UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GERENCIAMENTO DE ANOTAÇÕES DE BIOSSEQÜÊNCIAS UTILIZANDO ASSOCIAÇÃO ENTRE ONTOLOGIAS E ESQUEMAS XML

Marcus Vinícius Carneiro Teixeira

São Carlos - SP
Maio/2008

Marcus Vinícius Carneiro Teixeira

Gerenciamento de Anotações de Biosseqüências Utilizando Associação entre Ontologias e Esquemas XML

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

Orientador: Mauro Biajiz.

Co-orientador: Ricardo Rodrigues Ciferri.

São Carlos - SP

Maio/2008

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

T266ga

Teixeira, Marcus Vinícius Carneiro.
Gerenciamento de anotações de biosseqüências utilizando associações entre ontologias e esquemas XML / Marcus Vinícius Carneiro Teixeira. -- São Carlos : UFSCar, 2008.
106 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2008.

1. Banco de dados. 2. Bioinformática. 3. Anotação de genoma. 4. Projeto de banco de dados. 5. Ontologias. I. Título.

CDD: 005.74 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

"Gerenciamento de Anotações de Biosseqüências utilizando Associação entre Ontologias e Esquemas XML"

MARCUS VINÍCIUS CARNEIRO TEIXEIRA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Membros da Banca:

Prof.Dr. Mauro Biajiz
(Orientador - DC/UFSCar)

Prof. Dr. Ricardo Rodrigues Ciferri
(Co-orientador – DC/UFSCar)

Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho (ICMC/USP)

Prof. Dr. Sérgio Lifschitz
(INF/PUC-Rio)

São Carlos
Maio/2008

À minha família, por tanto carinho, amizade e respeito que sempre me dedicaram.

Em especial, aos meus pais, que sempre batalharam pela minha felicidade.

AGRADECIMENTOS

Aos meus pais e aos meus irmãos, que sempre me apoiaram e torceram em todos os desafios na minha vida.

À minha namorada Mariana, pelo companheirismo, incentivo, por tantos momentos felizes, por tantos ensinamentos, por sua alegria.

Ao meu amigo Iuri Rizzi, pela inestimável amizade e companheirismo ao longo de seis anos.

Aos amigos que fiz durante os anos de estudos na UFSCar (Mario, Crisinha, Cris, Josi, Honasses, Fabinho, Fábio, Luciano, Luciana, Mary, Pablão, Fabiano, Thaisinha, Carol, Marcelo, Vanessa, Sara, Dalci, Daniel, Tati, Caco, Carlos CB, Dani Toledo, Hapu, Laine, Roberta, Thaís, Fernandinha, Sônia, Sol e tantos outros) os quais sempre me proporcionam momentos de alegria e que me acompanharam em muitas aventuras e momentos marcantes. Sem dúvida fizeram parte de uma fase de ouro, me ensinaram muita coisa e me fizeram crescer como ser humano. Uns continuam aqui por perto, outros agora estão distantes, mas todos serão sempre lembrados.

Aos amigos mais antigos, dos tempos de ensino médio, a quem devo muito do que sou hoje, pois muitas coisas aprendi, e ainda aprendo, com eles e por muitas coisas vividas e compartilhadas. Obrigado por tudo.

Aos meus orientadores, Mauro e Ricardo, pela dedicação, conversas, paciência, ensinamentos e amizade. Em especial ao Mauro, por tantos anos de convívio, as mais variadas conversas, sobre todo tipo de assunto, e pelas orientações enquanto professor e amigo.

Aos demais professores do Departamento de Computação, pela dedicação e ensinamentos em todos esses anos de estudos na UFSCar.

Aos companheiros do Grupo de Banco de Dados da UFSCar pelas conversas e cooperação.

Às pessoas que contribuíram respondendo ao questionário usado na realização dos testes.

Ao CNPq, pelo auxílio financeiro.

“Viva!

Bom mesmo é ir à luta com determinação,

abraçar a vida com paixão,

perder com classe e vencer com ousadia

porque o mundo pertence a quem se atreve

e a vida é muito para ser insignificante.”

(Charles Chaplin)

“A vida é uma peça de teatro que não permite ensaios.

Por isso, cante, chore, ria e viva intensamente antes que a

cortina se feche e a peça termine sem aplausos.”

(Charles Chaplin)

“É melhor ser alegre que ser triste

Alegria é a melhor coisa que existe

É assim como a luz no coração”

(Vinícius de Moraes)

RESUMO

A Bioinformática é uma área da ciência que visa suprir pesquisas de genomas com ferramentas computacionais que permitam o seu desenvolvimento tecnológico. Dentre essas ferramentas estão os ambientes de anotação e os Sistemas Gerenciadores de Bancos de Dados (SGBDs) que, associados a ontologias, permitem a formalização de conceitos do domínio e também dos esquemas de dados. Os dados produzidos em projetos genoma são geralmente textuais e sem uma estrutura de tipo regular, além de requerer evolução de esquemas. Por suas características, SGBDs semi-estruturados oferecem enorme potencial para tratar tais dados. Assim, este trabalho propõe uma arquitetura para um ambiente de anotação de biosseqüências baseada na persistência dos dados anotados em bancos de dados XML. Neste trabalho, priorizou-se o projeto de bancos de dados e também o apoio à anotação manual realizada por pesquisadores. Assim, foi desenvolvida uma interface que utiliza ontologias para guiar a modelagem de dados e a geração de esquemas XML. Adicionalmente, um protótipo de interface de anotação manual foi desenvolvido, o qual faz uso de ontologias do domínio de biologia molecular, como a Gene Ontology e a Sequence Ontology. Essas interfaces foram testadas por usuários com experiências nas áreas de Bioinformática e Banco de Dados, os quais responderam a questionários para avaliá-las. O resultado apresentou qualificações muito boas em diversos quesitos avaliados, como exemplo agilidade e utilidade das ferramentas. A arquitetura proposta visa estender e aperfeiçoar o ambiente de anotação Bio-TIM, desenvolvido pelo grupo de Banco de Dados do Departamento de Computação da Universidade Federal de São Carlos (UFSCar).

Palavras-chave: Bioinformática, Anotação de Genoma, Bancos de Dados de Biologia Molecular, XML, Bancos de Dados XML, Projeto de Banco de Dados, Ontologias.

ABSTRACT

Bioinformatics aims at providing computational tools to the development of genome researches. Among those tools are the annotations systems and the Database Management Systems (DBMS) that, associated to ontologies, allow the formalization of both domain conceptual and the data scheme. The data yielded by genome researches are often textual and with no regular structures and also requires scheme evolution. Due to these aspects, semi-structured DBMS might offer great potential to manipulate those data. Thus, this work presents architecture for biosequence annotation based on XML databases. Considering this architecture, a special attention was given to the database design and also to the manual annotation task performed by researchers. Hence, this architecture presents an interface that uses an ontology-driven model for XML schemas modeling and generation, and also a manual annotation interface prototype that uses molecular biology domain ontologies, such as Gene Ontology and Sequence Ontology. These interfaces were proven by Bioinformatics and Database experienced users, who answered questionnaires to evaluate them. The answers presented good assessments to issues like utility and speeding up the database design. The proposed architecture aims at extending and improving the Bio-TIM, an annotation system developed by the Database Group from the Computer Science Department of the Federal University from São Carlos (UFSCar).

Keywords: Bioinformatics, Genome Annotation, Biological Databases, XML, XML Databases, Database Design, Ontologies.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AOM – Asset Oriented Modeling
BDBM – Banco de Dados de Biologia Molecular
DER – Diagrama Entidade-Relacionamento
DNA – Deoxyribonucleic Acid (Ácido Desoxirribonucléico)
DW – Data Warehouse
EST – Expressed Sequence Tag
GO – Gene Ontology
HTML – HyperText Markup Language
OWL – Web Ontology Language
RNA – Ribonucleic Acid (Ácido Ribonucléico)
SGBD – Sistema Gerenciador de Bancos de Dados
SO – Sequence Ontology
UML – Unified Modeling Language
XML – eXtensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

Projetos genoma buscam codificar biosseqüências (e.g., seqüências de nucleotídeos e seqüências de aminoácidos), no processo conhecido por seqüenciamento de genoma, além de identificar e caracterizar seus genes. Outro objetivo é gerar observações que auxiliem o entendimento das funções desenvolvidas pelos genes dentro dos organismos, incluindo a produção de proteínas e as suas funções moleculares. Tais projetos são importantes, pois contribuem para o desenvolvimento de outras áreas de pesquisa, tais como medicina, agricultura e pecuária.

O processo de seqüenciamento de genomas (STEIN, 2001; OKURA, 2002; LEMOS, 2004) inicia-se em laboratórios (*in vitro*) onde biólogos utilizam técnicas avançadas e máquinas seqüenciadoras para recolhimento do material genético e a sua análise. Nesta fase, o objetivo principal é identificar as seqüências de bases de nucleotídeos que compõem um DNA, que podem ser adenina (A), citosina (C), guanina (G) ou timina (T). Este processo é realizado por meio de gel-eletrosforese, que envolve o uso de corantes fluorescentes que, excitados por lasers, variam seu comprimento de onda de acordo com a base de nucleotídeo encontrada (OKURA, 2002). Como ainda não é possível seqüenciar um genoma completo de uma única vez, as seqüências de nucleotídeos são quebradas em partes menores chamados *reads* para então serem lidas pelas máquinas seqüenciadoras. Assim, duas outras etapas seguintes devem ser realizadas, chamadas de montagem e fechamento do genoma. Na primeira, esses *reads* são realinhados para obter-se a cadeia de nucleotídeos original, e na segunda é realizada uma análise minuciosa para verificar se não sobraram regiões sem serem lidas, chamadas buracos (ou *gaps*). Ao fim dessa fase, ainda nenhuma informação que possa caracterizar esses dados é conhecida, assim estes são considerados dados brutos.

