# Enriquecimento de Dados de Proveniência de Análises Filogenéticas com Dados do NCBI: uma Abordagem Prática *

**Lucas S. Tito**[1], **Kary A. C. S. Ocaña**[2], **Daniel de Oliveira**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF) – Niterói – RJ – Brasil

[2]Laboratório Nacional de Computação Científica (LNCC) Petrópolis – RJ – Brasil

ltito@id.uff.br, karyann@lncc.br, danielcmo@ic.uff.br

***Abstract.*** *This paper proposes an approach called BioIntegrator, that aims at integrating and enriching provenance databases from phylogenetic analyzes using metadata present in external sources. Such approach aims at providing more analytical skills to scientists in their daily duties. Although it is a work in progress, the proposed approach has a clear potential regarding the analysis and evaluation of the results of experiments.*

***Resumo.*** *Esse artigo apresenta uma proposta de abordagem, chamada BioIntegrator, para integração e enriquecimento de bases de dados de proveniência de análises filogenéticas com metadados presentes em fontes externas. Tal abordagem visa oferecer maior capacidade analítica aos cientistas em suas tarefas diárias. Apesar de ser um trabalho em andamento, a abordagem proposta tem um potencial claro no que tange a análise e validação de resultados dos experimentos.*

## 1. Introdução

O volume de dados genômicos passível de análise pela comunidade científica cresce em um ritmo acelerado, devido às recentes tecnologias tanto na área biológica quanto na computação. *e.g.*, as tecnologias de sequenciamento de nova geração (SNG) e processamento de alto desempenho (PAD). Uma das áreas da bioinformática que mais se beneficia dessas tecnologias é a análise evolutiva filogenética. Essa área de pesquisa tem como objetivo gerar conhecimento sobre processos evolutivos ou relações filogenéticas entre espécies. Experimentos filogenéticos recebem como entrada sequências ou até mesmo genomas inteiros, produzindo árvores e diversas estatísticas utilizadas para inferir a história evolutiva ou a filodinâmica e a filogeografia de processos infecciosos entre espécies [Felsenstein 1996] [Ocaña et al. 2011]. Diversos experimentos filogenéticos já foram propostos na literatura, muitos deles modelados como *workflows* científicos e executados em Sistemas de *Workflows* (SWf) em ambientes de PAD, como o SciPhy [Ocaña et al. 2011]. Os *workflows* geram dados de proveniência que representam o conjunto de informações relacionadas à execução de um experimento [Freire et al. 2008]. Tais informações auxiliam os pesquisadores a relacionar a sequência das etapas do experimento, interpretar resultados, estudar a derivação dos dados envolvidos e *etc.*. Apesar de representarem um fator fundamental nos experimentos, as bases de proveniência isoladas, nem sempre fornecem todo o conhecimento necessário para os cientistas analisarem os resultados das suas pesquisas. O poder analítico de dados de proveniência depende muito se os mesmos se encontram associados a dados de domínio [de Oliveira et al. 2015]. Por exemplo, no caso do SciPhy, o *workflow* é composto por programas que alinham sequências, além de gerar árvores filogenéticas. Cada uma dessas sequências representa um gene/genoma de um organismo de interesse, e requere-se que as

*Os autores agradecem à CAPES, CNPq e FAPERJ por financiarem parcialmente este trabalho

