Gerência de Incerteza em Bancos de Dados de Proveniência de *Workflows* de Bioinformática*

Gustavo Tallarida[1], Kary Ocaña[2], Aline Paes[1], Vanessa Braganholo[1], Daniel de Oliveira[1]

[1]Instituto de Computação – Universidade Federal Fluminense (IC/UFF)
[2]Laboratório Nacional de Computação Científica (LNCC)

gustavotallarida@id.uff.br,
{alinepaes, vanessa, danielcmo}@ic.uff.br, karyann@lncc.br

Resumo. *Bancos de dados de proveniência de experimentos científicos desempenham um papel fundamental na ciência. Os modelos utilizados para representar esses dados assumem que existe uma certeza nos relacionamentos de proveniência. Entretanto, diversos experimentos não são determinísticos e seus resultados estão associados a incertezas. Realizar a análise dos dados de proveniência com tais incertezas não é trivial. Nesse artigo é proposta uma abordagem para gerência de incertezas em dados de proveniência baseada em um componente extrator que armazena os dados de proveniência e a incerteza associada em um banco de dados probabilístico. Experimentos mostraram um overhead aceitável da abordagem de cerca de 3% no tempo total de execução do workflow e 16% no tempo de processamento da consulta.*

Abstract. *Provenance databases play an essential role in scientific experiments. The models considered to represent such data assume that there is a certainty in all the provenance relations. However, several experiments are not deterministic, which makes their results to be associated with uncertainties. Analyze provenance data in the presence of such uncertainties is not trivial. In this paper, we address the management of non-deterministic provenance data by relying on an extractor component that stores both provenance data and its corresponding uncertainty values in a probabilistic database. Experiments show an acceptable overhead of 3% in the workflow runtime and 16% in the time spent to process a query.*

1. Introdução

Os *workflows* científicos são abstrações capazes de modelar o encadeamento de programas que fazem parte de um experimento. Esses *workflows* possuem dados de proveniência associados, que têm se tornado cada vez mais relevantes tanto na questão da reprodutibilidade quanto na depuração dos experimentos [Mattoso *et al*. 2010]. Os dados de proveniência podem ser capturados por sistemas especializados como o Karma [Simmhan *et al*. 2008] ou por Sistemas de Gerência de *Workflows* Científicos (SGWfC).

[1] Os autores gostariam de agradecer ao CNPq e a FAPERJ por financiarem esse trabalho

Dados de proveniência podem ser caracterizados como proveniência prospectiva (*p-prov*), associada à especificação do *workflow* e proveniência retrospectiva (*r-prov*), associada à execução de um determinado *workflow* [Freire *et al*. 2008]. Artigos recentes [De Oliveira *et al*. 2015; Gonçalves *et al*. 2012] defendem que Bancos de Dados (BD) de proveniência, além de armazenarem *p-prov* e *r-prov*, armazenem também dados de domínio, de modo a prover informações do experimento e informações específicas de domínio que em conjunto poderão levar a novas descobertas em e-Ciência. BDs de proveniência seguem modelos de proveniência padrão como o PROV do W3C [Moreau e Missier 2013] ou as extensões ProvONE e PROV-Wf. Cabe ressaltar que o PROV-Wf [Costa *et al*. 2013] também é capaz de representar os dados de domínio.

Muitos dos BDs de proveniência de SGWfCs que armazenam *p-prov*, *r-prov* e dados de domínio assumem que os relacionamentos de proveniência são absolutamente certos, ou seja, são determinísticos. Porém, em alguns casos essa premissa não se sustenta, já que existem *workflows* com atividades não determinísticas (*i.e.,* nem sempre produzem o mesmo dado de saída) como é o caso de experimentos científicos de áreas na bioinformática. Nesse artigo vamos considerar o *workflow* de filogenia denominado SciPhy [Ocaña et al. 2011], que será usado como exemplo ao longo do texto.