O próximo passo de um projeto genoma é a caracterização das seqüências obtidas, chamado de processo de anotação. Este processo geralmente ocorre em laboratórios computacionais (*in silico*) e permite a identificação de regiões de interesse e a definição de funções para essas regiões, que contribuem para o entendimento do genoma. Por exemplo, podem-se confirmar processos biológicos já conhecidos ou gerar novas descobertas que venham a tratar questões ainda não esclarecidas aos biólogos. Assim, a anotação de um projeto genoma é considerada de grande importância para as pesquisas nessa área, exigindo-se muita atenção em sua execução. Em especial, Stein (STEIN, 2001) define um projeto genoma como sendo somente tão bom quanto sua anotação. Pode-se destacar o mapeamento genético,

a atribuição funcional a genes e a interação entre eles, como as principais contribuições de projetos genoma.

1.1 Motivação

A automatização do processo de seqüenciamento teve como conseqüências a agilização das pesquisas de genoma e a obtenção de um enorme volume de dados genômicos a cada análise feita sobre uma biosseqüência. Hoje é possível obter seqüências de nucleotídeos de cromossomos eucarióticos (i.e., organismos com núcleo celular) e procarióticos (i.e., organismos sem núcleo celular) inteiros. Muito dessa evolução surgiu com o desenvolvimento de uma nova disciplina, a Bioinformática. Ela tem o propósito de estudar e apoiar as pesquisas de genoma com o desenvolvimento de ferramentas computacionais para processamento, armazenamento, recuperação e análise de dados relevantes, o que tem se tornado essencial aos projetos genoma (OKURA, 2002; MEYER *et al.*, 2003). Essas ferramentas são específicas da área e geralmente envolvem algoritmos complexos para tratar e analisar dados genômicos. Alguns sistemas combinam diversas dessas ferramentas para análise de dados genômicos e serão chamados, no contexto deste trabalho, de ambientes de anotação.

Outro fator que contribui para o avanço em pesquisas de genoma é a disponibilidade de bancos de dados públicos que permitem o compartilhamento de informações entre diferentes grupos de pesquisa. Entre os principais bancos de dados públicos estão GenBank (BENSON *et al.*, 2005), EMBL (EUROPEAN BIOINFORMATICS INSTITUTE, 2008a), DDBJ (DNA DATA BANK OF JAPAN, 2006) e SWISS-PROT (EUROPEAN BIOINFORMATICS INSTITUTE, 2008b). Os três primeiros armazenam seqüências de genomas e anotações, enquanto o último é um banco de dados de seqüências de aminoácidos (proteínas) curado, ou seja, seus dados são validados por pesquisadores e não contêm redundâncias. Uma conseqüência direta do enorme volume de dados produzido foi a implantação de bancos de dados via Sistemas Gerenciadores de Bancos de Dados (SGBDs) em projetos genoma, necessários por serem mais apropriados para o armazenamento e consulta aos dados produzidos. Antes disso, os biólogos moleculares tratavam os dados de biosseqüências como textos simples, coletando e armazenando-os em arquivos do tipo texto. Entre as questões que os bancos de dados de genoma, também conhecidos como bancos de

dados de biologia molecular (BDBMs), tentam solucionar, estão: armazenamento e representação dos dados biológicos; integração das diversas fontes de dados biológicas; e interfaces de acesso intuitivas para uso pelos cientistas (SEIBEL et al., 2000; LEMOS *et al.*, 2003b). Essas são questões que não são fáceis de serem solucionadas.

Existem diversos ambientes de anotação descritos na literatura (alguns dos principais estão detalhados no capítulo 5) que disponibilizam ferramentas para que pesquisadores realizem suas anotações. Cada um desses ambientes foi desenvolvido em função de projetos de Bioinformática aos quais se propuseram a auxiliar e, assim, possuem propriedades particulares, geralmente com esquemas de dados heterogêneos e sem qualquer conexão semântica. Esse fator leva à criação de esquemas de dados de difícil integração.

O Grupo de Banco de Dados (GBD) do Departamento de Computação (DC) da Universidade Federal de São Carlos (UFSCar) tem oferecido serviços voltados à manutenção de dados de projetos genoma. Para o auxílio a esse serviço, foi desenvolvido um ambiente denominado Bio-TIM (*Bioinformatics – Transparent Information Management*) (OLIVEIRA, 2005), o qual permite o recebimento de cromatogramas e os processa até a geração das biosseqüências e de uma série de relatórios. Entretanto, o ambiente Bio-TIM ainda apresenta uma série de restrições, as quais tornam a funcionalidade de anotação bastante incompleta e parcial . Assim, este trabalho pretende estender as funcionalidades de anotação do Bio-TIM para que este se adeqüe melhor às necessidades de uso dos biólogos.

1.2 Objetivo

Este trabalho de Mestrado visa propor uma arquitetura para um ambiente de anotação de projetos genoma, chamado BioFOX, capaz de organizar esses dados com o uso de ontologias e bancos de dados semi-estruturados, possibilitando, assim, melhorias como agregação de semântica aos dados, padronização de conceitos e criação de esquemas de dados flexíveis. Essa arquitetura é composta por quatro módulos, os quais visam tratar separadamente as ferramentas de análise de biosseqüências, os bancos de dados e as anotações manuais. Ela será usada para a melhoria do Bio-TIM, estendendo as suas funcionalidades de anotação e tornando-o útil a diferentes projetos genoma. Enfoque é dado ao tratamento dos bancos de dados e das anotações manuais, principalmente ao primeiro, o qual deve tratar a semântica relacionada aos dados.

Este trabalho se insere no contexto de Bioinformática e visa estudar e investigar a manutenção de anotações produzidas por projetos genoma. Isto compreende a efetiva definição das anotações, seu armazenamento e posterior recuperação. Mais especificamente, este trabalho propõe o uso de bancos de dados semi-estruturados XML em um ambiente de anotações e apresenta uma interface de apoio à definição de esquemas do banco de dados. A escolha pelo modelo semi-estruturado se deve à natureza dos dados genômicos, que apresentam uma estrutura irregular e com constante ausência de informações. Além disso, a integração de fontes de dados heterogêneas pode ser melhor representada por um modelo de dados interoperável baseado em XML, uma vez que cada uma dessas fontes pode ter um modelo de dados próprio (LEMOS *et al.*, 2003b). Um terceiro fator, que torna atrativo o uso do modelo semi-estruturado, é a flexibilidade que ele apresenta para a evolução de esquemas, muito requerido por BDBMs (SEIBEL et al., 2000; LEMOS *et al.*, 2003b).

O apoio à definição do esquema do banco de dados XML pode ser realizado por meio da associação entre os dados do domínio da aplicação e um componente que permita o raciocínio automatizado sobre eles. Esse componente é uma ontologia, termo utilizado para o compartilhamento daquilo que se entende por um determinado domínio e seus conceitos (USCHOLD e GRUNINGER, 1996). Uma ontologia provê um vocabulário de termos específicos para a representação de conceitos presentes em um domínio de conhecimento e ainda cria uma rede de relacionamentos entre esses termos (GRUBER, 1995; STOFFEL *et al.*, 1997; GUARINO, 1998). Associada aos esquemas do banco de dados XML, uma ontologia também possibilita determinar a semântica dos dados a serem armazenados. É um meio poderoso para a análise e integração de dados biológicos.

Assim, este trabalho proporciona meios para apoiar o projetista do banco de dados quanto à definição de esquemas de bancos de dados semi-estruturados, que devem ser definidos de acordo com o domínio de aplicação do projeto genoma associado, e também auxiliar os pesquisadores em suas anotações manuais e busca por novos conhecimentos. Para o projetista, a idéia fundamental é que ele seja guiado por uma ontologia para a definição de esquemas de bancos de dados XML válidos, que tenham semântica agregada a seus dados, com flexibilidade para evoluir e compatíveis com outros esquemas de mesmo domínio, facilitando a integração de dados providos por ambos. Aos anotadores, devem ser criadas ferramentas de apoio à padronização de suas observações, com o objetivo de gerar anotações com maior qualidade.

Dentre as contribuições deste trabalho, destaca-se o desenvolvimento de uma interface para a criação de esquemas XML, baseada em uma ontologia. Essa interface contribui para a

criação de esquemas de dados e para a padronização e compartilhamento de esquemas de mesmo domínio. Além disso, ela também contribui para a criação de esquemas independentes estruturalmente sem, contudo, perder a semântica associada. Essa interface foi avaliada por um grupo de 14 (quatorze) usuários, com experiências em Bioinformática e/ou Banco de Dados. Dentre os quesitos avaliados constam questões sobre a experiência de cada usuário, sobre o uso da interface e sobre a relação com a Bioinformática.

