores padronizados e definições em linguagem natural (BODENREIDER, STEVENS, 2006).

Além da padronização sintática, outros fatores como total disponibilidade das ontologias, tanto para uso como para colaboração no desenvolvimento, evolução contínua e curagem constante são características dos vocabulários desenvolvidos no âmbito do consórcio OBO. Desta forma, são estabelecidos princípios que têm como intuito principal evitar a heterogeneidade semântica e sintática ainda comum aos recursos de bioinformática, possibilitando a multiplicidade de ontologias sem que isso gere redundância de esforços e multiplicação desordenada de vocabulários inconsistentes entre si.

[1] http://www.mged.org/Workgroups/MAGE/mage-om.html
[2] http://sig.biostr.washington.edu/projects/fm/

3 ALINHAMENTO DE ONTOLOGIAS

A integração de diferentes fontes de informação é um problema antigo, que tem desafiado a Ciência da Computação há muitos anos. Segundo van Harmelen (2005), a partir do momento que temos dois computadores, queremos trocar informações entre eles, e, da mesma forma, assim que temos duas bases de dados, queremos integrá-las, fazendo com que elas se comuniquem.

A interoperabilidade necessária para essa comunicação se apresenta em vários níveis. No nível *físico*, soluções como protocolos de rede (TCP/IP, HTTP, etc.) praticamente resolveram o problema da conectividade física. Porém, a conectividade apenas no nível físico não é suficiente. Também são necessários padrões que definam a forma como as informações serão trocadas, ou seja, que forneçam interoperabilidade *sintática*, e iniciativas como HTML e XML representam importantes progressos nesta área. Além de concordar com o formato da mensagem, é importante também estabelecer um acordo sobre o seu significado. Esta interoperabilidade *semântica* é um problema ainda em aberto, apesar de muito trabalho já ter sido desenvolvido nesta área (VAN HARMELEN, 2005).

O crescente número de ontologias disponibilizadas publicamente, bem como a quantidade de aplicações que as utilizam, torna mais evidente o problema da interoperabilidade semântica. Apesar de serem idealizadas para fornecer um vocabulário comum dentro de um determinado domínio, o que existe na realidade são diversas ontologias desenvolvidas por diferentes grupos, e diversas aplicações usando estas diferentes ontologias dentro do mesmo domínio. Estes grupos podem utilizar nomes distintos para o mesmo conceito, e o mesmo nome para representar conceitos diferentes. O modo como os conceitos se relacionam e são classificados varia de acordo com a visão do grupo sobre o domínio, e também com a finalidade da construção da ontologia.

Apesar das diferenças estruturais e de nomenclatura entre as ontologias, é necessário que os sistemas que as utilizam se comuniquem. Essas aplicações precisam determinar se conceitos compartilhados são semanticamente equivalentes, se conceitos diferentes têm o mesmo significado e se suas ontologias pertencem ao mesmo domínio (BREITMAN, 2005).

O alinhamento de ontologias é uma condição necessária para tornar possível a interoperabilidade entre sistemas semânticos, identificando relações entre elementos individuais de múltiplas ontologias (EHRIG, 2007). Além de permitir comunicação e reúso, o processo de

alinhamento permite a combinação de informação e conhecimento contido em diversas ontologias, enriquecendo as descrições de dados anotados através destes recursos.

3.1 Definição formal

O termo *alinhamento* possui diversas interpretações e ainda não existe um consenso sobre seu significado. Ehrig (2007) define o alinhamento de duas ontologias como "[...] para cada entidade (conceito, relação ou instância) na primeira ontologia, tentar encontrar uma entidade correspondente, com o mesmo significado pretendido, na segunda ontologia." Já Euzenat e Shvaiko (2007) consideram alinhamentos os resultados do *casamento (matching) de ontologias*, definido como a tarefa de encontrar correspondências entre entidades semanticamente relacionadas de diferentes ontologias, sendo que estas relações podem ser, além de equivalência, de consequência, de disjunção, etc. Segundo esta última definição, o alinhamento é visto como um produto, e não como um processo, como enfatiza a primeira. De Buijn *et al.* (2006) também reforçam o conceito de processo, definindo alinhamento como a descoberta de correspondências entre ontologias, as quais são representadas pelo *mapeamento de ontologias*. Os conceitos relacionados, como casamento, mapeamento, integração e outros serão melhor definidos na próxima seção.

Para o propósito deste trabalho, será adotada definição de Ehrig (2007), e o alinhamento será encarado como o processo de estabelecer relações de igualdade um-para-um entre os termos de duas ontologias de um mesmo domínio, que possuem algum tipo de interseção. Formalmente, temos:

Definição 1 (Alinhamento de Ontologias): Uma função de alinhamento de ontologias, aqui chamada de *align*, baseada no conjunto E de todas as entidades $e \in E$ e baseada no conjunto de possíveis ontologias O, é uma função parcial

$$align: E \times O \times E \rightarrow O$$

Para um dado elemento e temos $\text{align}_{O1,O2}(e)$, ou simplesmente $align(e)$. Assim que um alinhamento (parcial) entre duas ontologias $O1$ e $O2$ é estabelecido, pode-se dizer que a entidade

e foi alinhada com a entidade *f* quando *align*(*e*) = *f*. Um par de entidades (*e*, *f*) para o qual os critérios apropriados para o alinhamento ainda precisam ser testados é chamado de alinhamento candidato. Embora outras relações também possam ser consideradas para o alinhamento, apenas a igualdade está incluída nesta definição.

3.2 Conceitos relacionados

Além de alinhamento, existem vários outros conceitos que denotam operações relacionadas à interoperabilidade de ontologias. Todas estas operações têm como essência o estabelecimento de relações entre ontologias, mas podem diferir no tipo de entrada que recebem, no resultado que produzem ou no tipo de relação considerada. Também com relação a estes termos não existe consenso, sendo que diversos autores adotam definições consideravelmente distintas. A seguir são apresentadas algumas destas definições (KLEIN, 2001; DING *et al.*, 2002; DE BRUIJN *et al.*, 2005):

- **Combinação**: duas ou mais ontologias diferentes são utilizadas para uma determinada tarefa, sendo que as relações de combinação podem ser de qualquer tipo, e não apenas de identidade. A combinação está ilustrada na Figura 1.

Figura 1. Combinação de ontologias

- **Integração**: duas ou mais ontologias geram uma única ontologia final, mas é possível identificar a origem de cada conceito, através do *namespace*. Este processo pode ser interessante quando as ontologias são de domínios diferentes possivelmente complementares. A ontologia resultante pode então cobrir um domínio maior. A Figura 2 representa a integração.

Figura 2. Integração de ontologias

- **Casamento (*Matching*)**: a partir de duas ontologias, tentamos encontrar duas entidades correspondentes. A relação não precisa ser necessariamente de igualdade, mas apenas uma relação específica é utilizada. A Figura 3 mostra o casamento de ontologias.

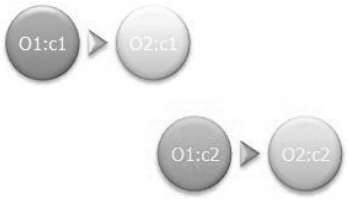

Figura 3. Casamento (*matching*) de ontologias

- **Mapeamento**: são criados axiomas de mapeamento que expressam as correspondências entre classes, relações e instâncias das duas ontologias. Estes axiomas são armazenados separadamente e descrevem como expressa os conceitos, relacionamentos e instâncias de uma ontologia em termos da outra. As ontologias de origem não são alteradas. O mapeamento está ilustrado na Figura 4.

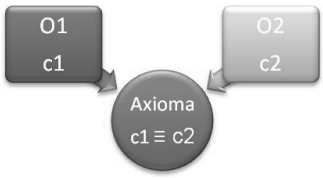

Figura 4. Mapeamento entre ontologias

- **Mediação**: processo de reconciliação de alto nível das diferenças entre ontologias para permitir interoperabilidade entre fontes de dados e sistemas que as utilizam, como ilustrado na Figura 5. Pode incluir alinhamento, mapeamento ou outras operações.

Figura 5. Mediação entre ontologias

- **Fusão (*Merge*)**: duas ontologias são unidas para criar uma única ontologia final, não sendo possível identificar a origem de cada elemento copiado para a nova ontologia. A fusão de ontologias está representada na Figura 6.

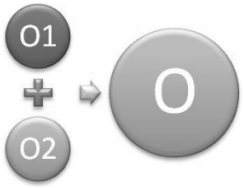

Figura 6. Fusão (*merge*) de ontologias

- **Transformação**: a semântica da ontologia é alterada para que ela possa ser utilizada com outra finalidade, diferente daquela para a qual foi desenvolvida, como mostrado na Figura 7.

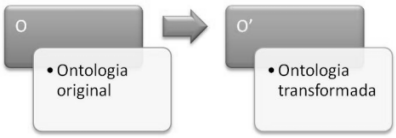

Figura 7. Transformação de ontologias

- **Tradução**: apenas a sintaxe da ontologia é alterada, transformando-se sua representação enquanto a semântica é preservada. A Figura 8 representa uma tradução onde a forma de representação (a linguagem em que a ontologia é expressa) é alterada de RDF(S) para OWL.

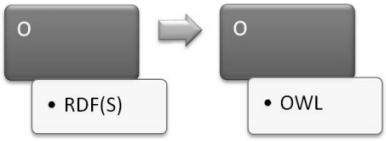

Figura 8. Tradução de ontologias

Como tem por finalidade encontrar entidades sobrepostas entre ontologias, o alinhamento pode ser parte integrante de muitas destas tarefas, geralmente consistindo em um passo anterior, cuja saída pode ser utilizada para, posteriormente, gerar diferentes resultados.

3.3 Aplicações

O alinhamento de ontologias tem se mostrado uma solução razoável para problemas presentes tanto em aplicações tradicionais, como integração de esquemas e bases de dados, como nos emergentes sistemas de compartilhamento de informação *peer-to-peer* e composição de serviços Web. Áreas como engenharia de ontologias, integração de informações, navegação e consulta na Web Semântica, entre outras, cada vez mais tiram proveito de operações de alinhamento para conciliar modelos estruturalmente heterogêneos.

Ehrig (2007) e Euzenat e Shvaiko (2007) listam diversas aplicações que refletem casos de uso concretos, onde o alinhamento de ontologias tem um papel importante na comunicação de sistemas e na manutenção da integridade e consistência de modelos e bases de dados. Alguns destes cenários são descritos a seguir.

3.3.1 Edição e importação de ontologias

O reuso de ontologias é uma prática importante não só para evitar a multiplicação de ontologias sobre o mesmo tópico, como também para manter um vocabulário uniforme entre diferentes grupos. Assim, ainda na fase de projeto da ontologia, quando pode ser necessário que várias ontologias de fontes externas relativas ao mesmo domínio sejam colocadas juntas, é necessário lidar com a heterogeneidade de recursos.

Identificar ontologias relevantes e relações entre as entidades destas ontologias são tarefas importantes para a engenharia de ontologias. Estas ontologias podem ser importadas e unidas em uma única ontologia final, onde o alinhamento auxiliará na descoberta de áreas de sobreposição entre elas. Por ser estático, este é um cenário simples, pois as ontologias são previamente selecionadas (em tempo de projeto) e a conciliação pode contar com entradas manuais, além do auxílio de ferramentas, para a execução das operações necessárias.

3.3.2 Evolução e versionamento de ontologias

Ontologias, principalmente aquelas de grande porte, podem evoluir ao longo do tempo. Além disso, podem também ser desenvolvidas de forma distribuída e colaborativa, gerando várias versões da mesma ontologia. A Gene Ontology é um bom exemplo de ontologia que, por estar em constante evolução, possui múltiplas versões, sendo que algumas aplicações que a utilizam podem manter uma versão atualizadas enquanto outras continua a usar uma antiga, possivelmente fazendo suas próprias atualizações localmente. O versionamento permite acompanhar todas estas mudanças e manter as relações entre novas ontologias que são criadas e as já existentes.

O alinhamento pode fornecer as informações necessárias sobre ligações entre diferentes versões de ontologias. Neste caso, o foco é na descoberta de diferenças, como entidades adicionadas, removidas ou renomeadas, entre duas versões. A qualidade do alinhamento para evolução e versionamento deve ser alta, requisito facilitado pelo fato de que diferentes versões sequenciais de uma ontologia provavelmente terão grandes áreas de sobreposição. Entradas manuais também são possíveis neste caso, o que ajuda a garantir a qualidade.

3.3.3 Integração de esquemas e dados

A integração de esquemas é um dos cenários mais antigos, comum quando duas bases de dados necessitam ser integradas em uma única (por exemplo, quando há uma aquisição ou fusão entre duas empresas, cada uma com a sua própria base de dados). Mesmo que as bases pertençam ao mesmo domínio, por terem sido desenvolvidas de forma independente, o alinhamento é necessário para identificar correspondências entre entidades semanticamente relacionadas, apontando sobreposições entre os esquemas antes da efetiva integração das bases de dados.

A integração de dados é uma tarefa similar, porém, neste caso, as informações provenientes de diversas fontes devem ser integradas sem que elas sejam unidas em uma única base. Para as aplicações que usam estes dados, instâncias iguais não devem ser distinguíveis, mesmo que originárias de diferentes repositórios, ou seja, para o usuário as fontes dos dados devem ser transparentes. O alinhamento permite a descoberta de correspondências entre os dados para que sejam apresentados ao usuário de maneira uniforme.

3.3.4 Composição de serviços web

Serviços web são processos que expõem sua interface na Web para que possam ser invocados por usuários. Serviços web podem ser incorporados em *workflows*, onde a saída de um é direcionada para a entrada de outro. Além disso, são projetados para serem substituíveis, de modo que o usuário possa sempre escolher novos e melhores serviços dinamicamente. Linguagens de representação de conhecimento e ontologias podem ser usadas para enriquecer a descrição destes serviços, gerando os chamados serviços web semânticos. Porém, mesmo que sejam modeladas através de ontologias, as descrições dos serviços diferem muito, bem como o formato de suas entradas e saídas.

Os diversos serviços disponíveis na web devem colaborar entre si, independente de suas representações heterogêneas, e cumprir as tarefas para as quais foram projetados. O alinhamento é essencial neste cenário tanto para relacionar descrições de serviços, para determinar se são equivalentes, podendo um substituir o outro, quanto para traduzir a saída de um serviço para a entrada de outro em um *workflow*. Identificar equivalências entre o que é produzido por um serviço e o que é esperado por outro possibilita a composição de inúmeros serviços para atingir um objetivo em particular, como o planejamento de uma viagem, onde o serviço de reservas de uma companhia aérea pode ser integrado ao serviço de reservas de uma rede de hotéis.

3.3.5 Comunicação e negociação entre agentes

Agentes são entidades de software caracterizadas por sua autonomia e capacidade de interação, que podem planejar suas ações e negociar com outros agentes para atingir seus objetivos. Agentes se comunicam trocando mensagens expressas em linguagens específicas para comunicação de agentes, que determinam o "envelope" da mensagem. O conteúdo, entretanto, é expresso em termos de uma ontologia. Quando dois agentes se encontram, eles têm a oportunidade de trocar mensagens, mas, como podem usar ontologias diferentes, podem não compreender um ao outro.

O alinhamento permite que agentes que precisam lidar com ontologias heterogêneas estabeleçam correspondências entre elas, identificando entidades equivalentes, e compreendam as

mensagens uns dos outros. O alinhamento pode ser realizado pelos próprios agentes, ou podem ser utilizadas bibliotecas ou serviços externos de alinhamento.

3.3.6 Consultas na Web

Cada vez mais páginas na Web são enriquecidas com anotações semânticas baseadas em ontologias. Porém, um grande número de ontologias diferentes são utilizadas nestas anotações, e mesmo páginas relativas ao mesmo domínio podem utilizar ontologias distintas. Além disso, usuários que fazem consultas na Web também utilizam sua própria terminologia.

Quando uma consulta de usuário é enviada para um mecanismo de busca, é necessário que ela seja reescrita em termos das ontologias que descrevem os recursos. O alinhamento é utilizado aqui para relacionar os termos da consulta às entidades das ontologias de interesse, identificando equivalências. Os resultados obtidos na consulta devem novamente ser submetidos ao alinhamento, para que a resposta seja traduzida de volta para a forma de representação original, a qual será apresentada ao usuário.

Estes são apenas alguns exemplos de casos de uso, existem muitos outros cenários possíveis, pois a heterogeneidade é um problema presente em qualquer processo que envolva comunicação entre aplicações e fontes de dados. Com um número cada vez maior de aplicações que adotam o uso de ontologias, no futuro a necessidade de alinhamento e outros processos que proporcionem interoperabilidade tende a crescer ainda mais.

3.4 Técnicas de alinhamento de ontologias

Para concretizar a tarefa de identificar entidades equivalentes entre ontologias, diversas ferramentas de alinhamento já foram desenvolvidas. Por trás destas ferramentas, existe também uma grande variedade de técnicas, utilizadas isoladamente ou combinadas para calcular o grau de similaridade entre conceitos, relacionamentos e instâncias. Euzenat e Shvaiko (2007) apresentam uma classificação detalhada destas abordagens. Embora este levantamento tenha como foco o que os autores chamam de mapeamento, que consiste em um processo mais genérico de identificação

de correspondências entre entidades, no qual as relações consideradas podem ser de qualquer natureza, as classificações resultantes também se aplicam à definição de alinhamento aqui adotada.

As técnicas básicas são dividas em duas categorias: *nível de elemento* e *nível de estrutura*. As técnicas do nível de elemento computam correspondências analisando entidades isoladamente, ignorando suas relações com outras entidades. Já as técnicas do nível de estrutura calculam correspondências analisando o modo como as entidades se apresentam juntas dentro da estrutura da ontologia.

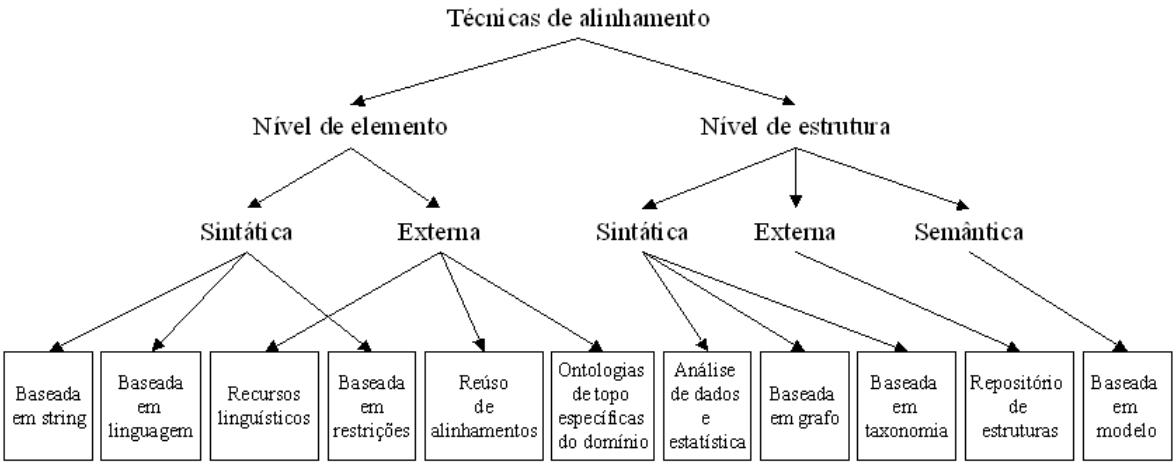

Figura 9. Classificação das abordagens elementares de alinhamento. Adaptado de Euzenat e Shvaiko (2007)

Na Figura 9, notamos um segundo nível na árvore de classificação das técnicas de alinhamento. A característica principal das técnicas sintáticas é a interpretação da entrada de acordo com sua estrutura, seguindo algum algoritmo claramente definido. As técnicas externas são aquelas que exploram recursos auxiliares externos de um domínio para interpretar a entrada, como tesauros ou mesmo entradas adicionais de usuários. Técnicas semânticas usam algum tipo de semântica formal, como, por exemplo, teoria de modelos, para interpretar a entrada e justificar seus resultados. O último nível da árvore apresenta as técnicas básicas de alinhamento, detalhadas a seguir.

3.4.1 Técnicas do nível de elemento

Técnicas do nível de elemento consideram as entidades da ontologia de forma isolada, sem analisar seus relacionamentos com as demais entidades. As principais técnicas deste nível são:

- **Técnicas baseadas em string:** geralmente são utilizadas para comparar e associar nomes e *labels* das entidades das ontologias. Strings são vistas como uma sequência de letras em um alfabeto, e, quanto mais similares as strings, maior a probabilidade de que os conceitos que elas representam sejam similares.
- **Técnicas baseadas em linguagem:** consideram nomes como palavras em uma linguagem natural, como português ou inglês. São baseadas em técnicas de processamento de linguagem natural, explorando propriedades morfológicas das palavras fornecidas como entrada.
- **Técnicas baseadas em restrições:** são algoritmos que lidam com as restrições internas aplicadas às definições das entidades, como tipo de dado e cardinalidade de atributos, entre outras.
- **Recursos linguísticos:** recursos como tesauros específicos do domínio são utilizados para auxiliar na tarefa de associação de palavras (neste caso, nomes de entidades são considerados palavras de uma linguagem natural), baseando-se nas relações linguísticas entre elas, como sinônimos, generalizações e especializações.
- **Reuso de alinhamentos:** forma alternativa de explorar recursos externos, que armazenam resultados de alinhamento anteriores entre ontologias. Por exemplo, se temos armazenados os resultados do alinhamento entre o e o', e entre o e o'', podemos reutilizar estes resultados no alinhamento de o' e o''. Este tipo de técnica é particularmente interessante quando todas as ontologias envolvidas são do mesmo domínio e possuem um número muito grande de entidades, pois o reuso de resultados de alinhamento anteriores pode ser aplicado a fragmentos menores da ontologia, dividindo grandes tarefas de alinhamento em pequenas sub-tarefas.
- **Ontologias de topo e ontologias formais específicas do domínio:** ontologias de topo, como a *Suggested Upper Merged Ontology* (SUMO) (PEASE, NILES, LI, 2002) e a *Descriptive Ontology for Linguistic and Cognitive Engineering* (DOLCE) (GANGEMI *et al.*, 2002), podem ser utilizadas como fontes externas de conhecimento comum. A

principal característica destas ontologias é que elas são baseadas em lógica, e deste modo, técnicas que exploram estes recursos são baseados em semântica. Ontologias específicas do domínio também podem ser usadas como recursos externos, pois usam termos com sentidos relevantes apenas para este domínio, e que não estão relacionados a conceitos similares em outros domínios. Como exemplo, no alinhamento de ontologias da área médica, uma ontologia como a *Foundational Model of Anatomy* (FMA) (ROSSE, MEJINO JR., 2003) pode ser usada como contexto, considerando-se o domínio da anatomia. Como a abordagem proposta neste trabalho apoia-se no uso de ontologias de topo, a técnica de alinhamento baseada em ontologia de fundamentação a ser utilizada será mais bem detalhada na Seção 3.5.

3.4.2 Técnicas do nível de estrutura

Técnicas do nível de estrutura, ao contrário das técnicas do nível de elemento, comparam as entidades das ontologias a serem alinhadas considerando seus relacionamentos com as demais entidades. Algumas destas técnicas são apresentadas a seguir:

- **Técnicas baseadas em grafo:** algoritmos que consideram as ontologias como grafos rotulados. A análise de similaridade entre um par de nós representando conceitos nas duas ontologias baseia-se em suas posições dentro do grafo, considerando-se que, se dois nós de duas ontologias são similares, suas vizinhanças, ou seja, seus nós adjacentes, também devem apresentar um certo grau de similaridade.
- **Técnicas baseadas em taxonomia:** também consistem em algoritmos baseados em grafo, que levam em consideração apenas a relação de especialização. Da mesma forma que as técnicas baseadas em grafo, se uma ligação do tipo é-um (*is-a*) conecta termos já considerados similares, é possível que seus nós adjacentes também sejam similares.
- **Repositórios de estruturas:** repositórios de estruturas armazenam ontologia e seus fragmentos, juntamente com medidas de similaridade relativas a pares de entidades, por exemplo, coeficientes no intervalo 0..1. Ao contrário do reuso de ontologias, repositórios de estruturas armazenam apenas similaridades entre ontologias, e não alinhamentos. Quando novas estruturas (ontologias ou fragmentos) vão ser alinhadas, primeiro busca-se por similaridades checando-se as estruturas disponíveis no repositório. O objetivo é

identificar estruturas suficientemente similares que indiquem que vale a pena prosseguir com o alinhamento, evitando operações sobre estruturas dissimilares.

