Ontologia Probabilística para Auxiliar na Recuperação de Modelos Biológicos[1]

Wladimir Pereira, Kate Revoredo

Programa de Pós-Graduação em Informática

Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

Av. Pasteur, 296 – Urca – Cep 22290-240 – Rio de Janeiro – RJ – Brazil

{wladimir.pereira, katerevoredo}@uniriotec.br

Abstract. *The Cell Component Ontology (CelO), an ontology expressed in OWL-DL that describes semantically biological models associated with the context of electrophysiology, has no support for dealing with uncertainty. It is demonstrated in this paper that a computational environment based on ontologies (CelO) and Bayesian Networks can help researchers in the modeling phase of the cycle of experimental knowledge of Biology, retrieving accurately biological models.*

Resumo. *A Cell Component Ontology (CelO), uma ontologia expressa em OWL-DL que possibilita expressar a semântica de modelos biológicos associados ao contexto da eletrofisiologia, não possui suporte para lidar com a incerteza. É demonstrado neste trabalho que um ambiente computacional baseado em ontologias (CelO) e Redes Bayesianas é capaz de auxiliar o pesquisador na fase de modelagem do ciclo experimental de conhecimento da Biologia, recuperando modelos biológicos de uma maneira mais precisa.*

1. Introdução

Em [Matos et al. 2010] foi apresentada a Cell Component Ontology (CelO), uma ontologia expressa em OWL-DL que é derivada da CellML [Cuellar et al. 2003], uma linguagem de marcação baseada em XML (eXtensible Markup Language) [Bray et al. 2000] criada especificamente com o propósito de descrever variáveis, equações e componentes de modelos biológicos de maneira formal, sem ambiguidades, legível por humanos e processável por máquinas.

Cada modelo CellML é composto por uma rede de componentes interconectados, que é a menor unidade funcional do modelo, e por variáveis, que são entidades que têm como propósito representar quantidades usadas nas equações. Além disso, há as conexões, que mapeiam variáveis entre componentes, permitindo a troca de informações entre eles. A Figura 1 mostra um esquema dos elementos que compõem um modelo CellML.

[1] Esse trabalho faz parte do escopo do projeto "Infraestrutura de apoio a Gerência de experimentos científicos em Modelagem Computacional" com apoio do CNPQ (número 559998/2010-4)

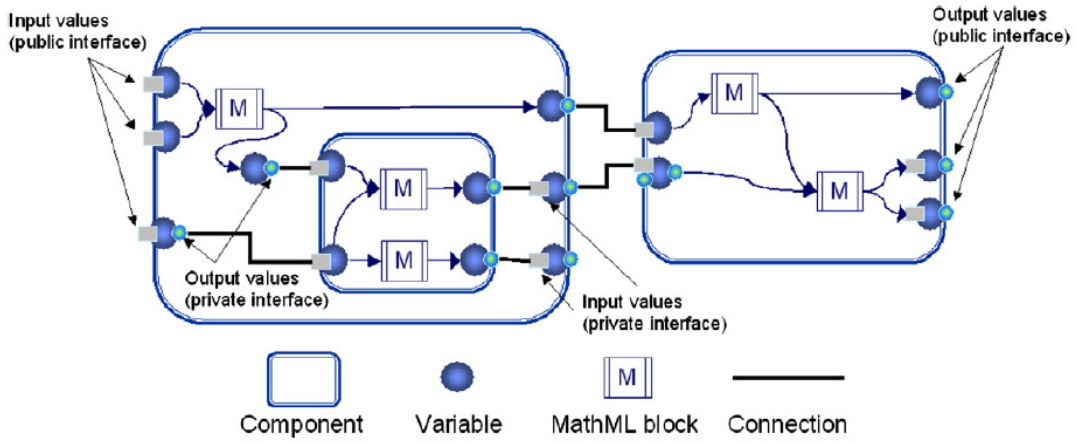

Figura 1. Representação em esquema dos elementos de um modelo CellML [Matos et al. 2010]

O objetivo da ontologia CelO é acrescentar semântica a modelos biológicos descritos em CellML, associados ao contexto da eletrofisiologia, possibilitando expressar o conhecimento intrínseco do modelo, possibilitar a validação semântica de novos modelos, reusar componentes de outros modelos, automatizar processos de composição de modelos e possibilitar que a procura de modelos seja realizada de forma semântica.