1.3 Organização da Dissertação

Esta dissertação está organizada em 8 capítulos, sendo resumidos a seguir.

O Capítulo 2, Anotação de Projetos Genoma, define o conceito de anotação de genoma e sua importância dentro de projetos genoma. São mostrados os três níveis de anotação existentes e como as anotações são tratadas em um ambiente de anotação.

O Capítulo 3 define ontologias e como elas podem ser incorporadas a um ambiente de anotação. Muitas ontologias do domínio de biologia molecular têm sido criadas e utilizadas para a integração de fontes de dados heterogêneas, como a *Gene Ontology* (ASHBURNER *et al.*, 2000) e a *Sequence Ontology* (EILBECK *et al.*, 2005), apresentadas nesse capítulo.

O Capítulo 4 introduz o modelo de dados semi-estruturado e como ele tem sido utilizado na Bioinformática. São apresentadas as linguagens XML e XML Schema, bases para o desenvolvimento deste trabalho, e as vantagens de tal modelo para a representação de dados de biosseqüências.

O Capítulo 5 apresenta alguns dos principais ambientes de anotação existentes. Foram levantadas suas principais características e abordagens utilizadas para auxiliar os biólogos em suas pesquisas. A última seção desse capítulo contém uma análise comparativa entre eles, considerando-se características que desejamos implementar neste trabalho.

O Capítulo 6 apresenta a arquitetura do ambiente de anotação BioFOX e cada um de seus quatro módulos. Duas subseções apresentam também o *Namespace* de Anotações Genômicas contendo vocabulários XML e a ontologia de aplicação desenvolvida para apoio à definição de esquemas XML.

O Capítulo 7 descreve o desenvolvimento e os testes sobre a interface XML Database Design. Baseada em uma ontologia de aplicação, ela tem como objetivo auxiliar o projetista do banco de dados na modelagem de esquemas XML.

Por fim, o Capítulo 8 apresenta a conclusão deste trabalho, seus resultados, contribuições e trabalhos futuros.

2 ANOTAÇÃO DE PROJETOS GENOMA

2.1 Considerações Iniciais

Lincoln Stein (STEIN, 2001) descreve detalhadamente o significado e a importância de uma anotação genômica. Ele faz uma advertência segundo a qual a seqüência do genoma de um organismo é uma fonte de informação diferente de qualquer outra à qual biólogos tenham tido acesso, sendo seu valor somente tão bom quanto a sua anotação. Ou seja, sem uma boa anotação para descrever uma seqüência, esta pode não ter qualquer valor. O objetivo de uma anotação de alta qualidade é identificar as características chaves do genoma, em particular os genes e os seus produtos. Devido à sua importância, diversas ferramentas e recursos para anotação são constantemente desenvolvidos. A evolução de genes e a caracterização das funções moleculares exercidas pelas proteínas são assuntos ainda pouco desvendados, e a esperança por encontrar tais respostas está sobre a análise das anotações genômicas.

A anotação de genoma tem como principal objetivo adicionar informações a uma biosseqüência, através de sua análise e interpretação, que possam determinar um significado biológico a ela e relacioná-las a conhecimentos biológicos prévios. Uma anotação genômica pode ser separada em três níveis: nível de nucleotídeo, nível de proteína e nível de processo. A Figura 1, extraída de (STEIN, 2001), apresenta esses níveis e suas relações, segundo este autor.

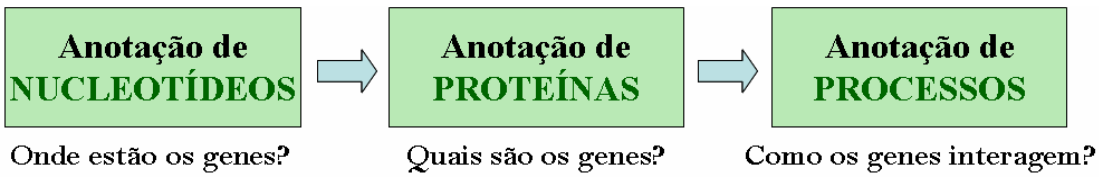


Figura 1 Níveis de anotações genômicas e as perguntas que se deseja responder em cada uma delas.

As seções 2.2 a 2.4 apresentam um breve resumo desses níveis, também com base em (STEIN, 2001). Por fim, a seção 2.5 aborda as fontes de dados com as quais os ambientes de anotação trabalham.

2.2 Anotação de Nucleotídeos

Na anotação em nível de nucleotídeos procura-se mapear regiões previamente conhecidas, caracterizadas e identificadas pela genética, citogenética ou por mapeamento híbrido por radiação. Nessa fase, genes, seqüências de RNAs (i.e., transportador, ribossômico e outros não codificadores), regiões regulatórias, marcações genômicas (ex. *primer*), seqüências repetitivas e até mesmo evidências que indiquem duplicação ancestral no genoma são identificadas. Assim, pontos de referência são marcados em uma biosseqüência, facilitando o reconhecimento por parte dos pesquisadores.

Vários algoritmos *ab initio* já foram criados com o objetivo de encontrar genes em genomas eucarióticos (GENSCAN (BURLIN e KARLIN, 1997), Genie (BESEMER *et al.*, 2000), GeneMark.hmm (BESEMER e BORODOVSKY, 1999), Grail (UBERACHER e MURAL, 1991), HEXON (SOLOVYEV *et al.*, 1994), MZEF (ZHANG, 1997), Fgenes (SOLOVYEV *et al.*, 1995) e HMMGene (KROGH, 1997)). Esses algoritmos buscam identificar genes em biosseqüências sem o uso de conhecimento prévio sobre similaridades com outros genes. Porém a sua própria natureza os torna não muito confiáveis e por isso eles não têm sido muito utilizados. Outros algoritmos verificam a similaridade de biosseqüências com a de organismos previamente seqüenciados. Uma coincidência de um nucleotídeo com seqüências de cDNA ou *expressed sequence tags* (ESTs), mesmo que de outras espécies, é uma boa evidência do fato de que determinada região pertença a um gene. Assim, algoritmos de similaridade como o BLAST (ALTSCHUL *et al.*, 1990) e o FAST (PELLEGRINI *et al.*) têm sido muito utilizados por sua eficiência quando comparados com a execução manual da pesquisa. Uma abordagem mais recente com relação à predição de genes é a que combina predições *ab initio* com dados de similaridade em um único modelo probabilístico (STEIN, 2001).

A anotação genômica somente não é mais simples devido a processos ainda não decifrados, tais como *splicing* alternativo, que podem gerar anotações não confiáveis (STEIN, 2001). *Splicing* é o processo de remoção dos íntrons e ligação dos éxons para formar uma seqüência contígua no RNA, e que em seguida será responsável pela produção de proteínas. Em um *splicing* alternativo, essas seqüências contíguas podem ser recombinadas com a troca de posições entre os éxons, assumindo diferentes formas. Esse processo valida a teoria de que uma mesma seqüência de DNA codificadora pode gerar diferentes proteínas (STAMMA *et al.*, 2005).

2.3 Anotação de Proteínas

O passo seguinte à anotação de nucleotídeos é determinar as proteínas de organismos e suas funções derivadas do gene. A obtenção de um catálogo definitivo das proteínas dos organismos é um passo crucial nas pretensões da comunidade científica. Além disso, determinar a relação de homologia entre proteínas também é importante. Uma homologia pode ser classificada como paralogia ou ortologia, ambas relacionadas ao processo evolutivo de espécies (JENSEN, 2001). A paralogia entre duas proteínas significa que, durante algum processo de divisão celular em uma determinada espécie, um gene foi duplicado sendo que cada um pode ou não desempenhar diferentes funções em um mesmo organismo. Já a relação de ortologia indica que durante um processo de especiação um mesmo gene se manteve nas duas novas espécies geradas, desenvolvendo as mesmas funções. Mas também é possível haver paralogia entre dois genes em diferentes espécies se previamente tiver ocorrido ortologia. Assim, nem sempre um gene presente em uma espécie, quando comparado a outro similar, mas pertencente a outra espécie, pode ser considerado como tendo a mesma função.

Dentre os milhares de genes encontrados em um organismo, apenas uma pequena fração corresponde a uma proteína bem caracterizada e conhecida. Devido ao grande número de proteínas com função desconhecida, anotadores geralmente começam por classificá-las em grupos ou famílias de proteínas, e por usar similaridades com proteínas melhor caracterizadas de outras espécies. Entretanto, devido ao modo como o processo de evolução acontece, nem sempre é possível associar tais proteínas por similaridades, justamente pelas questões de homologia explicitadas no parágrafo anterior.

Uma forma comum de se realizar a anotação de proteínas é procurar similaridades utilizando ferramentas computacionais como o BLASTP (ALTSCHUL *et al.*, 1990) ou PSI-BLAST (ALTSCHUL e KOONIN, 1998), utilizando diferentes bancos de dados de proteínas. O banco de dados de seqüências de proteínas mais precioso é o SWISS-PROT, o qual apresenta uma coleção de seqüências de proteínas confirmadas e extensivamente anotadas. Ele contém ainda referências a outros bancos de dados de biosseqüências e estruturas, referências bibliográficas, identificação da família protéica e descrições sobre a provável função e papel biológico da proteína.