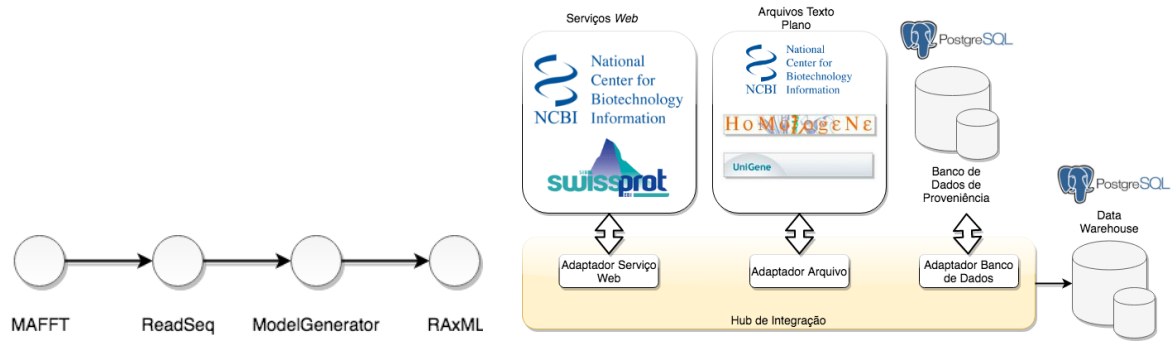


**Figura 1. (a) O *Workflow* SciPhy (b) Arquitetura da Abordagem Proposta**

informações e metadados sobre tal organismo estejam integrados, para produzir e analisar resultados respaldados por informações científicas relevantes. Muitas vezes, esses dados de domínio necessários encontram-se desassociados dos dados de proveniência do *workflow*, o que requer a manipulação do especialista, que é laboriosa, podendo levar a erros. Enriquecer uma base de proveniência com dados de domínio não é uma tarefa simples, mas é uma abordagem usada atualmente de maneira manual pelos cientistas (biólogos e geneticistas). Abordagens existentes já propuseram essa integração [de Oliveira et al. 2017], porém ou elas assumem que os dados de domínio a serem integrados são definidos *a priori* ou assumem que o acesso é sempre realizado por *Web Services* e nem sempre essas abordagens são possíveis. Além disso, a granularidade da informação pode ser específica para uma área (determinar uma doença genética) que pode ser necessário a manipulação de diferentes dados o que leva a integrar muitos bancos de dados. Este artigo propõe uma abordagem, chamada $BioIntegrator$, que visa a integração entre dados de domínio de diferentes fontes e dados de proveniência gerados por experimentos científicos modelados como *workflows*. Tal abordagem pode ser executada *a priori* ou *a posteriori*, dependendo da necessidade do cientista. Além disso, ela é adaptativa no que tange o acesso aos dados, podendo ser via *Web Services*, programas próprios, extratores de dados, *etc.*. Atualmente, essa integração é focada para análises filogenéticas, mas a mesma tecnologia pode ser extrapolada para outras áreas biológicas.

## 2. Motivação: o *Workflow* SciPhy

O SciPhy é um *workflow* de bioinformática, que gerencia de forma distribuída e paralela, sequências genéticas e constrói árvores filogenéticas evolutivas entre organismos [Ocaña et al. 2011]. O SciPhy (Figura 1(a)) é composto de quatro atividades: (I) alinhamento de sequências (MAFFT), (II) conversão de alinhamento (ReadSeq), (III) eleição do modelo evolutivo (ModelGenerator) e (IV) geração de árvores (RAxML).

Sendo assim, o SciPhy pode ser executado para múltiplos objetivos, como por exemplo comparar diversas árvores de parasitas, identificar drogas que sejam efetivas contra eles ou realizar estudos de filogeografia e filodinâmica para estudar a propagação entre continentes (*e.g.* Ebola e Zika). O SciPhy foi implementado no SciCumulus [de Oliveira et al. 2010]. A base de proveniência do SciCumulus contém informações do *Workflow*, suas atividades, ativações (execuções de atividades), arquivos produzidos e parâmetros consumidos [de Oliveira et al. 2017]. Entretanto, os dados de domínio não se encontram integrados de forma natural a essa base, pois depende do cientista a escolha de quais bases de dados serão as mais informativas dependendo da pesquisa. Desta forma, a proposta consiste na importação de tais informações para a base de proveniência para enriquecê-la e tornar as análises dos cientistas mais completas. A seguir, apresentamos

a proposta de um arcabouço genérico para enriquecimento de bases de proveniência de *workflows* filogenéticos, que pode ser utilizado por diferentes SWfs.

## 3. Abordagem Proposta: $BioIntegrator$

De acordo com o que foi apresentado anteriormente, podemos perceber que o cientista necessita de um acesso integrado a múltiplas fontes de dados, desde bancos de dados tradicionais (bancos de dados de proveniência) até dados semiestruturados ou não estruturados, como domínio biológico. Nesse sentido, uma abordagem de enriquecimento de bancos de dados de proveniência visa oferecer um acesso uniforme a fontes de dados distribuídas e heterogêneas. Essa abordagem pode se basear em duas diferentes arquiteturas: (I) Abordagem virtual [Chawathe et al. ], onde os dados de proveniência e de domínio permanecem isolados e são integrados via consultas, e (II) Abordagem materializada, onde todos os dados são acessados, limpos, integrados e armazenados em um *Data Warehouse* [Widom 1995] e as consultas analíticas são submetidas ao mesmo *Data Warehouse*.

A Figura 1(b) apresenta a arquitetura da abordagem proposta composta de quatro componentes principais: (I) Fontes externas de dados que podem ser serviços *Web*, bancos de dados ou arquivos, (II) Base de dados de proveniência, que contém todo o histórico de execução do *workflow*, (III) Hub de integração, que contém componentes adaptadores que extraem informações das mais diversas fontes e as integram no (IV) *Data Warehouse* de proveniência que contém dados históricos e de domínio. Nesse contexto, diferentes fontes de dados externas podem ser importadas para o banco de dados de proveniência para o enriquecimento de informações relevantes durante a análise dos dados. *A priori*, utilizaremos três diferentes fontes de dados em nossa abordagem: (I) a base de dados de proveniência do SciPhy no SciCumulus, (II) o NCBI *Taxonomy database*[1], e (III) o *HomoloGene database*[2]. O NCBI *Taxonomy Database* é uma base de dados de informações taxonômicas e filogenéticas, entre outras fontes [Federhen 2011]. A base *HomoloGene* contém informações do gene, proteína, etc. e, juntamente com o NCBI *Taxonomy Database*, pode ser uma fonte rica de informações para a base de proveniência. Para integrar tais bases, optou-se por utilizar a abordagem *Global-As-View* (GAV) [Halevy 2000] que requer que cada objeto do esquema global (o *Data Warehouse*) seja expressado como uma visão de banco de dados a partir das fontes externas. Apesar de ser um trabalho em andamento, a integração foi realizada com sucesso e a abordagem está em processo de validação e mostra-se promissora no que tange oferecer capacidade analítica aos cientistas.