- **Técnicas baseadas em modelo:** algoritmos baseados em modelo (ou semanticamente fundamentados) manipulam entradas baseando-se na sua interpretação semântica. A ideia por trás deste tipo de técnica é que, se duas entidades são similares, elas compartilham a mesma interpretação lógica. Técnicas de inferência de lógica descritiva são exemplos de técnicas baseadas em modelo.
- **Técnicas estatísticas e de análise de dados:** técnicas que aproveitam amostras (preferencialmente grandes) de uma população com o objetivo de encontrar regularidades e discrepâncias. Isto ajuda a agrupar itens ou computar distâncias entre eles. Análise de correspondência e distribuições de frequência são exemplos destes tipos de técnicas.

3.5 Técnica baseada em ontologias de topo

Ontologias de topo, ou de fundamentação, correspondem a um conjunto de categorias de alto nível, independentes de domínio, que englobam noções como objetos, eventos, atributos, conexões espaço-temporais, dependências e outros conceitos comuns a qualquer área do conhecimento. Ontologias de fundamentação fornecem uma semântica formal rigorosa para estas noções de alto nível, e servem como uma base conceitual para ontologias de domínio, que vão, por sua vez, prover uma visão especializada sobre uma determinada área (PROBST, 2006).

Guizzardi (2009) reforça essa definição, estabelecendo que uma ontologia de fundamentação "é um arcabouço formal de conceitos genéricos (isto é, independentes de domínio) do mundo real que podem ser usados para falar sobre domínio materiais". Em geral, esse tipo de ontologia é utilizado como modelo de referência que define os conceitos permitidos em uma linguagem de modelagem conceitual bem fundamentada, possibilitando que ela capture a semântica do mundo real.

Além dos benefícios na construção de modelos conceituais de um domínio, a utilização de ontologias de fundamentação pode ser útil também no alinhamento de ontologias. Identificando as meta-categorias das quais os conceitos são derivados é possível estabelecer sua natureza, diferenciando, por exemplo, *objetos* de *processos*, ou *tipos* de coisas de *papéis* desempenhados por estas. Essa distinção pode ajudar a evitar associações incorretas no processo de alinhamento,

restringindo a indicação de termos equivalentes àqueles derivados da mesma meta-categoria, ou seja, que possuem a mesma natureza conceitual. A Figura 10 exemplifica o uso de ontologias de fundamentação no alinhamento de ontologias.

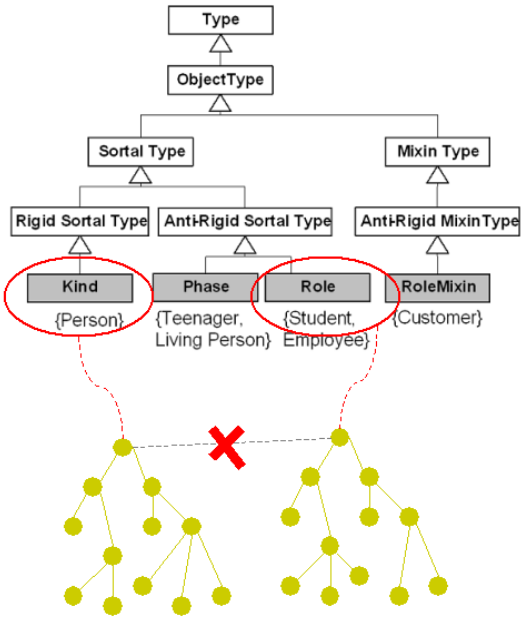

Figura 10. Uso de ontologia de fundamentação no alinhamento de ontologias

Na figura acima, temos um fragmento de ontologia de fundamentação, introduzido por Guizzardi (2009), onde são modelados apenas objetos (*endurants*). Nessa ontologia, a classe "Tipo" (*"Kind"*), que se caracteriza por ser um sortal (que fornece um princípio de identidade) rígido, é disjunta da classe "Papel" (*"Role"*), que é um sortal não rígido. De forma simplificada, um sortal rígido é aquele que não varia ao longo do tempo (por exemplo, uma "Pessoa" é sempre uma "Pessoa"), ao contrário do sortal não rígido, que estabelece uma condição que pode ser válida apenas em um determinado momento no tempo (por exemplo, um "Estudante" pode deixar de ser um "Estudante"). No exemplo da Figura 10, temos duas ontologias sendo alinhadas e um par de classes onde uma delas descreve um "Tipo" e a outra um "Papel". Identificando a origem dessas classes com a intermediação da ontologia de fundamentação, podemos evitar este tipo de associação, ainda que alguma medida de similaridade retorne um valor alto, pois conceitos de naturezas distintas jamais podem ser considerados equivalentes.

3.6 Similaridade entre ontologias

A identificação de similaridades entre duas ontologias é uma etapa fundamental no processo de alinhamento. A similaridade, que corresponde a um valor numérico que indica o quanto dois elementos são similares ou diferentes, pode ser obtida a partir da comparação de ontologias inteiras ou apenas de subelementos destas. As medidas de similaridade aqui apresentadas baseiam-se no trabalho de Ehrig (2007), que adota uma classificação dividida em camadas (Seção 3.6.1), porém a maioria dessas medidas apresenta relação direta com as técnicas básicas apresentadas na seção anterior.

Ehrig (2007) define formalmente similaridade como "[...] uma função que mapeia um par de conjuntos de entidades e suas ontologias correspondentes para um número real expressando a similaridade entre os dois conjuntos." Esta função é expressa como:

$$sim: \mathcal{B}(E) \times \mathcal{B}(E) \times O \times O \rightarrow [0, 1]$$

onde $\mathcal{B}(E)$ representa o conjunto de entidades e O, as ontologias correspondentes, sendo que:

- $\forall e, f \in \mathcal{B}(E), O_1, O_2 \in O, sim(e, f, O_1, O_2) \geq 0$ *(positividade)*
- $\forall e, f, g \in \mathcal{B}(E), O_1, O_2 \in O, sim(e, e, O_1, O_2) \geq sim(f, g)$ *(maximalidade)*
- $\forall e, f \in \mathcal{B}(E), O_1, O_2 \in O, sim(e, f, O_1, O_2) = sim(f, e, O_2, O_1)$ *(simetria)*
- $\forall e, f \in \mathcal{B}(E), O_1, O_2 \in O, sim(e, f, O_1, O_2) = 1 \Leftrightarrow e = f$: *Dois conjuntos de entidades são idênticos.*
- $\forall e, f \in \mathcal{B}(E), O_1, O_2 \in O, \in 0\ sim(e, f, O_1, O_2) < 1$: *Dois conjuntos de entidades são similares/diferentes até certo ponto.*
- $\forall e, f \in \mathcal{B}(E), O_1, O_2 \in O, sim(e, f, O_1, O_2) = 0 \Leftrightarrow e \neq f$: *Dois conjuntos de entidades são diferentes e não têm nenhuma característica comum.*

A relação de similaridade pode ser estabelecida entre conceitos, relacionamentos ou instâncias, sendo que e e f podem representar apenas subárvores ou ontologias inteiras. Um conjunto também pode ser constituído de apenas uma entidade, o que reduz a relação à similaridade entre dois elementos individuais.

3.6.1 Camadas de similaridade

Várias abordagens para alinhamento de ontologias usam comparações léxicas para determinar a similaridade entre entidades (por exemplo, MCGUINNESS *et al.*(2000)), ou seja, consideram apenas a representação dos elementos e calculam a similaridade com base no nome de classes, relacionamentos e instâncias. Como estes elementos representam conceitos de um domínio real, sua comparação deve ir além da representação, levando também em consideração seus significados e o uso dentro do contexto do qual fazem parte (EHRIG, 2007). Ehrig e Sure (2005a) definem três camadas que permitem o cálculo da similaridade entre entidades a partir de diferentes características: camada de dados, camada de ontologia e camada de contexto.

A Figura 11 mostra a organização dessas três camadas, além de uma quarta camada ortogonal a elas que representa o conhecimento específico do domínio ao qual pertencem as ontologias.

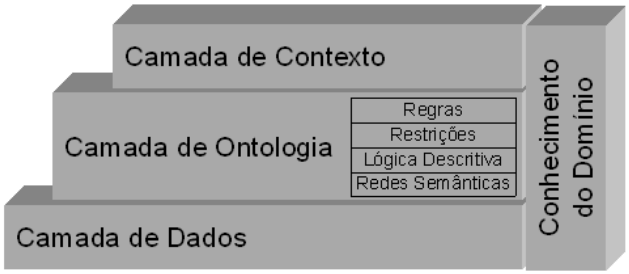

Figura 11. Camadas de similaridade. Adaptado de Ehrig (2007)

Na camada de dados, as entidades são comparadas considerando-se apenas valores, como strings ou inteiros, usando medidas de similaridade genéricas para esses tipos de dados. Na camada de ontologia, são considerados também os relacionamentos entre as entidades. Essa camada pode ser dividida em vários níveis: no nível mais baixo, ontologias podem ser tratadas simplesmente como grafos compostos de conceitos e relacionamentos (redes semânticas). Essas redes podem ser enriquecidas através de lógica descritiva, restrições e regras, sobre as quais também podem ser aplicadas medidas de similaridade. Já na camada de contexto, é considerado o modo como as entidades da ontologia são usadas em um contexto externo, ou seja, são utilizadas informações externas à ontologia, principalmente a respeito da aplicação onde ela é utilizada. Por último, alguns domínios podem ter seu próprio vocabulário adicional (por exemplo, o domínio bibliográfico geralmente usa elementos do Dublin Core (CAPLAN, 1995) como `dc:author` ou

`dc:title`) o qual é tratado de maneira especial, retornando suas próprias similaridades. O conhecimento específico do domínio pode estar situado em qualquer nível de complexidade da ontologia, como pode ser notado na Figura 11. As medidas de similaridade específicas de cada camada serão apresentadas a seguir.

3.6.2 Medidas de similaridade específicas

Para que tanto a representação como o significado e o uso de uma entidade possam ser levados em consideração no cálculo da similaridade, existem várias medidas específicas, cada uma contemplando uma característica distinta dos conceitos comparados. A seguir são listadas, para cada uma das camadas citadas anteriormente, algumas destas medidas.

Camada de dados

Considerando-se os conceitos da ontologia como tipos de dados, temos que as instâncias são seus valores concretos. Com base no valor dessas instâncias podemos utilizar operadores que permitem comparações baseadas no conteúdo de strings ou em valores numéricos. Exemplos de medidas dentro da camada de dados são:

- **Igualdade:** Em alguns casos pode ser necessária a igualdade entre os valores de dados para que duas entidades sejam consideradas similares (por exemplo, valores usados como identificadores). Considerando v_1 e v_2 valores de dados, a igualdade é definida como:

$$sim_{igualdade}(v_1, v_2) \coloneqq \begin{cases} 1, \text{se } v1 = v2, \\ 0, \text{caso contrário} \end{cases}$$

- **Similaridade sintática:** Para o cálculo da similaridade sintática entre strings é utilizada a distância de edição (LEVENSHTEIN, 1965). O objetivo desta medida é determinar quantas operações, que podem ser adição, remoção ou substituição de caracteres, são necessárias para transformar uma string em outra. Maedche e Staab (2002) definem a similaridade sintática como o inverso da distância de edição, expressa como *ed*. Esta

medida é muito utilizada na comparação de nomes de conceitos, onde diferenças no uso de minúsculas e maiúsculas, abreviações e erros de digitação geram valores baixos de distância de edição e, consequentemente, alta similaridade sintática.

$$sim_{sintática}(v_1, v_2) \coloneqq \max\ (0, \frac{\min(|v_1|, |v_2|) - ed(v_1, v_2)}{\min(|v_1|, |v_2|)})$$

- **Similaridade Baseada em Distância (para Valores Numéricos):** quando o tipo de dado assume um valor numérico, é utilizada uma medida de similaridade baseada na diferença aritmética entre os dois valores comparados (BERGMANN, 2002). Esta medida é aplicável quando os valores estão dentro de um limite definido, como, por exemplo, um subconjunto de números inteiros, sendo que *maxdif* é a diferença entre os valores máximo e mínimo desse intervalo. γ é um número real, usado como parâmetro para ajustar a influência de valores altos de diferença no cálculo da similaridade.

$$sim_{dif}(v_1, v_2) \coloneqq 1 - \left(\frac{|v_1 - v_2|}{maxdif}\right)^{\gamma}$$

Objetos

Enquanto na camada de dados são comparados simplesmente valores de dados, aqui serão definidas as medidas de similaridade para comparação de objetos, ainda sem considerar seu tipo semântico. Algumas medidas possíveis são:

- **Igualdade de objetos:** a igualdade entre objetos de ontologia é baseada na existência de declarações lógicas, explicitamente incluídas na ontologia (por exemplo, em OWL através da propriedade `owl:sameAs`), ou identificada anteriormente, de forma manual ou automática. Considerando *e* e *f* duas entidades de ontologias e *align* uma função de alinhamento como definido na Seção 3.1, temos:

$$sim_{objeto}(e,f) \coloneqq \begin{cases} 1, \text{se } align(e) = f, \\ 0, \text{caso contrário} \end{cases}$$

- **Coeficiente de Dice:** medida utilizada para comparar dois conjuntos de entidades *E* e *F*. Dois conjuntos de entidades podem ser comparados com base na sobreposição dos indivíduos dos conjuntos ($e \in E, f \in F$). A desvantagem desse coeficiente é que só são retornados resultados se cada indivíduo possui um identificador único, onde indivíduos iguais têm o mesmo identificador, mesmo considerando-se várias ontologias.

$$sim_{dice}(E,F) \coloneqq \frac{2 \cdot |E \cap F|}{|E| + |F|}$$

- **Coeficiente de Jacquard:** coeficiente relacionado à similaridade de conjuntos anterior. Ele calcula a fração de elementos sobrepostos comparados ao número total de elementos existentes nos dois conjuntos.

$$sim_{jacquard}(E,F) \coloneqq \frac{|E \cap F|}{|E \cup F|}$$

- **Ligação única:** a similaridade máxima é um operador básico para comparações. A ligação única lida com similaridades de entidades individuais, considerando o valor máximo encontrado como a similaridade entre os conjuntos. Para o cálculo dos valores individuais são usadas previamente outras medidas, que retornam um valor no intervalo 0..1 expressando a similaridade entre as entidades.

$$sim_{\text{única}}(E,F) = \max_{(e,f)|e \in E, f \in F}(sim(e,f))$$

- **Ligação média:** o valor médio também é uma medida possível para determinar a similaridade entre conjuntos. A ligação média calcula a média entre as similaridades dos indivíduos para determinar a similaridade total entre os dois conjuntos comparados.

$$sim_{completa}(E,F) = \frac{\sum_{\forall(e,f)|e \in E, f \in F} sim\,(e,f)}{|E| \cdot |F|}$$

- **Multi similaridade:** compara conjuntos representando cada um deles através de um elemento médio. Cada entidade é descrita por um vetor que representa sua similaridade em relação a todos os outros elementos contidos nos dois conjuntos. Um vetor representativo, calculado através da média de todos os vetores individuais, é criado para cada conjunto. A similaridade é então determinada pelo cosseno entre os vetores dos dois conjuntos. **e** = (sim(e, e_1), sim(e, e_2), ..., sim(e, f_1), sim(e, f_2), ...), **f** idem.

$$sim_{multi}(E,F) = \frac{\sum_{e \in E} \mathbf{e}}{|\sum_{e \in E} \mathbf{e}|} \cdot \frac{\sum_{f \in F} \mathbf{f}}{|\sum_{f \in F} \mathbf{f}|}$$

Camada de Ontologia

Nesta camada, são apresentadas medidas de similaridade específicas derivadas da estrutura característica de uma ontologia. Algumas dessas medidas são:

- **Similaridade de *label*:** o *label* é um atributo básico de qualquer entidade em uma ontologia, que consiste de um identificador compreensível por humanos, expresso em uma linguagem natural. *Labels* são comparados utilizando-se a medida de similaridade sintática. Dicionários, como o WordNet, podem ser utilizados na comparação. Uma desvantagem desse tipo de medida é o fato de homônimos poderem retornar valores altos de similaridade erroneamente.

$$sim_{label}(e,f) \coloneqq sim_{sintática}(label(e), label(f))$$

- **Similaridade taxonômica para conceitos:** determina a similaridade de conceitos em uma hierarquia. $\alpha \geq 0$, $\beta \geq 0$ são parâmetros que dimensionam respectivamente a contribuição do tamanho l e da profundidade h do menor caminho na hierarquia de conceitos. O tamanho do menor caminho é uma métrica para medir a distância conceitual entre c_1 e c_2. Essa medida pode também ser utilizada para comparação de relações,

também organizadas em uma hierarquia. Aqui, *e* é a constante matemática (base do sistema de logaritmos neperianos) igual a aproximadamente 2,7182818.

$$sim_{taxonômica}(c_1, c_2) \coloneqq \begin{cases} e^{-\alpha l \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}}, \text{se } c_1 \neq c_2, \\ 1, \text{caso contrário} \end{cases}$$

- **Similaridade extensional de conceitos:** dois conceitos são considerados similares se possuem instâncias (extensões) similares. A similaridade entre os conjuntos de instâncias ι_C de cada conceito pode ser calculada usando-se a multi similaridade.

$$sim_{extensão}(c_1, c_2) \coloneqq sim_{multi}(\iota_C(c_1), \iota_C(c_2))$$

- **Similaridade de domínio e escopo:** medida de similaridade para relações, baseando-se em suas definições de domínio e escopo (*domain* e *range*). *esc* e *dom* referem-se ao domínio e escopo, respectivamente, das relações r_1 e r_2.

$$sim_{domEsc}(r_1, r_2) \coloneqq 0{,}5 \cdot (sim_{objeto}(esc(r_1), esc(r_2)) + sim_{objeto}(dom(r_1), dom(r_2)))$$

- **Similaridade de conceito para instâncias:** medida de similaridade para instâncias. Duas instâncias i_1 e i_2 têm um certo grau de similaridade se pertencem à mesma classe.

$$sim_{conceito}(i_1, i_2) = sim_{objeto}(c_1, c_2)\ sendo\ \ i_1 \in \iota_C(c_1), i_2 \in \iota_C(c_2)$$

Camada de contexto

Nesta camada, o contexto, que pode ser, por exemplo, a aplicação onde as entidades são de fato utilizadas, é analisado para calcular a similaridade entre elas. Duas entidades são similares se são utilizadas no mesmo contexto. No modelo de ontologia considerado, *e* e *f* podem ser conceitos ou instâncias, e Uso (*e*, *con*) corresponde à frequência do uso de *e* no contexto *con*.

$$sim_{uso}(e, f) \coloneqq sim_{dif}(\text{Uso}(e, con), \text{Uso}(f, con))$$

As medidas listadas são apenas alguns exemplos, muitos outros são possíveis em cada uma das camadas de similaridade mencionadas. Bernstein *et al.* (2005) apresentam um trabalho mais completo sobre medidas de similaridade, que inclui resultados de experimentos realizados com o auxílio de um *framework* que as implementa. O subconjunto de medidas aqui apresentadas inclui aquelas que serão utilizadas neste trabalho, dentre elas serão escolhidas aquelas que mais se adaptam às características das ontologias biomédicas a serem alinhadas.

4 FERRAMENTAS DE ALINHAMENTO

Motivadas pela expansão do uso de ontologias em diversas aplicações, as pesquisas na área de alinhamento de ontologias têm se multiplicado nos últimos anos. O desenvolvimento de diversas ontologias sobre o mesmo domínio e a necessidade de comunicação e interoperabilidade entre os sistemas que as utilizam têm levado à criação de dezenas de ferramentas que realizam tarefas de alinhamento, *merge* e integração, entre outras. O *Ontology Alignment Evaluation Initiative* (OAEI)[3] avalia anualmente um número expressivo de novas ferramentas, que apresentam melhorias crescentes nos resultados, através da utilização de técnicas cada vez mais elaboradas.

As primeiras ferramentas desenvolvidas utilizavam apenas técnicas simples, como comparação sintática de nomes de conceitos. O *Chimaera* (MCGUINNES *et al*., 2000), por exemplo, indica que dois termos são equivalentes quando eles têm o mesmo nome, quando os nomes são muito parecidos (apresentam algum trecho comum) ou quando um nome é prefixo do outro. Já o PROMPT (NOY, MUSEN, 2000) vai um pouco além da comparação de nomes. Utilizando um algoritmo denominado *AnchorPROMPT* (NOY, MUSEN, 2003), embora de forma ainda pouco flexível, a ferramenta desenvolvida como um *plug-in* para o editor de ontologias *Protégé*[4] usa os relacionamentos (não hierárquicos) para procurar por possíveis similaridades entre as classes. Nas ferramentas desenvolvidas mais recentemente, podemos notar, além da comparação sintática de nomes e estruturas, a agregação de muitas das técnicas apresentadas na Seção 3.4, como a utilização de recursos externos e métodos semânticos.

Algumas dessas ferramentas serão apresentadas e analisadas a seguir. O objetivo deste levantamento é analisar as abordagens adotadas por cada uma delas, o que servirá de base para a elaboração da abordagem proposta neste trabalho. Essas ferramentas foram selecionadas entre as participantes das edições de 2007 e 2008 do OAEI, escolhendo-se aquelas que obtiveram os melhores resultados, além de APIs amplamente utilizadas no desenvolvimento destas aplicações. A comparação entre elas levará em consideração as técnicas que empregam para realizar a tarefa de alinhamento. Além disso, será apresentado também um levantamento da disponibilidade, facilidade de uso e viabilidade de extensão das ferramentas.

[3] http://oaei.ontologymatching.org/
[4] http://protege.stanford.edu/

4.1 ASMOV

ASMOV (*Automated Semantic Matching of Ontologies with Verification*) é a implementação de um algoritmo que usa características léxicas e estruturais das ontologias para calcular a similaridade entre elas iterativamente, gerando alinhamentos que serão posteriormente verificados com a finalidade de assegurar que não possuem inconsistências semânticas (JEAN-MARY, SHIRONOSHITA, KABUKA, 2009).

O processo seguido pela ferramenta é dividido em dois componentes principais: cálculo de similaridade e verificação semântica. Além das duas ontologias a serem alinhadas, o algoritmo aceita também como entrada opcional um conjunto de alinhamentos pré-determinados. Para o cálculo de similaridade, o algoritmo iterativamente obtém uma média ponderada entre quatro medidas de similaridade: *similaridade léxica* (ou terminológica), usando comparação de strings ou recursos externos, como tesauros, duas *similaridades estruturais*, sendo a primeira baseada nas relações de especialização (*similaridade relacional* ou hierárquica), e a segunda nas restrições entre classes e propriedades (*similaridade interna* ou de restrição), e *similaridade extensional*, que utiliza as instâncias da ontologia na comparação.

O pré-alinhamento resultante desta primeira etapa é então submetido ao processo de verificação semântica, que confere se determinados axiomas inferidos a partir de um alinhamento estão realmente declarados na ontologia, removendo correspondências que levam a inferências que não podem ser verificadas. Para esta tarefa, são utilizados axiomas de definição de domínio e escopo, e relações de especialização, equivalência e disjunção. A realização da verificação semântica, aliada ao uso de tesauros específicos do domínio, melhora significativamente o desempenho do sistema, porém o fato de ser uma ferramenta totalmente automatizada ainda leva à apresentação de alinhamentos incorretos, resultado que se acentua ainda mais quando as ontologias são pouco similares.

4.2 DSSim

DSSim (NAGY, VARGAS-VERA, MOTTA, 2006) é um sistema que incorpora a teoria da evidência de Dempster-Shafer (SHAFER, 1976) no processo de mapeamento de ontologias com o objetivo de aumentar o número de mapeamentos corretos, avaliando o modo como a

aplicação da função de crença (SMETS, 1988) pode melhorar os resultados do mapeamento através da combinação de similaridades, cujos valores são calculados por algoritmos de similaridade sintática e semântica.