Uma análise complementar consiste na procura de domínios funcionais e as bases de dados mais utilizadas nesse processo são: PFAM (BUTEMAN *et al.*, 2000), PRINTS (ATTWOOD *et al.*, 2000), PROSITE (HOFFMAN *et al.*, 1999), ProDom (CORPET *et al.*,

1999), SMART (PONTING *et al.*, 1999) e BLOCKS (HENIKOFF *et al.*, 2000). Novamente aqui, esses diversos bancos de dados de famílias de proteínas, domínios e padrões não têm padronizadas suas nomenclaturas, seus métodos de busca e suas adequações a diversas tarefas, sendo que muitas vezes um mesmo dado é representado de formas diferentes entre elas. Isso torna difícil interpretar os resultados quando uma proteína predita tem entradas similares em diversos desses bancos, já que não se tem certeza se elas se referem ao mesmo conceito. Por isso, foi desenvolvido um banco integrado de assinaturas de proteínas, conhecido como InterPro (APWEILER *et al.*, 2001), que procura integrar as informações dos bancos anteriormente citados. Cada entrada do InterPro contém uma breve descrição da família ou domínio, uma lista de proteínas do SWISS-PROT ou TrEMBL (EUROPEAN BIOINFORMATICS INSTITUTE, 2008b) que o contém, referências bibliográficas e *links* para cada um dos bancos de dados membros.

2.4 Anotação de Processos

Finalmente, a última etapa de uma anotação genômica objetiva relacionar o genoma a processos biológicos. Assim, determinar como os genes e proteínas relacionam-se com processos tais como o ciclo celular, a morte celular, a embriogênese, o metabolismo e a manutenção da saúde, torna-se uma característica da análise de seqüências de genoma e um passo fundamental na obtenção de conhecimentos aprofundados. Esse passo também é conhecido como anotação funcional (STEIN, 2001).

Para a anotação em nível de processo é necessário mais do que trabalho computacional. Técnicas biológicas de alta produção como mutagênese mediada por *transposons*, análise de expressão em *microarrays*, ensaio de expressão de proteínas por espectroscopia de massa, ensaios baseados em *green-fluorescent-protein* para determinar localização e padrões temporais de expressão e estudos de duplo-híbrido em leveduras têm sido de fundamental importância para identificar o papel de genes e proteínas nos processos biológicos (STEIN, 2001).

2.5 Fontes de Dados em Ambientes de Anotação

Ambientes de anotação podem prover acesso a diferentes tipos de anotação, que podem ser classificados segundo sua fonte, a listar:

- **Anotação importada:** a fonte de seus dados é externa ao ambiente de anotação, geralmente bancos de dados públicos de genoma.
- **Anotação automática:** a fonte de seus dados são ferramentas computacionais utilizadas para análise de biosseqüências; os dados produzidos por elas são considerados anotações automáticas.
- **Anotação manual:** são observações realizadas por pesquisadores do projeto, geralmente biólogos, com base tanto em seu próprio conhecimento quanto na literatura ou em anotações pré-existentes.

Atualmente existem diversos bancos de dados públicos no domínio de biologia molecular, dos quais alguns já foram citados em capítulos anteriores (ex: GenBank, DDBJ e Swiss-Prot). Cada um desses bancos de dados mantém seus dados em formatos que consideram mais adequados, provendo livre acesso a eles para a comunidade científica. O GenBank, por exemplo, mantém suas anotações sobre seqüências de DNA em um formato texto de acordo com o padrão ASN.1 (INTERNATIONAL ORGANIZATION FOR STANDARTIZATION, 1987) e, atualmente, disponibiliza-os também em formato XML (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2008). Outros exemplos são o AceDB (SANGER INSTITUTE, 2008), que utiliza um esquema orientado a objetos, mas com os dados armazenados em um formato próprio, o .ACE, com sintaxe semelhante à XML; e o Swiss-Prot, que mantém seus dados em um banco de dados relacional. Alguns ambientes de anotação trabalham com a integração de diferentes fontes de dados, podendo-se utilizar abordagens tais como SGBDs Distribuídos e Heterogêneos, *Multidatabase* ou através de um *Data Warehouse* (DW) (SEIBEL et al., 2000). A integração de anotações permite disponibilizá-las para processamento de consultas e comparações mais ágeis com os dados produzidos pelo projeto.

As anotações automáticas são de grande importância para os projetos genoma, pois são responsáveis por acelerar a produção de dados em relação à anotação puramente manual, como era feito nos primórdios desses projetos, uma vez que são geradas por ferramentas computacionais de Bioinformática. As ferramentas de Bioinformática são geralmente desenvolvidas para a resolução de problemas específicos desse domínio. Stein (STEIN, 2001)

lista ferramentas comumente utilizadas, suas funções e vantagens. Um critério importante para uso dessas ferramentas é que elas devem gerar anotações confiáveis, com graus de confiabilidade que possam ser definidos por seus usuários.

A terceira fonte de dados de um ambiente de anotação são as anotações manuais produzidas pelos próprios pesquisadores. Estes são responsáveis por analisar os dados gerados pelo projeto e anotar suas observações, podendo ser gerados novos conhecimentos dessas observações. Um controle de versão sobre as anotações manuais é importante, pois permite que os pesquisadores atualizem as suas próprias anotações (LEMOS *et al.*, 2004). As anotações manuais devem ser apoiadas por interfaces com recursos que possibilitem ao pesquisador fazer suas observações de forma mais ágil e com maior qualidade. Um exemplo é o uso de vocabulários controlados ou ontologias do domínio de biologia molecular, os quais permitem maior controle e qualidade sobre as anotações (LEMOS *et al.*, 2004).

Um ambiente de anotação capaz de fornecer esses três tipos de anotação pode auxiliar os pesquisadores a fazerem um trabalho de maior qualidade, uma vez que lhes fornece diversas informações. No entanto, deve-se atentar para a qualidade dessas informações, para que erros não sejam propagados a anotações subseqüentes. A arquitetura BioFOX, proposta neste trabalho, visa armazenar e manipular dados provindos de qualquer um dos três tipos de fontes de dados de anotações.

2.6 Considerações Finais

Neste capítulo foi descrito o que são anotações em projetos genoma, seu papel e importância nesse contexto. A anotação pode ser classificada em três níveis: nucleotídeos, proteínas e processos ,. Além disso, também foram descritas as fontes de dados dos ambientes de anotação, ou seja, de onde provêm as anotações de um projeto genoma (i.e., fontes externas, automáticas e manuais). Estes conceitos são importantes para o entendimento da natureza dos dados gerados ou trabalhados no domínio de projetos genoma e também da importância de uma boa anotação.

O capítulo 3 discute o significado de ontologias e destaca algumas ontologias do domínio de biologia molecular. As ontologias têm exercido papel importante para a padronização e compartilhamento de anotações.

3 ONTOLOGIAS

3.1 Considerações Iniciais

Com o grande volume de dados de projetos genoma espalhados em diversas fontes de dados, integrá-los e interpretá-los se tornou um grande desafio para os pesquisadores e um passo importante para que novos conhecimentos sejam inferidos e também novas descobertas biológicas sejam feitas. O maior problema se deve ao fato de não haver uma padronização quanto à forma de representação dos dados nessas diferentes fontes de dados biológicos. Dessa forma, cada uma delas utiliza um modelo de dados particular e, não raramente, utilizam terminologias diferentes para representar um mesmo conceito, o que pode confundir os pesquisadores em suas interpretações.

As ontologias têm sido utilizadas com o intuito de resolver esse problema de integração de dados biológicos. Suas definições de conceitos e seus relacionamentos dentro de um domínio permitem que esse conhecimento seja compartilhado por uma mesma comunidade. A definição de uma ontologia requer longos estudos a respeito do domínio de interesse, sendo geralmente sua formulação deixada para pesquisadores e especialistas na área. A biologia é uma ciência onde novos conhecimentos são gerados a todo tempo e, por isso, ontologias desse domínio devem acompanhar sua evolução. Uma ontologia só será bem sucedida caso a comunidade científica faça uso de suas definições, pois só então será possível atingir seus objetivos principais.

Assim, o propósito deste capítulo é abordar ontologias e seus usos, especialmente no domínio de biologia molecular. Desta forma, a seção 3.2 descreve o conceito de ontologia, detalhando suas definições e usos. Em seguida, nas seções 3.3 e 3.4, são descritas duas importantes ontologias desenvolvidas para a biologia molecular, a *Gene Ontology* e a *Sequence Ontology*.

3.2 Definições

Tradicionalmente uma disciplina de estudos em filosofia, as ontologias são agora um tópico chave para o desenvolvimento de bancos de dados biológicos. Ontologia é o termo utilizado para o compartilhamento daquilo que se entende por um determinado domínio e seus conceitos (USCHOLD e GRUNINGER, 1996). Assim, ela provê um vocabulário de termos específicos para a representação de conceitos presentes em um domínio de conhecimento e ainda cria uma rede de relacionamentos entre esses termos (GRUBER, 1995; STOFFEL *et al.*, 1997; GUARINO, 1998).

