Unificando a Comparação e Busca de Fenótipos em *Model Organism Databases*

Luana Loubet Borges[1], **André Santanchè**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)

luanaloubet@gmail.com, santanche@ic.unicamp.br

Resumo. *Model Organism Databases (MODs) são largamente utilizados em pesquisas nas áreas médica e biológica. Como cada MOD é usualmente especializado em um tipo de organismo – e.g., peixe-zebra, rato, humano, camundongo – torna-se difícil a busca da mesma característica em organismos distintos para fins de correlação e comparação. Este trabalho apresenta um framework chamado Unified MOD Discovery Engine, cujo objetivo é permitir a correlação e busca de dados de vários MODs, a partir da unificação da sua representação dos dados. Este artigo apresenta o primeiro passo nesta direção, em que foram analisados e comparados os modelos de dados de dois MODs, o ZFIN (peixa-zebra) e MGI (camundongo), como base para a concepção de um modelo unificado. Tal modelo é a base de um grafo interligado, que permitirá ao usuário fazer buscas e comparações de forma unificada.*

1. Introdução e Motivação

Model Organism Databases (MODs) são repositórios específicos para conhecimento biológico [Hedges 2002], cuja definição não é estritamente estabelecida. Consideramos que cada MOD armazena dados sobre um *organismo modelo*, podendo conter seu genótipo e fenótipo, permitindo realizar pesquisas de conhecimento biológico, como genética, desenvolvimento e evolução. Nas últimas décadas o termo "*organismo modelo*" se referia a um pequeno e seleto grupo de espécies, estudadas profundamente em laboratório e ricamente documentadas [Hedges 2002]. Na medida em que os mecanismos para mapeamento genético se tornaram mais acessíveis, o conceito de organismo modelo se expandiu para um conjunto mais amplo de espécies [Hedges 2002].

A comparação de organismos modelo a partir dos seus fenótipos tem um grande potencial na análise e descoberta de correlações entre organismos e fornecerá uma forma eficiente, por exemplo, de identificar genes correlatos candidatos a causar doenças nos diversos modelos [Washington et al. 2009]. Fenótipo é um conjunto de características físicas e comportamentais de um indivíduo, resultante da interação do seu genótipo com o ambiente. Genótipo refere-se à composição genética do indivíduo. Para que esse cruzamento de dados seja possível entre MODs é preciso que eles estejam unificados. No entanto, organismos modelo não são registrados homogeneamente, tendo corriqueiramente, seus dados armazenados em forma de texto livre, além de não ter um modelo unificado, dificultando buscas e comparações automatizadas.

Outro conceito fundamental neste contexto são os *profiles*, que consistem em definir um foco das informações relevantes para realizar buscas, análises e analogia entre

organismos. No contexto de doenças, por exemplo, um *profile* pode ser composto por elementos de descrição do fenótipo da doença e seu genótipo associado. O *profile* torna-se a unidade de busca, isto é, a comparação é feita entre o *profile* buscado – e.g., olho ausente – e aquele recuperado da base de dados. Os fenótipos podem ser associados a ontologias no método Entidade-Qualidade (EQ) [Balhoff et al. 2010], em que a Entidade está contida em uma ontologia específica de organismos, associada a um termo de Qualidade usualmente da ontologia *Phenotype and Trait Ontology* (PATO) [Washington et al. 2009], e.g., *entidade* (olho) e *qualidade* (ausente).

O nosso trabalho visa contribuir neste contexto, através de um framework para unificar MODs heterogêneos e subsidiar a criação de *profiles* que propiciem a comparação de organismos. Ele parte da proposta de um modelo de organismo genérico – criado a partir da análise de modelos para a descrição de fenótipos – que contém dados relevantes para o pesquisador.