A integração da ontologia CelO com a CellML possibilita que o pesquisador modele em um nível alto de abstração e execute computacionalmente o modelo sem necessidade de conhecimento da linguagem em XML.

De acordo com [Macedo 2005], o ciclo experimental do conhecimento da Biologia passa por três fases, que podem ser vistas na Figura 2: na primeira, modelos biológicos são propostos e hipóteses são apresentadas; na segunda, simulações computacionais são executadas com os modelos biológicos propostos, combinando dados de diferentes experimentos físicos, gerando predições sobre o comportamento do sistema, provendo uma visão mais acurada dos fenômenos estudados; na terceira, o resultado de cada simulação é analisado, podendo surgir novas hipóteses desta análise, o que reiniciaria o ciclo.

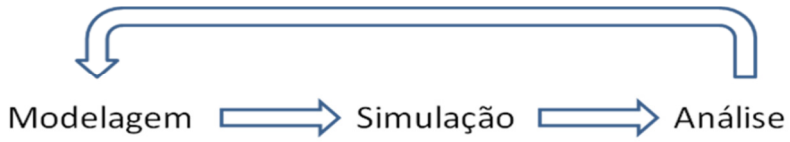

Figura 2. Ciclo Experimental do Conhecimento da Biologia [Macedo 2005]

Na fase de modelagem, que é o foco deste trabalho, o pesquisador pode obter na ontologia CelO a representação semântica do conceito ou fenômeno de interesse (por exemplo, o "potencial da membrana" e "canal iônico de sódio") e pesquisar quais modelos biológicos estão de alguma forma associados ao conceito ou ao fenômeno pesquisado. Em seguida, o pesquisador pode escolher um dos modelos biológicos listados para executar as simulações.

Dentro deste ciclo, a etapa de recuperação de um modelo biológico a ser tomado como ponto de partida deve ser precisa e retornar o modelo biológico mais adequado à necessidade do pesquisador, já que novos modelos biológicos são desenvolvidos a partir

de componentes de um modelo biológico existente. Um novo componente pode ser inserido e o modelo biológico ajustado, estabelecendo a conexão deste com os demais componentes. Após a simulação, dependendo dos resultados obtidos, a inclusão deste novo componente é confirmada ou o mesmo é substituído. Este processo pode se repetir por diversas vezes, o que torna o processo trabalhoso e sujeito a erros.

A CelO não possui suporte para lidar com a incerteza, ou seja, não é possível definir um grau intermediário de pertinência dos modelos biológicos existentes no repositório à consulta realizada. Como exemplo, ao pesquisar por "potencial da membrana" e "canal iônico de sódio", o agente responsável pela pesquisa, caso não consiga encontrar uma resposta categórica, deveria agir com um grau de incerteza, informando os modelos biológicos com maior probabilidade de atender às necessidades do pesquisador.

Por outro lado, a pesquisa feita por Ding e Peng [2004] e o trabalho de Ding et al. [2006], que gerou a linguagem BayesOWL, tiveram o objetivo de estender a OWL para representar a incerteza por meio do uso de redes bayesianas [Charniak 1991]. Os autores apresentam o conceito de probabilidade dentro da OWL, isto é, a semântica da OWL é ampliada através de marcações adicionais visando representar a incerteza. O resultado é uma ontologia que pode ser traduzida em uma rede Bayesiana, porém, em ambos os casos, o uso de anotações particulares do domínio limitam a capacidade de expressar modelos probabilísticos mais complexos ou genéricos, restringindo as soluções para classes de problemas muito específicos. No caso da BayesOWL, o foco é o mapeamento de ontologias, desta forma, a estrutura da linguagem é adequada para que este objetivo seja alcançado.

Visando a interoperabilidade com ontologias não probabilísticas, a linguagem PR-OWL foi proposta por [Costa e Laskey 2006]. A linguagem também é uma extensão para a linguagem OWL e o modelador pode obter uma ontologia em OWL padrão e utilizar os recursos da PR-OWL apenas para as partes da ontologia que necessitarem de suporte probabilístico. Em sua abordagem, ontologias OWL podem ser usadas para representar modelos probabilísticos complexos, de uma forma que é suficientemente flexível para ser usado por diversas ferramentas probabilísticas baseadas em redes Bayesianas. O problema desta abordagem é que, para lidar com a incerteza, é necessário modificar e reorganizar a base de conhecimento original, através da introdução de novas relações. Tarefa esta que pode ser trabalhosa e normalmente requer um bom conhecimento em redes Bayesianas. Além disso, requer a participação de um especialista para criar as tabelas de probabilidades condicionais.