Diferentes tipos de ontologias podem ser desenvolvidas, de acordo com o seu grau de generalização, como mostrado na Figura 2 abaixo, retirada de (GUARINO, 1998):

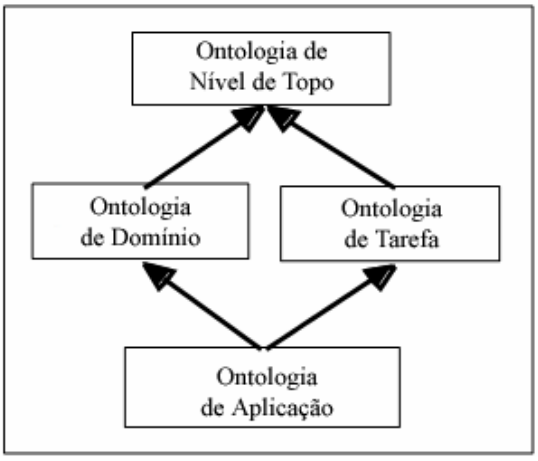


Figura 2 Tipos de ontologias, de acordo com seu nível de dependência de uma tarefa em particular ou de um ponto de vista. As flechas representam relacionamentos de especialização.

- Ontologias de nível de topo descrevem conceitos bem gerais, como espaço, tempo, matéria, objeto, evento, ação, etc., as quais são dependentes de um problema ou domínio particular.
- Ontologias de domínio e de tarefa descrevem, respectivamente, o vocabulário relacionado a um domínio genérico (como medicina ou música) e uma tarefa ou atividade genérica (como diagnose ou vendas), especializando-se os termos apresentados na ontologia de nível de topo.
- Ontologias de aplicação descrevem conceitos dependentes tanto de um domínio quanto de uma tarefa particulares, sendo, portanto, geralmente uma especialização de ambas as ontologias relacionadas. Esses conceitos

> geralmente correspondem a papéis desempenhados por entidades de domínios quando da execução de certa atividade.

Sistemas de informação mais recentes têm abordado o uso de ontologias, sendo chamados de Sistemas de Informação Guiados por Ontologias (*Ontology-Driven Information Systems*). Uma ontologia pode servir como base para a construção do sistema, em tempo de projeto, ou como base para cada um dos componentes de um sistema de informação, a serem listados: programas de aplicação, bancos de dados e interfaces de usuários (GUARINO, 1998).

Em bancos de dados, as ontologias podem servir como linguagens de definição de esquemas, podendo executar um papel importante na análise de requisitos e na modelagem conceitual. Uma ontologia também pode ser usada como um esquema conceitual global para a integração de informações (GUARINO, 1998). No entanto, não permitem fazer inferências sobre os dados, sendo assim consideradas estáticas. Mesmo assim, as ontologias associadas a um banco de dados permitem maior eficiência no processamento de consultas (SEIBEL et al., 2000).

Uma das linguagens mais utilizadas para a definição de uma ontologia é a OWL (*Web Ontology Language*), a qual foi utilizada neste trabalho para a criação de uma ontologia de aplicação. A definição dessa ontologia seguiu uma classificação para os conceitos, definida em (DORNELES, 2000), a qual considera os conceitos como léxicos ou não-léxicos. Conceitos léxicos são aqueles que podem ser diretamente representados por um computador como uma cadeia de bits (números, caracteres e assim por diante). Os conceitos não-léxicos não têm representação direta em um computador. Essa ontologia de aplicação é apresentada no capítulo 6.

Dentro do contexto de biologia molecular, as ontologias permitem um raciocínio automatizado sobre os dados biológicos por meio de sua rede de relacionamentos e também por meio de operadores mereológicos, os quais são utilizados para a análise de relacionamentos parte-todo. Assim, sistemas de *software* de inferência (motores de inferência) podem navegar pela rede de relacionamentos da ontologia para fornecer acesso a informações implícitas na ontologia.

A importância da adoção de ontologias em projetos genoma tem sido reconhecida pela comunidade científica, como pode ser notado em (SCHULZE-KREMER, 1998) ou mesmo em projetos como o da *Gene Ontology* (ASHBURNER *et al.*, 2000) e da *Sequence Ontology* (EILBECK *et al.*, 2005). Essas duas ontologias são do domínio de biologia molecular, sendo exemplos de esforços realizados por diversos grupos de trabalho para padronizar os termos

utilizados nesse domínio. Essas duas ontologias foram escolhidas para integrar a proposta deste trabalho e serão analisadas nas próximas seções.

3.3 Gene Ontology

A *Gene Ontology* (GO) foi criada com o objetivo de definir um único conjunto de termos que representasse todo tipo de dados sobre genes e seus produtos e que pudesse ser compartilhado por diferentes fontes de dados. Sua principal motivação foi a evidência observada após o seqüenciamento de diversos genomas de que uma grande fração de genes responsáveis por funções biológicas vitais é compartilhada entre todos os organismos eucarióticos. Portanto, analisar um determinado gene em um único organismo poderia ser determinante para o conhecimento de outros organismos. Assim foi criado o GO Consortium (ASHBURNER *et al.*, 2000).

O projeto GO envolve o desenvolvimento de três ontologias para descrever, respectivamente, função molecular, processo biológico e componente celular, e provê um banco de dados comunitário para apoiar o uso dessas ontologias. A função molecular descreve o que um produto de gene faz, em nível bioquímico (ex.: enzima, transportador). O processo biológico descreve um amplo objetivo biológico ao qual um produto de gene está relacionado, seja por participação direta ou por alguma contribuição no processo (ex.: crescimento e manutenção celular, metabolismo da pirimidina). O componente celular descreve a localização de um produto de gene dentro de estruturas celulares e complexos macromoleculares (ex.: complexo de Golgi, ribossomo). Essas classificações particulares foram escolhidas porque elas representam conjuntos de informações comuns a toda forma de vida e são básicas para anotação de informação sobre genes e produtos de genes. Essas três ontologias são independentes umas das outras. Portanto, um gene pode ser anotado em relação a qualquer uma das três, sem necessariamente ter que ser anotado nas demais ontologias.

O GO Consortium foi iniciado por cientistas associados a três bancos de dados de organismos: SGD (BALL *et al.*, 2000), banco de dados do Genoma *Saccharomyces*; FlyBase (THE FLYBASE, 1999), banco de dados do Genoma *Drosophila*; e MGD/GXD (BLAKE *et al.*, 2000), banco de dados do Genoma Rato. Atualmente, mais grupos participam do projeto. Cada um deles está anotando genes e produtos de genes utilizando termos da GO e incorporando essas anotações ao seu respectivo banco de dados.

Para fazer anotações de acordo com a GO, alguns atributos devem ser documentados, para prover suporte à anotação. Toda anotação deve ser atribuída a uma fonte (que pode ser uma referência literária, outro banco de dados ou uma análise computacional) e indicar os tipos de evidências que a citada fonte provê para permitir a associação entre o produto de gene e o termo da GO.

Os bancos de dados colaboradores do projeto GO enviam arquivos de anotação, delimitados por tabulações, contendo os campos descritos na Tabela 1.

Tabela 1 Tabela com os campos representados em um arquivo de anotação GO. A coluna 'Requerido?' indica se o campo é obrigatório ou opcional.

Coluna	Conteúdo	Requerido?	Exemplo
1	DB	obrigatório	SGD
2	DB_Object_ID	obrigatório	S000000296
3	DB_Object_Symbol	obrigatório	PHO3
4	Qualifier	opcional	NOT
5	GO ID	obrigatório	GO:0003993
6	DB:Reference (|DB:Reference)	obrigatório	PMID:2676709
7	Evidence Code	obrigatório	IMP
8	With (or) From	opcional	GO:0000346
9	Aspect	obrigatório	F
10	DB_Object_Name	opcional	acid phosphatase
11	DB_Object_Synonym (|Synonym)	opcional	YBR092C
12	DB_Object_Type	obrigatório	gene
13	taxon (|taxon)	obrigatório	taxon:4932
14	Date	obrigatório	20010118
15	Assigned_by	obrigatório	SGD

A primeira coluna, *DB*, se refere ao código de um dos bancos de dados associados ao projeto. A coluna cinco (5) indica a qual termo da GO a anotação está sendo associada. A coluna seis (6), *DB:Reference*, é a que indica quais as referências que provêem suporte à evidência descrita na coluna sete (7), *Evidence Code*. Esta coluna deve ser preenchida com um código pré-determinado que indique as possíveis evidências. No exemplo da tabela, IMP

significa "*Inferred from Mutant Phenotype*". A coluna nove (9) indica a qual das ontologias a anotação está sendo associada: P para processo biológico, F para função molecular e C para componente celular. Maiores descrições sobre os campos e seus conteúdos podem ser encontrados em (THE GENE ONTOLOGY, 2008).

O projeto GO ainda provê um banco de dados relacional com a definição da ontologia (com todos os seus termos e relacionamentos), além de anotações de produtos de genes de diferentes organismos. Esses dados são disponibilizados em diferentes formatos, entre eles o XML. Alguns sistemas de *software* foram desenvolvidos para a manipulação desses dados. Podem ser citados o *software* AmiGO, que permite a navegação sobre a ontologia para pesquisas sobre termos da ontologia; e o GOst, o servidor GO BLAST, que permite aos usuários submeter consultas e obter as biosseqüências e anotações GO de todos os produtos de genes similares no banco de dados GO (THE GENE ONTOLOGY, 2004).