Este trabalho está organizado da seguinte maneira: a Seção 2 apresenta trabalhos relacionados; a Seção 3 descreve o modelo unificado; a Seção 4 apresenta como será feita a busca; a Seção 5 apresenta as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

[Washington et al. 2009] utilizaram vários MODs para realizar a integração de genótipos com seus respectivos fenótipos e descobrir genes ortólogos[1] que sofreram mutação em diferentes espécies, resultando em cegueira nos seus portadores. Para este estudo foi preciso gerar um modelo unificado de vários MODs heterogêneos contendo os genes que seriam considerados na comparação, foram escolhidos 11 genes humanos que possuem genes ortólogos em camundongos, peixe-zebra e drosófila, contidos no *Online Mendelian Inheritance in Man* (OMIM), além de genes de camundongos, peixes-zebra e drosófilas obtidos de bases diferentes.

[Washington et al. 2009] obtiveram os seguintes resultados: (i) alelos variantes contém fenótipos mais similares que os demais alelos do mesmo gene; (ii) é possível recuperar genes mutantes responsáveis por fenótipos anômalos a partir da análise de similaridade destes fenótipos; (iii) identificação de genes ortólogos pelo cruzamento de dados de fenótipos em diferentes espécies. Estes resultados não seriam obtidos se fosse feita a comparação apenas com o genótipo, pois esta abordagem apresenta dois problemas principais: (1) as bases genéticas de grande parte das doenças normalmente são desconhecidas; (2) ainda que a base genética seja conhecida, algoritmos de comparação de genes e/ou genótipos são feitos através do alinhamento de sequências; no caso de doenças ocorre uma mutação no gene causador da mesma, tornando tais algoritmos inadequados, pois essa comparação trata genes a partir da similaridade entre as cadeias. Por esta razão, a comparação é feita através dos fenótipos das doenças, neste caso, os sintomas da doença.

[Washington et al. 2009] enfrentaram duas grandes dificuldades: (1) tiveram que criar manualmente um modelo homogêneo de vários MODs utilizados apenas para o *profile* analisado; (2) criaram um *profile* a partir de várias ontologias, selecionando os termos relevantes para a pesquisa. Da mesma forma, vários pesquisadores enfrentam as mesmas dificuldades, tendo que integrar MODs e definir *profiles* manualmente, pois não existe

[1] genes derivados de um ancestral comum que possuem a mesma função em espécies diferentes

ferramenta computacional que construa um modelo unificado a partir de vários MODs distintos e que suporte profiles associados a ontologias.

Phenomicdb (`http://phenomicdb.info/`) é uma ferramenta que realiza a integração de vários MODs para pesquisas com fenótipos [Kahraman et al. 2005]. Comparado com a nossa proposta, a busca realizada é limitada a apenas uma descrição de um item de fenótipo. O diferencial do nosso trabalho é que ele suportará buscas por *profiles* com vários itens descritivos, utilizando diferentes formatos para a representação de fenótipos.

3. Modelo Unificado

Com o objetivo de sanar a dificuldade relatada na seção anterior, este trabalho propõe um framework para realizar a busca e comparação de *profiles* definidos pelo usuário em um conjunto de MODs de forma transparente. O ponto de partida foi analisar dois MODs de referência amplamente usados e citados em trabalhos relacionados – o ZFIN e o MGI – como bases para a proposta de um modelo unificado.

ZFIN é um MOD que contém tanto dados de genótipos quanto fenótipos do peixe-zebra, em que os fenótipos são descritos pelo método EQ citado anteriormente [Sprague et al. 2006, Washington et al. 2009]. O modelo parcial do banco de dados referente a fenótipos do ZFIN é apresentado na Figura 1(a). Uma descrição de fenótipo é formada por um conjunto de declarações (`Phenotype_statement`) envolvendo uma Entidade (`ZFA_term`) e uma Qualidade (`PATO_term`) ligadas a ontologias externas: ZFA (Zebrafish Anatomy Ontology), GO (*Gene Ontology*) e PATO. Entidades e qualidades são generalizadas como termos (`term`) que têm um auto-relacionamento com tipo (e.g., is-part-of), pois pode-se construir uma taxonomia de termos.