Em [Devitt et al. 2006], os autores apresentam um algoritmo para automatizar a construção de Redes Bayesianas e representar com precisão um domínio de interesse. As tarefas envolvidas neste processo exigem a introdução de um especialista na definição de quais propriedades da ontologia ou quais relações entre os conceitos correspondem aos relacionamentos da rede bayesiana. É uma abordagem muito interessante, porque as dependências entre os nós que correspondem as classes da ontologia que não estão explicitadas na ontologia podem ser identificadas por este especialista. A tarefa de estimar as probabilidades condicionais não foi tratada nesse trabalho.

O objetivo deste trabalho é demonstrar que um ambiente computacional baseado em ontologias (CelO) e Redes Bayesianas é capaz de auxiliar o pesquisador na fase de

modelagem do ciclo experimental de conhecimento da Biologia, recuperando modelos biológicos de uma maneira mais precisa.

2. Proposta

Conforme pôde ser visto nos trabalhos citados na seção anterior, uma abordagem frequentemente utilizada para a gestão do conhecimento e da incerteza é a combinação de Ontologias e Redes Bayesianas.

Rede Bayesiana (RB) é um grafo direcionado acíclico, onde cada nó é uma variável identificada a partir do domínio de aplicação e cada arco representa a dependência direta entre as variáveis. Cada variável tem um domínio de valores possíveis que ela pode assumir e associada a ela há uma tabela de probabilidades condicionais (CPT) que fornece a probabilidade para cada valor possível desta variável [Charniak 1991].

A Figura 3 mostra um exemplo de RB onde é possível perceber que a variável *dog-out* é influenciada diretamente tanto pela variável *family-out* como pela variável *bowel-problem* e que a mesma possui uma CPT associada a ela que pode ser definida como P(*dog-out*) = <0.99, 0.90, 0.97, 0.30>.

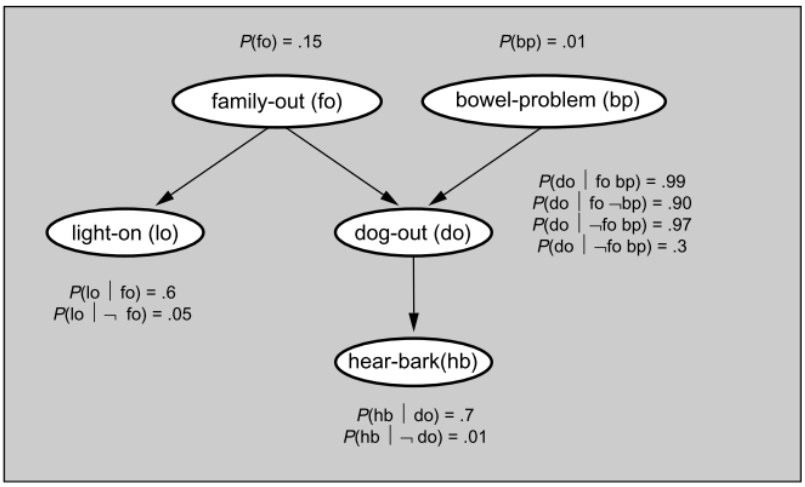

Figura 3. Exemplo de Rede Bayesiana [Charniak 1991]

O uso de ontologias foi descrito em [Guarino 1995] como um meio para adicionar semântica à web. Ele define ontologias como uma representação formal de um conhecimento compartilhado, processável por máquinas. Uma ontologia representa as classes de entidades de um domínio de aplicação, as propriedades das classes, as relações entre as classes e os papéis que as classes podem desempenhar.

O conhecimento pode ser extraído de uma ontologia usando o raciocínio lógico, explorando as relações entre as classes (conceitos) e os fatos armazenados nele (as instâncias das classes). Isto é, ontologias consistem em duas partes: uma parte referida como TBox, que contém o conhecimento sobre os conceitos (classes, por exemplo) e as relações entre eles (ou seja, papéis); e uma outra parte referida como ABox, que contém conhecimento sobre as entidades (ou seja, indivíduos) e como eles se relacionam com as classes [Andrea e Franco 2011].