3.4 Sequence Ontology

A *Sequence Ontology* (SO) surgiu do mesmo grupo de pesquisa do projeto GO, com colaboração dos mesmos projetos genoma e outros mais que foram se juntando ao grupo inicial. O escopo do projeto SO está na descrição de características e propriedades de uma biosseqüência. As características se referem a uma região da seqüência, tais como *éxon* e *íntron*; já as propriedades descrevem atributos dessas características.

A SO é um vocabulário controlado e estruturado para descrição precisa de anotações genômicas. Seus conjuntos de termos e definições visam facilitar a troca, análise e gerenciamento de dados genômicos. A SO trata relacionamentos do tipo parte-todo, possibilitando um raciocínio automatizado sobre dados descritos com ela. Além disso, alguns operadores mereológicos são aplicáveis às instâncias de dados. Cinco operadores mereológicos são definidos pela SO: sobreposição, disjunção, produto binário, diferença e soma binária. Exemplos de aplicação desses operadores podem ser encontrados em (EILBECK *et al.*, 2005).

Alguns trabalhos para a representação da SO como um formato de troca de dados ou como um componente integral de um banco de dados já foram ou estão sendo desenvolvidos. Entre eles, a última versão do *generic feature format* (GFF) (THE SEQUENCE ONTOLOGY, 2008), a GFF3, utiliza as definições de termos da SO para identificar as

anotações linha a linha. O formato GFF é um arquivo texto com suas estruturas separadas por linhas.

Outra aplicação mais elaborada para a representação da SO é a desenvolvida pelo projeto CHADO (GENERIC MODEL ORGANISM DATABASE, 2008a). Neste projeto, foi desenvolvido um esquema de banco de dados modular, com o objetivo de integrar dados moleculares e genéticos. O CHADO é parte do projeto *Generic Model Organism Database* (GMOD) (2008b). Um de seus módulos visa a representação de uma ontologia, com entidades para a representação de termos e relacionamentos entre os mesmos. Além disso, outro módulo compreende as características genômicas. A forma como o CHADO relaciona uma determinada característica anotada à ontologia é através de uma chave estrangeira identificando o termo na SO.

Um terceiro trabalho é o Chaos-XML (BERKELEY DROSOPHILA GENOME PROJECT, 2008a). Assim como o formato GFF3, Chaos-XML é um formato de arquivo que utiliza a SO para classificar e estruturar dados, mas está mais intimamente ligado ao projeto CHADO. Chaos-XML é um mapeamento XML hierárquico do esquema relacional CHADO. Sendo assim, não tem uma representação direta da ontologia sobre o documento XML.

Atualmente, a SO utiliza três tipos básicos de relacionamento entre seus termos: **kind_of** (especialização), **derives_from** (derivação) e **part_of** (composição). Esses relacionamentos estão definidos em (EILBECK, LEWIS *et al.*, 2005). O parágrafo seguinte baseia-se na Figura 3, encontrada em (EILBECK, LEWIS *et al.*, 2005).

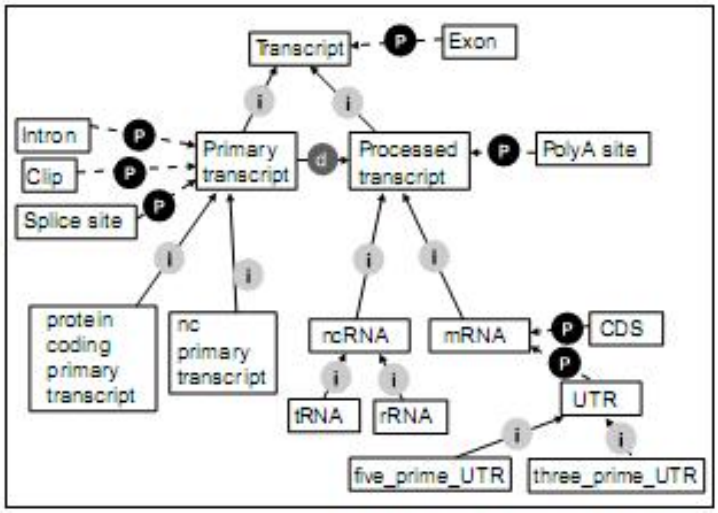


Figura 3 Parte da *Sequence Ontology* mostrando como os termos e relacionamentos são usados em conjunto para descrever um conhecimento sobre biosseqüências. O relacionamento **kind_of** é representado por setas marcadas com ‘i’; o relacionamento **part_of** é marcado com ‘P’; e o relacionamento **derives_from** é marcado com ‘d’ (EILBECK, LEWIS *et al.*, 2005).

Para melhor compreensão de como os programas de computadores poderão atingir níveis de inferência de conhecimento sobre um banco de dados, os sistemas de *software* baseados na SO (ou seja, que estiverem de acordo com suas especificações) precisam somente ser providos com uma versão atualizada da ontologia e tudo mais seguirá normalmente. Isso porque esses programas não precisam codificar o fato de que um tRNA é um tipo de transcrito (relacionamento **kind_of**); ele precisa simplesmente saber que relacionamentos **kind_of** são transitivos e hierárquicos e ser capaz de navegar internamente a rede de relacionamentos especificada pela ontologia para inferir logicamente esse fato.

3.5 Considerações Finais

Este capítulo abordou os principais conceitos que envolvem uma ontologia. Foram descritos os tipos de ontologias e como elas podem ser utilizadas dentro de sistemas de informação. Além disso, também foi abordado seu uso no domínio de biologia molecular, sendo descritas duas ontologias muito utilizadas por biólogos, a *Gene Ontology* e a *Sequence Ontology*.

O capítulo 4 aborda o uso do modelo semi-estruturado para a representação de dados provenientes de projetos genoma, assim como os aspectos relacionados à associação entre esse modelo de dados e ontologias.

4 MODELO DE DADOS SEMI-ESTRUTURADO PARA PROJETOS GENOMA

4.1 Considerações Iniciais

Projetos genoma produzem dados sobre biosseqüências e informações a elas associadas, ou seja, as suas anotações. Como descrito em (LEMOS *et al.*, 2003b), as biosseqüências são dados puramente textuais representando suas seqüências de nucleotídeos e aminoácidos. Porém, as suas anotações são heterogêneas, sem uniformidade em suas representações, sendo estas consideradas irregulares, ou ainda semi-estruturadas.

Neste capítulo são descritos sucintamente o modelo de dados semi-estruturado e sua representação em XML (seção 4.2), um modelo para a representação conceitual de objetos para XML (seção 4.3), assim como o suporte do modelo de dados semi-estruturado ao armazenamento e gerenciamento de dados produzidos em projetos genoma (seção 4.4).

4.2 XML e XML Schema

A XML (*eXtensible Markup Language*) é uma linguagem de marcação simples e de fácil portabilidade, para documentos que contenham informações estruturadas ou semi-estruturadas. Atualmente, já existem muitos sistemas de *software* para sua criação, manipulação e visualização.

A principal característica da XML é que um usuário é livre para criar suas próprias marcações, o que possibilita a representação de qualquer abstração real (seja ela um objeto tal como um organismo vivo, ou mesmo um processo tal como procedimentos de pesquisa). Documentos XML podem conter semântica extremamente rica relacionada aos seus dados, uma vez que o esquema dos dados está embutido com o próprio dado, tornando fácil a sua compreensão e intercâmbio. A XML tornou-se padrão para a representação de dados semi-estruturados devido a algumas de suas características, tais como flexibilidade para modificação de esquema, interconexão de fontes através de *links*, fácil compreensão e por se tornar um "*framework*" para a definição de especificações-padrões (GUERRINI e JACKSON, 2000; MELLO *et al.*, 2000; ACHARD *et al.*, 2001; BRAGANHOLO e HEUSER, 2001).

Alguns de seus principais usos atualmente são a integração e a comunicação de diferentes sistemas de *software* e também a criação de bancos de dados semi-estruturados.

Nos bancos de dados relacionais, cada relação (ou tabela) é composta por atributos, tuplas representando um conjunto de valores para os atributos (ou seja, são os dados armazenados no banco), e participa de relacionamentos com outras relações. O esquema da relação é que determina sua estrutura e suas características próprias, enquanto o esquema do banco de dados é um conjunto de esquemas de relações mais o conjunto de restrições de integridade. Durante a fase de projeto do banco de dados, um esquema de relação deve definir restrições de atributos como, por exemplo, seu tipo, tamanho, valor padrão, etc., fazendo com que todo dado armazenado siga a estas regras.

Da mesma forma, um banco de dados XML nativo deve definir restrições aos dados a serem armazenados para que estes sejam coerentes e mantenham a consistência do banco de dados. Um documento XML pode ter seu esquema definido através da linguagem XML Schema, a qual provê meios para a definição de sua estrutura, seu conteúdo e sua semântica. Essas definições (regras) de armazenamento de dados são fundamentais para os bancos de dados, uma vez que conferem integridade e facilitam a manutenção dos dados. Assim, bancos de dados XML nativos utilizam XML Schema para a composição de suas estruturas e regras para a persistência dos dados.

Um banco de dados XML nativo é um conjunto de documentos XML persistentes que podem ser manipulados. Há duas orientações possíveis a um documento XML, segundo (GRAVES, 2003), orientado ao processamento de texto ou ao processamento de dados. Documentos XML orientados ao processamento de texto são usados por sua habilidade em capturar linguagens naturais (humanas) como em manuais de usuários e páginas *Web*. Os orientados ao processamento de dados são utilizados principalmente para a transferência de dados. Eles são caracterizados por estruturas altamente regulares que se repetem muitas vezes. Nestes, é importante definir a estrutura física dos documentos, ordem e composição dos elementos quando se prioriza o desempenho no processamento de consultas do banco de dados.