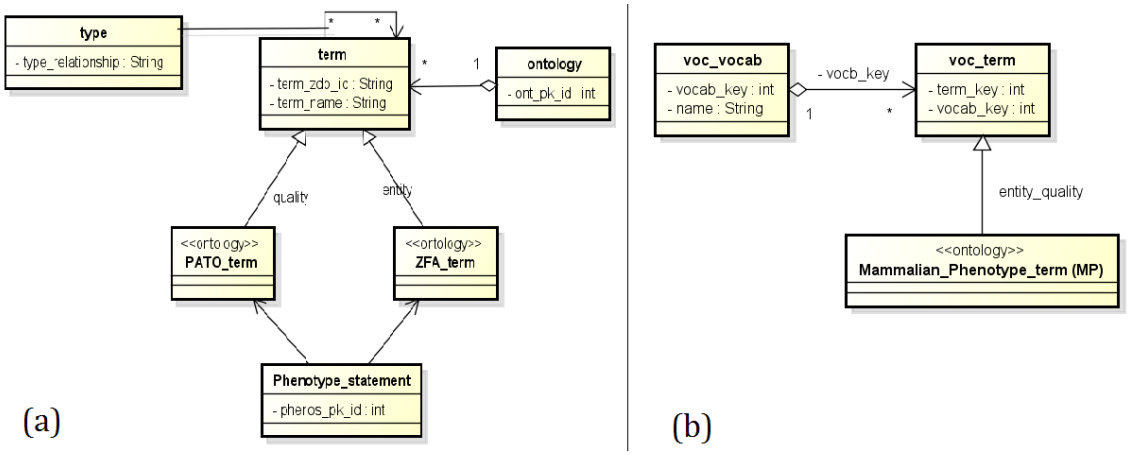

Figura 1. Modelo do banco de dados do ZFIN e do MGI.

MGI é um MOD com dados de genótipos e fenótipos de camundongos [Blake et al. 2003]. A Figura 1(b) retrata um modelo parcial do banco de dados de fenótipos do MGI. A descrição do fenótipo, assim como no ZFIN, é tratada como um conjunto de declarações. Cada declaração corresponde no MGI a um termo (`voc_term`). Cada termo é associado à ontologia *Mammalian Phenotype* que é uma variante da abordagem EQ, pois cada conceito da ontologia já é a composição da Entidade mais a Qualidade

[Smith et al. 2004]. A classe `voc_vocab` correspondente à classe `ontology` do modelo do ZFIN e possibilita o uso de termos de várias ontologias.

A Figura 2 apresenta o nosso modelo unificado, em que um fenótipo (`Phenotype`) é composto por um conjunto de declarações (`Statement`) que correspondem à composição de Entidades e Qualidades, como acontece no `voc_term` do MGI. A classe `Statement_EQ` especializa o `Statement` e é capaz de representar a entidade e a qualidade de forma discriminada como faz o ZFIN (classe `term`). A classe `voc_vocab` do MGI e `ontology` do ZFIN correspondem à classe `Ontology` no modelo proposto. Além disso, as classes `Statement`, `Entity` e `Quality` possuem um auto-relacionamento para registrar sinônimos. A classe `Profile` é formada por um `Phenotype`. Futuramente o `Profile` será integrado com informações de genótipos também.

Os modelos apresentados do ZFIN e do MGI refletem o banco de dados relacional original de ambos. Entretanto, nosso modelo unificado é baseado em uma estrutura de grafos e por isso mapearemos os modelos para um banco de dados de grafos de propriedades [Robinson et al. 2013] fazendo com que cada classe vire um nó, os relacionamentos serão arestas e os atributos das classes viram propriedades dos nós e/ou arestas. O mesmo acontece com o modelo proposto neste trabalho.

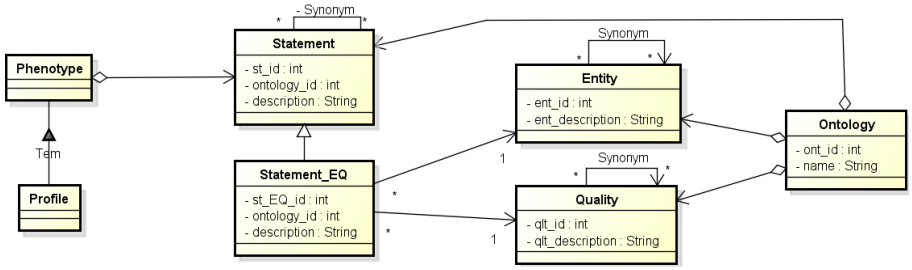

Figura 2. Modelo proposto para a ferramenta *Unified MOD Discovery Engine*.

