ISSN 1981-2841

APLICAÇÕES DE XML NA BIOINFORMÁTICA[1]

XML APLICATIONS ON BIOINFORMATICS

Leandro Silveira[2] e Giovani Rubert Librelotto[3]

RESUMO

A representação da informação genômica, obtida pelos esforços de sequenciamento de nucleotídeos do DNA, é essencial para o tratamento computacional pós-genômico, que visa à análise funcional e estrutural de moléculas e processos biológicos. Com essa perspectiva, foram estudados dois padrões de armazenamento da informação genômica, a *Bioinformatic Sequence Markup Language* (BSML) e a *RNA Markup Language* (RNAML). Neste estudo, o objetivo é verificar a importância da linguagem XML, aos pesquisadores da área de bioinformática.

Palavras-chave: pós-genômico, *Bioinformatic Sequence Markup Language* (BSML)*, RNA Markup Language* (RNAML).

ABSTRACT

The genomic information representation obtained by the sequencing of DNA cores is essential for the post-genomic computational treatment that aims the functional and structural analysis of molecules and biological processes. In this perspective, two standards of storage to the genomic information were studied: Bioinformatics Sequence Markup Language (BSML) and RNA Markup Language (RNAML). This paper aims to verify the importance of XML language in the bio-informatics field.

Keywords: *post-genomic, Bioinformatic Sequence Markup Language (BSML), RNA Markup Language (RNAML).*

---

[1] Trabalho de Iniciação Científica - UNIFRA.
[2] Acadêmico do Curso de Sistemas de Informação - UNIFRA - E-mail: lsilveira@gmail.com.
[3] Orientador - UNIFRA. E-mail: giovani@unifra.br

INTRODUÇÃO

Os dados gerados pelos processos automatizados de sequenciamento genômico precisam ser representados e armazenados. No armazenamento, a preocupação ocorre com a eficiência do acesso e com a manutenção da consistência da informação. Por outro lado, na tarefa de representação, criam-se modelos computacionais processáveis que irão influenciar na complexidade de análise dos dados representados. O último processo, descrito de forma genérica, é a essência do estudo da biologia computacional que, diante dos dados biológicos, desenvolve ferramentas de interpretação, como predição de estruturas e funções de proteínas, a partir de uma sequência de nucleotídeos. Isso mostra que a representação desses dados é de suma importância para o tratamento computacional pós-genômico que visa à análise funcional e estrutural de moléculas e processos de interesse da biotecnologia (AMABIS; RODRIGUES, 2002).

A modelagem em *eXtensible Markup Language* (XML) (RAMALHO; HENRIQUES, 2002), além de ser adequada à representação de conteúdos e constituir padrão para a troca de informação, permite a manipulação dos dados por meio de consultas e modificações com linguagens apropriadas. A modelagem, por meio de XML Schema, é uma abordagem nova que sinaliza a possibilidade de uma padronização da informação genômica. Além disso, ela também otimiza o tratamento por parte de *software* que manipula esse tipo de dado (FALLSIDE et al., 2004).

Dessa forma, pode-se perceber a necessidade da transformação entre os dialetos utilizados para a representação da informação genômica, transformando-as para os dialetos que as representem de maneira mais eficiente e consistente.

O objetivo geral, neste artigo, é mostrar a utilização da linguagem XML na bioinformática, comparar dois dialetos utilizados, transformando um documento do formato de BSML para um documento no formato RNAML, sem a perda de informações que são essenciais para a representação de um *ribonucleic acid* (RNA).

Outros objetivos da pesquisa são: apresentar o contexto biológico para a comunidade acadêmica da área da informática, indicando as características referentes à biologia molecular; apresentar as principais características e aplicações da linguagem XML na bioinformática por meio da criação de dialetos; estudar e apresentar alguns modelos utilizados atualmente para representação de dados na bioinformática; criar uma folha de estilos XSL (CLARK, 1999) para a transformação de documentos XML utilizados na bioinformática.

Este artigo inicia com uma apresentação dos conceitos básicos sobre *topic maps* na seção 2. Na seção 3, descreve-se o processo do desenvolvimento do esquema relacional e do banco de dados para o armazenamento de *topic maps*. Os trabalhos relacionados estão na seção 4 e uma síntese do artigo é apresentada na seção 5.