A ferramenta atua no contexto de resposta a consultas, onde um usuário formula questões contendo conceitos e propriedades de um determinado domínio. Para responder a estas questões, são utilizadas diferentes ontologias, sendo necessários mapeamentos entre seus conceitos e propriedades, que serão então utilizados para responder à consulta original. O primeiro passo do algoritmo é o cálculo da similaridade entre nomes, feita através de comparação de strings e recuperação de sinônimos usando o WordNet. O próximo passo, chamado de similaridade sintática, na verdade é uma comparação estrutural, onde um par de nós de duas ontologias representadas como grafos é analisado de acordo com sua posição dentro da estrutura, sendo considerados similares se suas vizinhanças também forem similares.

O diferencial da ferramenta é a manipulação de incerteza, através da aplicação da teoria da evidência de Dempster-Shafer, que provê um mecanismo para modelagem e inferência de informações incertas, além de permitir que evidências de diversas fontes (no caso, as diversas medidas de similaridade utilizadas) sejam combinadas usando a regra de combinação de Dempster. A vantagem desta técnica é que ela permite uma distinção clara entre o que é incerto e o que é ignorado, porém, como cada crença (similaridade) é calculada individualmente por um agente diferente, no momento da combinação pode haver conflitos, resultantes de hipóteses incompatíveis assumidas por estes agentes, que podem levar a mapeamentos incorretos.

4.3 OntoDNA

OntoDNA é uma ferramenta que utiliza a técnica de Análise Formal de Conceitos (*Formal Concept Analysis* - FCA) como parte do processo de mapeamento e *merge* de ontologias, capturando os atributos e a estrutura inerente às ontologias, formada pelos relacionamentos entre conceitos, para resolver problemas semânticos como sinônimos e polissemia (KIU, LEE, 2006). Nesse sistema, a taxonomia da ontologia é utilizada pela FCA como conhecimento adicional para resolver interpretações semânticas em diferentes contextos.

A utilização de técnicas de clusterização não supervisionada faz do OntoDNA uma ferramenta totalmente automatizada, gerando uma ontologia final, resultado da fusão das duas

ontologias originais, sem a intervenção do usuário. As duas etapas mais importantes neste processo são a aplicação da técnica de clusterização conhecida como *Self-Organizing Map* (SOM), que organiza os dados agrupando objetos similares dentro de um contexto formal pré-definido, aproveitando-se das características naturais do conjunto de dados, seguida da comparação de strings através da distância de edição de Levenshtein (LEVENSHTEIN, 1965).

O principal problema que o OntoDNA tenta contornar é a falta de conhecimento prévio da estrutura das ontologias. Esse conhecimento é necessário em técnicas de mineração de dados supervisionadas, mas nem sempre está disponível. A utilização da FCA como técnica não supervisionada tem como objetivo categorizar os conceitos de forma a facilitar a identificação de entidades equivalentes na ausência de conhecimento prévio, além de lidar com possíveis complexidades estruturais e diferenças semânticas entre as ontologias. Porém, a forte dependência em relação aos nomes de conceitos para resolver a heterogeneidade entre as ontologias, conferindo um peso elevado à similaridade léxica em detrimento de outras abordagens, representa uma limitação do sistema.

4.4 Falcon

Falcon também é uma ferramenta de alinhamento automático, que conta com dois *matchers* integrados: um baseado em comparação linguística para ontologias, chamado de LMO, e outro baseado em comparação de estruturas de grafos, chamado de GMO. Nesse sistema, o GMO toma como entrada os alinhamentos obtidos pelo LMO e gera como saída alinhamentos adicionais (JIAN *et al.*, 2005).

Além da comparação léxica entre nomes de conceitos, o LMO também utiliza uma abordagem baseada em análise estatística para estabelecer similaridades linguísticas entre as entidades. Após o cálculo da similaridade léxica, que usa a distância de edição como base, as ontologias são submetidas à análise estatística através do algoritmo VSM (*Vector Space Model*), em uma abordagem semelhante à adotada pelos mecanismos de recuperação de informação, representando cada entidade da ontologia (classes, propriedades e instâncias) como um documento virtual populado com termos retirados de seus nomes, *labels* e comentários. As similaridades entre esses documentos virtuais calculadas pelo VSM fornecem os alinhamentos que servirão de entrada para o GMO.

Na comparação estrutural, as ontologias são representadas como modelos de grafo. Os pares de entidades previamente relacionados pelo LMO são analisados comparando-se os papéis (sujeito, predicado ou objeto) que assumem nas triplas das quais fazem parte. O cálculo da similaridade estrutural pode gerar novos alinhamentos e é de grande importância quando a comparação léxica gera resultados pouco expressivos. Contudo, o GMO não consegue lidar com ontologias muito grandes, o que faz com que o Falcon ignore a informação estrutural durante o alinhamento de ontologias de grande porte.

4.5 FOAM

FOAM (*Framework for Ontology Aligning and Mapping*) é uma ferramenta de alinhamento e mapeamento implementada com base em um processo que faz uso da semântica codificada nas ontologias, onde diferentes instanciações desse processo focam em vários aspectos da tarefa de alinhamento (EHRIG, SURE, 2005b). Desenvolvida a partir da experiência obtida com a ferramenta QOM – *Quick Ontology Mapping* (EHRIG, STAAB, 2004), pode ser utilizada tanto como um sistema independente quanto como uma API Java integrada a outras aplicações.

A abordagem adotada, chamada de NOM (*Naïve Ontology Mapping*), tem seis passos principais: (1) seleção de recursos, onde pequenos fragmentos da ontologia, como identificadores, *labels*, etc., são escolhidos para descrever uma entidade específica; (2) seleção de pares, onde duas entidades das duas ontologias são eleitas para serem comparadas; (3) cálculo da similaridade, que indica a similaridade entre as duas entidades de acordo com um conjunto de medidas, cada uma relativa a um dado recurso utilizado na descrição das entidades (similaridade de identificadores, de *labels*, etc.); (4) agregação de similaridades, onde os valores obtidos a partir das vária medidas de similaridade são agregados em um valor único, que representará a similaridade entre as duas entidades; (5) interpretação, que, usando os valores agregados, um limite pré-definido e estratégias de interpretação, decide se as entidades devem ou não ser consideradas equivalentes; e (6) iteração, onde as equivalências já obtidas são propagadas pelas ontologias, para que a similaridade de um par possa influenciar no cálculo da similaridade de seus vizinhos. Em cada iteração, todos os valores de similaridades são recalculados e as iterações se repetem até que os valores agregados se estabilizem.

Os alinhamentos resultantes são devolvidos em um arquivo texto, que registra os pares de termos considerados equivalentes e o valor da similaridade entre eles. A interação com o usuário é opcional e pode ser configurada juntamente com outros parâmetros. Estes parâmetros também podem ser ajustados automaticamente pelo sistema, de acordo com a tarefa a ser executada (alinhamento, *merge*, versionamento, etc.), e com o tamanho das ontologias envolvidas. Apesar de utilizar vários recursos das ontologias, a similaridade entre os *labels* das entidades ainda é o fator de maior importância na abordagem, sendo que a ferramenta não lida bem com alterações nestes elementos, como traduções, sinônimos, etc.

4.6 SAMBO

SAMBO (*System for Aligning and Merging Biomedical Ontologies*) é uma ferramenta de alinhamento e *merge* de ontologias, desenvolvida de acordo com um *framework* que prevê interação com o usuário: vários *matchers*, que calculam similaridades entre os termos de duas ontologias usando diferentes estratégias, são combinados para filtrar os resultados e oferecer sugestões de alinhamento ao usuário, que pode então aceitá-las ou rejeitá-las, influenciando futuras sugestões (LAMBRIX, TAN, 2006).

A ferramenta é voltada para o alinhamento de ontologias biomédicas, e tem como foco a adoção de estratégias que se apliquem às ontologias atualmente disponíveis neste domínio. Combinando as estratégias implementadas por um *matcher* linguístico, um *matcher* estrutural, conhecimento específico do domínio (através do uso do tesauro *Unified Medical Language System* - UMLS[5]), e um *matcher* de aprendizagem (que faz uso da literatura da área, como o MEDLINE[6] e outros periódicos biomédicos), o sistema devolve ao usuário sugestões de alinhamento. Essas sugestões consistem de pares de termos com um valor de similaridade acima de um limite pré-estabelecido. Tanto este limite como os pesos atribuídos a cada um dos *matchers* podem ser definidos pelo usuário, bem como quais *matchers* serão utilizados no processo de alinhamento.

Por fim, o usuário pode decidir se os termos são equivalentes, se possuem uma relação do tipo "é-um", ou se a sugestão deve ser rejeitada. A ferramenta também oferece um modo manual,

[5] http://www.nlm.nih.gov/research/umls/
[6] http://medline.cos.com/

onde o usuário pode visualizar as duas ontologias e escolher os termos a serem alinhados manualmente. Apesar do fato de a combinação de estratégias melhorar significativamente os resultados, nem sempre fica claro como obter os melhores alinhamentos, pois, como a escolha dos parâmetros é deixada a cargo do usuário, muitas vezes é difícil para este definir quais são os melhores algoritmos a serem utilizados em cada caso, e que pesos devem ser atribuídos a cada um.

4.7 RiMOM

RiMOM (*Risk Minimization based Ontology Mapping*) é uma ferramenta de alinhamento que também faz uso da combinação de múltiplas estratégias, porém, para uma determinada tarefa de alinhamento, ela tenta encontrar uma combinação "ótima" de estratégias a serem aplicadas a esta situação específica. Além de ontologias, o RiMOM também aceita como entrada esquemas de banco de dados (LI, LI, TANG, 2007).

A ferramenta possui três partes: um repositório de ontologias, para onde são importados os esquemas e ontologias; um gerenciador de projetos, que permite ao usuário criar diversas tarefas de alinhamento e atribuir diferentes parâmetros a cada uma delas; e o processo de alinhamento propriamente dito, que controla a execução de uma tarefa de alinhamento baseado nas características das ontologias. A primeira etapa é a estimativa dos fatores de similaridade de *label* e estrutural entre as duas ontologias, valores que servirão de base para a seleção de estratégias, onde são decididos quais algoritmos linguísticos e características estruturais das ontologias serão utilizados no alinhamento. Após a execução de cada uma das estratégias selecionadas e da combinação dos alinhamentos obtidos, caso o fator de similaridade estrutural calculado anteriormente seja alto, é executada a propagação de similaridades, que refina os resultados e pode também encontrar novos alinhamentos ainda não descobertos.

As estratégias adotadas pela ferramenta incluem a distância de edição, similaridade de vetores, similaridade de caminhos, entre outras. A vantagem do RiMOM é sua capacidade de selecionar dinamicamente as estratégias mais adequadas às ontologias a serem alinhadas, porém os parâmetros (pesos) a serem atribuídos a cada estratégia ainda precisam ser inseridos manualmente, o que pode influenciar negativamente nos resultados, já que combinações

diferentes de parâmetros para o mesmo conjunto de estratégias podem gerar alinhamentos distintos, podendo ser difícil para o usuário em alguns casos definir os valores mais adequados.

4.8 Lily

Lily é um sistema de mapeamento de ontologias que possui quatro funcionalidades principais: alinhamento genérico de ontologias, alinhamento de ontologias de grande porte, alinhamento semântico de ontologias e depuração de alinhamento, usando estratégias híbridas na execução destas tarefas. A ferramenta tem como princípio fundamental utilizar as informações úteis para o processo de alinhamento presentes nas ontologias de forma eficiente e correta (WANG, XU, 2008).

O processo de alinhamento consiste de três passos: pré-processamento, onde os dados necessários para os próximos passos são preparados, cálculo de alinhamentos, onde os métodos mais adequados são utilizados para computar a similaridade entre elementos de diferentes ontologias, e pós-processamento, responsável por extrair, depurar e avaliar os alinhamentos encontrados. Entre as técnicas utilizadas pela ferramenta estão: extração de subgrafos semânticos, que tentam descrever o significado de cada entidade na ontologia, *matcher* baseado em Documentos de Descrição Semântica (*Semantic Description Document – SDD*), que mede a similaridade literal entre as ontologias, e *matcher* baseado em propagação de similaridade (*Strong Similarity Propagation – SSP*), para os casos onde a informação literal é escassa. Também é utilizado conhecimento extraído da Web, que, através de mecanismos de busca, procura por padrões léxico-sintáticos para encontrar relações semânticas entre as ontologias, além de uma técnica própria para lidar com ontologias de grande porte, baseada em âncoras positivas e negativas que ajudam a predizer quais são os possíveis pares de entidades similares, evitando a segmentação ou modularização das ontologias.

A ferramenta é capaz de selecionar automaticamente qual técnica deve ser utilizada de acordo com o tamanho das ontologias a serem alinhadas. Uma desvantagem é o fato de ser necessário extrair um subgrafo semântico para cada conceito e propriedade na ontologia, o que consome muito tempo, reduzindo o desempenho do sistema.

4.9 CIDER

CIDER (*Context and Inference baseD alignER*) é definida como uma ferramenta de alinhamento baseada em esquema, conferindo um peso maior às informações do nível do esquema, em detrimento dos dados do nível de instância. Como a maioria das ferramentas, utiliza uma combinação de técnicas básicas no processo de alinhamento, como comparação léxica, taxonômica e de relações, entre outras. (GRACIA, MENA, 2008)

O algoritmo utilizado é na verdade parte de um sistema maior, e foi elaborado para descobrir similaridades entre possíveis sentidos de uma mesma palavra-chave utilizada por um usuário em uma consulta. Estes sentidos poderiam então ser posteriormente integrados se fossem suficientemente similares. Como já fazia consultas a ontologias para coletar os possíveis significados da palavra-chave, o algoritmo pôde ser generalizado para aceitar como entrada duas ontologias quaisquer e comparar todos os pares de termos possíveis entre elas. Inicialmente, a ferramenta extrai o contexto ontológico (sinônimos, descrições textuais, hiperônimos, hipônimos, propriedades, etc.) de cada termo envolvido, até uma certa profundidade. Um mecanismo de inferência transitiva complementa esta etapa. O passo seguinte é o cálculo da similaridade entre todos os pares de termos, que considera similaridade linguística (*labels* e descrições) e similaridade estrutural (usando o contexto ontológico previamente extraído para cada termo, comparando taxonomias e relacionamentos não hierárquicos), atribuindo um peso a cada uma destas contribuições.

A saída é um documento RDF contendo os alinhamentos obtidos, cujos valores de similaridade se encontram acima de um limite pré-estabelecido. Segundo avaliação dos autores, a ferramenta ainda apresenta um tempo de resposta longo e prioriza a precisão dos resultados, embora afirmem que a revocação manteve um valor “aceitável” durante as avaliações realizadas.

4.10 Aroma

Aroma (*Association Rule Ontology Matching Approach*) é uma abordagem híbrida, extensional e assimétrica desenvolvida para encontrar relações semânticas entre taxonomias textuais. Essas relações podem ser de equivalência ou de generalização-especialização

(*subsumption*), e são identificadas com base no paradigma de regras de associação entre entidades (DAVID, GUILLET, BRIAND, 2007).

Diferentemente de outras ferramentas de alinhamento, o Aroma não se baseia somente em medidas de similaridade. Seu diferencial é a mineração de regras de associação, que têm como objetivo encontrar correlações entre elementos de um conjunto, e podem ser definidas como proposições da forma "*Se antecedente então consequente*" (*antecedente* → *consequente*, na notação formal). Como os algoritmos de mineração normalmente produzem uma grande quantidade de regras, a ferramenta utiliza medidas de interesse (*Interestingness Measures – IMs*) para avaliar a qualidade e auxiliar na escolha das melhores regras, reduzindo o conjunto a ser utilizado no processo de alinhamento. Esse processo é dividido em duas partes: a representação de cada entidade da ontologia por um conjunto de termos e dados relevantes, e a descoberta de regras de associação binárias entre estas entidades. Para ontologias expressas em OWL, a primeira etapa consiste na extração do vocabulário de entidades a partir de suas anotações (*label*, descrição, etc.) e instâncias. A segunda parte usa o modelo de regras de associação e uma medida de interesse chamada *intensidade de implicação* para inferir implicações entre os vocabulários previamente extraídos. O uso de uma medida de similaridade entre strings complementa a abordagem, buscando possíveis alinhamentos ainda não descobertos nas etapas anteriores.

O uso de técnicas mais sofisticadas e o fato do considerar relações de generalização-especialização além das de equivalência constituem vantagens importantes do Aroma, que pode obter resultados melhores quando as ontologias a serem alinhadas apresentam grande diferença estrutural, pois pode ser comum que se escolha incluir termos mais específicos em uma ontologia, enquanto outra pode ser desenvolvida sobre o mesmo domínio, mas com termos mais gerais. Como é fortemente dependente de informação textual, a ferramenta não apresenta bons resultados quando *labels* e comentários estão ausentes.

4.11 Análise e avaliação das ferramentas

O OAEI vem se mostrando uma iniciativa importante, estimulando desenvolvimentos e melhorias na área de alinhamento de ontologias. Os resultados das últimas edições mostram que as ferramentas desenvolvidas recentemente têm evoluído significativamente, aumentando a

qualidade dos alinhamentos que retornam e empregando um número cada vez maior de recursos para identificar equivalências entre as ontologias.

Em relação às técnicas descritas na Seção 3.4, podemos identificar muitas delas nas abordagens adotadas pelas ferramentas apresentadas nas seções anteriores. Dentre elas se destacam: (i) ASMOV e FOAM, que usam um conjunto considerável de características das ontologias na comparação, aproveitando grande parte das medidas de similaridade descritas na Seção 3.6.2; e (ii) Aroma, que também considera vários recursos das ontologias, mas emprega técnicas diferenciadas - utiliza regras de associação e usa medidas de similaridade apenas para complementar o processo de alinhamento, que também se diferencia pelo fato de buscar não só por equivalências, mas também por relações de generalização-especialização.

Todas as ferramentas analisadas empregam pelo menos uma técnica do nível de elemento, sendo a mais comum a comparação de strings, e uma do nível de estrutura, sendo a comparação de grafos a mais utilizada. Falcon, SAMBO, RiMOM e Aroma vão um pouco além da comparação de strings, utilizando técnicas baseadas em linguagem, que consideram os termos como uma palavra de um vocabulário ou de um idioma e não simplesmente como uma sequência de caracteres. Esses termos podem ser identificadores, *labels*, ou qualquer outro recurso que represente as entidades sendo comparadas.

Esta comparação pode ou não contar com o auxílio de recursos linguísticos externos. SAMBO e ASMOV, por exemplo, usam tesauros específicos do domínio. Algumas ferramentas utilizam este tipo de recurso com outras finalidades, como DSSim e FOAM, que utilizam o WordNet para recuperar sinônimos dos termos envolvidos na comparação. O WordNet também é utilizado pelo CIDER, mas com o intuito de identificar os diferentes sentidos que uma palavra pode assumir e eleger os mais adequados ao contexto da consulta do usuário. Ainda no nível de elemento, técnicas adicionais também são possíveis, como é o caso do Falcon, que estabelece similaridade linguística entre entidades com o auxílio de análise estatística. O uso de instâncias para comparar os conceitos das quais são derivadas é outra possibilidade; ASMOV e FOAM são algumas das ferramentas que utilizam esse recurso.

Em relação à comparação estrutural, embora as técnicas baseadas em grafo, utilizadas por todas as ferramentas para esta tarefa, sejam as formas mais simples e de mais fácil implementação, elas utilizam muito pouco da informação disponível nas ontologias, pois as consideram simplesmente como conjuntos de nós ligados por arestas, sem se preocupar com o

significado dessas ligações. DSSim e Falcon são exemplos de ferramentas que se apoiam exclusivamente nesse tipo de técnica para o cálculo da similaridade estrutural. Já sistemas como ASMOV, FOAM e CIDER vão um pouco além, analisando os conceitos dentro de uma hierarquia e usando as relações do tipo "é-um" para auxiliar na identificação de equivalências. Além disso, ASMOV e CIDER usam também os relacionamentos não hierárquicos, aproveitando a semântica das restrições entre conceitos, que são recursos importantes em uma ontologia. Ainda na comparação estrutural, o CIDER utiliza outras informações, como sinônimos, propriedades, etc. Também usando a semântica codificada nas ontologias, mas com uma técnica mais sofisticada, o OntoDNA utiliza a Análise Formal de Conceitos como forma de comparação estrutural, e o Lily usa mecanismos de busca para extrair conhecimento da Web e reconhecer relações semânticas entre os conceitos. Aroma e ASMOV completam suas abordagens de comparação estrutural semântica com técnicas baseadas em modelo, sendo que esta última usa axiomas de definição de domínio e escopo para realizar a verificação semântica dos alinhamentos obtidos.

Uma outra forma de aproveitar as informações estruturais é realizar a propagação de similaridades pelas ontologias. FOAM e ASMOV adotam uma abordagem iterativa, onde os valores de similaridade obtidos para cada par de entidades influenciam no cálculo destes valores para entidades vizinhas nas iterações seguintes. Assim é possível refinar os resultados, aumentando a similaridade de conceitos com nomes diferentes, mas em contextos muito semelhantes, e diminuindo este valor para conceitos homônimos que, apesar de possuírem o mesmo nome, apresentam-se em contextos muito distintos. A propagação pode ser realizada também em um único passo, como é o caso do RiMOM. Porém, a ferramenta só executa a propagação se a similaridade estrutural, calculada previamente, for considerada alta. Lily também só utiliza a propagação de similaridade se julgar que as informações literais presentes nas ontologias são insuficientes.

A Tabela 2 resume as principais características das ferramentas apresentadas, em relação às técnicas de alinhamento de ontologias descritas na Seção 3.4. Apesar de agregarem várias técnicas, nem sempre as ferramentas utilizam todas elas em conjunto. Uma análise preliminar das características das ontologias envolvidas para definir quais estratégias devem ser usadas pode poupar tempo de execução, evitando passos desnecessários que não ajudarão no cálculo da similaridade. RiMOM, por exemplo, seleciona as estratégias a serem combinadas dinamicamente,

de acordo com uma análise inicial da similaridade entre as duas ontologias. Lily também seleciona técnicas dinamicamente, porém de acordo com o tamanho das ontologias a serem alinhadas.

Tabela 2. Técnicas empregadas pelas ferramentas de alinhamento de ontologias analisadas

Ferramenta	***Técnicas***										
	Nível de elemento						*Nível de estrutura*				
	Baseada em string	Baseada em linguagem	Recursos linguísticos	Baseada em restrições	Reuso de alinhamentos	Ontologias de topo	Análise de dados	Baseada em grafo	Baseada em taxonomia	Repositório de estruturas	Baseada em modelo
ASMOV	✓		✓	✓				✓	✓		✓
DSSim	✓		✓					✓			
OntoDNA	✓						✓	✓	✓		
Falcon	✓	✓						✓			
FOAM	✓		✓	✓				✓	✓		
SAMBO	✓	✓	✓					✓	✓		
RIMOM	✓	✓						✓			
Lily	✓		✓	✓				✓			
CIDER	✓		✓					✓	✓		
Aroma	✓	✓		✓				✓	✓		✓

Independentemente de selecionarem ou não quais técnicas serão efetivamente empregadas no alinhamento, as ferramentas geralmente associam pesos a cada uma destas estratégias. Algumas delas definem esses parâmetros automaticamente, enquanto outras deixam esta tarefa a cargo do usuário. RiMOM e SAMBO são exemplos de ferramentas que pedem que o usuário defina os pesos de cada *matcher*, o que pode ser difícil e, consequentemente, interferir no resultado do alinhamento, caso não se possua o conhecimento necessário para ajustar esses parâmetros. Disponibilizar valores padrão, baseados em experimentos prévios, para definir os pesos mais adequados a cada estratégia, pode ser uma boa opção para não comprometer a eficácia do sistema. FOAM, por exemplo, disponibiliza uma série de parâmetros já com valores padrão pré-definidos, que podem opcionalmente ser configurados pelo usuário.