O SciPhy tem como objetivo gerar árvores filogenéticas a partir de uma série de sequências de DNA e RNA e é composto por cinco atividades. Uma das atividades principais é o alinhamento múltiplo de sequências (AMS) que compara sequências biológicas para identificar regiões similares, que possam ser consequência de um mesmo evento de evolução. O AMS é um pré-requisito para atividades subsequentes como a escolha do modelo evolutivo e geração da árvore filogenética. Porém os métodos de AMS não são determinísticos dado que, na biologia de sistemas, a história evolutiva de um conjunto de sequências biológicas pode ser representada por um ou vários alinhamentos. Assim, é importante que consideremos a incerteza associada ao alinhamento, porque a análise dos resultados do SciPhy é dependente do grau de incerteza de cada alinhamento produzido. Uma das métricas existentes para se avaliar a incerteza de um alinhamento é o *SP-Score* [Ahola *et al*. 2008], que fornece uma probabilidade para cada alinhamento e representa o grau de incerteza do mesmo. Entretanto, nos SGWfCs existentes não é possível representar tal incerteza pois os BDs de proveniência não representam esse tipo de informação.

O objetivo desse artigo é introduzir uma abordagem para a gerência de dados de proveniência que considere incertezas, chamada de *Probabilistic Provenance Analyzer* (PPA). O PPA é baseado em um componente extrator que armazena os dados de proveniência e sua incerteza associada em BDs probabilísticos (BDPs). O componente do PPA pode ser acoplado a um SGWfC de forma a enriquecer seu BD de proveniência. Nos experimentos apresentados nesse artigo, utilizamos o SGWfC SciCumulus e o *workflow* SciPhy como estudo de caso para comprovar a viabilidade da proposta.

O artigo está organizado em quatro seções, além da introdução. A Seção 2 apresenta o referencial teórico acerca de bancos de dados probabilísticos e os trabalhos relacionados. A Seção 3 apresenta a abordagem proposta. A Seção 4 apresenta a avaliação experimental, e, finalmente, a Seção 5 conclui esse artigo.

2. Bancos de Dados Probabilísticos e Trabalhos Relacionados

Um Banco de Dados Probabilístico (BDP) é capaz de armazenar dados que apresentam um grau de incerteza associado [Re e Suciu 2007]. Ele permite associar a cada atributo ou tupla uma probabilidade $\in (0,1]$, onde 0 representa dados que são incorretos e 1 os corretos. Dessa forma, para valores entre 0 e 1, não podemos afirmar com certeza se um dado (valor do atributo ou tupla) é correto ou não.

Em um BDP não existe apenas uma instância do BD, e sim diversas instâncias possíveis, cada uma com uma probabilidade associada. Esse tipo de BD é baseado na semântica dos mundos possíveis [Re and Suciu 2007]. Cada subconjunto de tuplas de um BDP representa um mundo possível com uma probabilidade associada. A soma das probabilidades de todos os mundos possíveis é sempre igual a 1. Assim, o BDP é uma distribuição de probabilidade sobre um conjunto de mundos possíveis e segue um esquema relacional $R = \langle R_1, ..., R_k \rangle$ que consiste de k relações, onde cada relação está associada a uma aridade $r_j \geq 0$. Um mundo possível w é definido por $w = \langle R_1^w, ..., R_k^w \rangle$ onde R_j^w é uma relação de aridade r_j sobre um universo U e o conjunto de mundos possíveis $W = (\langle R_1^1, ..., R_k^1, p^{[1]} \rangle, ..., \langle R_1^n, ..., R_k^n, p^{[n]} \rangle)$ de relações $R_1, ..., R_k$ e $0 < p^{[i]} \leq 1$ de forma que $\sum_{w \in W} p^{[w]} = 1$. Cada $w \in W$ é um mundo possível e $p^{[i]}$ sua probabilidade associada. Chamamos uma relação de completa ou certa se todas as suas instanciações são as mesmas em todos os mundos possíveis.

As operações da álgebra relacional (seleção, projeção, produto cartesiano, união, diferença e renomeação) podem ser aplicadas em cada mundo possível de forma independente. Existem diversos sistemas de gerência de bancos de dados probabilísticos (SGBDP) como o MystiQ [Boulos *et al*. 2005] e o MayBMS [Huang *et al*. 2009]. Na solução apresentada nesse artigo utilizamos o SGBDP MayBMS que apresenta para cada tabela e tupla armazenada, os valores dos atributos e sua probabilidade associada. Dessa forma, essa representação se torna atrativa para armazenar dados de proveniência e de domínio que possuem um grau de incerteza associada, cenário típico na bioinformática que foi eleita como caso de estudo no presente artigo.

A gerência de incerteza em BDs de proveniência tem sido pouco explorada em pesquisas na literatura. Idika *et al*. [2013] discutem um modelo de proveniência genérico que considera relacionamentos com incerteza envolvida. Entretanto, o modelo apresentado não é instanciado em um BD real e nem aplicado a experimentos científicos. Chapman *et al*. [2010] discutem como adicionar um nível de confiança aos dados de proveniência para realizar melhores inferências. Apesar de considerar níveis de incerteza, Chapman *et al*. não utilizam BDPs, pois se baseiam em uma representação direta do *Open Provenance Model* [Moreau *et al*. 2011].