4. Busca baseada em Profile

Esta seção descreve a arquitetura que projetamos para a realização de uma busca unificando diferentes MODs, em que há um esforço extra para tratar a representação heterogênea dos dados de cada base, já que eles não são homogêneos. Descrições de fenótipos podem ser encontradas em formatos distintos, como textos livres (o que dificulta o uso computacional), C/CS (que é uma forma de descrição semi-estruturada), Entidade-Qualidade (EQ) e uma variante dele que chamaremos de *EQ composto* (tal como no MGI). Como exemplo das formas de descrições, temos que no OMIM as descrições são em texto livre, no MGI são em EQ *composto* e no ZFIN são em EQ.

O nosso sistema propõe a unificação da busca e comparação em MODs distintos. A busca/comparação é feita a partir de uma interface unificada, que fornecerá uma visão homogênea das informações, independentemente de como elas estão armazenadas nos seus MODs de origem.

Tomando o caso descrito por [Washington et al. 2009] como base de pesquisa em vários MODs, apresentaremos a nossa arquitetura através de um exemplo de uma consulta feita no ZFIN e MGI. Ao fazer uma busca no ZFIN pelo fenótipo *lens decreased size* são

retornados vários genes associados a esse fenótipo, entre eles, o gene Pax6b. Esse fenótipo é descrito por meio de sua entidade (*lens*) separada de sua qualidade (*decreased size*).

Ao realizar a mesma busca pelo fenótipo *lens decreased size* no MGI são retornados vários genes, entre eles o gene pax6 que causa microftalmia, que refere-se ao olho pequeno. Mas a interpretação não é tão trivial pois o sistema não retorna o fenótipo exatamente como ele foi buscado. O fenótipo microftalmia tem o sinônimo *lens decreased size* que foi buscado anteriormente. Essas descrições de fenótipos no MGI estão em EQ *composto*.

Ao interligar essas informações do ZFIN e MGI obtemos os genes que causam doenças que levam a cegueira no zebrafish e no camundongo. Essas informações são úteis para realizar pesquisas sobre essa doença também em humanos, já que o gene causador da cegueira em humanos é o PAX6 ortólogo aos genes do peixe-zebra e camundongo.

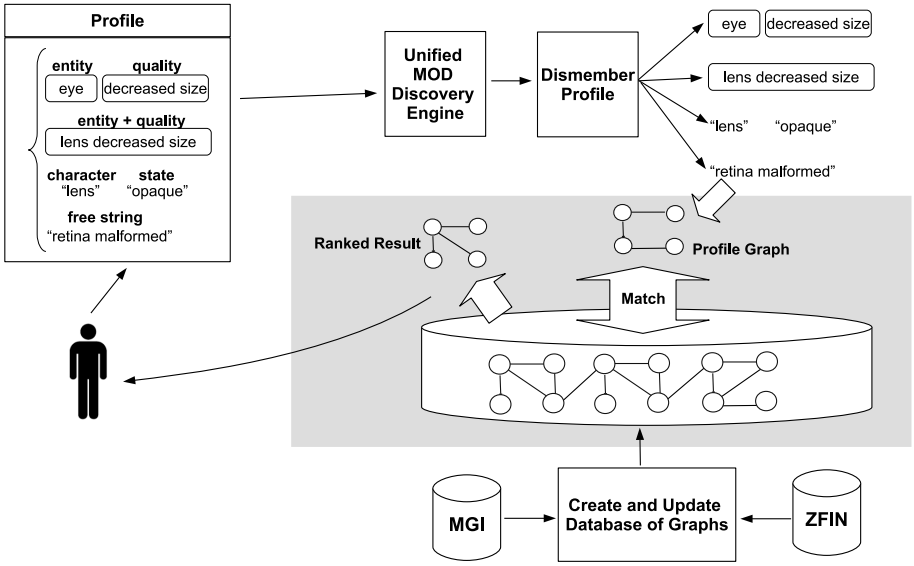

Figura 3. Arquitetura da nossa proposta.