BIOINFORMÁTICA

Segundo Gibas e Jambeck (2001), até meados do século passado, profissionais das áreas relacionadas à genética mantinham um questionamento a respeito da estrutura química do material genético. No ano de 1953, essa estrutura foi decodificada por Watson e Crick; com o acesso a essa informação, os estudos relacionados à biologia molecular intensificam-se e, a partir da segunda metade da década de 90, surgiram os sequenciadores automáticos para a organização e a interpretação desses dados que aumentam cada vez mais. Necessita-se, então, da otimização do processo de sequenciamento e ordenação dos dados obtidos e de um processo seguro e confiável de armazenamento.

Conforme o aumento das pesquisas realizadas nos laboratórios de biologia, a informática passa a ser indispensável e novas técnicas e ferramentas computacionais são desenvolvidas para acelerarem a interpretação de dados contidos no DNA (CHI, 2006).

Um dos fatos que despertam curiosidade na área da biologia e que ajudam a intensificar e justificar o avanço das pesquisas, experimentos e investimentos nessa área pode ser exemplificado da seguinte forma: ao realizar uma comparação física ou comportamental de um ser humano com uma mosca, não se consegue identificar uma semelhança, porém ao se comparar os genes de um humano com o da mosca-das-frutas (*Drosophila melanogaster*), um biólogo consegue verificar que a mosca possui um gene chamado *eyeless,* que desempenha uma função no desenvolvimento dos olhos, pois ao eliminar esse gene do genoma, por métodos de biologia molecular, verifica-se que a mosca se desenvolverá sem os olhos. Por sua vez, o humano possui um gene que é responsável por uma condição chamada *aniridia*. Nos seres humanos que não têm esse gene, os olhos desenvolvem-se sem a íris. Ao inserir o gene da *aniridia* em uma mosca-das-frutas sem os olhos, ele promoverá a produção de olhos normais em uma *Drosophila*.

Na compararação das sequências do DNA de *eyeless* e *aniridia*, é possível verificar a semelhança citada, mas para realizar tal comparação manualmente seria muito demorado e desgastante, consequentemente, o uso da tecnologia da informação agiliza o processo de análise de sequências.

Até o final dos anos 80, o sequenciamento de DNA era realizado manualmente e exigia um tempo maior do que os sequenciadores automáticos atuais. Desse modo, a quantidade de dados aumentou rapidamente e surgiu, com isso, a necessidade de manter esses dados acessíveis e organizados. Os sequenciadores automáticos surgiram na segunda metade da década de 90, a partir da rapidez com que as sequências passaram a ser analisadas. Assim, precisou-se de uma área que abrangesse diversas outras - a bioinformática, pois essa aborda conceitos matemáticos e estatísticos, bem como faz uso da química e da informática, sendo todos conceitos voltados à organização, ao acesso e à interpretação de dados relacionados à biologia molecular que, para desvendar a complexidade das informações contidas no material genético dos seres vivos, necessita da interdisciplinaridade de diversas áreas.

Uma lista de métodos computacionais utilizados por pesquisadores dessa nova área é definida a seguir: uso de banco de dados públicos e formatos de dados, alinhamento e busca de sequência, predição de genes, alinhamento múltiplo de sequências, análise filogenética, análise da sequência de proteínas, análise das propriedades da estrutura protéica e análise de microarrays de DNA, entre outros. (GIBAS; JAMBECK, 2001)

XML NA BIOINFORMÁTICA

O XML atraiu a atenção para essa área por sua forma de representação de dados. Atualmente, existem vários dialetos oriundos do XML na área da bioinformática. Essa linguagem foi utilizada com um grande êxito no armazenamento dos dados biológicos, como as sequências de DNA, RNA e de proteínas. Com essa linguagem, pode-se gerar um documento contendo todos os dados relacionados à sequência de bases de aminoácidos e também todo o processo de obtenção e de outros dados de suma importância para os pesquisadores.

Outra utilização, dentre outras muito importantes, está na troca das informações entre laboratórios e pesquisadores que utilizam a *internet* como meio de troca pela facilidade pelo grande suporte que ela vem recebendo. O uso do XML tem favorecido a bioinformática, devido às seguintes características:

• Flexibilidade: é simples de modificar o DTD (*Document Type Definition* – que descreve a estrutura dos elementos de um arquivo em XML); os dados contidos no XML também são muito fáceis de serem alterados; os arquivos em XML e DTD são compreendidos pelo ser humano e são de fácil leitura e edição.

• Sua estrutura é muito simples, pois não passa de texto puro e isso torna mais simples a sua transmissão pela *internet*.

• Trabalha em multiplataforma (*Windows*, *Unix*, etc.), ou seja, não depende da plataforma para ser acessada e usada. É utilizada por várias companhias, como a *Microsoft*, *Oracle*, IBM, entre outras.