Bancos de dados XML operam com um conjunto de documentos orientados ao processamento de dados, onde as operações fornecidas são mais destinadas à manipulação de dados do que ao processamento de texto. Um SGBD XML nativo deve fornecer acesso direto aos documentos XML e a trechos dele (i.e., elementos ou atributos), e a possibilidade de consulta a eles. Usar um SGBD XML é especialmente apropriado para a captura de uma área

com relacionamentos hierárquicos complexos, como nos bancos de dados científicos, em grandes empresas e em sistemas de produção (GRAVES, 2003).

Geralmente os documentos XML são modelados segundo grafos direcionados rotulados, os quais contêm vértices representando objetos identificáveis e arestas que conectam um objeto a outros que componham sua estrutura, a qual é hierárquica (MELLO *et al.*, 2000). Outras abordagens estendem o modelo Entidade-Relacionamento (ER), como a ERX (PSAILA, 2000), a XER (SENGUPTA *et al.*, 2003) e a XSEM (NECASKY, 2007). Porém, este trabalho faz uso de uma parte da *Asset Oriented Modeling* (ASSET ORIENTED MODELING, 2008) para a modelagem conceitual de esquemas XML, um método de modelagem desenvolvido especialmente para documentos XML, pois captura suas principais características. A vantagem da AOM sobre modelos tradicionais, como o Diagrama Entidade-Relacionamento (DER) ou a *Unified Modeling Language* (UML) (OBJECT MANAGEMENT GROUP, 2006), é que, ao final, ela produz uma modelagem quase direta em relação ao documento XML.

A Figura 4 apresenta um exemplo de documento XML com dados sobre seqüências de nucleotídeos do DNA de um camarão, e a Figura 5 apresenta um exemplo de documento XML Schema, o qual valida os dados contidos no documento XML. Esses exemplos são resumos adaptados do projeto genoma EST do camarão *Litopenaeus vannamei* (SHEST, 2008).

```
<?xml version="1.0" encoding="UTF-8"?>
<sequence xmlns = "http://shrimp.dc.ufscar.br"
  xmlns:xsi = "http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation = "http://shrimp.dc.ufscar.br/sequence.xsd"
  ID="1" name="COORD-MC2-090304-1-001-F06.b">
        <fasta>
        AAATGGTCTAGAAAGCTTCTCGAGGGCCGAGGCGGCCGACATGTNATTAT
        CTTTTATTCTTCTCAATTTTTTCCACTTGATTGTAGTATATATCTTTGAA
        TCCAAGAAAACGGCAATGTGGAGGCCATTACTCATCTGGGACTTCAATCT
        CCTCCGCCGTAATGGCCGAATTCCCGGGCATATGTCCGGTACCGTCGACT
        GATAACTTCGTATAATGTATGCTATACCAATTTATGCGGCCGATTCCTCC
        CTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCTACCGCCCT
        ATAGTGAGTCGTATTACAGATCTACTGGCCGTCGTTTTACACGCGTAAGG
        CGGTAATACGGTTATTCACANAATCAGGGGATAACGCAAGAAAGAACATG
        TGAGCAAAAGGGCATCAAGAGGTCACGAACCGTAAAAAGGCCGCGTTGCT
        GGCGTATTACCATAGGCTCCGCCCCCCTGAGGAGCCTCCCAAAAATCGAC
        GCTCAAGTTACAAGTGGCGAAACCCGACAGGACTATAAAGATTCCAGGCG
        TTTCCCCCTGGAAGCTCCCTCGTACGCTCTCCTGTTCCGACCCTGCCGGT
        TACCGGATACCTGTTCGGCTTTATTCGCTTCGGGAAGGGTGGCGCTTTCT
        CATAGCCCACGCCCTAAGGATCTCCATTGCGGGGTAGGCGTTTCGGCCAA
        AGCCGGGCTGTGTGTACGGAACCCCAGCTTTA
        </fasta>
        <hq_start>66</hq_start>
        <hq_end>361</hq_end>
        <length>732</length>
        <date>090304</date>
        <experiment>1</experiment>
        <board>1</board>
</sequence>
```

Figura 4 Documento XML contendo dados de biosseqüências. Este documento é validado por um documento XML Schema definido em *xsi:schemaLocation*.

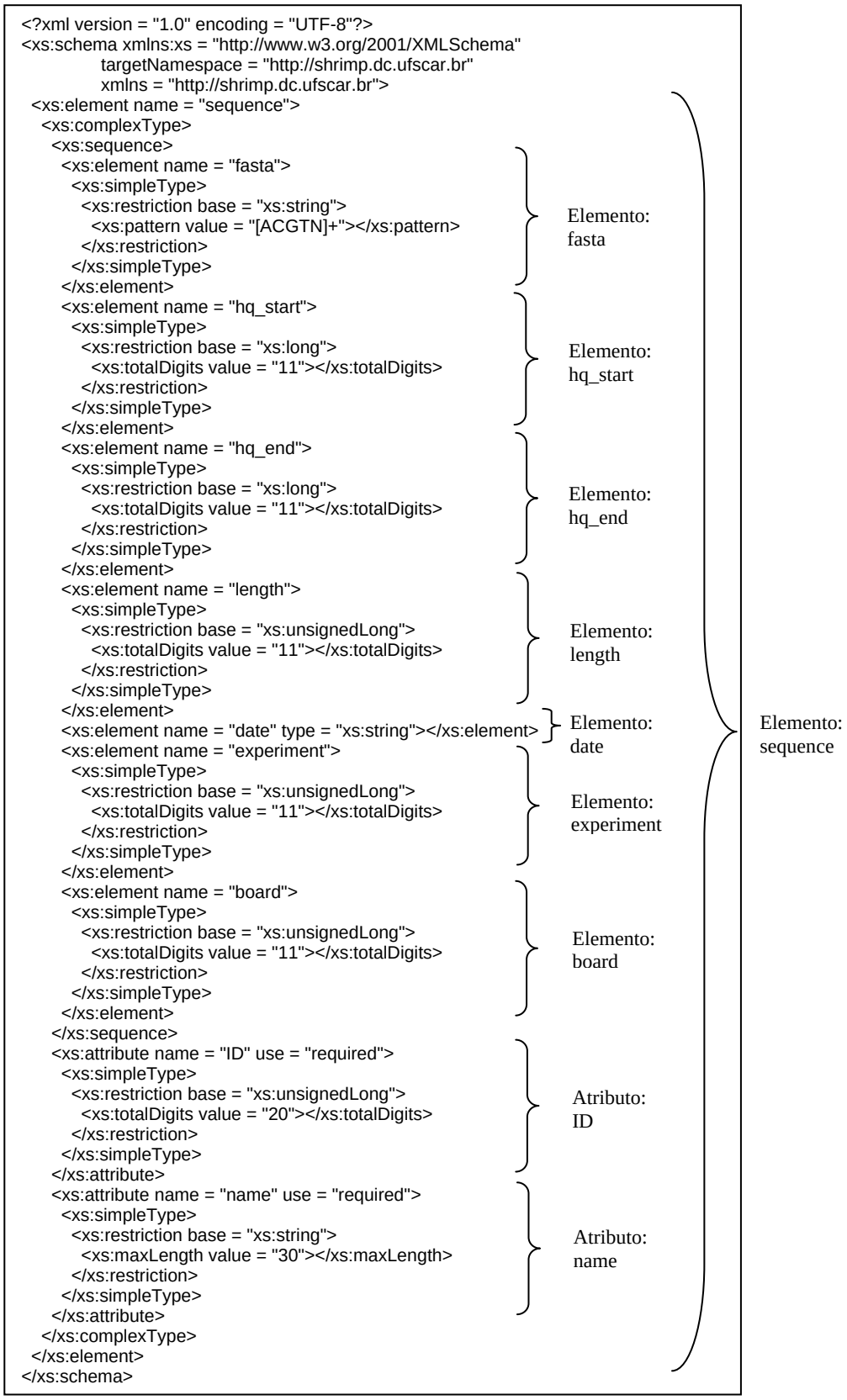


Figura 5 XML Schema responsável por validar os dados contidos em um documento XML. Determina a estrutura do documento e o tipo do seu conteúdo.

A Figura 4 apresenta o elemento ‘sequence’ como raiz do documento XML, o qual tem os atributos ‘ID’ e ‘name’, além de atributos de *namespace* e instância (‘xmlns’, ‘xmlns:xsi’ e ‘xsi:schemaLocation’). Para uma determinada ‘sequence’, esse documento representa ainda os elementos ‘fasta’, ‘hq_start’, ‘hq_end’, ‘length’, ‘date’, ‘experiment’ e ‘board’.