3. Abordagem Proposta: *Probabilistic Provenance Analyzer*

Com o objetivo de fornecer uma solução para o problema da gerência de incertezas em BDs de proveniência, propomos o uso do *Probabilistic Provenance Analyzer* (PPA) para extração e cálculo da probabilidade associada a cada dado consumido ou produzido pelo *workflow*. O PPA é uma atividade artificial que pode ser inserida no *workflow*. A execução de um PPA tem duas fases atribuídas a dois componentes principais (Figura 1): (i) o *extrator de dados*, encarregado da extração da proveniência e dos dados de domínio (o componente acessa cada arquivo produzido pela atividade e extrai valores

determinados pelo cientista) e (ii) a função de probabilidade *P: D → (0,1]*, que calcula a probabilidade associada para cada conjunto de dados de domínio no BD de proveniência (formula fornecida pelo cientista, que varia dependendo do experimento). Um PPA é incluído preferencialmente depois de uma atividade de um *workflow* (mas pode ser incluído em qualquer posição do *workflow*) de forma que possa extrair dados de domínio e calcular a probabilidade associada aos mesmos.

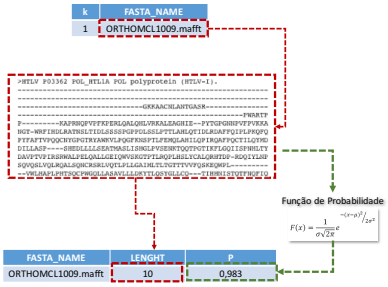

Figura 1 Execução do PPA

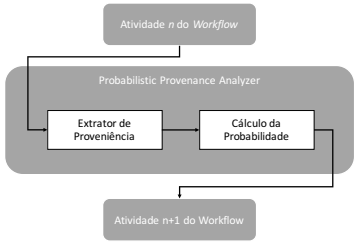

Figura 2 Implementação do SCCPPA

A proposta dos PPA foi concebida para SGWfCs que seguem uma álgebra de workflows como a proposta por Ogasawara et al. [2011]. Nessa álgebra, todos os dados consumidos e produzidos pelos workflows são representados como relações (cada atributo da relação é um parâmetro consumido/produzido pela atividade do workflow) e cada atividade é regida por operadores básicos de manipulação de dados. Cada um dos operadores (*Map*, *Reduce*, etc.) possui relações como um dos seus operandos. A definição do workflow científico é então mapeada para um conjunto de expressões algébricas. A abordagem PPA foi implementada no SGWfC SciCumulus e demandou duas modificações na versão do SciCumulus original: (i) a inserção de um componente que implementa o PPA (Figura 2) para extração de dados de proveniência e cálculo de probabilidades e (ii) a extensão do esquema do BD de proveniência do SciCumulus para considerar incertezas. Em relação ao esquema do BD, cada tabela do esquema original (disponível em https://s3.amazonaws.com/jidm/relational-database-schema-SCC.png) foi estendida para que cada tupla de cada relação associada aos dados de domínio extraídos (i.e., dlmafft, dlreadseq, dlmodelgenerator e dlraxml1) possuísse uma probabilidade associada. O componente foi implementado em Java 7 e a seleção dinâmica da função de probabilidade é baseada no padrão de projeto Strategy.

4. Avaliação Experimental

Nesta seção, apresentamos uma avaliação inicial da abordagem proposta por meio de consultas executadas no BD de proveniência relacional e no BDP de proveniência. Para a consulta escolhida avaliamos o resultado obtido, o *overhead* imposto na extração e cálculo das probabilidades, e o processamento da consulta propriamente dito. No experimento utilizamos a base de proveniência do *workflow* SciPhy executado no SGWfC SciCumulus. O arquivo XML do SciCumulus que define o *workflow* SciPhy foi instrumentado manualmente para inserir os PPAs necessários ao longo do *workflow* (após as atividades AMS e ModelGenerator). O BD de proveniência do SciPhy contém dados de uma execução com um conjunto de 200 arquivos multifasta de sequências de proteínas extraídas do banco biológico RefSeq e cada arquivo é constituído em média por 10 sequências biológicas (*length*). O BD de proveniência contém 1.000 tuplas de

execuções de atividades, 1.000 tuplas de dados de domínio, e mais de 19.000 tuplas representando arquivos produzidos.