A seguir, abordam-se alguns dos dialetos utilizados na bioinformática (GIBAS; JAMBECK, 2001):

BIOPOLYMER MARKUP LANGUAGE (BIOML)

O objetivo com o *Biopolymer Markup Language* (BIOML) é fornecer uma estrutura extensível para essa anotação e um veículo comum para trocar essa informação entre os cientistas que usam o *World Wide Web*.

O BIOML foi projetado para ser disponibilizado livremente, ou seja, sem direitos autorais e é uma linguagem ligeiramente diferente das outras, pois seu documento original descreve um objeto físico, por exemplo, uma proteína em particular, de tal modo que toda a informação experimental conhecida sobre esse objeto pode ser associada ao objeto de uma maneira lógica e significativa (BIOML, 2006).

BIOINFORMATIC SEQUENCE MARKUP LANGUAGE (BSML)

A criação inicial do *Bioinformatic Sequence Markup Language* (BSML)[3] teve como objetivo resolver o problema da representação de dados de sequências biológicas. Sua criação coincidiu com o aumento do número de base de dados internacionais de sequência e com a sofisticação de métodos computacionais para a análise de sequências. A BSML descreve as moléculas biológicas (sequências do DNA, do RNA e da proteína), incluindo dados da sequência, tabelas da característica, referências da literatura e dados tabulares associados (por exemplo, valores da expressão do gene).

[3] BSML: Linguagem de Anotação de Sequências de Bioinformática (ou, em inglês, BSML: *Bioinformatics Sequence Markup Language*).

O objetivo da BSML foi criar um modelo para representação de dados que permitisse "renderizar" (tornar) gráficos de objetos biológicos, bem como conectar esses objetos visuais aos dados subjacentes (ou seja, para ligar o comportamento do objeto da exposição às sequências, às anotações e às ligações que representa).

BSML é um modelo detalhado dos dados para as sequências e suas anotações e codifica a semântica dos dados que representa. Ele serve tanto para a interação humana quanto às máquinas no processamento de sequências e anotações. Essa linguagem pode ser usada para descrever fenômenos de sequências relacionadas a todos os níveis, do nível biomolecular ao nível do organismo completo e, assim, fornece um meio excelente para a pesquisa dos genomas.

Esse modelo suporta a execução de métodos genéricos de *software* (mapas lineares e circulares, anotações, posições, etc.) e traz o genoma ao computador do biólogo que não é um especialista em computadores (BSML, 2006).

CHEMICAL MARKUP LANGUAGE (CML)

A *Chemical Markup Language* (CML) surgiu em 1998 utilizando a XML, pois essa permite o desenvolvimento de linguagens de editoração (ou marcação) específicas de domínio de aplicação. Inicialmente, a CML foi desenvolvida com base no padrão de uma aplicação SGML, tendo migrado para XML devido aos novos recursos adicionados ao XML, de modo que o tornou superior ao SGML. Ela permite a descrição das informações de estrutura moleculares, bem como a reutilização dessas informações sem a alteração dos dados no processo (CML, 2006).

RNA MARKUP LANGUAGE (RNAML)

A necessidade de novos métodos informatizados na ciência do *RNA Markup Language* (RNA) tem aumentado dramaticamente nesses últimos anos, devido à grande quantidade de dados gerados no sequenciamento do projeto genoma e nas análises *microarray*.

O BIOML fornece uma estrutura extensível da informação experimental sobre as entidades moleculares, como as proteínas e os genes, e seu conteúdo é utilizado na representação e na visualização da informação genética.

Embora ambas as linguagens, BIOML e BSML, incluam a informação do RNA, suas sintaxes são focalizadas nos elementos gene-relacionado do RNA, com locais de início e de término da transcrição de outras, e nenhuma delas dirige-se à informação da estrutura do RNA.

Os formatos existentes não capturaram detalhes da biologia do RNA adequadamente. Com base na necessidade de uma sintaxe padrão, um grupo de cientistas do RNA, em 1998, decidiu que deveria ser criado um formato para a troca de dados do RNA, que fosse adotado como um formato padrão para fornecer um único formato, representando informação específica às moléculas do RNA. Essa nova linguagem escolheu o padrão de representação para suas informações: a linguagem XML, pela flexibilidade, pela adaptação infinita, pela manutenção, pela simplicidade e pela portabilidade (RNAML, 2006).

COMPARANDO BSML E RNAML

A integração de informações tem sido amplamente estudada. Desse modo, vários sistemas têm sido propostos e desenvolvidos para integrar múltiplas fontes, o que consiste em uma tarefa complexa. Uma das razões para tal complexidade é a heterogeneidade na estrutura das fontes. Uma forma de minimizar a complexidade da integração, adotada pela maioria dos sistemas de integração, é definir um modelo de dados comum para representar a estrutura e o conteúdo das fontes.