Por fim, uma característica importante a ser observada em uma ferramenta de alinhamento de ontologias é seu grau de automação. Muitas preferem realizar todo o processo de forma automática, enquanto outras oferecem interação com o usuário. Ambas as abordagens possuem

vantagens e desvantagens e se aplicam a situações diferentes. O alinhamento automático, embora apresente uma probabilidade maior de conter resultados incorretos, é a melhor opção quando agregado a sistema onde o tempo de resposta deve ser o menor possível, como consultas de usuários na Web. Já as ferramentas semi-automáticas podem apresentar um nível maior de correção nos resultados, já que estes passam pela avaliação do usuário antes de serem consolidados, mas podem exigir grande esforço, dependendo do volume de dados a ser avaliado e da qualidade da interface de visualização disponibilizada pela ferramenta. O ideal é o equilíbrio entre estas duas opções, onde a ferramenta consiga automatizar a maior parte do processo e exija a intervenção do usuário apenas em casos de incerteza, diminuindo assim seu esforço. Esta é a abordagem adotada pelo FOAM, que apresenta apenas os alinhamentos considerados duvidosos, para que sejam confirmados ou rejeitados pelo usuário, mas esta é apenas uma das possibilidades do sistema, já que a intervenção do usuário é opcional, sendo um parâmetro configurável. ASMOV, DSSim, OntoDNA, Falcon, RiMOM, CIDER e Aroma são totalmente automatizadas, enquanto no SAMBO o alinhamento pode ser semi-automático, sendo que as decisões do usuário influenciam as sugestões que serão apresentadas futuramente, ou totalmente manual. Lily opta pela interação com o usuário somente na fase final de depuração, apenas para eliminar erros como redundâncias e inconsistências.

A partir desta análise, podemos notar que algumas ferramentas se destacam por adotarem abordagens mais completas, utilizando um conjunto amplo de técnicas de alinhamento, sendo boas candidatas a ponto de partida para a elaboração de um mecanismo de alinhamento voltado para a área de Bioinformática. Porém, além dessas características, outros pontos importantes também devem ser avaliados para verificar a viabilidade de reutilização destas ferramentas. Essa avaliação tem como objetivo definir qual dentre os sistemas estudados será estendido, dando origem a um aplicativo destinado ao alinhamento de ontologias biomédicas para o processo de anotação genômica. Assim, será possível tirar proveito das melhores abordagens e adaptá-las aos objetivos propostos. Para definir qual ferramenta poderia ser reutilizada e adaptada aos objetivos propostos, além das qualidades técnicas, foi necessário verificar a disponibilidade e a possibilidade de extensão de cada uma delas.

Para esta avaliação, algumas questões práticas foram levantadas, dando origem a um conjunto de critérios que ajudarão a verificar a possibilidade de reuso das ferramentas. Esses critérios são descritos a seguir:

- **Disponibilidade:** é necessário que a ferramenta esteja disponível e que possa ser instalada e testada localmente. A maioria dos autores disponibiliza suas aplicações para *download*, mas alguns preferem apenas descrever o sistema e apresentar resultados de testes, sem oferecer acesso direto ao programa;
- **Código aberto:** os autores devem fornecer também o código fonte e permitir que ele seja alterado por terceiros. Muitos desenvolvedores que disponibilizam o código de suas aplicações utilizam licenças de *software* livre, como a GNU GPL[7], que garante o direito a modificações, desde que o produto resultante também seja *software* livre passível de futuras alterações por outros usuários;
- **Documentação:** o código deve vir acompanhado de documentação que permita a fácil compreensão de variáveis e rotinas, bem como da utilização e do funcionamento geral do sistema. A qualidade da documentação pode facilitar ou inviabilizar a extensão da aplicação;
- **APIs associadas:** a maioria das ferramentas utiliza APIs de terceiros no seu desenvolvimento. É importante que essas APIs também sejam de fácil acesso, bem documentadas, atualizadas e que ofereçam suporte aos desenvolvedores que as utilizam;
- **Linguagem:** ferramentas desenvolvidas em linguagens de programação mais populares, como C ou Java, que também contam com farta documentação e suporte na Web, são de mais fácil compreensão e extensão;
- **Suporte a OWL-DL:** todas as ontologias da OBO, além do formato próprio adotado pelo consórcio (*.obo*), também são disponibilizadas em OWL-DL. É necessário que a ferramenta de alinhamento a ser adaptada dê suporte a este tipo de codificação;
- **Atualização:** como as pesquisas nesta área têm se desenvolvido rapidamente, é necessário que a ferramenta seja recente, tendo sido desenvolvida nos últimos anos. Alguns sistemas continuam em desenvolvimento ativo mesmo após a apresentação de resultados, pois estão inseridos no âmbito de projetos mais longos, que garantem atualização constante.

A Tabela 3 mostra o resumo da avaliação das ferramentas analisada em relação aos critérios descritos acima. Esses dados foram obtidos a partir de informações fornecidas pelos

[7] http://www.gnu.org/licenses/licenses.pt-br.html

autores e equipes de desenvolvimento, extraídos tanto de publicações quanto de páginas Web dedicadas à divulgação dos projetos.

Tabela 3. Características das ferramentas analisadas em relação à viabilidade de extensão

Ferramenta	*Critérios*							
	Site	**Disponível para Download**	**Código aberto**	**Documentação**	**APIs associada**	**Linguagem**	**Suporte a OWL-DL**	**Última Atualização**
ASMOV	http://support.infotechsoft.com/integration/ASMOV/OAEI-2008/	sim[8]	não	indisponível	Jena, Thesaurus adapter API	Java	sim	2009
DSSim	http://kmi.open.ac.uk/people/miklos/OAEI2008/tools/DSSim.zip	sim	não	indisponível	Alignment API	Java	sim	2009
OntoDNA	http://pesona.mmu.edu.my/~cckiu/OAEI2007.htm	sim	não	indisponível	não informado	não informado	sim	2007
Falcon	http://iws.seu.edu.cn/projects/matching/	sim	sim	indisponível	Jena	Java	sim	2008
FOAM	http://www.aifb.uni-karlsruhe.de/WBS/meh/foam/download.htm	sim	sim	boa	KAON2, WEKA	Java	sim	2006
SAMBO	http://www.ida.liu.se/~iislab/projects/SAMBO/	não	não	indisponível	não informado	não informado	sim	2008
RIMOM	http://keg.cs.tsinghua.edu.cn/project/RiMOM/	sim	não	pequena	OWL-API	Java	sim	2009
Lily	http://ontomappinglab.googlepages.com/lily.htm	sim	não	pequena	Jena, Simpack	Java	sim	2009
CIDER	http://sid.cps.unizar.es/SEMANTICWEB/ALIGNMENT/OAEI08/	sim[9]	não	boa	Alignment API	Java	sim	2008
Aroma	http://www.inrialpes.fr/exmo/people/jdavid/oaei2008/AROMA_oaei2008.jar.	sim	sim	pequena	Jena	Java	sim	2009

Além do levantamento dessas características, foi realizada também uma pequena experiência em relação à utilização de cada uma das ferramentas, para verificar a facilidade de instalação, uso e o tipo de resultado retornado por cada uma. Os testes foram feitos com duas

[8] Requisição para download deve ser feita por e-mail ao autor
[9] Link para download é disponibilizado após requisição ao autor

ontologias simples e de pequeno porte (com apenas 10 classes cada uma): animalsA.owl[10] e animalsB.owl[11], e não tiveram como objetivo avaliar a qualidade dos alinhamentos retornados; a finalidade dessa investigação foi apenas observar o comportamento de cada sistema. A Tabela 4 resume os seguintes pontos: se foi possível fazer o *download* e instalar a ferramenta, se o alinhamento das duas ontologias foi concluído com sucesso, e algumas observações gerais sobre a utilização de cada aplicação.

Tabela 4. Observações sobre instalação e uso das ferramentas analisadas

Ferramenta	***Download e Instalação***	***Teste concluído***	***Observações***
ASMOV	não	não	Requisição para download deve ser feita diretamente para o autor; email enviado não obteve resposta
DSSim	sim	não	Só executa um exemplo fornecido com a ferramenta, o comando indicado dispara uma exceção, não há documentação
OntoDNA	sim	sim	Resultado no formato *.cex*, exige outra ferramenta (ConExp[12]) para ser visualizado
Falcon	sim	sim	Sem documentação
FOAM	sim	sim	Resultado em arquivo *.txt*, leitura difícil
SAMBO	não	não	Não está disponível para *download*
RIMOM	sim	sim	Requer a instalação de três módulos Perl, a compilação de um deles falhou (alguns arquivos necessários não foram encontrados). O alinhamento pode ser feito mesmo assim
Lily	sim	não	Aparentemente voltada para os testes do OAEI, pois pede um alinhamento de referência, que, apesar de poder ser deixado em branco, dispara uma exceção se nenhum arquivo for informado
CIDER	sim	sim	Utiliza o WordNet, mas retorna erro de conexão com esse aplicativo. O alinhamento prossegue mesmo assim
Aroma	sim	não	Não encontra uma classe da API Jena, mesmo com todos os arquivos *.jar* necessários adicionados ao *classpath*

[10] http://www.aifb.uni-karlsruhe.de/WBS/meh/foam/ontologies/animalsA.owl
[11] http://www.aifb.uni-karlsruhe.de/WBS/meh/foam/ontologies/animalsB.owl
[12] http://conexp.sourceforge.net/

A partir dessas análises, é possível traçar algumas conclusões. Na Tabela 2, podemos notar que ASMOV, SAMBO, FOAM e Aroma são as ferramentas que empregam o maior número de técnicas combinadas, apresentando assim as abordagens mais completas dentre os sistemas analisados, sendo fortes candidatas ao reuso. Porém, na Tabela 3, verificamos que o ASMOV não possui código aberto, o que inviabiliza sua reutilização. Além disso, o fato de não estar disponível para *download*, mesmo após requisição ao autor, como mostrado na Tabela 4, exclui essa aplicação da lista de candidatas.

O SAMBO, apesar da vantagem de já ser um sistema voltado para o alinhamento de ontologias biomédicas, não está acessível, pois não é disponibilizado pelos autores. O Aroma, apesar de estar disponível para *download* e ter código aberto, apresentou problemas na execução, como mostra a Tabela 4. Como possui uma documentação pouco expressiva, conforme registrado na Tabela 3, torna-se difícil não só sua extensão, como também a simples utilização e compreensão de seu funcionamento.

O FOAM, além das vantagens de estar disponível e possuir código aberto, por também ser uma API Java, possui farta documentação voltada para desenvolvedores, descrevendo todas as classes e métodos nela contidos. O fato de retornar os resultados em um arquivo texto de leitura um pouco difícil, como observado na Tabela 4, não chega a ser um problema, já que a saída pode ser tratada e apresentada em um formato mais amigável para o usuário final. Além disso, esse formato adotado para a saída (a URL das duas entidades consideradas equivalentes e o valor da similaridade entre elas, separados por ponto e vírgula) é também o formato aceito como alinhamento de referência, que pode ser fornecido como entrada auxiliando no cálculo da similaridade em alinhamentos futuros. Como já foi desenvolvida com o intuito de ser flexível e extensível, possibilitando a adição de novas medidas de similaridade, e se mostrou como a melhor opção nas análises realizadas, o FOAM foi a ferramenta escolhida, e servirá como base para a abordagem proposta neste trabalho, sendo adaptada para realizar o alinhamento de ontologias da OBO para auxiliar no processo de anotação genômica.

5 ABORDAGEM

No capítulo anterior, vimos como a área de alinhamento de ontologias vem avançando nos últimos anos, com o desenvolvimento de novas ferramentas que empregam um conjunto cada vez mais variado de técnicas e medidas de similaridade que aproveitam as características das ontologias para identificar equivalências entre suas entidades.

No entanto, a maioria dessas ferramentas é de propósito geral, e considera as características comuns a qualquer ontologia, independente do domínio para o qual foram desenvolvidas. Porém, dentro de alguns domínios, as ontologias podem apresentar algumas particularidades, como hierarquias mais (ou menos) profundas, ausência de instâncias ou ênfase na criação de restrições, entre outras. Este é o caso das ontologias do domínio biomédico, em especial aquelas desenvolvidas pelo consórcio OBO, onde a ênfase está na padronização do vocabulário e na definição da hierarquia, sendo que instâncias não são incluídas e há um número bem pequeno de propriedades (*Object Properties*).

Um dos objetivos deste trabalho é elaborar uma abordagem de alinhamento que leve em consideração as características específicas das ontologias da OBO, selecionando as medidas de similaridade mais adequadas, priorizando as que medem a similaridade dos recursos disponíveis e descartando aquelas que têm como foco características pouco expressivas nessas ontologias. Assim, é possível tirar proveito das informações codificadas nas ontologias biomédicas sem onerar o processo de alinhamento com cálculos desnecessários.

Entretanto, como as ferramentas analisadas apresentam abordagens elaboradas e bons resultados, ignorar esses avanços e iniciar uma nova abordagem do zero não seria a melhor opção, pois muito do que já foi implementado pode ser reaproveitado, apenas adequando alguns pontos às necessidades do domínio biomédico. Assim, a proposta deste trabalho é selecionar a melhor abordagem entre as apresentadas no capítulo anterior e complementá-la, realizando os ajustes necessários para que a ferramenta resultante seja dedicada às ontologias da OBO, usando apenas medidas de similaridade que trabalham com as informações presentes nestes vocabulários específicos. Como foi discutido na Seção 4.11, a ferramenta FOAM se mostrou a melhor opção, devido tanto à abordagem adotada quanto às facilidades de manipulação e adaptação, sendo escolhida para servir como base da abordagem proposta.

Além das medidas já apresentadas e utilizadas pelas ferramentas, a identificação de equivalência também conta com o auxílio da utilização de ontologias de fundamentação. Essas ontologias permitem identificar a natureza dos conceitos sendo comparados, e ajudam a aumentar a qualidade dos alinhamentos, provendo mais um parâmetro no cálculo de similaridades. Deste modo, desenvolve-se um mecanismo que pode futuramente ser integrado ao ambiente de anotação genômica, permitindo ao anotador utilizar várias ontologias, identificando facilmente termos equivalentes entre elas, e escolher aquele que apresenta a descrição mais detalhada, enriquecendo assim a anotação final.

5.1 Etapas da abordagem

Definida a ferramenta a ser reutilizada, será traçada agora a estrutura da abordagem a ser adotada no alinhamento de ontologias biomédicas dentro do processo de anotação genômica. Essa abordagem tem como finalidade principal identificar termos da GO, já utilizados no processo de anotação genômica, em outras ontologias da OBO, para que o anotador possa assim verificar qual desses termos possui a descrição mais detalhada, podendo consequentemente trazer mais detalhes também à anotação.

Antes do alinhamento propriamente dito, são necessários alguns procedimentos, tanto para selecionar e preparar os recursos a serem utilizados, como para facilitar a manipulação das ontologias e reduzir o tempo de execução do futuro sistema. Deste modo, a abordagem proposta contará com os seguintes passos:

- **Preparação:** antes de ter início o processo de anotação genômica, que contará com o auxílio do alinhamento de ontologias para que múltiplas ontologias biomédicas possam ser utilizadas, é necessária uma etapa de preparação. Nessa etapa, são definidas as ontologias envolvidas no alinhamento: *fonte*, *alvo* e *de fundamentação*. A ontologia fonte é aquela a partir da qual buscamos termos equivalentes, neste caso, a GO, e a ontologia alvo é aquela que será alinhada com a GO. A ontologia de fundamentação que auxiliará o processo de alinhamento também é selecionada nessa etapa. A associação entre os termos das ontologias de fundamentação e de domínio (fonte e alvo), que será descrita na Seção 5.2.2, também é realizada nessa etapa, preparando assim as ontologias e deixando-as disponíveis para a etapa de alinhamento. Mais de uma ontologia biomédica pode ser

disponibilizada como alvo para o alinhamento com a GO; neste caso, todas elas devem ser preparadas nessa etapa, ou seja, associadas à ontologia de fundamentação, ficando a cargo do anotador escolher posteriormente qual delas será de fato utilizada no alinhamento;

- **Anotação semi-automática:** como descrito no Capítulo 2**,** existem várias ferramentas que podem automatizar parcial ou totalmente o processo de anotação. Na anotação semi-automática, as sequências são pré-anotadas automaticamente, a partir de sequências similares identificadas em fontes de dados como o GOA (*Gene Ontology Annotation*) Database (BARREL *et al*., 2009). Neste caso, são recuperadas sequências consideradas similares àquela que está sendo anotada, e a descrição é copiada para ser complementada manualmente pelo anotador. A anotação semi-automática é a forma como a tarefa já é realizada atualmente e servirá como ponto de partida da abordagem;
- **Identificação do termo da GO:** como resultado da anotação semi-automática, especialmente quando o GOA é utilizado como fonte de pesquisa, além da anotação é também destacado o termo da GO que foi identificado na descrição da sequência. Essa identificação também já é realizada atualmente e o termo da GO recuperado servirá como entrada para o próximo passo;
- **Extração do fragmento da GO:** com o intuito de reduzir o tempo de processamento, são extraídos fragmentos da ontologia fonte, a GO. Essa ontologia conta atualmente com mais de 29000 termos e submetê-la completa ao alinhamento teria um custo muito alto. Sendo assim, optou-se por dividi-la em ontologias menores e alinhar cada um desses fragmentos com a ontologia alvo, escolhida na etapa de preparação. Com o termo da GO utilizado na anotação semi-automática em mãos, é extraído então um fragmento dessa ontologia. O termo identificado anteriormente é utilizado como parâmetro e a ontologia resultante será composta pela classe que o representa e pelas demais classes no seu entorno;
- **Limpeza e tratamento das ontologias:** a ontologia obtida no passo anterior é então submetida à limpeza e tratamento de nomes. Como as ontologias da OBO geralmente são de grande porte, é necessário um pré-processamento, onde todas as instâncias são removidas. Essas instâncias são na verdade metadados como definições, subconjuntos de origem, recursos externos, etc., e não indivíduos das classes que representam termos biológicos. A remoção dessas instâncias não representa impacto no processo de alinhamento (em termos de perda de informação) e reduz o tamanho das ontologias em

aproximadamente 65%. Outro tratamento necessário antes da realização do alinhamento é em relação aos nomes das classes. Por padrão todos os nomes de classes nas ontologias da OBO são identificadores numéricos, precedidos da sigla da ontologia. Assim, na GO os nomes de classes são da forma GO_XXXXXXX e o mesmo ocorre para qualquer outra ontologia, alterando-se somente o prefixo, sendo que o termo biológico propriamente dito fica registrado no *label* da classe. Assim, para facilitar a comparação sintática, para cada classe das ontologias, o *label* é convertido no novo nome da classe. Esse passo precisa ser realizado em tempo de execução para o fragmento extraído da ontologia fonte. Para a ontologia alvo, definida na etapa de preparação, pode ser mantida uma cópia local já limpa e tratada;

- **Alinhamento:** com as duas ontologias preparadas, é realizado então o alinhamento entre elas. O processo de alinhamento é baseado na abordagem NOM (*Naïve Ontology Mapping*), descrita na Seção 4.5, utilizando as medidas de similaridade selecionadas, que serão relacionadas na Seção 5.2.1, e alinhamentos anteriores, que foram recuperados e armazenados a partir de execuções anteriores e que servirão tanto de alinhamento de referência, indicando os alinhamentos já avaliados como corretos, como também para identificar os alinhamentos incorretos já descartados anteriormente, evitando que eles voltem a ser apresentados ao usuário. Além disso, o alinhamento conta com o auxílio adicional de uma ontologia de fundamentação. O processo de alinhamento e os principais recursos utilizados nesta etapa são descritos na Seção 5.2;
- **Apresentação dos termos equivalentes:** concluído o alinhamento, os resultados obtidos são apresentados ao usuário. Dentre todos os pares de termos retornados, são selecionados apenas aqueles que contêm o termo da ontologia fonte (GO) identificado anteriormente na anotação semi-automática, e o(s) termo(s) da ontologia alvo equivalente(s) a ele, encontrado(s) no alinhamento, é(são) mostrado(s) ao anotador;
- **Validação:** com a relação dos termos equivalentes aos termos da GO em mão, o anotador, que possui conhecimento suficiente do domínio, pode julgar se os alinhamentos retornados estão corretos, ou seja, se os termos realmente se referem ao mesmo conceito biológico. Após descartar os alinhamentos incorretos, o anotador pode então comparar as descrições dos termos nas duas (ou mais) ontologias e verificar qual possui a melhor descrição, selecionando o mais adequado para ser utilizado na anotação final.

Assim, a anotação final contará com três componentes na sua formação: a descrição proveniente da anotação semi-automática, o termo da GO também identificado automaticamente e os termos da ontologia alvo resultantes da validação do anotador, que poderão ser utilizados na complementação manual da anotação. A utilização desses termos ficará a critério do pesquisador, pois o objetivo desta abordagem é apenas oferecer-lhe mais opções, através de uma funcionalidade que pode facilmente ser integrada à ferramenta já utilizada, permitindo que outras ontologias possam ser usadas além da GO, sem que isso exija um esforço adicional excessivo ou a utilização de diversas ferramentas externas ao ambiente de anotação genômica. A Figura 12 esquematiza a abordagem descrita acima, usando como exemplo as ontologias GO como fonte e INOH Event, cujo prefixo é IEV, como alvo.

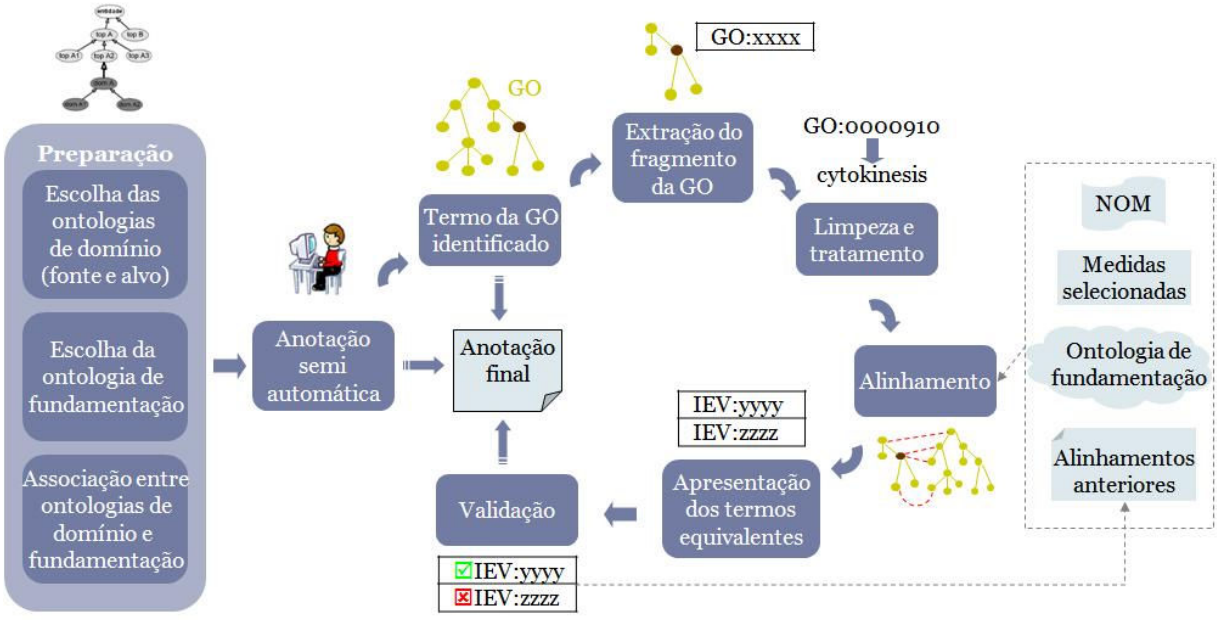

Figura 12. Abordagem proposta para alinhamento de ontologias biomédicas dentro da anotação genômica

5.2 Alinhamento

Dentro da estrutura apresentada na seção anterior, a etapa correspondente ao alinhamento é o ponto principal da abordagem proposta. Em relação às abordagens utilizadas atualmente, representadas pelas ferramentas desenvolvidas nos últimos anos, algumas delas descritas no Capítulo 4, este trabalho se diferencia por dois aspectos principais: a seleção de medidas de

similaridade de acordo com as características das ontologias a serem alinhadas, neste caso, as ontologias biomédicas desenvolvidas pelo consórcio OBO, e a utilização de ontologias de fundamentação como mais um recurso no cálculo de similaridades durante o processo de alinhamento. A seguir, esses dois pontos são detalhados, justificando a seleção do subconjunto de medidas de similaridade e apresentando a forma como ontologias de topo são integradas ao alinhamento de ontologias biomédicas dentro do processo de anotação genômica.