Para exemplificar o apoio oferecido pela abordagem aos cientistas em suas análise e levando em conta as bases de dados supramencionadas, uma possível consulta que seria facilitada é: quais são as categorias, proteínas e nucleotídeos de sequências homólogas, cuja árvore filogenética foi gerada pelo *Sciphy*? Sem a abordagem proposta, os cientistas deveriam buscar no *Homologeny Database* sequências homólogas (que não se encontram alinhadas), as usaria como *input* para o *Sciphy* que por sua vez as alinharia e geraria uma árvore filogenética, e para cada sequência os pesquisadores em questão buscariam no *Taxonomy Database* as categorias, as proteínas e os nucleotídeos das sequências contidas na árvore filogenética já alinhadas. Como mencionado, estas etapas manuais são trabalhosas e podem levar a erros.

## 4. Trabalhos Relacionados

O problema de integração entre bases de dados biológicas não é novo [Thiam Yui et al. 2011]. Em [Thiam Yui et al. 2011], três soluções são apresentadas: um banco de dados federado, a

[1]https://www.ncbi.nlm.nih.gov/taxonomy

[2]https://www.ncbi.nlm.nih.gov/homologene

abordagem de *data warehousing* (utilizada nesse artigo) e a abordagem baseada em links. [Hernandez and Kambhampati 2004] também discutem diversas abordagens de integração e definem que *data warehousing* é a que oferece mais vantagens.

## 5. Discussões e Trabalhos Futuros

Neste artigo apresentamos a proposta do $BioIntegrator$, uma abordagem para integração de dados de proveniência com dados de domínio que podem ser armazenados e consultados de diferentes maneiras. O objetivo do $BioIntegrator$ é integrar diferentes bases de dados científicas de domínio com bases de dados de proveniência de SWf, para que juntas aumentem e facilitem a extração do conhecimento de uma determinada área científica. Inicialmente foram incorporadas as bases do NCBI *Taxonomy Database* e do *HomoloGene*, porém planejamos integrar outras bases de dados como o UniProt *database*[3] e de vias metabólicas como o KEGG textitdatabase[4]. Planejamos também exportar o *Data Warehouse* para um banco de dados noSQL de forma a aumentar a escalabilidade da abordagem proposta e desta maneira aprimorar a arquitetura de integração apresentada na Seção 3.

## Referências


Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., and Widom, J. The tsimmis project: Integration of heterogeneous information sources.

de Oliveira, D., Ogasawara, E., Baião, F., and Mattoso, M. (2010). Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows. In *2010 IEEE CLOUD*, pages 378–385.

de Oliveira, D., Silva, V., and Mattoso, M. (2015). How much domain data should be in provenance databases? In *TaPP 15*, Scotland.

de Oliveira, W. M., Ocaña, K. A. C. S., de Oliveira, D., and Braganholo, V. (2017). Querying provenance along with external domain data using prolog. *JIDM*, 8(1):3–18.

Federhen, S. (2011). The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143.

Felsenstein, J. (1996). [24] inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. In *M. in enzym.*, volume 266, pages 418–427.

Freire, J., Koop, D., Santos, E., and Silva, C. T. (2008). Provenance for computational tasks: A survey. *Computing in Science and Engg.*, 10(3):11–21.

Halevy, A. Y. (2000). Theory of answering queries using views. *SIGMOD Rec.*, 29(4):40–47.

Hernandez, T. and Kambhampati, S. (2004). Integration of biological sources: Current systems and challenges ahead. *SIGMOD Rec.*, 33(3):51–60.

Ocaña, K. A. C. S., de Oliveira, D., Ogasawara, E., Dávila, A. M. R., Lima, A. A. B., and Mattoso, M. (2011). Sciphy: A cloud-based workflow for phylogenetic analysis of drug targets in protozoan genomes. In *BSB*, pages 66–70.

Thiam Yui, C., Liang, L. J., Jik Soon, W., and Husain, W. (2011). A survey on data integration in bioinformatics. In *Inf. Eng. and Inf. Sci.*, pages 16–28.

Widom, J. (1995). Research problems in data warehousing. In *CIKM'95*, CIKM '95, pages 25–30, New York, NY, USA. ACM.


[3] http://www.uniprot.org/statistics/Swiss-Prot

[4] http://www.genome.jp/kegg/pathway.html