NOME	USADO PARA	SCHEMA	COMENTÁRIOS
BSML	Alinhamento e anotação de sequências.	DTD	• Não possui XML Schema (xsd); • Muito extenso; • Bem documentado.
RNAML	Sequências de RNA, estruturas em 2D e 3D(dimensões).	DTD e XML Schema	• XML Schema não possui NameSpaces; • Muito extenso.

VANTAGENS E DESVANTAGENS

A vantagem do BSML sobre o RNAML é que sua estrutura pode descrever sequências de DNA, RNA e proteínas, incluindo dados da sequência, tabelas da característica, referências da literatura e dados tabular associados (por exemplo, valores da expressão do gene). A desvantagem é que ainda utiliza um DTD como esquema, seu conteúdo é muito extenso, o que torna sua transformação para outros dialetos mais complexa, apesar de estar bem documentado.

RNAML, por sua vez, tem como vantagem a representação da estrutura do RNA, a qual não é bem armazenada no dialeto do BSML. Por se tratar de um dialeto para um tipo específico de informação, este pode ser mais objetivo, dando prioridade às informações da biologia do RNA, sem a perda de informações. Outra vantagem, como se pode observar na tabela anterior, é que ele já possui seu esquema no formato de um XML Schema. A desvantagem está justamente nesse ponto: a limitação para somente um tipo de informação.

ANÁLISE DOS DIALETOS

Desenvolveu-se uma XSL referente às informações contidas na tabela 1, que são informações básicas contidas em um documento do dialeto BSML. Devido ao conteúdo a ser transformado tratar de um ácido ribonucléico (RNA), deve-se, primeiramente, verificar se o documento que está no formato BSML se trata do mesmo tipo. Para isso, serão analisados os trechos dos DTDs dos dois dialetos.

A partir da linha 12 até a 23 tem-se os atributos de uma sequência do dialeto BSML. No atributo *molecule,* que pertence ao elemento <Sequence>, obtém-se a confirmação de que se trata do ácido ribonucléico (RNA). Após a confirmação, é iniciada a transformação, copiando as informações contidas no referido atributo para o atributo *type,* filho do elemento <molecule> do dialeto RNAML.

```
<!ELEMENT Sequence (
        Attribute*,
        Feature-tables?,
        (Seq-data |
        Seq-data-import)?,
        Numbering?,
        Modification*,
        Segment*,
        ..
        ...)>

<!ATTLIST Sequence
        length CDATA #IMPLIED
        molecule (
        mol-not-set |
        dna | rna | aa | na |
        other-mol) “dna”
        topology (
        top-not-set | linear |
        circular | tandem |
        top-other) “linear”
        refs IDREFS #IMPLIED
>

<!ELEMENT
        Seq-data (#PCDATA)>


...
```

(a) – bsml.dtd

```
<!ELEMENT molecule (
        identity?,
        sequence*,
        structure?)>

<!ATTLIST molecule
        id ID #REQUIRED
        type (rna | dna) “rna”
        comment CDATA #IMPLIED
        ...
>

<!ELEMENT sequence (
        numbering-system*,
        seq-data?,
        seq-annotation?)>

<!ATTLIST sequence
        strand CDATA #IMPLIED
        length CDATA #IMPLIED
        circular(true|false)#IMPLIED
        ...
>

<!ELEMENT seq-data (#PCDATA)>
<!ATTLIST seq-data
        comment CDATA #IMPLIED
        ...
>

...
```

(b) – rnaml.dtd]

Tabela 1 – comando parcial de XPath, indicando quais os campos que se relacionam do dialeto BSML com o RNAML.

ATRIBUTO	BSML	RNAML	COMENTÁRIO
length	Sequence/@length	sequence/@length	sem nenhuma condição
molecule	Sequence/@molecule	molecule/@type	"rna" (obrigatoriamente)
topology	Sequence/@topology	sequence/@circular	Se for circular o atributo do dialeto RNAML recebe "true", caso contrário "false"
Id	Sequence/@id	molecule/@id	sem nenhuma condição
comment	Sequence/@comment	seq-data/@comment	sem nenhuma condição
ELEMENTO	**BSML**	**RNAML**	**COMENTÁRIO**
seq-data	Seq-data	seq-data	sem nenhuma condição

O modelo proposto por Graauw baseia-se na explicitação dos tipos de tópicos, ocorrências e associações em tabelas separadas, permitindo a adição de novos dados também no modelo ER, pela simples inserção de dados nessas tabelas, já que tudo é representado por tópicos e a diferença está apenas no tipo (instâncias). Isso traduz o paradigma *topic maps* para o modelo ER, porém este oferece a possibilidade de identificar quais os tipos, por meio da observação da tabela original, o que não é possível no padrão *topic maps*.