5.2.1 Escolha das medidas de similaridade

O primeiro passo na adaptação do FOAM é a escolha das medidas de similaridade. Essa ferramenta adota um conjunto de medidas que tem como objetivo aproveitar o máximo possível de características das ontologias no processo de alinhamento. Dentre as medidas descritas na Seção 3.6.2, as seguintes são utilizadas pelo FOAM: igualdade, similaridade sintática, igualdade de objetos, multi similaridade, similaridade de *label*, similaridade taxonômica, similaridade extensional, similaridade de domínio e escopo e similaridade de conceito.

Como as ontologias da OBO apresentam características próprias, como número reduzido de propriedades e ausência de instâncias, entre outras, apenas um subconjunto dessas medidas será utilizado. Na escolha, também deve ser levada em conta a utilidade de algumas medidas em relação ao uso de ontologias de fundamentação, como será discutido mais adiante. Após estudar o formato da GO e de algumas outras ontologias biomédicas, as seguintes medidas foram excluídas:

- **Similaridade extensional:** essa medida utiliza instâncias na comparação de conceitos, ou seja, se duas instâncias, ou dois conjuntos de instâncias, são similares, então é provável que os conceitos dos quais são derivadas também sejam similares. Para a OBO, são consideradas instâncias as sequências anotadas com termos de suas ontologias biomédicas. Deste modo, as instâncias ficam armazenadas nos bancos de dados genômicos e não nas ontologias em si. Como essa informação não está disponível na ontologia, o uso dessa medida torna-se desnecessário.
- **Similaridade de domínio e escopo:** essa medida tem como finalidade alinhar propriedades. No âmbito da anotação genômica, a identificação de termos equivalentes é o ponto mais importante, sendo assim apenas conceitos serão alinhados. O fato de as

ontologias da OBO apresentarem um número pouco expressivo de propriedades e relações (não hierárquicas) diminui a necessidade do uso desta medida.

- **Similaridade de conceito:** essa medida se destina à comparação de instâncias e equivale ao inverso da similaridade extensional, ou seja, se dois conceitos são similares então é provável que suas instâncias também sejam similares. Como já foi dito anteriormente, as ontologias da OBO não possuem instâncias, fazendo com que a aplicação dessa medida tenha menor importância.

Desta forma, definimos um subconjunto de medidas de similaridade a serem aplicadas no alinhamento de ontologias biomédicas dentro do processo de anotação genômica, sendo incorporadas à abordagem descrita na Seção 5.1. A Figura 13 ilustra as medidas selecionadas.

Figura 13. Medidas de similaridade selecionadas para o alinhamento de ontologias da OBO

5.2.2 Utilização de ontologias de fundamentação

Um conjunto de categorias cuja existência está atrelada a um domínio específico é chamado de ontologia material. Assim, uma ontologia desenvolvida para descrever, por exemplo, o domínio da biologia molecular é considerada uma ontologia material. Porém, este conjunto de categorias pode ser criado com base em um outro conjunto de (meta-) categorias, conhecido como ontologia de fundamentação (GUIZZARDI, 2009).

Como descrito na Seção 3.5, associar essas meta-categorias aos conceitos de uma ontologia material pode ser útil no alinhamento pelo fato de tornar possível identificar a natureza desses conceitos, transformando esta informação em mais um parâmetro no cálculo de

similaridades, diminuindo assim as chances de serem retornadas associações entre termos de naturezas conceituais distintas.

Para isso, é necessário tornar explícita essa relação entre os conceitos das ontologias de domínio e de fundamentação. Vários trabalhos (MIKA *et al*., 2004; FALLAHI *et al*., 2008; DAMJANOVIĆ *et al*., 2007; PROBST, 2006) mostram que o procedimento mais adequado para esta tarefa é a associação feita de forma manual, dada a necessidade de conhecimento do domínio e de interpretação dos elementos deste domínio dentro do contexto oferecido pela ontologia de fundamentação para o estabelecimento de associações corretas, o que requer intervenção humana. Assim, para que seja possível identificar a natureza conceitual dos termos das ontologias da OBO, provendo mais um parâmetro para o cálculo de similaridades no processo de alinhamento, será realizada a associação manual destes termos com as categorias de uma ontologia de fundamentação, completando desta forma a abordagem proposta apresentada anteriormente.

Segundo Probst (2006), na execução desse tipo de associação, existem três situações possíveis, ilustradas na Figura 14:

A) Os conceitos mais gerais da ontologia de domínio são alinhados aos conceitos mais específicos da ontologia de fundamentação;
B) Relações taxonômicas da ontologia de domínio se sobrepõem a relações taxonômicas da ontologia de fundamentação. Neste caso, a ontologia de domínio pode conter conceitos que são mais gerais que os conceitos mais específicos da ontologia de fundamentação;
C) Um conceito da ontologia de domínio pode atuar como superclasse de vários conceitos da ontologia de fundamentação (conceito *"dom A"* na Figura 14C), ou um conceito da ontologia de domínio é especificado como subclasse de mais de um conceito de alto nível.

Na Figura 14, as setas mais grossas representam relações "subtipo-de" entre um conceito da ontologia de domínio (*dom*) e um conceito da ontologia de topo (*top*). Nos cenários mostrados, a situação A é apresentada como o caso ótimo, onde não é necessário alinhar cada conceito da ontologia de domínio com conceitos da ontologia de topo, mas apenas os conceitos mais gerais, ou seja, aqueles que estão no primeiro nível, sendo que as demais classes herdam recursivamente a associação atribuída à sua superclasse. O esforço necessário para a realização do alinhamento

manual nessa situação é pequeno, pois geralmente há um número reduzido de conceitos no primeiro nível de uma ontologia.

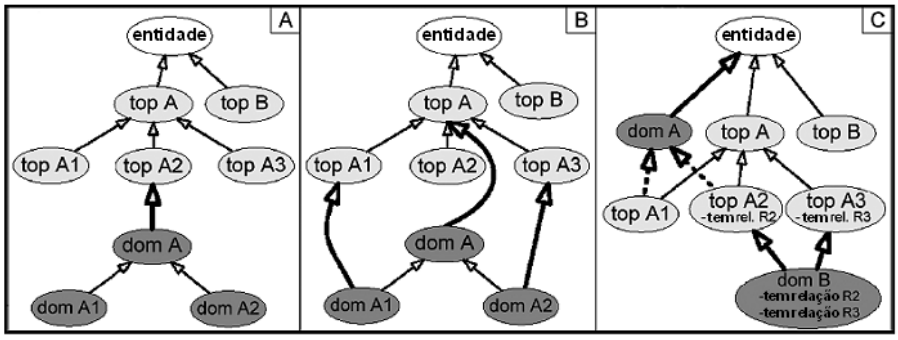

Figura 14. Cenários possíveis no alinhamento entre ontologias de domínio e de topo. Adaptado de Probst (2006)

Já nas situações B e C, pode ser necessária uma remodelagem da ontologia do domínio, removendo ou adicionando conceitos para que a situação A seja atingida. Estas situações podem ocorrer quando as ontologias de domínio não passam por uma modelagem rigorosa no momento em que são desenvolvidas, deixando falhas que precisam ser corrigidas para que se atenham à conceitualização formal subjacente a uma ontologia de fundamentação.

Como as ontologias da OBO são desenvolvidas com base em métodos científicos apoiados por princípios sólidos, com filosofias de projeto compartilhadas por todos os seus colaboradores (OBO FOUNDRY, 2009), e tarefas de remodelagem de ontologias de domínio estão fora do escopo deste trabalho, será assumido o caso ótimo, e as classes mais gerais das ontologias biomédicas selecionadas para o alinhamento (ontologias fonte e alvo) serão associadas às classes mais específicas de uma ontologia de fundamentação.

Com esta associação em mãos, será possível integrar a ontologia de fundamentação à hierarquia das ontologias biomédicas, ou seja, as ontologias da OBO serão editadas para que as classes mais específicas da ontologia de topo passem a ser superclasses de seus conceitos mais gerais, segundo as associações previamente estabelecidas. A vantagem de se ter a meta-categoria de cada conceito integrada à hierarquia da ontologia é que podemos utilizar esta informação durante o processo de alinhamento, através da medida de similaridade taxonômica, descrita na seção 3.6.2. Assim, podemos identificar a natureza conceitual de dois termos candidatos ao alinhamento apenas subindo na hierarquia, e compará-las a fim de determinar se estes termos

podem ser considerados equivalentes ou não. Como o FOAM, ferramenta selecionada para servir como base para a abordagem proposta, já utiliza a medida de similaridade taxonômica, a utilização da ontologia de fundamentação como parâmetro adicional no cálculo de similaridades será incorporada automaticamente ao processo de alinhamento.

5.3 Trabalhos relacionados

Segundo Euzenat e Shvaiko (2007), como ontologias de fundamentação são sistemas baseados em lógica, consequentemente as técnicas de alinhamento que as exploram são baseadas em semântica. Assim, sua utilização em conjunto com outras técnicas representaria uma vantagem em relação a métodos puramente sintáticos. Apesar disso, quase não há relatos de sistemas que utilizem este tipo de abordagem. Até o momento, o único trabalho de que se tem conhecimento foi desenvolvido por Mascardi, Locoro e Rosso (2010), que apresentam um conjunto de algoritmos que exploram ontologias de topo como "pontes semânticas" no processo de alinhamento. Esses algoritmos realizam o alinhamento das duas ontologias de domínio com a ontologia de topo e das duas ontologias de domínio entre si, combinando posteriormente os resultados obtidos nos dois processos.

Essa abordagem mostra bons resultados, aumentando a revocação e mantendo uma precisão comparável, em relação ao alinhamento direto de duas ontologias (sem o intermédio da ontologia de fundamentação). Além disso, nos testes realizados, esses algoritmos conseguiram recuperar associações corretas entre conceitos que possuíam o mesmo significado, mas tinham baixa similaridade léxica. Porém, o fato de implementar um processo totalmente automático aumenta a probabilidade de associações incorretas. A automação completa também se mostra inadequada quando se analisam os resultados obtidos de acordo com as ontologias de topo utilizadas: SUMO-OWL, OpenCyc e DOLCE. Destas, SUMO-OWL e OpenCyc proporcionam bons resultados, pois possuem, além dos conceitos de alto nível, alguns conceitos específicos de vários domínios, o que facilita a identificação automática de similaridades quando são comparadas com ontologias de domínio. Já a DOLCE, por ser uma ontologia de fundamentação "pura", leva a uma degradação dos resultados, quando comparados aos obtidos com o alinhamento direto. O fato de o alinhamento automático necessitar da presença de conceitos

específicos de domínio na ontologia de fundamentação utilizada distorce a finalidade deste recurso, que é prover um conjunto formal de categorias totalmente independente de domínio.

Um outro tipo de trabalho relacionado, realizado com mais frequência, é o alinhamento de ontologias de domínio com ontologias de fundamentação, ou seja, a associação de conceitos de uma ontologias a categorias de mais alto nível com a finalidade resolver problemas de ambiguidade conceitual e de heterogeneidade semântica. Mika *et al*. (2004) e Fallahi *et al*. (2008) realizam o alinhamento de ontologias de descrição de serviços web com a ontologia de fundamentação DOLCE, com o objetivo de eliminar ambiguidades aumentando assim a precisão na descoberta de serviços. A abordagem adotada pelos autores inclui ontologias intermediárias, sendo que o alinhamento é realizado em vários estágios. Uma estratégia semelhante é utilizada por Damjanović *et al*. (2007), que também usa ontologias intermediárias para alinhar os conceitos de ontologias desenvolvidas para a colaboração na área de mecatrônica às categorias da DOLCE. Probst (2006), por outro lado, realiza uma cuidadosa análise dos conceitos de uma ontologia de observações e medidas, que descreve informações geográficas, para alinhá-los diretamente aos conceitos da DOLCE, especificando de forma precisa quais entidades do mundo real são denotadas pelos objetos de informação descritos por esta ontologia de domínio, e restringindo assim as possíveis interpretações dos elementos do modelo.

6 EXPERIMENTO

A anotação genômica baseada em ontologia permite a utilização de um vocabulário comum, compartilhado por diferentes bases de dados, mesmo que estas tenham sido criadas por grupos de pesquisa distintos, permitindo que possam ser combinadas de forma consistente, diminuindo a possibilidade de ambiguidades na interpretação dos dados. A utilização de diversas ontologias vem contribuir com o enriquecimento das anotações geradas, oferecendo ao anotador uma gama maior de recursos para a descrição das sequências.

A abordagem para alinhamento de ontologias proposta neste trabalho visa a permitir a identificação correta dos termos mais adequados em outras ontologias além da Gene Ontology, atualmente a mais (e frequentemente a única) utilizada no processo de anotação. Isto deve se traduzir em resultados úteis para o anotador, ou seja, em alinhamentos corretos. Para validar a estratégia proposta, verificando se conseguimos com ela um número maior de alinhamentos considerados válidos pelo anotador, foi realizado um experimento cujo planejamento e passos de execução são descritos a seguir.

6.1 Planejamento do experimento

Para que fosse possível comparar os resultados obtidos originalmente com uma ferramenta de alinhamento de propósito geral com aqueles obtidos através do processo adaptado às características das ontologias da OBO, fazendo uso apenas do subconjunto de medidas de similaridade selecionadas, além do apoio de uma ontologia de fundamentação, o experimento foi dividido em duas etapas. Na primeira etapa foi realizado o alinhamento entre a GO e outra ontologia da OBO utilizando somente a ferramenta FOAM, e na segunda parte foi utilizada esta mesma ferramenta, porém alterada para contemplar somente as medidas de similaridade selecionadas, contando também com o auxílio de uma ontologia de fundamentação, como descrito na Seção 5.2.2, seguindo assim a abordagem proposta.

Em ambas as etapas foram utilizados os seguintes parâmetros na execução: alinhamento totalmente automático, número de iterações igual a 10 e valor de corte (valor mínimo de similaridade para que o par de entidades seja considerado equivalente) igual a 0,95. Nos dois casos, a GO foi dividida em fragmentos, e cada fragmento foi alinhado com a segunda ontologia,

gerando assim subconjuntos parciais de resultados. Esses subconjuntos foram então consolidados, eliminando-se redundâncias e gerando, para cada etapa, uma única lista de pares de termos considerados equivalentes.

Assim, foram obtidos dois conjuntos de resultados, cada um relativo a uma etapa do experimento. Os resultados de cada parte foram então submetidos à validação de um profissional do domínio, capaz de avaliar se os pares de termos retornados são realmente equivalentes. Isto permitiu uma comparação quantitativa, o que possibilita avaliar de forma objetiva se os recursos introduzidos no processo de alinhamento trouxeram melhorias aos resultados. Os passos principais seguidos na realização deste experimento são descritos a seguir.

6.2 Preparação

A preparação corresponde ao primeiro passo da abordagem, como pode ser verificado na Figura 12. Nesta etapa são definidas as ontologias fonte, alvo e a ontologia de fundamentação que auxiliará o processo de alinhamento. A associação entre as ontologias de fundamentação e de domínio também é realizada nesta etapa. O modo como cada uma destas tarefas foi executada para o experimento é descrito a seguir.

6.2.1 Escolha das ontologias biomédicas

O primeiro passo foi a escolha das ontologias a serem alinhadas, sendo que a ontologia fonte eleita foi a Gene Ontology. Trata-se de uma escolha natural, pois grande parte das pesquisas genômicas que adotam anotação baseada em ontologia utiliza a GO para esta tarefa, em grande parte devido ao uso do GOA, como explicado no Capítulo 2, e o objetivo desta pesquisa é permitir a localização dos termos deste vocabulário em especial em outras ontologias da OBO. Apesar do grande número de ontologias biomédicas disponíveis atualmente, muitas delas possuindo áreas de sobreposição com a GO, foi decidido em um primeiro momento utilizar apenas uma delas como alvo, para que não fosse gerado um volume muito grande de dados, o que dificultaria sua posterior análise.

INOH (*Integrating Network Objects with Hierarchies*) Event Ontology, BRENDA (*BRaunschweig ENzyme DAtabase*) Tissue Ontology, Pathway Ontology, Sequence Ontology e

Systems Biology Ontology foram algumas das ontologias candidatas, por apresentarem termos em comum com a GO. Dentre elas, a INOH Event se mostrou a melhor opção pelo fato de já haver um alinhamento manual entre essa ontologia e a GO. Esse alinhamento é proveniente de uma lista de 800 termos da GO utilizados no sequenciamento do organismo *T. rangeli*, realizado por um grupo de pesquisa do Instituto Oswaldo Cruz (WAGNER, 2006). Desses 800 termos, verificou-se que 26 possuíam coincidência verbal com termos da INOH Event. A presença de termos com coincidência verbal não é uma condição necessária para o alinhamento, porém a identificação dessa área de sobreposição entre as ontologias fonte e alvo ajuda a delimitar o contexto do experimento. Esses termos, bem como as classes equivalentes na GO e na INOH Event estão relacionados na Tabela 5.

Tabela 5. Termos com coincidência verbal presentes na GO e na INOH Event

Termo	***Classe na GO***	***Classe na INOH Event***
actin cytoskeleton organization and biogenesis	GO_0030036	IEV_0001323
Binding	GO_0005488	IEV_0000004
BMP signaling pathway	GO_0030509	IEV_0000863
cell adhesion	GO_0007155	IEV_0000105
cell differentiation	GO_0030154	IEV_0000002
cell growth	GO_0016049	IEV_0000091
cell motility	GO_0048870	IEV_0000103
cell proliferation	GO_0008283	IEV_0000076
cell-cell signaling	GO_0007267	IEV_0002611
Cytokinesis	GO_0000910	IEV_0000079
cytoskeleton organization and biogenesis	GO_0007010	IEV_0001323
DNA repair	GO_0006281	IEV_0000872
DNA replication	GO_0006260	IEV_0000393
embryonic development	GO_0009790	IEV_0000397
immune response	GO_0006955	IEV_0000098
lamellipodium biogenesis	GO_0030032	IEV_0000093
MAPKKK cascade	GO_0000165	IEV_0000445
nervous system development	GO_0007399	IEV_0002710
organ morphogenesis	GO_0009887	IEV_0000576
oxidative phosphorylation	GO_0006119	IEV_0001431
Phosphorylation	GO_0016310	IEV_0000005
regulation of transcription	GO_0045449	IEV_0001833
response to stress	GO_0006950	IEV_0001337
response to wounding	GO_0009611	IEV_0001341
secretory pathway	GO_0045055	IEV_0002539
Translation	GO_0006412	IEV_0000219

Outra vantagem da INOH é o fato de também possuir, assim como a GO, uma página Web[13] onde é possível fazer buscas por palavras-chave, navegar na hierarquia e consultar as descrições dos termos da ontologia. Essa é uma das razões pelas quais a GO muitas vezes é a única ontologia utilizada para anotação, pois seu mecanismo de busca e navegação, chamado AmiGO[14], facilita o trabalho do anotador, evitando que ele percorra a hierarquia manualmente para encontrar o termo que deseja, funcionalidade a qual quase nenhuma outra ontologia dispõe. Como apresenta um mecanismo semelhante ao da GO, a utilização da INOH em um primeiro momento pode ser mais conveniente, até que ferramentas apropriadas para visualização de ontologias sejam incorporadas ao ambiente de anotação genômica.

Assim como a maioria das ontologias da OBO, GO e INOH estão em constante evolução, sendo atualizadas periodicamente. As versões utilizadas neste experimento correspondem àquelas disponíveis no dia 26 de junho de 2009, e foram obtidas a partir do site da OBO Foundry[15].

6.2.2 Escolha da ontologia de fundamentação

O segundo passo seguido foi escolher qual dentre as ontologias de topo disponíveis atualmente representa o melhor recurso a ser utilizado para a validação da abordagem proposta. Esta ontologia foi utilizada apenas na segunda parte do experimento.

O crescente interesse na utilização deste tipo de ontologia, que torna possível negociar significados e estabelecer consensos a respeito de conceitos compartilhados (GANGEMI *et al.*, 2002), tem levado ao desenvolvimento de diversas ontologias de fundamentação, dentre as quais se destacam: Cyc (MATUSZEK *et al.*, 2006), SUMO (PEASE, NILES, LI, 2002), BFO (GRENON, SMITH, GOLDBERG, 2004), DOLCE (GANGEMI *et al.*, 2002), Ontologia de Sowa (SOWA, 1999) e UFO (GUIZZARDI, WAGNER, 2004).

Embora todas essas ontologias se caracterizem por conterem conceitos de alto nível, independentes de domínio, Cyc e SUMO se diferenciam por conterem também muitos conceitos específicos de domínio, o que as torna ontologias de grande porte, sendo que a primeira contém cerca de 300000 conceitos, e a última, cerca de 1000 conceitos, o que inviabiliza a utilização destes recursos dentro da abordagem proposta. Já a ontologia de Sowa e a UFO, apesar de serem

[13] http://www.inoh.org:8083/ontology-viewer/
[14] http://amigo.geneontology.org/cgi-bin/amigo/go.cgi
[15] http://www.obofoundry.org/

modelos formais bem fundamentados e totalmente independentes de domínio, não possuem representação em OWL, necessária para que a ontologia de fundamentação possa ser integrada à ontologia de domínio, o que também as exclui da lista de candidatas. Desta forma, DOLCE e BFO apresentam-se como as ontologias de fundamentação mais adequadas aos propósitos deste trabalho.

Mascardi, Cordì e Rosso (2006) apresentam uma comparação detalhada entre estas e outras ontologias de topo. A seguir são descritas brevemente apenas as características das duas ontologias de interesse: DOLCE e BFO. Uma comparação mais criteriosa entre as duas também pode ser encontrada no trabalho de Grenon (2003).

DOLCE

DOLCE (*Descriptive Ontology for Linguistic and Cognitive Engineering*) é uma ontologia de topo desenvolvida como um módulo da Biblioteca de Ontologias de Fundamentação (*Foundational Ontologies Library*) do projeto WonderWeb[16]. É uma ontologia de *particulares*, ou seja, o domínio de discurso está restrito a estas entidades, e que possui um claro viés cognitivo, pretendendo capturar as categorias ontológicas subjacentes à linguagem natural e ao senso comum humano (GANGEMI *et al.*, 2002). As categorias introduzidas por esta ontologia são tidas como artefatos cognitivos dependentes da percepção humana, de aspectos culturais e de convenções sociais.

DOLCE é baseada na distinção entre *endurants* (entidades duradouras) e *perdurants* (eventos), que se diferenciam em relação ao seu comportamento no tempo. Enquanto *endurants* estão completamente presentes (isto é, todas as suas partes estão presentes) em qualquer momento em que eles estejam presentes, *perdurants* se estendem no tempo, acumulando diferentes partes temporais, sendo que, em qualquer momento em que eles estejam presentes, eles estão apenas parcialmente presentes, pois algumas de suas partes temporais (fases anteriores ou futuras, por exemplo) podem não estar presentes.

Algumas das categorias assumidas na DOLCE compreendem qualidades (*qualities* e *quality regions*), substanciais (*substantials*), agregados (*aggregates*), objetos (*objects*), recursos (*features*) e ocorrências (*occurrences*). Estas categorias são consideradas propriedades rígidas de

[16] http://wonderweb.semanticweb.org/

acordo com a metodologia utilizada no seu desenvolvimento, a OntoClean (GUARINO, WELTY, 2002), que reforça a necessidade de, a princípio, priorizar estes tipos de propriedades.

BFO

BFO (*Basic Formal Ontology*) é definida como "uma teoria das estruturas básicas da realidade", desenvolvida no *Institute for Formal Ontology and Medical Information Science* (IFOMIS), na Universidade de Leipzig. É uma ontologia formal baseada nos princípios metodológicos do realismo (sustenta que a realidade e seus componentes existem independentemente de qualquer representação disso), falibilismo (aceita que qualquer teoria ou classificação está sujeita a revisão), perspectivismo (defende que existe uma pluralidade de alternativas, sendo as várias perspectivas sobre a realidade igualmente legítimas), e adequatismo (defende que estas visões alternativas não podem ser reduzidas a uma única visão básica). Sua característica principal é a tentativa de ser fiel à realidade, aceitando, ao mesmo tempo, a multiplicidade de perspectivas possíveis sobre esta realidade (GRENON, SMITH, GOLDBERG, 2004).