A seguinte consulta (*baseline*) retorna os arquivos utilizados como entrada (arquivos multifasta), o modelo evolutivo escolhido pela atividade ModelGenerator a partir de um alinhamento (programa Mafft), a incerteza associada ao alinhamento (valor *SP-Score*, atributo *predictedsp*), e a probabilidade do modelo evolutivo (atributo *prob1*). A consulta no PostgreSQL (puramente relacional) retorna somente os valores dos atributos (servindo como *baseline*), entretanto o MayBMS (probabilístico) retorna a probabilidade do modelo evolutivo final calculada com a operação *conf()*. As consultas para as versões do BD de proveniência, PostgreSQL (puramente relacional) e MayBMS (probabilístico) foram executadas 10 vezes em um computador com processador Intel Core i5 1.6ghz, 8GB de RAM e 1TB de disco rodando MacOS *el Capitan*.

```
SELECT dlmafft.fasta_name, dlmafft.length, dlmafft.predictedsp, dlmg.model1, dlmg.prob1
FROM dlmafft dlm INNER JOIN eactivation taskmafft ON (dlm.taskid = taskmafft.taskid)
INNER JOIN eactivity actmafft ON (taskmafft.actid = actmafft.actid) INNER JOIN cactivity
cactmafft ON (actmafft.cactid = cactmafft.actid) INNER JOIN cmapping mapmafftmg ON
(mapmaffmg.previousid = cactmafft.actid) INNER JOIN cactivity cactmg ON
(mapmafftmg.nextid = cactmg.actid) INNER JOIN eactivity actmg ON (cactmg.actid =
actmg.cactid) INNER JOIN eactivation taskmg ON (taskmg.actid = actmg.actid) INNER JOIN
dlmodelgenerator dlmg ON (dlmg.taskid = taskmg.taskid)
```

Para realizar a análise do tempo de processamento da consulta, a base de dados foi replicada 100 vezes (base artificial). A Tabela 1 apresenta um fragmento do resultado com os dados de domínio extraídos e suas probabilidades no MayBMS. A Figura 3 apresenta a média do tempo de execução das consultas no PostgreSQL (*baseline*) e no MayBMS. Como observado na Figura 3, a consulta no MayBMS demora cerca de 16,3% a mais que a consulta *baseline* no PostgreSQL. Porém conforme apresentado na Tabela 1, mesmo com o tempo de execução maior, o MayBMS oferece resultados mais ricos para o cientista, pois apresentam dados de domínio específicos do experimento da área de aplicação e seu grau de incerteza associado que podem levar a descobertas biológicas interessantes. Assim, acreditamos que o *overhead* imposto nesse caso é aceitável.

Tabela 1 Resultado da consulta

fasta_name	length	model	P
ORTHOMCL 1000	10	RtREV	0,916
ORTHOMCL 1002	10	Blosum62	0,877
ORTHOMCL 1003	10	RtREV	0,884
ORTHOMCL 1005	10	Blosum62	0,908
ORTHOMCL 1006	10	Blosum62	0,795
ORTHOMCL 1007	10	RtREV	0,896
ORTHOMCL 1008	10	Blosum62	0,947
ORTHOMCL 1009	10	Blosum62	0,983
ORTHOMCL 1018	10	RtREV	0,817

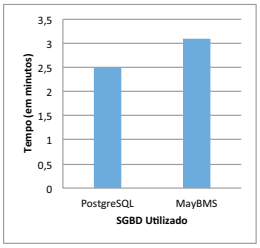

Figura 3 Média do tempo de Execução da Consulta

Além disso, o tempo de extração e carga dos dados no BD de proveniência foi calculado. Para extração dos dados de domínio de arquivos, cálculo da probabilidade e carga no BD de proveniência foram necessários 16,5 minutos. Isso equivale a aproximadamente 3,1% do tempo total de execução do *workflow*, o que é aceitável frente aos ganhos que o cientista pode obter. É importante ressaltar que a avaliação foi realizada sobre um único *workflow* com um volume pequeno de dados. Mais experimentos são necessários para verificar se o desempenho observado é verificado em outros cenários.