Segundo [Devitt et al. 2006], a tarefa de construção da estrutura da RB é dependente do conhecimento de um especialista e possui as seguintes etapas:

1. Identificar os conceitos relevantes definidos no TBox da ontologia e mapear cada um deles como uma variável da RB.
2. Especificar os valores possíveis para cada uma destas variáveis.
3. Identificar as relações de influência entre as variáveis.

A etapa de obtenção dos parâmetros das distribuições de probabilidade para cada variável (as CPTs) consiste na aprendizagem das distribuições de probabilidade inicial, que são calculadas diretamente das instâncias de ontologia (ABox).

A ideia é que a RB gerada após estas etapas represente o conhecimento probabilístico codificado por uma ontologia tanto em nível de conceito como em nível de instância e, quando associado à ontologia CelO, torne a recuperação de modelos biológicos mais precisa, o que auxiliará o pesquisador na fase de modelagem.

3. Considerações Finais

Neste trabalho é proposta uma abordagem que visa auxiliar o pesquisador na fase de modelagem do ciclo experimental de conhecimento da Biologia, recuperando modelos biológicos de uma maneira mais precisa. Além de detalhar a proposta, foram apresentados os conceitos de RB e de Ontologias, além de trabalhos relacionados ao tema.

Ao contrário de algumas das pesquisas citadas, esta abordagem tem como grande vantagem o fato de existir uma separação entre o conhecimento do domínio e o conhecimento probabilístico, isto é, os conceitos de probabilidade não são representados dentro da ontologia e a base de conhecimentos não é alterada. Desta forma, a proposta não exige que a OWL seja estendida.

Além disso, consideramos a abordagem proposta neste artigo mais vantajosa em um contexto geral já que propõe aprender uma RB a partir das instâncias da ontologia, diminuindo a necessidade de um especialista na definição das distribuições de probabilidade condicional.

Para a avaliação da proposta, será realizado um experimento, utilizando um repositório de modelos biológicos representados através da CelO, com foco no processo de recuperação de modelos. Visando confirmar o ganho da proposta, serão comparados os resultados obtidos com os apresentados em [Matos et al. 2010].

Referências

Andrea, B., e Franco, T. (2011). Mining Bayesian networks out of ontologies. Journal of Intelligent Information Systems. Published online first, 13 June 2011. doi:10.1007/s10844-011-0165-4.

Bray, T., Paoli, J. e Sperberg-McQueen, C. M. (2000). Extensible Markup Language (XML). W3C recommendation. World Wide Web Consortium. http://www.w3.org/XML/.

Charniak, E. (1991). Baysean Networks without Tears. AI Magazine, v. 12, n. 4, p. 50-63.

Costa, P. C. G. e Laskey, K. B. (2006). PR-OWL: A framework for probabilistic ontologies. In Proceedings of the 2006 conference on Formal Ontology in

Information Systems: Proceedings of the Fourth International Conference (FOIS 2006), pages 237-249. IOS Press, 2006. Available at http://portal.acm.org/citation.cfm?id=1566107.

Cuellar, A. A., Lloyd, C. M., Nielsen, P. F., Bullivant, D.P., Nickerson, D.P., Hunter, P.J. (2003). An Overview of CellML 1.1, a Biological Model Description Language. Simulation, v. 79, n. 12, p. 740-747.

Devitt, A., Danev, B. e Matusikova, K. (2006). Constructing Bayesian Networks Automatically using Ontologies. In Proceedings of Second Workshop on Formal Ontologies Meets Industry (FOMI 2006).

Ding, Z. e Peng, Y. (2004). A Probabilistic Extension to The Web Ontology Language OWL. In Thirty Seventh Hawaii International Conference on System Sciences (HICSS 04), IEEE CS Press, 2004, pp. 40111.1.

Ding, Z., Peng, Y. e Pan, R. (2006). BayesOWL: Uncertainty modeling in semantic web ontologies. Soft Computing in Ontologies and Semantic Web, p. 3–29.

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. International Journal of Human Computer Studies, v. 43, n. 5, p. 625–640.

Macedo, J. A. F. (2005). Um Modelo Conceitual para Biologia Molecular. PhD thesis, Departamento de Informática da PUC-Rio. Available at http://www.maxwell.lambda.ele.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=7939

Matos, E. E., Campos, F., Braga, R. e Palazzi, D. (2010). CelOWS: an ontology based framework for the provision of semantic web services related to biological models. Journal of Biomedical Informatics, v. 43, n. 1, p. 125-136.