Ao contrário da DOLCE, a BFO é uma ontologia de *universais*. A principal diferença entre particulares e universais é em relação à instanciação: particulares não têm instâncias, enquanto universais podem ser instanciados. Apesar de ser uma ontologia de fundamentação totalmente independente de domínio, a BFO foi elaborada com o intuito de servir como base para o desenvolvimento de novas ontologias dentro do domínio biomédico. Desta forma, ela se divide em dois componentes: um para objetos biológicos (correspondente à anatomia em geral), e outro para processos biológicos (correspondente à fisiologia em geral). Esses componentes correspondem às duas ontologias que compõem a BFO: SNAP, que abrange continuantes (*continuants*, equivalentes aos *endurants* da DOLCE), e SPAN, que engloba ocorrentes (*occurrents*, equivalentes aos *perdurants* na DOLCE).

Os universais descritos pela BFO abrangem, entre outros, regiões espaciais, entidades substanciais e entidades dependentes na SNAP, e entidades processuais, regiões temporais e espaço-temporais na SPAN. O objetivo de se deixar clara a divisão entre as ontologias SNAP e SPAN é permitir que seja possível focar, de forma alternada, tanto nos processos como nas entidades substanciais neles envolvidas, sejam como agentes ou como afetadas.

DOLCE e BFO apresentam vários pontos em comum, como a divisão entre *endurants* e *perdurants* (ou *continuants* e *occurrents*), teoria da relação parte-todo e teoria da dependência, entre outros, e algumas discrepâncias, como, por exemplo, o tratamento de particulares e universais e a descrição de elementos espaço-temporais (GRENON, 2003). As hierarquias das duas ontologias estão ilustradas na Figura 15.

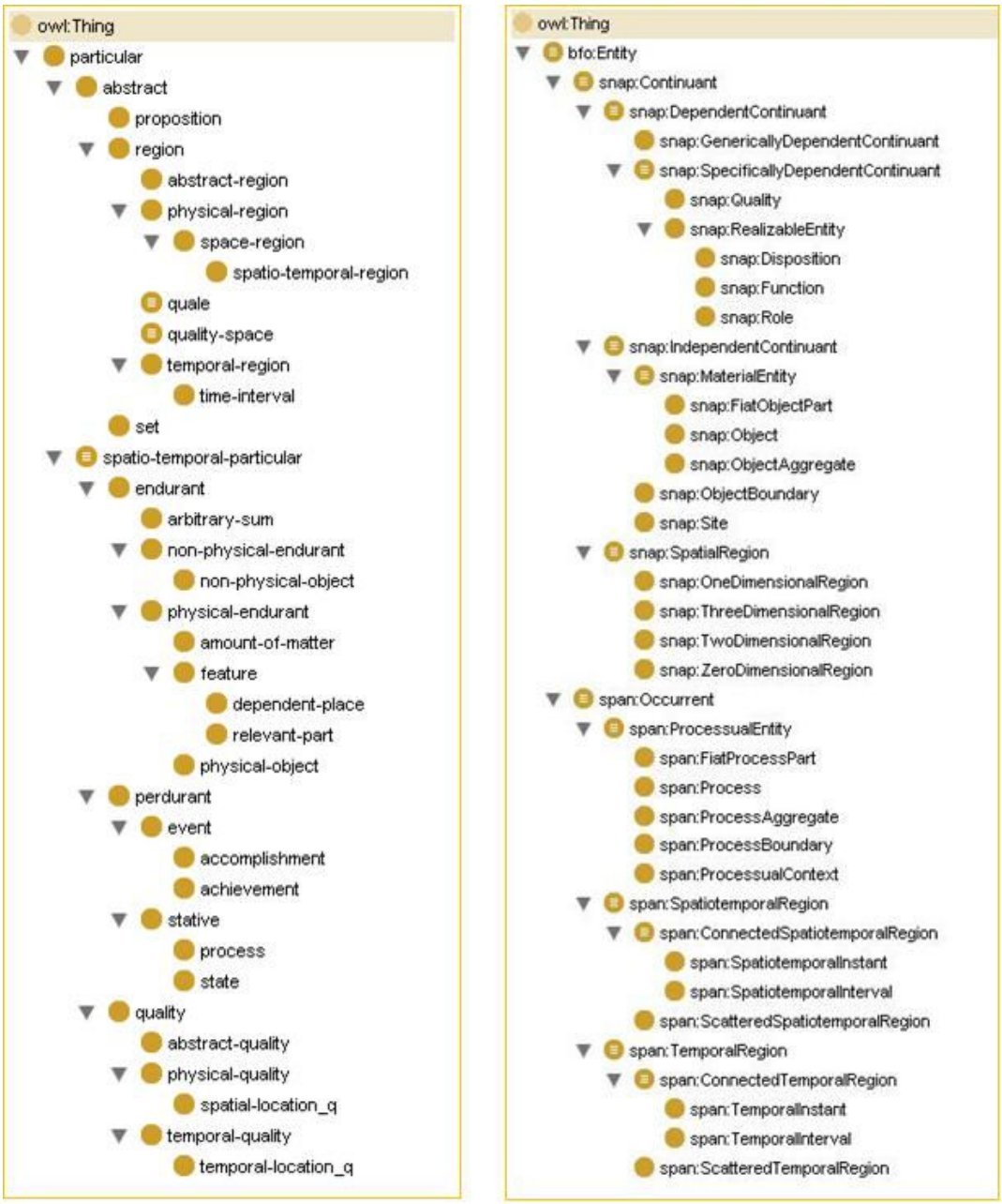

Figura 15. Hierarquias das ontologias de fundamentação DOLCE (à esquerda) e BFO (à direita)

Apesar do enfoque distinto em alguns aspectos, as duas ontologias se mostram adequadas e ambas poderiam ser utilizadas no processo de alinhamento de ontologias biomédicas dentro da anotação genômica. Para este experimento, optou-se por, a princípio, selecionar apenas uma delas, tendo sido escolhida a BFO. A escolha se justifica pelo fato de esta ontologia já ter sido desenvolvida dentro do domínio biomédico, com o objetivo de nortear o desenvolvimento de novas ontologias (materiais) nesta área, o que pode vir a facilitar a tarefa de associação entre conceitos de topo e de domínio, dados que os primeiros já foram elaborados para categorizar especificamente conceitos biológicos. Esta associação será descrita na próxima seção.

Vale ressaltar que a abordagem proposta foi elaborada de modo a aceitar quaisquer ontologias, sejam elas de domínio ou de fundamentação. Assim, da mesma forma que a ontologia INOH Event pode ser substituída por qualquer outra ontologia da OBO como alvo, também a BFO pode ser substituída por qualquer outra ontologia de fundamentação, sem alteração da estrutura proposta.

6.2.3 Associação entre as ontologias de domínio e de fundamentação

Escolhida a ontologia de fundamentação a ser utilizada, foi feita então a associação entre os conceitos mais específicos dessa ontologia com os termos mais gerais das ontologias Biological Process, proveniente da GO, e INOH Event, conforme descrito na Seção 5.2.2. Descartando-se a classe raiz (*"biological process"* na GO e *"Event"* na INOH Event) no primeiro nível das duas ontologias, existem 21 classes no segundo nível da Biological Process, que correspondem aos termos mais gerais dessa ontologia. Na INOH Event, existem apenas 2 classes no segundo nível, motivo pelo qual optou-se por utilizar as classes no terceiro nível, totalizando 6 conceitos considerados mais gerais para essa ontologia. Esses conceitos estão ilustrados na Figura 16 e os conceitos mais específicos da BFO podem ser conferidos na Figura 15 (à direita).

Como já foi mencionado anteriormente, o número reduzido de classes nos primeiros níveis facilita o trabalho de associação manual. Essa associação foi feita com o auxílio de uma bióloga, analisando-se as definições e exemplos fornecidos pela BFO para cada um de seus conceitos, e também as definições de termos da Gene Ontology e da INOH Event, quando estes

estavam disponíveis. A utilização de conhecimento do domínio, através da participação de um profissional da área, foi de fundamental importância, permitindo uma análise criteriosa e, consequentemente, uma categorização mais precisa. A associação entre os conceitos da Biological Process e da BFO está registrada na Tabela 6, e entre a INOH Event e a BFO, na Tabela 7.

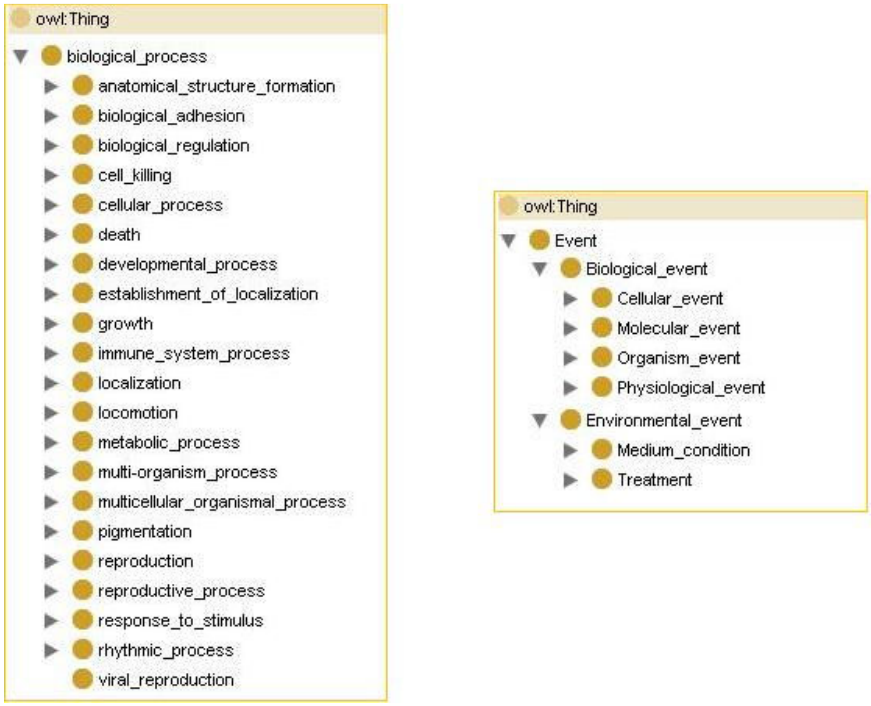

Figura 16. Conceitos mais gerais das ontologias Biological Process (à esq.) e INOH Event (à dir.)

Como pode ser observado, por se tratarem de ontologias relativas a processos biológicos, a associação concentrou-se predominantemente nas categorias da segunda parte da BFO, relativa à ontologia SPAN, que descreve eventos temporais, como explicado anteriormente.

Como há um número maior de classes no segundo nível da Biological Process, em comparação com a INOH Event, há uma variedade maior de tipos de conceitos, o que não ocorre nesta última, pois quando há um número pequeno de conceitos nos níveis mais altos da ontologia estes tendem a ser mais gerais, fornecendo pouca informação que permita diferenciá-los uns dos outros com mais precisão.

Tabela 6. Associação entre conceitos da GO (ramo Biological Process) e BFO

Gene Ontology (Biological Process)	**Conceitos BFO**																								
	GenericallyDependentContinuant	**Quality**	**Disposition**	**Function**	**Role**	**FiatObjectPart**	**Object**	**ObjectAggregate**	**ObjectBoundary**	**Site**	**ZeroDimensionalRegion**	**OneDimensionalRegion**	**TwoDimensionalRegion**	**ThreeDimensionalRegion**	**FiatProcessPart**	**Process**	**ProcessAggregate**	**ProcessBoundary**	**ProcessualContext**	**SpatiotemporalInstant**	**SpatiotemporalInterval**	**ScatterdSpatiotemporalRegion**	**TemporalInstant**	**TemporalInterval**	**Scattered TemporalRegion**
biological process																									
anatomical structure																✓									
biological adhesion																		✓							
biological regulation																			✓						
cell killing															✓										
cellular process																			✓						
death																		✓							
developmental process															✓										
establishment of localization																✓									
growth															✓										
immune system process																			✓						
localization																		✓							
locomotion																✓									
metabolic process																			✓						
multi-organism process																✓									
multicellular organismal																✓									
pigmentation																✓									
reproduction																		✓							
reproductive process																✓									
response to stimulus																			✓						
rhythmic process																			✓						
viral reproduction																✓									

Nota-se também na Tabela 7 que, apesar de a INOH Event ser uma ontologia de processos, a classe *"Medium condition"* foi classificada como *"Site"*, categoria pertencente à SNAP, que descreve entidades duradouras. Isto acontece porque esta classe (e toda a sub-árvore da qual é raiz) possivelmente está mal posicionada na hierarquia, sendo subclasse de uma classe de natureza distinta da sua, pois *"Environmental event"*, analisando-se a uniformidade obtida na classificação, também seria um *"ProcessualContext"*, assim como as demais classes. Este fato reforça a escolha das classes no terceiro nível desta ontologia como uma boa opção para a

associação, já que, caso esta fosse realizada com as classes do segundo nível, a classe *"Medium condition"* poderia herdar uma categoria não condizente com a sua real natureza conceitual.

Tabela 7. Associação entre conceitos da INOH Event e BFO

INOH Event	**Conceitos BFO**																							
	GenericallyDependentContinuant	**Quality**	**Disposition**	**Function**	**Role**	**FiatObjectPart**	**Object**	**ObjectAggregate**	**ObjectBoundary**	**Site**	**ZeroDimensionalRegion**	**OneDimensionalRegion**	**TwoDimensionalRegion**	**ThreeDimensionalRegion**	**FiatProcessPart**	**Process**	**ProcessAggregate**	**ProcessBoundary**	**ProcessualContext**	**SpatiotemporalInstant**	**SpatiotemporalInterval**	**ScatterdSpatiotemporalRegion**	**TemporalInstant**	**TemporalInterval**	**Scattered TemporalRegion**
Event																									
Biological event																									
Cellular event																			✓						
Molecular event																			✓						
Organism event																			✓						
Physiological event																			✓						
Environmental event																									
Medium condition										✓															
Treatment																			✓						

Concluída a associação, foi feita então a integração física das ontologias. Utilizando o editor de ontologias Protégé, Biological Process e INOH Event foram editadas para que a ontologia de fundamentação passasse a fazer parte de suas hierarquias. Após a importação da BFO, as categorias desta foram adicionadas como superclasses das classes mais gerais das duas ontologias de domínio, segundo as relações obtidas anteriormente. Na Figura 17 podemos ver a ontologia BFO sendo importada para dentro da ontologia Biological Process (1), permitindo que as classes da primeira sejam superclasses da segunda, como, por exemplo, *"locomotion"*, que além de ter como superclasse *"biological_process"*, agora também é subclasse de *"Process"* (2).

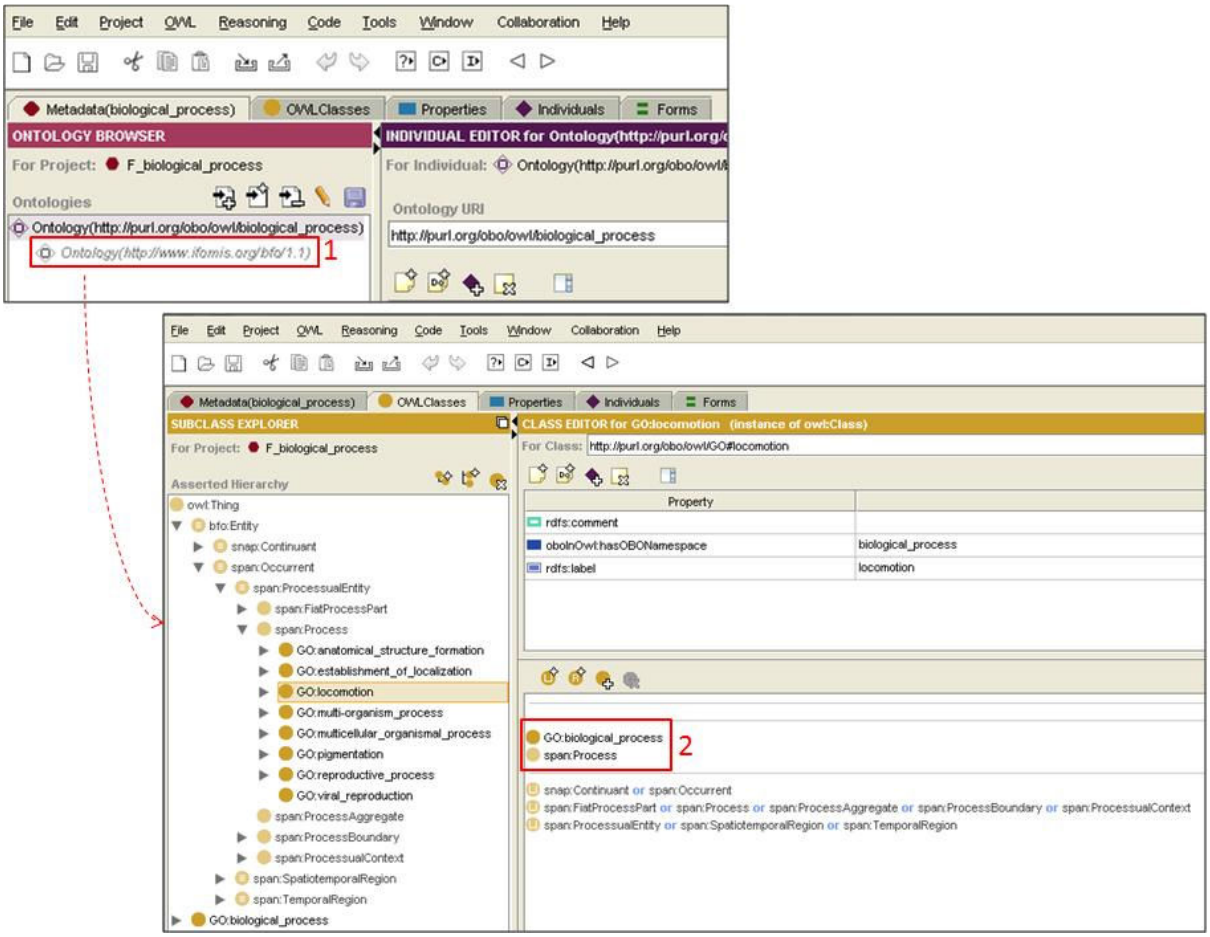

Figura 17. Integração entre as ontologias de domínio e de topo no Protégé

O mesmo processo foi repetido para cada uma das classes listadas nas tabelas 6 e 7. Desta forma, as demais classes herdam a categoria atribuída à sua superclasse, e a informação sobre a natureza conceitual, fornecida pela ontologia de fundamentação, fica disponível para cada conceito da ontologia de domínio, e pode agora ser utilizada durante o alinhamento como mais um parâmetro no cálculo de similaridades.

6.3 Seleção de termos, extração de fragmentos, limpeza e tratamento

Como descrito na Seção 5.1, devido ao grande porte da GO, é necessário segmentá-la, gerando fragmentos que são alinhados com a ontologia INOH Event. Para efetuar a divisão foi

utilizado o Galen Segmenter[17], um aplicativo que segmenta qualquer ontologia tendo como foco uma classe fornecida como parâmetro. A ontologia resultante é composta por essa classe e suas superclasses e subclasses, que são extraídas até uma certa profundidade. A fragmentação foi feita com base na lista de 26 termos citada na Seção 6.2.1, ressaltando novamente que a coincidência verbal não é um requisito obrigatório, e foi utilizada aqui como critério de seleção dos termos apenas para reduzir o tamanho do conjunto a ser utilizado no experimento. Desses termos, 25 pertenciam ao ramo Biological Process da GO e apenas um (*"binding"*) fazia parte da Molecular Function. Como a INOH Event também se refere a processos biológicos, decidiu-se trabalhar apenas com a Biological Process, sendo então geradas 25 pequenas ontologias, todas elas possuindo área de sobreposição com a INOH Event.

Para a etapa de limpeza e tratamento, foram desenvolvidos pequenos aplicativos para automatizar estas tarefas. Uma classe Java realiza a remoção das instâncias de metadados e outra classe faz a troca dos identificadores numéricos pelos *labels*, que passam a ser os novos nomes das classes. A limpeza, relacionada às instâncias, facilita bastante a manipulação das ontologias pela ferramenta, pois o tamanho excessivo pode comprometer o desempenho do sistema, exigindo muito tempo para o processamento.

Em relação ao tratamento de nomes, é importante frisar que, apesar de o FOAM considerar diversas características das ontologias na comparação, o nome da classe ainda é o ponto de partida e tem grande influência no valor final da similaridade. O aplicativo desenvolvido, que converte o *label* no novo nome da classe, ajuda a evitar associações do tipo "GO_0002064" e "IEV_0002064", classes cujos *labels* são, respectivamente, *"epithelial cell development"* e *"Binding of GPCR and G alpha q"*, que possuem coincidência verbal de nomes, mas se referem a termos distintos. Como exemplo de resultado deste aplicativo, a classe "GO_0000910" passou a se chamar *"cytokinesis"*, e, de forma semelhante, todas as classes das duas ontologias tiveram seu nome original substituído por seu *label*. Vale salientar que o FOAM realiza comparação de *labels*, incluindo essa informação no cálculo da similaridade, porém esse tratamento prévio nos nomes das classes pode facilitar bastante o trabalho da ferramenta, já que esta é uma das informações mais importantes no processo de alinhamento.

[17] http://www.co-ode.org/galen/

6.4 Alinhamento

Terminada a fase de limpeza e tratamento das ontologias, foi executada a primeira parte do experimento, utilizando somente a ferramenta FOAM, sem nenhum recurso adicional. Foram realizados então 25 alinhamentos, combinando a ontologia INOH Event com cada um dos 25 fragmentos extraídos da GO. A Figura 18 mostra um exemplo de execução do FOAM, que é chamado através de linha de comando, passando-se como argumento o arquivo de configuração que define os parâmetros do alinhamento.

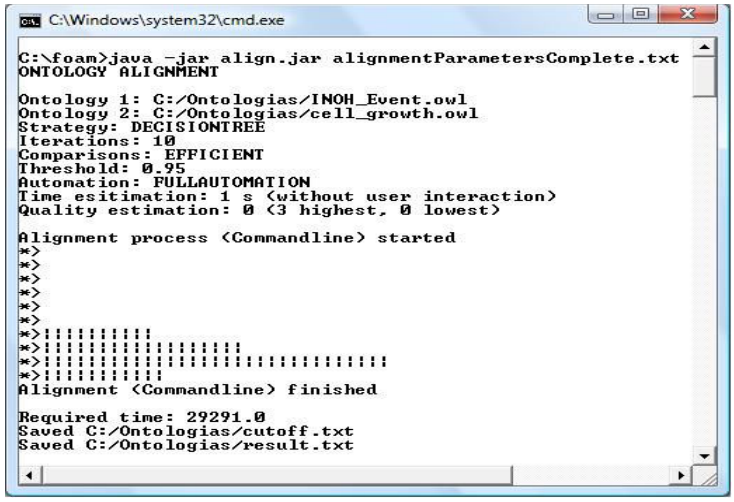

Figura 18. Alinhamento das ontologias cell_growth.owl e INOH_Event.owl no FOAM

Após repetir o processo para todos os 25 fragmentos, foram geradas 25 planilhas contendo os resultados parciais do alinhamento entre GO e INOH Event. Esses resultados foram então consolidados em uma única planilha, de onde foram removidas as redundâncias, sendo obtidos 57 alinhamentos, ou seja, 57 pares de termos considerados equivalentes. A lista completa dos alinhamentos obtidos após a consolidação dos resultados está no Apêndice A[18].

Para a segunda parte do experimento, antes de realizar o alinhamento entre as ontologias, foram feitas algumas adaptações no FOAM. Essas alterações foram realizadas com o intuito de adequar o funcionamento da ferramenta às características das ontologias do OBO, refletidas no conjunto de medidas de similaridade selecionadas.

[18] A primeira parte deste experimento foi realizada em colaboração com a professora Maria Cláudia Reis Cavalcanti, do Instituto Militar de Engenharia (IME)

Como o FOAM realiza o alinhamento tanto de classes como de propriedades e instâncias, o primeiro passo foi eliminar estas duas últimas entidades do processo de comparação, pois as ontologias da OBO não possuem instâncias e têm um número bem pequeno de propriedades. Além disso, apenas as classes que representam os termos utilizados na anotação interessam ao anotador, devendo este ser o único resultado retornado.