A Figura 3 representa a nossa proposta. O usuário interagirá com a ferramenta na criação do *profile* que é dado como entrada. Neste caso, cada linha corresponde a uma descrição de fenótipo dada pelo usuário, podendo ser em texto livre, EQ, entre outras. Em seguida, a nossa ferramenta terá acesso a um banco de dados de grafos criado previamente que importa as informações contidas no ZFIN e MGI referentes a fenótipo. O nosso framework *Discovery Engine* executará algoritmos de *match* para comparar e analisar profiles. Para tornar possível essa comparação é necessário desmembrar o *profile* em unidades básicas que descrevem o fenótipo (*dismember profile* na Figura 3). Sobre estes itens serão aplicados algoritmos para análise de similaridade para busca e comparação de *profiles*. Como resultado da busca, a ferramenta gera um grafo contendo resultados com informações do ZFIN e MGI ranqueadas por similaridade. O *Profile Graph* da Figura 3 corresponde à representação do profile na forma de grafo, a ser confrontado com as descrições de fenótipos em banco de dados de grafos. Além de importar dados do ZFIN e MGI o banco de dados de grafos também será usado para interligá-las e melhorar o resultado das comparações.

Para realizar a busca no banco de dados através do *profile* utilizaremos métricas

de similaridade também usadas por [Washington et al. 2009]: *Information Content* (IC), métricas semânticas de similaridade e análise de sobreposição [Mistry and Pavlidis 2008].

5. Conclusões

Pesquisadores precisam cruzar dados de vários organismos e recorrem a diversos MODs, contendo diferentes representações de dados, dificultando a interligação dos mesmos. Neste trabalho nós apresentamos um modelo unificado para representação de fenótipos – baseado na análise de dois MODs, o ZFIN e o MGI – bem como o projeto do framework *Unified MOD Discovery Engine*, que permitirá ao usuário realizar buscas por descrições de profiles de organismos em MODs distintos de forma unificada.

Como trabalhos futuros pretendemos implementar o *engine* cujo projeto foi apresentado neste artigo e estender a proposta para outros MODs, como OMIM (humanos), RGD (ratos), Flybase (moscas), entre outros. Além de integrar informações de genótipos que ainda não estão sendo consideradas.

Agradecimentos. Este trabalho foi parcialmente financiado pela Capes 01P-3501-2014), FAPESP/Cepid em Engenharia e Ciência da Computação (2013/08293-7), o Instituto Microsoft Research FAPESP Virtual (NavScales project), CNPq (MuZOO Project), FAPESP-PRONEX (eScience project), , INCT em Web Science e subvenções individuais do CNPq.

Referências

Balhoff, J. P., Dahdul, W. M., Kothari, C. R., Lapp, H., Lundberg, J. G., Mabee, P., Midford, P. E., Westerfield, M., and Vision, T. J. (2010). Phenex: ontological annotation of phenotypic diversity. *PLoS One*, 5(5):e10500.

Blake, J. A., Richardson, J. E., Bult, C. J., Kadin, J. A., Eppig, J. T., Group, M. G. D., et al. (2003). Mgd: the mouse genome database. *Nucleic acids research*, 31(1):193–195.

Hedges, S. B. (2002). The origin and evolution of model organisms. *Nature Reviews Genetics*, 3(11):838–849.

Kahraman, A., Avramov, A., Nashev, L. G., Popov, D., Ternes, R., Pohlenz, H.-D., and Weiss, B. (2005). Phenomicdb: a multi-species genotype/phenotype database for comparative phenomics. *Bioinformatics*, 21(3):418–420.

Mistry, M. and Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity. *BMC bioinformatics*, 9(1):327.

Robinson, I., Webber, J., and Eifrem, E. (2013). *Graph databases*. O'Reilly.

Smith, C. L., Goldsmith, C.-A. W., and Eppig, J. T. (2004). The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology*, 6(1):R7.

Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D. G., Mani, P., Ramachandran, S., et al. (2006). The zebrafish information network: the zebrafish model organism database. *Nucleic acids research*, 34(suppl 1):D581–D585.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., and Lewis, S. E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS biology*, 7(11):e1000247.