5. Conclusão

O uso de *workflows* para modelar experimentos de bioinformática vem crescendo a cada ano. Diversos *workflows*, como o SciPhy, têm produzido um grande volume de dados que deve ser analisado. Tais dados são normalmente extraídos de arquivos e carregados em BDs de proveniência de forma que possam ser consultados. Porém, os BDs de proveniência utilizados pelos SGWfCs para armazenar os dados assumem que existe uma certeza nos relacionamentos de proveniência. Entretanto, essa premissa nem sempre é verdadeira para todos os tipos de experimentos. Por exemplo, o *workflow* SciPhy apresenta atividades não determinísticas, como o alinhamento que é a fase mais representativa na bioinformática. Este artigo propõe o uso de *Probabilistic Provenance Analyzers* com o objetivo de extrair dados de domínio e calcular a incerteza associada a cada dado extraído. Resultados mostraram que a consulta que considera a incerteza é aproximadamente 16,3% mais custosa que sua versão sem incerteza e que o processo de extração e cálculo da incerteza insere um *overhead* de cerca de 3,1% no tempo total de execução do *workflow*. Consideramos esses custos aceitáveis frente ao ganho que o cientista tem nas análises integradas de proveniência e domínio com incerteza associada.

Referências Bibliográficas

Ahola, V., Aittokallio, T., Vihinen, M. and Uusipaikka, E. (2008). Model-based prediction of sequence alignment quality. Bioinformatics (Oxford, England), v. 24, n. 19, p. 2165–2171.

Boulos, J., Dalvi, N., Mandhani, B., et al. (2005). MYSTIQ: A System for Finding More Answers by Using Probabilities. In Int. Conf. Management of Data (SIGMOD), pp. 891-893.

Chapman, A., Blaustein, B. and Elsaesser, C. (2010). Provenance-based Belief. In Workshop on the Theory and Practice of Provenance (TaPP). p. 11.

Costa, F., Silva, V., De Oliveira, D., et al. (2013). Capturing and Querying Workflow Runtime Provenance with PROV: A Practical Approach. In EDBT/ICDT Workshops, pp. 282-289.

De Oliveira, D., Silva, V. and Mattoso, M. (2015). How Much Domain Data Should Be in Provenance Databases? In Workshop on Theory and Practice of Provenance (TaPP).

Freire, J., Koop, D., Santos, E. and Silva, C. T. (2008). Provenance for Computational Tasks: A Survey. Computing in Science Engineering, v. 10, n. 3, p. 11–21.

Gonçalves, J. C. de A. R., Oliveira, D. De, Ocaña, K. A. C. S., Ogasawara, E. and Mattoso, M. (2012). Using Domain-Specific Data to Enhance Scientific Workflow Steering Queries. In International Provenance and Annotation Workshop (IPAW), pp. 152–167.

Huang, J., Antova, L., Koch, C. and Olteanu, D. (2009). MayBMS: A Probabilistic Database Management System. In Int. Conf. Management of Data (SIGMOD), pp. 1071-1071.

Idika, N., Varia, M. and Phan, H. (2013). The Probabilistic Provenance Graph. In IEEE Security and Privacy Workshops (SPW), pp 34-41.

Mattoso, M., Werner, C., Travassos, G. H., et al. (2010). Towards supporting the life cycle of large scale scientific experiments. Int. Journal of Business Process Integration and Management, v. 5, n. 1, p. 79.

Moreau, L., Clifford, B., Freire, J., et al. (2011). The Open Provenance Model core specification (v1.1). Future Generation Computer Systems, v. 27, n. 6, p. 743–756.

Moreau, L. and Missier, P. (2013). The PROV Data Model and Abstract Syntax Notation. W3C Recommendation.

Ocaña, K. A. C. S., Oliveira, D. De, Ogasawara, E., et al. (2011). SciPhy: A Cloud-Based Workflow for Phylogenetic Analysis of Drug Targets in Protozoan Genomes. In Advances in Bioinformatics and Computational Biology, pp. 66-70.

Ogasawara, E., Dias, J., Oliveira, D., et al. (2011). An Algebraic Approach for Data-Centric Scientific Workflows. Proc. of the Int. Conf. on Very Large Data Bases (PVLDB), v. 4, n. 12, p. 1328–1339.

Re, C. and Suciu, D. (2007). Management of Data with Uncertainties. In Conference on Information and Knowledge Management (CIKM), pp. 3-8.

Simmhan, Y. L., Plale, B. and Gannon, D. (2008). Query capabilities of the Karma provenance framework. Concurrency and Computation: Practice and Experience, v. 20, n. 5, p. 441–451.