Outra alteração necessária foi a seleção de regras. No FOAM, durante a comparação de duas entidades, é utilizado um conjunto de regras, sendo que cada regra implementa uma determinada medida de similaridade. Assim, o sistema foi adaptado para que fossem utilizadas apenas as regras relativas às medidas previamente selecionadas, como detalhado na Seção 5.2.1. Algumas outras alterações menores também foram efetuadas, apenas para ajustar a apresentação dos resultados, melhorando a legibilidade e facilitando sua manipulação.

Com a ferramenta adaptada e as ontologias preparadas foi executado então o alinhamento entre GO e INOH, contando agora com o auxílio da ontologia de fundamentação BFO, que foi associada às ontologias de domínio como descrito na Seção 6.2.3. Os 25 termos utilizados novamente deram origem a 25 fragmentos da GO, mas desta vez enriquecidos com as categorias da ontologia de fundamentação BFO. Foram novamente realizados 25 alinhamentos, gerando 25 planilhas de resultados. Após a consolidação desses 25 resultados parciais e eliminação de redundâncias, foram obtidos 64 pares de termos considerados equivalentes, desconsiderando-se as associações envolvendo termos da BFO. Os resultados desse alinhamento estão relacionados no Apêndice B.

6.5 Validação e análise dos resultados

Para que fosse possível comparar quantitativamente os resultados das duas partes do experimento, determinando o número de acertos e erros obtidos nos alinhamentos, e verificar se a abordagem proposta realmente proporciona um número maior de alinhamentos corretos, os dois conjuntos de pares de termos foram submetidos à validação de um biólogo, com experiência na área de sequenciamento e anotação genômica. A participação de um profissional do domínio nesta etapa é de grande importância, pois é necessário conhecimento específico suficiente para afirmar se os dois termos, retornados como resultado do processo de alinhamento, correspondem ao mesmo conceito biológico ou não, garantindo maior confiabilidade e precisão nos resultados.

Além de resultados totalmente corretos (os dois termos são equivalentes), ou totalmente incorretos (os dois termos representam conceitos biológicos distintos), o processo de alinhamento, por analisar também a estrutura das ontologias, buscando similaridades dentro dos contextos onde as classes estão inseridas, pode retornar termos que não são equivalentes, mas que possuem algum tipo de relação. Estas associações também podem vir a ser úteis, caso o anotador se interesse por relações do tipo "é-um", "parte-de" ou outras, e tenha facilidade para identificá-las, podendo agregar também estas informações à anotação. Sendo assim, para a análise dos alinhamentos obtidos, foi adotada uma classificação onde, para cada par de termos, foi atribuído um grau de 0 a 5, variando de totalmente incorreto (0) a totalmente correto (5). Esta classificação está relacionada na Tabela 8.

Tabela 8. Classificação adotada na validação dos alinhamentos

Grau	***Significado***
5	correto
4	relação forte
3	relação média
2	relação fraca
1	relação insignificante
0	incorreto

A Tabela 9 apresenta os resultados da validação para as duas partes do experimento: na coluna "Resultados", à esquerda, a primeira parte, com os 57 pares de termos obtidos com o FOAM, e à direita, os 64 pares obtidos com a aplicação da abordagem proposta, na segunda parte do experimento.

Tabela 9. Resultados da validação do experimento

Classificação		***Resultados***			
		Parte 1		**Parte 2**	
Grau	*Significado*	*Quantidade*	*Porcentagem*	*Quantidade*	*Porcentagem*
5	correto	27	47%	27	42%
4	relação forte	8	14%	12	19%
3	relação média	7	12%	7	11%
2	relação fraca	2	4%	1	2%
1	relação insignificante	8	14%	9	14%
0	incorreto	5	9%	8	13%
	Total	57	100%	64	100%

Se focarmos apenas em resultados corretos e incorretos, podemos notar na tabela acima que, dos 57 pares de termos considerados equivalentes pelo FOAM, sem recursos adicionais, 47% estão corretos e 9% totalmente incorretos. Já a abordagem proposta registra, para os 64 pares de termos retornados, 42% corretos e 13% totalmente incorretos. Nota-se uma diminuição de 5% na taxa de acertos e um aumento de 4% na taxa de erros. Porém, estes resultados correspondem ao conjunto de alinhamentos cujo valor de corte foi estabelecido como 0,95, como descrito na Seção 6.1, que estipula os parâmetros utilizados neste experimento. Isto significa que todos os pares de termos retornados possuem valor de similaridade igual ou superior a 0,95. Analisando-se esses valores de similaridade (que podem ser conferidos nos Apêndices A e B), notamos que há uma variação maior nos valores da segunda parte do experimento, em relação aos valores obtidos na primeira parte. Se subirmos o valor de corte, e considerarmos pares de termos com similaridade igual ou superior a 0,97, na primeira parte eliminamos apenas 3 dos 57 pares, ficando assim com 54 pares de termos. Já na segunda parte, correspondente à aplicação da abordagem proposta, dos 64 pares retornados, eliminamos 22, restando desta forma 42 pares de termos. Os resultados da validação para este subconjunto de pares obtidos a partir da utilização de um valor de corte mais alto estão listados na Tabela 10.

Tabela 10. Resultados da validação do experimento considerando o valor de corte igual a 0,97

Classificação		***Resultados***			
		Parte 1		**Parte 2**	
Grau	*Significado*	*Quantidade*	*Porcentagem*	*Quantidade*	*Porcentagem*
5	correto	27	50%	27	64%
4	relação forte	8	15%	8	19%
3	relação média	6	11%	2	5%
2	relação fraca	2	4%	0	0%
1	relação insignificante	7	13%	4	10%
0	incorreto	4	7%	1	2%
	Total	54	100%	42	100%

Na tabela acima verificamos que, utilizando valor de corte igual a 0,97, obtemos com a execução do FOAM 50% de pares corretos e 7% totalmente incorretos. Já com a aplicação da abordagem proposta, 64% dos pares retornados estavam corretos, enquanto 2% estavam totalmente incorretos. Notamos um aumento significativo de 14% na taxa de acertos e uma diminuição de 5% na taxa de erros quando a abordagem proposta é utilizada para o alinhamento, associada a um valor de corte mais rigoroso. Além disso, podemos notar também um aumento de

4% nos pares onde os termos possuem relação forte, que podem, de forma secundária, apresentar utilidade para o anotador, e uma diminuição na porcentagem de pares que apresentam relações mais fracas, que poderiam representar apenas ruídos para o pesquisador.

Podemos concluir, a partir destas análises que, apesar de a abordagem proposta retornar um número maior de erros, estes alinhamentos incorretos possuem um valor de similaridade mais baixo, sendo possível filtrá-los com um valor de corte adequado, que pode ser obtido experimentalmente. Por outro lado, sem os recursos oferecidos pela abordagem, o FOAM retorna alinhamento corretos, relacionados e totalmente incorretos com valores de similaridade muito próximos, tornando difícil a filtragem dos resultados.

Outro ponto a ser ressaltado é o ganho em resultados novos, obtido a partir da aplicação da abordagem. Na Tabela 10, podemos notar que, na segunda parte do experimento, foram retornados 27 pares corretos. Desses 27, 10 pares correspondem a resultados novos, ou seja, são pares não encontrados pelo FOAM na primeira parte do experimento, mas retornados como resultado quando a abordagem proposta é utilizada. Em relação aos pares onde os termos apresentam relação forte, dos 8 pares encontrados na segunda parte do experimento, 6 são resultados novos encontrados quando a abordagem foi utilizada. A Tabela 11 apresenta os ganhos em resultados novos, dividindo os resultados obtidos na segunda etapa do experimento em dois grupos: resultados antigos (que apresentam sobreposição com a primeira etapa), ou seja, resultados em comum com a execução do FOAM sem recursos adicionais, e resultados totalmente novos.

Tabela 11. Ganhos em resultados novos com a aplicação da abordagem

Classificação		***Resultados da aplicação da abordagem***			
		Antigos (sobreposição)		**Novos**	
Grau	*Significado*	*Quantidade*	*Porcentagem*	*Quantidade*	*Porcentagem*
5	correto	17	85%	10	45%
4	relação forte	2	10%	6	27%
3	relação média	0	0%	2	9%
2	relação fraca	0	0%	0	0%
1	relação insignificante	1	5%	3	14%
0	incorreto	0	0%	1	5%
	Total	20	100%	22	100%

Na Tabela 11 podemos observar que, dos 42 pares de termos retornados na segunda parte do experimento, 20 pares já haviam sido recuperados na primeira parte pelo FOAM. Desses, 17

foram avaliados como corretos, ou seja, dos resultados encontrados pelo FOAM e aproveitados pela abordagem, 85% são corretos. Os 22 pares restantes correspondem aos resultados encontrados exclusivamente pela abordagem e, dentre eles, 10 são totalmente corretos, o que representa 45% de resultados úteis para o anotador entre os resultados novos.

Podemos observar também que, apesar dos ganhos em resultados novos, alguns pares corretos foram perdidos. A Tabela 10 mostra que o FOAM retornou 27 pares corretos na primeira parte do experimento. Desses 27 pares, 17 também foram recuperados pela abordagem na segunda parte. Os 10 pares restantes, apesar de terem sido avaliados como corretos, não foram encontrados na segunda etapa. Essa perda pode ser decorrente da política de comparação da ferramenta associada à alteração estrutural realizada nas ontologias, em consequência da integração da ontologia de fundamentação à hierarquia das ontologias de domínio. Como a ferramenta compara todas as classes da primeira ontologia com todas as classes da segunda, retornando um valor de similaridade para cada par, a comparação entre classes da ontologia de fundamentação e das ontologias de domínio, que provavelmente retorna valores baixos de similaridade, pode estar introduzindo algum ruído, já que todos os valores influenciam na similaridade de pares vizinhos durante as iterações seguintes.

Quando comparamos esses resultados com aqueles obtidos por Mascardi, Locoro e Rosso (2010), descrito na Seção 5.3, notamos uma melhora significativa. Neste trabalho, que também trata do alinhamento utilizando ontologias de topo, combinando os resultados dos alinhamentos feitos de forma direta e com intermédio da ontologia de fundamentação, apesar de uma melhora na revocação, houve uma degradação na precisão. Quando consideramos apenas o uso da DOLCE, que é uma ontologia de fundamentação totalmente independente de domínio, assim como a BFO aqui utilizada, há um aumento de 2,5% na revocação e uma queda de 14% na precisão. O cálculo da revocação não é realizado aqui, pois seria necessário um alinhamento de referência feito de forma manual, porém podemos notar que, comparado à degradação na precisão observada por Mascardi, Locoro e Rosso, o aumento de 14% na precisão obtido com a aplicação da abordagem proposta é um ganho significativo, que justifica os esforços de sua implementação.

Desta forma, podemos afirmar que a abordagem proposta neste trabalho, que realiza o alinhamento de ontologias biomédicas utilizando um conjunto de medidas de similaridade adequadas às características desses recursos e conta com o auxílio de uma ontologia de

fundamentação, se aplicada em conjunto com um valor de corte mais rigoroso, apesar de uma pequena perda, apresenta resultados superiores àqueles obtidos com uma ferramenta de alinhamento convencional, aumentando a precisão dos resultados, retornando uma proporção maior de associações corretas e reduzindo a taxa de erros dos alinhamentos que serão apresentados ao usuário final.

7 CONCLUSÃO

O desenvolvimento de pesquisas genômicas tem avançado muito nos últimos anos, beneficiando diversas áreas da sociedade e exigindo cada vez mais recursos tecnológicos que deem suporte às tarefas de manipulação e armazenamento da grande quantidade de dados gerada. Neste cenário, as pesquisas na área de Bioinformática são de grande importância para apoiar a descoberta, sequenciamento e descrição de novos organismos, bem como a posterior disponibilização e acesso às informações resultantes.

A anotação genômica é uma tarefa importante no processo de sequenciamento, pois, dentre outros usos, permite que uma descrição seja atribuída a cada sequência identificada. Esses conjuntos *sequência-anotação* passam a compor grandes bases de dados genômicas, que podem ser acessadas por grupos de pesquisa em qualquer parte do mundo. Para isso, é necessário que seja utilizada uma linguagem comum, o que se torna possível com a utilização de ontologias. Atualmente, um grande número de ontologias biomédicas têm sido desenvolvidas, proporcionando um vasto vocabulário para a tarefa de anotação. A Gene Ontology é o exemplo mais expressivo, sendo a ontologia mais (e muitas vezes, a única) utilizada neste processo.

O objetivo deste trabalho é facilitar a utilização de diversas ontologias biomédicas no processo de anotação genômica. Através da elaboração de uma abordagem de alinhamento de ontologias voltado para o domínio biomédico, torna-se possível, a partir de um termo da GO utilizado na anotação e recuperado automaticamente, identificar termos equivalentes em outras ontologias biomédicas, permitindo ao anotador escolher qual dentre eles possui a descrição mais detalhada, tornando-se o mais adequado à anotação.

A finalidade principal dessa abordagem é melhorar a qualidade dos alinhamentos obtidos, ou seja, recuperar um número maior de alinhamentos corretos. Para isso, foi realizada uma avaliação envolvendo diversas ferramentas de alinhamento, com o intuito de determinar qual delas possuía a abordagem mais completa e apresentava maior potencial de reutilização. A abordagem adotada pela ferramenta FOAM, conhecida como NOM (*Naïve Ontology Mapping*), que calcula a similaridade entre entidades iterativamente considerando um amplo conjunto de características das ontologias, aliada a vantagens como código aberto, boa documentação e extensibilidade, são pontos fortes que confirmaram a reutilização da ferramenta como a melhor opção.

Além do reuso de uma abordagem iterativa que considera diversas características das ontologias no cálculo de equivalências, dois pontos principais nortearam a elaboração da abordagem: a seleção de um conjunto de medidas de similaridade considerando-se as características específicas das ontologias das OBO, e a utilização de uma ontologia de fundamentação, que fornece mais um parâmetro no cálculo de similaridades, permitindo a análise da natureza conceitual de cada termo, diminuindo assim a possibilidade de associações entre termos pertencentes a categorias distintas.

A realização de um experimento comparativo, que contrapôs os resultados retornados com a execução do FOAM, sem nenhum recurso adicional, e aqueles retornados com a aplicação da abordagem proposta, ambos avaliados após a validação de um profissional do domínio, mostrou que a abordagem, associada a um valor de corte mais rigoroso, proporcionou maior precisão no resultados, levando a um aumento de 14% nos alinhamentos corretos e uma diminuição de 5% na taxa de erros.

Além da melhoria obtida nos resultados, uma outra contribuição importante é a introdução de ontologias de fundamentação no processo de alinhamento, pois as ferramentas disponíveis atualmente não consideram a noção de natureza conceitual dos elementos no momento da comparação. A categoria de cada termo ainda é um aspecto pouco explorado; distingui-las e usar essa informação no cálculo de similaridades, através do uso de ontologias de fundamentação, cujo desenvolvimento se apoia em sólidas bases teóricas, é um passo que ajuda a aumentar a influência de fatores semânticos nesse processo, expandindo ainda mais o universo de informações a serem exploradas durante o alinhamento.

Também se inclui como contribuição da abordagem proposta a elaboração de um processo mais abrangente que, além do alinhamento em si, envolve etapas de preparação e tratamento das ontologias, que são importantes por estarmos lidando com ontologias de grande porte. O pré-processamento desses recursos é um passo essencial para garantir um bom desempenho na execução do sistema, pois, sem esse tratamento prévio, o alto consumo de tempo e memória poderia inviabilizar a realização do alinhamento. Dessa forma, o processo de alinhamento de ontologias é considerado em toda sua extensão, desde a preparação das ontologias para o alinhamento propriamente dito, até a interação com o usuário para a validação dos resultados.

Assim, a abordagem proposta para alinhamento de ontologias biomédicas dentro do processo de anotação genômica se mostra como uma boa opção para complementar esta atividade, permitindo ao pesquisador utilizar diversas ontologias, podendo escolher dentre elas qual, em determinado momento, apresenta o termo mais adequado às suas necessidades. A identificação de termos equivalentes, com maior precisão nos resultados apresentados, obtidos através de um aplicativo que pode facilmente ser integrado ao ambiente já utilizado pelo anotador, pode sem dúvida trazer contribuições positivas às pesquisas nessa área.

7.1 Trabalhos futuros

Como descrito na Seção 6.4, a ferramenta FOAM passou por algumas modificações para se adaptar aos objetivos e necessidades da abordagem proposta. Assim como o original, esse novo aplicativo resultante consiste em uma ferramenta de linha de comando configurável e independente de plataforma, podendo facilmente ser chamada a partir de aplicações maiores. Como sugestão de trabalho futuro, esta ferramenta poderia ser incorporada a um sistema de anotação genômica, permitindo que a abordagem possa ser avaliada em um ambiente real de pesquisa, sendo aplicada durante um projeto de sequenciamento e anotação. Assim será possível avaliar, além dos aspectos quantitativos já verificados no experimento descrito no Capítulo 6, a utilidade da abordagem do ponto de vista qualitativo, conferindo seu impacto no trabalho diário do anotador, e verificando se a disponibilização de mais recursos vem facilitar seu trabalho e contribuir com o aumento da qualidade das informações geradas.

Para diminuir o número de alinhamentos corretos não encontrados pela abordagem, como discutido na Seção 6.5, poderiam ser realizados alguns ajustes adicionais na ferramenta, que poderia ser alterada para que as classes da ontologia de fundamentação fossem comparadas somente entre si. Desta forma, seriam evitadas comparações entre conceitos de topo e de domínio, que provavelmente retornam valores baixos de similaridade, introduzindo ruídos nos resultados finais.

Outra sugestão de trabalho futuro seria em relação à visualização dos resultados. Quando o anotador valida um alinhamento, ele tem em mãos apenas o par de termos considerados equivalentes. Se, no momento da validação, estes termos pudessem ser visualizados dentro de sua hierarquia, permitindo identificar elementos próximos, como suas superclasses, o anotador

poderia aceitar ou descartar uma associação mais rapidamente e com mais segurança. A integração de uma ferramenta de visualização de ontologias ao ambiente de anotação genômica pode contribuir na validação do alinhamento, otimizando o trabalho do anotador ao passo que evita que ele necessite sair de seu ambiente para acessar uma ferramenta externa para manipular as ontologias que utiliza.

Por fim, a realização de mais experimentos, abrangendo um conjunto maior de ontologias e uma variedade maior de termos utilizados nas anotações, ajudaria a reforçar as conclusões traçadas anteriormente. Além disso, a validação por um grupo maior de biólogos ajudaria a avaliar o impacto da subjetividade dessas avaliações nos resultados, permitindo verificar como eles são afetados pela aplicação de diferentes pontos de vista.

REFERÊNCIAS

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W., LIPMAN, D. J. **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. Nucleic Acids Research, v.25, n.17, p.3389-3402. 1997.

BARRELL, D., DIMMER, E., HUNTLEY, R. P., BINNS, D., O'DONOVAN, C., APWEILER, R. **The GOA database in 2009 - an integrated Gene Ontology Annotation resource.** Nucleic Acids Research, v.37, Database Issue, p.396-403. 2009.

BBOP. **Berkeley Bioinformatics and Ontologies Project**. Disponível em <http://www.berkeleybop.org>. Acesso em 16/10/2009.

BELLOZE, K.T. **Uma Extensão do Processo de Anotação Genômica para Ampliar o Uso e a Evolução Colaborativa de Ontologias no Domínio da Biologia Molecular**. Tese de Mestrado. Instituto Militar de Engenharia. Rio de Janeiro. 2007.

BERGMANN, R. **Experience Management: Foundations, Development Methodology, and Internet-Based Application**. Springer. 2002.

BERNSTEIN, A., KAUFMANN, E., KIEFER, C., BÜRKI, C. **Simpack: A generic Java library for similarity measures in ontologies**. Technical report. University of Zurich. Zurique. 2005.

BODENREIDER, O., STEVENS, R. **Bio-ontologies: current trends and future directions.** Briefings in Bioinformatics, v.7, n.3, p.256–274. 2006.

BORODOVSKY, M., MCININCH, J. **Recognition of genes in DNA sequence with ambiguities**. Biosystems, v.30, n.1-3, p.161-171. 1993.

BREITMAN, K.K. **Web Semântica: A Internet do Futuro**. LTC. 2005.

CAPLAN, P. **You call it corn, we call it syntax-independent metadata for document-like objects (defining a standard for describing network accessible information resources)**. Public-Access Computer Systems Review, v. 6, n. 4, p.19-23. 1995.

CHUNG, S., WOOLEY, J. **Challenges Faced in the Integration of Biological Information**. In LACROIX, Z., CRITCHLOW, T. **Bioinformatics: Managing Scientific Data**. Morgan Kaufmann. 2003.

DAMJANOVIĆ, V., BEHRENDT, W., PLÖßNIG, M., HOLZAPFEL, M. **Developing Ontologies for Collaborative Engineering in Mechatronics.** In *Proceedings of the 4th European Conference on The Semantic Web: Research and Applications.* Innsbruck, Áustria. 2007.

DAVID, J., GUILLET, F., BRIAND, H. **Association Rule Ontology Matching Approach.** International Journal on Semantic Web and Information Systems, v.3, n.2, p. 27-49. 2007.

DÁVILA, A. M. R., LORENZINI, D.M., MENDES, P. N., SATAKE, T. S., SOUSA, G. R., CAMPOS, L. M., MAZZONI, C. J., WAGNER, G., PIRES, P. F., GRISARD, E. C., CAVALCANTI, M. C. R., CAMPOS, M. L. M. **GARSA: genomic analysis resources for sequence annotation.** Bioinformatics, v.21, p.4302-4303. 2005.

DE BRUIJN, J., EHRIG, M., FEIER, C., MARTÍN-RECUERDA, F., SCHARFFE, F., WEITEN, M., **Ontology mediation, merging and aligning**. In DAVIES, J., STUDER, R., WARREN, P. **Semantic Web Technologies. Trends and Research in Ontology-based Systems**. Wiley and Sons. 2006.

DE BRUIJN, J., EHRIG, M., MARTÍN-RECUERDA, F., POUERES, A., PREDOIU, L. **SEKT deliverable D4.4.1: Ontology mediation management.** Technical report. University of Innsbruck. Innsbruck. 2005.

DELCHER, A.L., HARMON, D., KASIF, S., WHITE, O., SALZBERG, S.L. **Improved microbial gene identification with Glimmer**. Nucleic Acids Research, v.27, n.23, p.4636–4641. 1999.

DING, Y., FENSEL, D., KLEIN, M., OMELAYENKO, B. **The Semantic Web: yet another hip?** Data Knowledge Engineering, v.41, n.2–3, p.205–227. 2002.

DOWELL, R., JOKERST, R., DAY, A., EDDY, S., STEIN, L. **The distributed annotation system.** BMC Bioinformatics, v.2, n.7. 2001.

EDDY, S. **HMMER - profile hidden Markov models for biological sequence analysis Version 2.3.2.** Howard Hughes Medical Institute and Dept. of Genetics Washington, University School of Medicine. St. Louis, EUA. 2003.

EHRIG, M. **Ontology Alignment: Bridging de Semantic Gap**. Springer. 2007.

EHRIG, M., STAAB, S. **QOM – Quick Ontology Mapping**. In *Proceedings of the Third International Semantic Web Conference (ISWC2004)*. Hiroshima, Japão. 2004.

EHRIG, M., SURE, Y. **Active ontology alignment**. In *Proceedings of the Collaboration Workshop for the Future Semantic Web at ESWC-2005*. Heraklion, Grécia. 2005.

EHRIG, M., SURE, Y. **FOAM - Framework for Ontology Alignment and Mapping - Results of the Ontology Alignment Evaluation Initiative.** In *Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies.* Banff, Canadá. 2005.

EUZENAT, J., SHVAIKO, P. **Ontology Matching**. Springer. 2007.

EWING, B., GREEN, P. **Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities.** Genome Research. v.8, n.3, p.186–194. 1998.

EWING, B., HILLIER, L., WENDL, M. C., GREEN, P. **Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment.** Genome Research, v.8, n.3, p.175–185. 1998.