CONCLUSÃO

A bioinformática é uma área interdisciplinar que está aumentando exponencialmente, a partir da grande quantidade de dados adquiridos pelos pesquisadores da área no que tange a biologia molecular. A área da informática está cada vez mais presente nesse campo e auxilia os biólogos nos processos de análise e na predição dos dados.

Com o avanço da tecnologia, nos últimos anos, essa área aumentou ainda mais. A utilização da linguagem XML na bioinformática para a estruturação das informações foi um passo muito importante para ambas as áreas, pois os dados relacionados à bioinformática devem ser armazenados de uma forma estruturada e ter relações, como na vida real. Esses dialetos, oriundos do XML, têm se mostrado bem utilizados e representam a informação de forma coerente com a informação desejada.

XML é a abreviação de *Extensible Markup Language*. Essa linguagem não pertence a nenhuma empresa ou instituição, pois foi desenvolvida por um grupo de pessoas independentes, sob o respaldo do W3C (*World Wide Web Consortium*), de acordo com experiências anteriores, como o SGML ou o HTML. Em termos mais formais, pode-se definir o XML como uma linguagem de anotação descritiva extensível, tendo, portanto, características que tornam desejáveis esse tipo de linguagem: independência relativamente às plataformas de *software* e de *hardware* utilizadas, longevidade, baixos custos de manutenção, facilidade na reutilização, entre outros.

Devido ao grande volume de dados e da troca constante das informações, fez-se necessário o uso do XML para estruturar, mantendo os dados consistentes, facilitando também o intercâmbio de informações pela portabilidade que essa tecnologia possui.

O XML possui uma linguagem, conhecida como XSL, que tem o poder de transformar um documento em outro tipo de documento, com o auxílio de outro mecanismo dessa linguagem, o XPath. Ambas as tecnologias andam juntas, o XSL necessita da XPath, pois ela é essencial para a realização da seleção dos nodos para a transformação. A XSL, gerada durante este estudo, transforma os dados necessários para a especificação de um RNA que estava armazenado em um documento escrito no dialeto BSML.

Dessa forma, pode-se concluir que essa linguagem é de suma importância na transformação de documentos. Neste estudo, mostrou-se o que uma única tecnologia é capaz de fazer com simples comandos, sem a necessidade da criação de *softwares* mais complexos para a realização dessa operação.

REFERÊNCIAS

AMABIS, José; RODRIGUES, Gilberto. **Fundamentos da biologia moderna**. 3. ed. São Paulo: Moderna, 23, 436 – 448, 2002.

BIOML, **Biopolymer Markup Language**, 2006. Disponível em: http://xml.coverpages.org/bioml.html. Acesso em: out. 2006.

BSML, **Bioinformatics Sequence Markup Language**, 2002. Disponível em: http://www.bsml.org. Acesso em: abr. 2006.

CLARK, James. W3C Recommendation - XSL Transformations (XSLT). **World Wide Web Consortium**, 1999**.** Disponível em: http://www.w3.org/TR/xslt#element-syntax-summary. Acesso em: nov. 2006

CHI, Cambrigdge Healthtec Institute Sequences. 2006. **DNA & beyond:** Evolving terminology for emerging technologies. Disponível em: http://www.genomicglossaries.com/content/sequencing_dna_beyond_gloss.asp. Acesso em: out. 2006.

CML, **Chemical Markup Language (CMLTM)**, 2006. Disponível em: http://www.xml-cml.org/. Acesso em: out. 2006.

FALLSIDE, D. et al; XML Schema Part 0: Primer, **World Wide Web Consortium**, 2004. Disponível em: http://www.w3.org/TR/xmlschema-0. Acesso em: out. 2006.

GIBAS, Cynthia. JAMBECK, Per. **Desenvolvendo bioinformática:** ferramentas de software para aplicações em biologia. Tradução Milarepa Ltda. Rio de Janeiro: Campus, 2001.

RAMALHO, José Carlos; HENRIQUES, Pedro. **XML & XSL** – da teoria à prática, 1. ed. Lisboa, PO: FCA, 2002.

RNAML. **RNA Markup Language,** 2006. Disponível em: http://wwwlbit.iro.umontreal.ca/rnaml/. Acesso em: ago. 2006.