FALLAHI, G. R., MESGARI, M. S., RAJABIFARD, A., FRANK, A. U. **A Methodology Based on Ontology for Geo-Service Discovery.** World Applied Sciences Journal, v.3, n.2, p.300–311. 2008.

FELSENSTEIN, J. **PHYLIP (Phylogenetic Inference Package), version 3.65.** Department of Genetics, University of Washington. Seattle, EUA. 2005.

FRISHMAN, D., VALENCIA, A. **Modern genome annotation: the BioSapiens Network**. Springer. 2008.

GANGEMI, A. GUARINO, N., MASOLO, C., OLTRAMARI, A., SCHNEIDER, L. **Sweetening Ontologies with DOLCE**. In BENJAMINS, V., GÓMEZ-PÉREZ, A. **Proceedings of the 13th European Conference on Knowledge Acquisition and Management (EKAW-2002)**, Lecture Notes in Computer Science. Springer. 2002.

GRACIA, J., MENA, E. **Ontology Matching with CIDER: Evaluation Report for the OAEI 2008.** In *Proceedings of the 3rd International Workshop on Ontology Matching (OM-2008).* Karlsruhe, Alemanha. 2008.

GRENON, P., SMITH, B., GOLDBERG, L. **Biodynamic Ontology: Applying BFO in the Biomedical Domain.** In PISANELLI, D. M. **Ontologies in Medicine**. IOS Press. 2004.

GRENON, P. **BFO in a nutshell: A bi-categorial axiomatization of BFO and comparison with DOLCE.** Relatório Técnico. Institute for Formal Ontology and Medical Information Science (IFOMIS), Faculdade de Medicina, Universidade de Leipzig, Alemanha. 2003.

GRUBER, T. **Ontology**. 2008. Disponível <http://tomgruber.org/writing/ontology-definition-2007.htm>. Acesso em 20/05/2009.

GUARINO, N. **The Ontological Level.** In CASATI, R., SMITH, B., WHITE, G. **Philosophy and the Cognitive Science.** Holder-Pivhler-Tempsky. 1994.

GUARINO, N., WELTY, C. **Evaluating Ontological Decisions with OntoClean**. Communications of the ACM, v.45, n.2, p.61–65. 2002.

GUIZZARDI, G., WAGNER, G. **A Unified Foundational Ontology and some Applications of it in Business Modeling.** In *Proceedings of the 16th International Conference on Advances in Information Systems Engineering (CAiSE 2004)*. Riga, Letônia. 2004.

GUIZZARDI, G. **Ontological Foundations for Structural Conceptual Models.** Universal Press. 2005.

GUIZZARDI, G. **Ontology-Driven Conceptual Modeling.** Material de curso. II Seminário de Pesquisa em Ontologia no Brasil. Rio de Janeiro, Brasil. 2009.

GUIMARÃES, M. P. **Uma Abordagem para Capturar a Proveniência de Dados na Área de Bioinformática**. Tese de Mestrado. Instituto Militar de Engenharia. Rio de Janeiro. 2009.

HUANG, X., MADAN, A. **A DNA Sequence Assembly Program**. Genome Research, v.9, n.9, p.868–877. 1999.

JEAN-MARY, Y., SHIRONOSHITA, E., KABUKA, M. **Ontology matching with semantic verification.** Web Semantics: Science, Services and Agents on the World Wide Web, v.7, n.3, p.235-251. 2009.

JIAN, N., HU, W., CHENG, G., QU, Y. **FalconAO: Aligning Ontologies with Falcon.** In *Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies.* Banff, Canadá. 2005.

KALFOGLOU, Y., SCHORLEMMER, M. **Ontology mapping: the state of the art**. Knowledge Engineering Review, v.18, n.1, p.1-31. 2003.

KANEHISA, M., GOTO, S. **KEGG: Kyoto Encyclopedia of Genes and Genomes**. Nucleic Acids Research, v.28, n.1, p.27-30. 2000.

KARPLUS, K., BARRETT, C., HUGHEY, R. **Hidden markov models for detecting remote proteins homologies**. Bioinformatics, v.14, n.10, p.846–856. 1998.

KIU, C.-C., LEE, C.-S. **Ontology Mapping and Merging through OntoDNA for Learning Object Reusability.** Educational Technology & Society, v.9, n.3, p.27-42. 2006.

KLEIN, M. **Combining and relating ontologies: an analysis of problems and solutions**. In GÓMEZ-PÉREZ, A., GRUNINGER, M., STUCKENSCHMIDT, H., USCHOLD, M. **Proceedings of Workshop on Ontologies and Information Sharing at IJCAI-01**. Seattle, EUA. 2001.

KUMAR ,S., TAMURA,. K, NEI, M. **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** Brief Bioinformatics, v.5, n.2, p.150–163. 2004.

LACROIX, Z., CRITCHLOW, T. **Bioinformatics: Managing Scientific Data**. Morgan Kaufmann. 2003.

LAMBRIX, P., TAN, H. **SAMBO - A System for Aligning and Merging Biomedical Ontologies** . Web Semantics: Science, Services and Agents on the World Wide Web, v.4, n.3, p.196-206. 2006.

LEMOS, M. **Workflow para bioinformática.** Tese de Doutorado. Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro. 2005.

LEVENSHTEIN, I. **Binary codes capable of correcting deletions, insertions, and reversals**. Doklady Akademii Nauk SSSR, v. 163, n. 4, p. 845-848. 1965.

LI, L., STOECKERT, C. J., ROOS, D. S. **Genomes OrthoMCL: Identification of Ortholog Groups for Eukaryotic.** Genome Research, v.13, n.9, p.2178–2189. 2003.

LI, Y., LI, J., TANG, J. **RiMOM – Ontology Alignment with Strategy Selection.** In *Proceedings of the 6th International Semantic Web Conference.* Busan, Coréia do Sul. 2007.

MAEDCHE, A., STAAB, S. **Measuring similarity between ontologies**. In BENJAMINS, V., GÓMEZ-PÉREZ, A. **Proceedings of the 13th European Conference on Knowledge Acquisition and Management (EKAW-2002)**, Lecture Notes in Computer Science. Springer. 2002.

MASCARDI, V., CORDÌ, V., ROSSO, P. **A Comparison of Upper Ontologies.** Relatório Técnico. Departamento de Ciência da Computação, Universidade de Gênova, Itália. 2006.

MASCARDI, V., LOCORO, A., ROSSO, P. **Automatic Ontology Matching Via Upper Ontologies: A Systematic Evaluation.** IEEE Transactions on Knowledge and Data Engineering, v.22, n.5, p.609–623. 2010.

MATUSZEK, C., CABRAL, J., WITBROCK, M., DEOLIVEIRA, J. **An Introduction to the Syntax and Content of Cyc.** In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*. Stanford, EUA. 2006.

MCGUINNESS, D.L., FIKES, R., RICE, J., WILDER, S. **An Environment for Merging and Testing Large Ontologies**. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000)*. Breckenridge, EUA. 2000.

MIKA, P., SABOU, M., GANGEMI, A., OBERLE, D. **Foundations for OWL-S: Aligning OWL-S to DOLCE.** Papers from 2004 AAAI Spring Symposium - Semantic Web Services, n.SS-04-06, p.52–60. 2004.

MULDER, N. J., APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BATEMAN, A., BINNS, D., BRADLEY, P., BORK, P., BUCHER, P., CERUTTI, L., COPLEY, R., COURCELLE, E., DAS, U., DURBIN, R., FLEISCHMANN, W., GOUGH, J., HAFT, D., HARTE, N., HULO, N., KAHN, D., KANAPIN, A., KRESTYANINOVA, M., LONSDALE, D., LOPEZ, R., LETUNIC, I., MADERA, M., MASLEN, J., MCDOWALL, J., MITCHELL, A., NIKOLSKAYA, A. N., ORCHARD, S., PAGNI, M., PONTING, C. P., QUEVILLON, E., SELENGUT, J., SIGRIST, C. J., SILVENTOINEN, V., STUDHOLME, D. J., VAUGHAN, R., WU, C. H. **InterPro, progress and status in 2005**. Nucleic Acids Research, v.33, p.D201–D205. 2005.

NAGY, M., VARGAS-VERA, M., MOTTA, E. **DSSim-ontology mapping with uncertainty.** In *Proceedings of the 1st International Workshop on Ontology Matching (OM-2006).* Athens, EUA. 2006.

NCBO. **The National Center for Biomedical Ontology**. Disponível em <http://bioontology.org>. Acesso em 16/10/2009.

NOTREDAME, C., HIGGINS, D. G., HERINGA, J. **T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment**. Journal of Molecular Biology, v.302, n.1, p.205–217. 2000.

NOY, N.F., MUSEN, M.A. **PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment**. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence(AAAI2000)*. Austin, EUA. 2000.

NOY, N.F., MUSEN, M.A. **The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping**. International Journal of Human-Computer Studies, v.59, n.6, p.983-1024. 2003.

OBO FOUNDRY. **The Open Biomedical Ontologies**. 2009. Disponível em <http://www.obofoundry.org>. Acesso em 20/05/2009.

ORG. **Ontology Research Group**. Disponível em <http://org.buffalo.edu>. Acesso em 16/10/2009.

PEARSON, W. R., LIPMAN, S. J. **Improved tools for biological sequence comparison.** Proceedings of the National Academy of Sciences, v.85, n.8, p.2444–2448. 1988.

PEASE, A., NILES, I., LI, J. **The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications**. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web.* Edmonton, Canadá. 2002.

PROBST, F. **Ontological Analysis of Observations and Measurements.** In *Proceedings of the 4th International Conference on Geographic Information Science (GIScience).* Münster, Alemanha. 2006.

RICE, P., LONGDEN, I., BLEASBY, A. **EMBOSS: The European Molecular Biology Open Software Suite.** Trends in Genetics, v.16, n.6, p.276–277. 2000.

ROSSE, C., MEJINO JR., J. **A reference ontology for biomedical informatics: the Foundational Model of Anatomy** . Journal of Biomedical Informatics, v.36, n.6, p.478-500. 2003.

RUBIN, D. L., SHAH, N. H., NOY, N. F. **Biomedical ontologies: a functional perspective.** Briefings in Bioinformatics, v.9, n.1, p.75–90. 2007.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M., BARRELL, B. **Artemis: sequence visualisation and annotation**. Bioinformatics, v.16, n.10, p.944–945. 2000.

SHAFER, G. **A Mathematical Theory of Evidence**. Princeton University Press. 1976.

SMETS, P. **Belief functions**. In SMETS, P., MAMDANI, E., DUBOIS, D., PRADE, H. **Non-Standard Logics for Automated Reasoning**. Academic Press. 1988.

SMITH, B., ASHBURNER, M., ROSSE, C., BARD, C., BUG, W., CEUSTERS, W., GOLDBERG, L. J., EILBECK, K., IRELAND, A., MUNGALL, C. J., THE OBI CONSORTIUM, LEONTIS, N., ROCCA-SERRA, P., RUTTENBERG, A., SANSONE, S-A., SCHEUERMANN, R. H., SHAH, N., WHETZEL, P. L., LEWIS, S. **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration**, Nature Biotechnology, v.25, n.11, p.1251–1255. 2007.

SOWA, J. F., **Knowledge Representation: Logical, Philosophical, and Computational Foundations**. Brooks Cole Publishing Co. 1999.

THE GENE ONTOLOGY CONSORTIUM. **The Gene Ontology Project in 2008**. Nucleic Acids Research, v. 36, p.440–444. 2008.

THOMPSON, J. D., HIGGINS, D. G., GIBSON, T. J. **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** Nucleic Acids Research, v.22, n.22, p.4673–4680. 1994.

VAN HARMELEN, F. **Ontology Mapping: A Way out of the Medical Tower of Babel?** Summary of Invited Talk. *10th Conference on Artificial Intelligence in Medicine (AIME 05)*. Aberdeen, Escócia. 2005.

WAGNER, G. **Geração e análise comparativa de seqüências genômicas de *Trypanosoma rangeli***. Tese de Mestrado. Instituto Osvaldo Cruz. Rio de Janeiro. 2006.

WANG, P., XU, B. **Lily: Ontology Alignment Results for OAEI 2008.** In *Proceedings of the 3rd International Workshop on Ontology Matching (OM-2008)*. Karlsruhe, Alemanha. 2008.

APÊNDICE A – Resultados da primeira parte do experimento

Resultado consolidado do alinhamento entre as ontologias GO (ramo *Biological Process*) e INOH Event, realizado como primeira parte do experimento, utilizando-se a ferramenta FOAM.

Classe GO (Biological Process)	*Classe INOH Event*	*Similaridade*
GO#BMP_signaling_pathway	IEV#BMP_signaling_pathway	1.0
GO#DNA_repair	IEV#DNA_repair	1.0
GO#JNK_cascade	IEV#JNK_cascade	1.0
GO#MAPKKK_cascade	IEV#MAPKKK_cascade	1.0
GO#Notch_signaling_pathway	IEV#Notch_signaling_pathway	1.0
GO#actin_filament_organization	IEV#Actin_filament_organization	0.9965405405405405
GO#brain_morphogenesis	IEV#Organ_morphogenesis	0.9965405405405405
GO#cell_communication	IEV#Cell_communication	0.9965405405405405
GO#cell_development	IEV#Seed_development	0.9965405405405405
GO#cell_differentiation	IEV#Cell_differentiation	0.9965405405405405
GO#cell_part_morphogenesis	IEV#Cellular_morphogenesis	0.9965405405405405
GO#embryonic_morphogenesis	IEV#Embryonic_morphogenesis	0.9965405405405405
GO#gene_expression	IEV#Gene_expression	0.9965405405405405
GO#growth	IEV#Growth	0.9965405405405405
GO#localization	IEV#colocalization	0.9965405405405405
GO#morphogenesis_of_an_epithelium	IEV#Morphogenesis_of_an_epithelium	0.9965405405405405
GO#morphogenesis_of_embryonic_epithelium	IEV#Morphogenesis_of_embryonic_epithelium	0.9965405405405405
GO#muscle_cell_differentiation	IEV#Muscle_cell_differentiation	0.9965405405405405
GO#neural_plate_morphogenesis	IEV#Neural_plate_morphogenesis	0.9965405405405405
GO#organ_development	IEV#Organ_development	0.9965405405405405
GO#response_to_external_stimulus	IEV#Response_to_external_stimulus	0.9965405405405405
GO#response_to_stimulus	IEV#Response_to_stimulus	0.9965405405405405
GO#response_to_stress	IEV#Response_to_stress	0.9965405405405405
GO#seed_development	IEV#Seed_development	0.9965405405405405
GO#sex_differentiation	IEV#Cell_differentiation	0.9965405405405405
GO#shoot_development	IEV#Wood_development	0.9965405405405405
GO#tube_development	IEV#Tube_development	0.9965405405405405
GO#wound_healing	IEV#Wound_healing	0.9965405405405405
GO#cardiac_cell_differentiation	IEV#Myoblast_differentiation	0.9897959183673469
GO#ectoderm_development	IEV#Embryonic_development	0.9897959183673469
GO#epidermis_development	IEV#Organ_development	0.9897959183673469
GO#JAK-STAT_cascade	IEV#JAK_STAT_pathway	0.9897959183673469
GO#positive_regulation_of_cellular_process	IEV#Positive_regulation_of_Toll_receptor_signaling	0.9897959183673469
GO#positive_regulation_of_developmental_process	IEV#Positive_regulation_of_gene_expression	0.9897959183673469
GO#system_development	IEV#Embryonic_development	0.9897959183673469
GO#lymph_gland_development	IEV#Negative_regulation_of_BAD_inactivation_signaling	0.989669702324317
GO#defense_response	IEV#Immune_response	0.9755409219190969
GO#DNA_replication	IEV#DRB_medication	0.9755409219190969
GO#epidermal_cell_differentiation	IEV#Smooth_muscle_cell_differentiation	0.9755409219190969
GO#IDP_phosphorylation	IEV#Autophosphorylation	0.9755409219190969
GO#male_sex_differentiation	IEV#Lymphocyte_differentiation	0.9755409219190969

Classe GO (Biological Process)	*Classe INOH Event*	*Similaridade*
GO#midgut_development	IEV#Flower_development	0.9755409219190969
GO#negative_regulation_of_cell_communication	IEV#Negative_regulation_of_cell_cycle_regulation	0.9755409219190969
GO#negative_regulation_of_cellular_process	IEV#Negative_regulation_of_gene_expression	0.9755409219190969
GO#negative_regulation_of_smoothened_signaling_pathway	IEV#Negative_regulation_of_G2-M-phase_transition_pathway	0.9755409219190969
GO#negative_regulation_of_transcription_termination	IEV#Negative_regulation_of_Lck_activation_signaling	0.9755409219190969
GO#neuron_differentiation	IEV#Transdifferentiation	0.9755409219190969
GO#osmosensory_signaling_pathway	IEV#Hedgehog_signaling_pathway	0.9755409219190969
GO#positive_regulation_of_cell_death	IEV#Positive_regulation_of_gene_expression	0.9755409219190969
GO#positive_regulation_of_cell_death	IEV#Positive_regulation_of_S_phase	0.9755409219190969
GO#response_to_chemical_stimulus	IEV#Response_to_endogenous_stimulus	0.9755409219190969
GO#response_to_DNA_damage_stimulus	IEV#Response_to_external_stimulus	0.9755409219190969
GO#spermatid_differentiation	IEV#Lymphocyte_differentiation	0.9755409219190969
GO#trichoblast_differentiation	IEV#Lymphocyte_differentiation	0.9755409219190969
GO#forebrain_neuron_development	IEV#Transcription_of_FGF8_target_gene	0.9575856443719413
GO#negative_regulation_of_signal_transduction	IEV#Positive_regulation_pathway	0.9575856443719413
GO#pyrimidine_dimer_repair	IEV#genetic_interaction	0.9575856443719413

APÊNDICE B – Resultados da segunda parte do experimento

Resultado consolidado do alinhamento entre as ontologias GO (ramo *Biological Process*) e INOH Event, realizado como segunda parte do experimento, aplicando-se a abordagem proposta (os alinhamentos envolvendo classes da ontologia de fundamentação foram desconsiderados).

Classe GO (Biological Process)	***Classe INOH Event***	***Similaridade***
GO#BMP_signaling_pathway	IEV#BMP_signaling_pathway	1.0
GO#DNA_repair	IEV#DNA_repair	1.0
GO#JNK_cascade	IEV#JNK_cascade	1.0
GO#MAPKKK_cascade	IEV#MAPKKK_cascade	1.0
GO#Notch_signaling_pathway	IEV#Notch_signaling_pathway	1.0
GO#actin_filament_organization	IEV#Actin_filament_organization	0.9965405405405405
GO#cell_communication	IEV#Cell_communication	0.9965405405405405
GO#cell_cycle	IEV#Cell_cycle	0.9965405405405405
GO#cell_differentiation	IEV#Cell_differentiation	0.9965405405405405
GO#cell_part_morphogenesis	IEV#Cellular_morphogenesis	0.9965405405405405
GO#cell_proliferation	IEV#Cell_proliferation	0.9965405405405405
GO#cell-cell_signaling	IEV#Cell-cell_signaling	0.9965405405405405
GO#cytokinesis	IEV#Cytokinesis	0.9965405405405405
GO#embryonic_morphogenesis	IEV#Embryonic_morphogenesis	0.9965405405405405
GO#gene_expression	IEV#Gene_expression	0.9965405405405405
GO#morphogenesis_of_an_epithelium	IEV#Morphogenesis_of_an_epithelium	0.9965405405405405
GO#morphogenesis_of_embryonic_epithelium	IEV#Morphogenesis_of_embryonic_epithelium	0.9965405405405405
GO#multicellular_organismal_process	IEV#Multicellular_organismal_process	0.9965405405405405
GO#negative_regulation_of_smoothened_signaling_pathway	IEV#Negative_regulation_of_Notch_signaling_pathway	0.9965405405405405
GO#phosphorylation	IEV#Phosphorylation	0.9965405405405405
GO#programmed_cell_death	IEV#Programmed_cell_death	0.9965405405405405
GO#response_to_abiotic_stimulus	IEV#Response_to_biotic_stimulus	0.9965405405405405
GO#response_to_stress	IEV#Response_to_stress	0.9965405405405405
GO#sex_differentiation	IEV#Cell_differentiation	0.9965405405405405
GO#stress_fiber_formation	IEV#Stress_fiber_formation	0.9965405405405405
GO#wound_healing	IEV#Wound_healing	0.9965405405405405
GO#JAK-STAT_cascade	IEV#JAK_STAT_pathway	0.9897959183673469
GO#phyllome_development	IEV#Tube_development	0.9897959183673469
GO#phyllome_development	IEV#Wood_development	0.9897959183673469
GO#smoothened_signaling_pathway	IEV#Wnt_signaling_pathway	0.9897959183673469
GO#system_development	IEV#Bud_development	0.9897959183673469
GO#brain_development	IEV#Wood_development	0.9755409219190969
GO#cell-cell_signaling_involved_in_cell_fate_specification	IEV#Cell-cell_signaling_involved_in_cell_fate_commitment	0.9755409219190969
GO#defense_response	IEV#Immune_response	0.9755409219190969
GO#hyperphosphorylation	IEV#Autophosphorylation	0.9755409219190969
GO#male_sex_differentiation	IEV#Osteoblast_differentiation	0.9755409219190969
GO#negative_regulation_of_cell_communication	IEV#Negative_regulation_of_cell_cycle_regulation	0.9755409219190969
GO#neuron_differentiation	IEV#Transdifferentiation	0.9755409219190969
GO#positive_regulation_of_caspase_activity	IEV#Positive_regulation_of_Caspase_cascade	0.9755409219190969

Classe GO (Biological Process)	*Classe INOH Event*	*Similaridade*
GO#prostate_gland_morphogenesis	IEV#Post-embryonic_morphogenesis	0.9755409219190969
GO#regulation_of_smoothened_signaling_pathway	IEV#Regulation_of_Fas_signaling_pathway	0.9755409219190969
GO#system_development	IEV#Seed_development	0.9755409219190969
GO#activation_of_immune_response	IEV#Molecular_event	0.9575856443719413
GO#cell_killing	IEV#Proteasome_degradation_in_cytosol	0.9575856443719413
GO#cell_killing	IEV#Proteasome_degradation_in_plasma_membrane	0.9575856443719413
GO#cell_killing	IEV#Proteasome_degradation_in_unidentified_cellular_location	0.9575856443719413
GO#establishment_of_localization	IEV#Proteasome_degradation_in_cytosol	0.9575856443719413
GO#humoral_immune_response	IEV#Response_to_stimulus	0.9575856443719413
GO#immune_effector_process	IEV#Translation	0.9575856443719413
GO#immune_response	IEV#Unknown_process	0.9575856443719413
GO#immune_system_process	IEV#Organism_event	0.9575856443719413
GO#immune_system_process	IEV#Treatment	0.9575856443719413
GO#leukocyte_mediated_cytotoxicity	IEV#Pathway	0.9575856443719413
GO#leukocyte_mediated_cytotoxicity	IEV#Temperature_treatment	0.9575856443719413
GO#leukocyte_mediated_immunity	IEV#Metabolic_pathway	0.9575856443719413
GO#leukocyte_mediated_immunity	IEV#Molecular_interaction	0.9575856443719413
GO#leukocyte_mediated_immunity	IEV#Plant_hormone_medication	0.9575856443719413
GO#lymphocyte_mediated_immunity	IEV#Protein_activation/inactivation_signaling	0.9575856443719413
GO#natural_killer_cell_mediated_immunity	IEV#Small_GTPase_activation_signaling	0.9575856443719413
GO#negative_regulation_of_signal_transduction	IEV#Regulation_of_Notch_signaling_pathway	0.9575856443719413
GO#production_of_molecular_mediator_of_immune_response	IEV#Low_temperature_treatment	0.9575856443719413
GO#production_of_molecular_mediator_of_immune_response	IEV#Signal_transduction_pathway	0.9575856443719413
GO#response_to_stimulus	IEV#Physiological_event	0.9575856443719413
GO#T_cell_mediated_immunity	IEV#Receptor_recycling	0.9575856443719413

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

CCMN - Bloco C - Cidade Universitária - Ilha do Fundão
Rio de Janeiro - RJ CEP: 21941-916
www.ppgi.ufrj.br