Ao se definir um documento de esquema para essa instância de documento XML, é preciso que cada um desses elementos e atributos seja representado. Assim, na Figura 5, é possível verificar que há definições correspondentes a cada um deles. Apenas o elemento ‘sequence’ contém sub-elementos e atributos (ele é composto por todos os outros elementos e atributos). Para cada elemento e atributo é possível definir restrições de domínio dos dados a serem armazenados, como o seu tipo de dado, padrões, limite de tamanho, valor máximo e mínimo, entre outros. Como exemplos, o elemento ‘fasta’ é do tipo *string* e deve obedecer ao padrão definido pela expressão regular [ACGTN]+; e o elemento ‘hq_start’ é do tipo *long* e composto por 11 (onze) dígitos.

Muitas especificações de padrões de comunicação têm sido criadas na Bioinformática, e a XML é o principal meio para desenvolver tal tarefa. Esses padrões têm como principal objetivo auxiliar o gerenciamento e a troca de informações entre pesquisas de genoma.

Alguns exemplos de seu uso na área biológica podem ser citados: RNAML (WAUGH *et al.*, 2002), criada para a representação de informações sobre RNA; e a *BIOpolymer Markup Language* (PROTEOMETRICS, 2008), que permite anotações complexas de informações de seqüências de nucleotídeos e de aminoácidos. Outros projetos buscaram na XML a solução para facilitar a transferência e publicação de informações biológicas. Assim surgiu a linguagem GAME (BERKELEY DROSOPHILA GENOME PROJECT, 2008b), com o intuito de promover trocas de dados entre membros do Projeto *Genoma Berkeley Drosophila* e Celera. Atualmente ela é utilizada como um padrão para a anotação de características de biosseqüências. Além dessas aplicações, a XML também tem sido utilizada para a representação de ontologias, inclusive a *Gene Ontology* (para mapeamento de seus conceitos e relacionamentos) e *Sequence Ontology* (projeto Chaos-XML, citado na seção 3.4).

Na seção seguinte a AOM é descrita, a qual é especialmente utilizado para a representação de dados em XML.

4.3 Asset Oriented Modeling

A Asset Oriented Modeling (AOM) foi criada com o objetivo de representar informações complexas, sendo especialmente expressiva para a representação de dados semi-estruturados. Algumas características importantes de seu modelo são:

- Abordagem única para entidades e relacionamentos (*assets*);
- Suporte a relacionamentos de maior ordem (relacionamentos entre relacionamentos);
- Suporte a estruturas de dados complexas baseadas em gramáticas regulares;
- Suporte a *namespaces* e visões de projeto.

Diferentemente de abordagens tradicionais, como o Modelo Entidade-Relacionamento, que modela substantivos como entidades (ou propriedade) e verbos como relacionamentos, esse modelo considera ambos *assets*. Dessa forma, relacionamentos clássicos e entidades são tratados da mesma forma, levando a uma considerável simplificação do modelo conceitual. Isso permite a representação de relacionamentos entre relacionamentos, os quais são chamados relacionamentos de maior ordem. Os relacionamentos de maior ordem são conceitos importantes em sistemas de informação, mas não podem ser apropriadamente representados por modelos como o ER e a UML. A AOM adotou relacionamentos de maior ordem da HERM (*Higher Order Entity Relationship Model*), de Bernhard Thalheim (THALHEIM, 2000). O conceito de *assets* é originário da RDF (*Resource Decription Framework*) (W3C, 2008).

Para cada *asset*, é possível definir propriedades complexas, as quais podem conter outras propriedades aninhadas e até mesmo utilizar sintaxe de expressão regular para determinar ordem e cardinalidade. A definição de propriedades com estruturas complexas permite que modelos AOM sejam geralmente muito menores que equivalentes em UML ou ER. A sintaxe de expressão regular permite a definição de estruturas de complexidade arbitrária. A possibilidade de definir tipos complexos (que também são *assets*) e usar essas definições de tipos juntamente com a definição de propriedades o torna mais compacto e permite o reuso de definições.

Outra característica da AOM é a adoção de identificadores de *namespaces* globais para modelos e *assets*, os quais passam a ser unicamente identificados. Isso permite criar um novo modelo a partir da fusão com modelos já existentes. Além disso, esses identificadores

únicos também possibilitam a separação de modelos em visões e o desenvolvimento distribuído dos mesmos.

A definição de *assets* (Figura 6.a) inclui um nome identificador, uma etiqueta, chaves e propriedades associadas ao *asset*, restrições, operações e anotações. Assim como em linguagens orientadas a objetos, onde classes abstratas podem ser definidas, a AOM também apresenta o conceito de *assets* abstratos. Tipicamente, definem um tipo criado pelo usuário e não podem ser instanciadas. *Assets* sem etiquetas devem ter instâncias com o mesmo nome do *asset*. Para um *asset* abstrato, sua etiqueta deve ser cinzenta (Figura 6.b).

A AOM Nível 2 introduz o conceito de Objeto de Negócio (ON), que é um objeto maior que agrega alguns *assets*. Os ONs estão diretamente ligados a objetos do mundo real e permitem um modelo intuitivo e natural do domínio do problema. Eles também são importantes guias para quando as estruturas de informação são implementadas em XML. Em geral, cada ON corresponde a um tipo de documento XML. Em particular, resulta em um modelo de fácil transformação para XML.

Em um *asset*, a definição de propriedades pode indicar: (i) propriedades aninhadas; (ii) seqüências (ordenadas ou não) e alternativas; (iii) repetições; e (iv) recursões. Arcos formam um relacionamento entre os *assets* e uma fronteira agrupando *assets* identifica um ON, o qual deve ter a mesma identidade do *asset* principal do conjunto que ela agrupa (Figura 7).

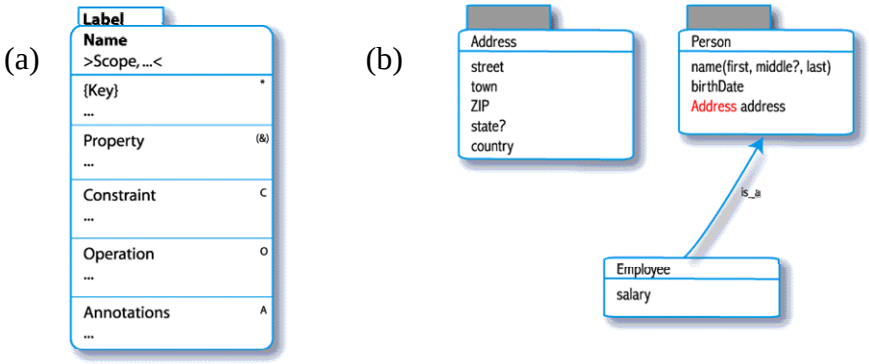


Figura 6 (a) Representação de um *Asset*; (b) Etiqueta cinza indica um *asset* abstrato, enquanto sem etiqueta indica que se referencia ao mesmo nome do *asset*.

Neste trabalho, algumas das definições opcionais não são consideradas, como as restrições, operações e anotações de um *asset*. Dentre os conceitos tratados para a modelagem de esquemas XML estão: *assets*, ONs, arcos e propriedades. Esses conceitos são abordados de forma simplificada em relação à original, forma a qual é detalhada no capítulo 7.

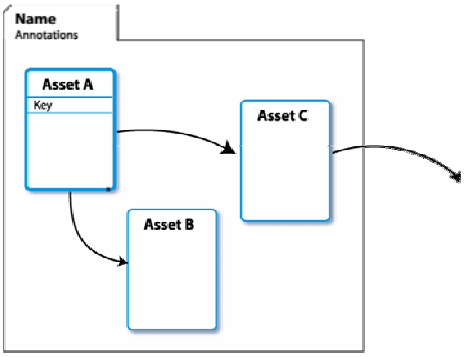


Figura 7 Objeto de Negócio encapsulando um conjunto de *assets* inter-relacionados.

A seção seguinte apresenta a motivação para o uso de bancos de dados semi-estruturados em projetos genoma.

4.4 Bancos de Dados de Genoma

Atualmente, a adoção de SGBDs em projetos genoma tem sido mais freqüente devido ao grande volume de dados com que lidam esses projetos e às facilidades de gerenciamento, consulta e atualização das informações. A seção 4.4.1 descreve os modelos de dados utilizados em BDBMs e a seção 4.4.2 aborda sucintamente alguns aspectos da integração desses dados.

4.4.1 Modelos de Dados

O modelo relacional é o mais utilizado, porém não facilita a compreensão do objeto biológico, podendo gerar um esquema com muitas tabelas. Apresenta ainda outros fatores negativos para a sua adoção em BDBMs, como a falta de flexibilidade para a evolução de esquemas e a falta de suporte para representação de atributos com múltiplos tipos de dados (SEIBEL et al., 2000). O modelo orientado a objetos consegue representar melhor objetos biológicos do que o relacional, pois permite mapeamento direto de conceitos complexos do mundo real em estruturas de dados do modelo. Algumas vantagens desse modelo são o conhecimento do objeto de forma completa e a coleção de métodos e de estruturas para

modelar, manter e consultar os dados. No entanto, uma de suas limitações reside em suas estruturas de dados fixas, fator que pode ser problemático para a evolução de esquemas, uma vez que pode acarretar na alteração da estrutura utilizada e mesmo na reprogramação dos métodos já implementados. Em (SEIBEL et al., 2000) é possível encontrar um estudo ressaltando as principais características, vantagens e desvantagens, na utilização de diferentes modelos de dados para a representação de dados da biologia molecular. Neste contexto, destaca-se o modelo semi-estruturado, por razões detalhadas a seguir.

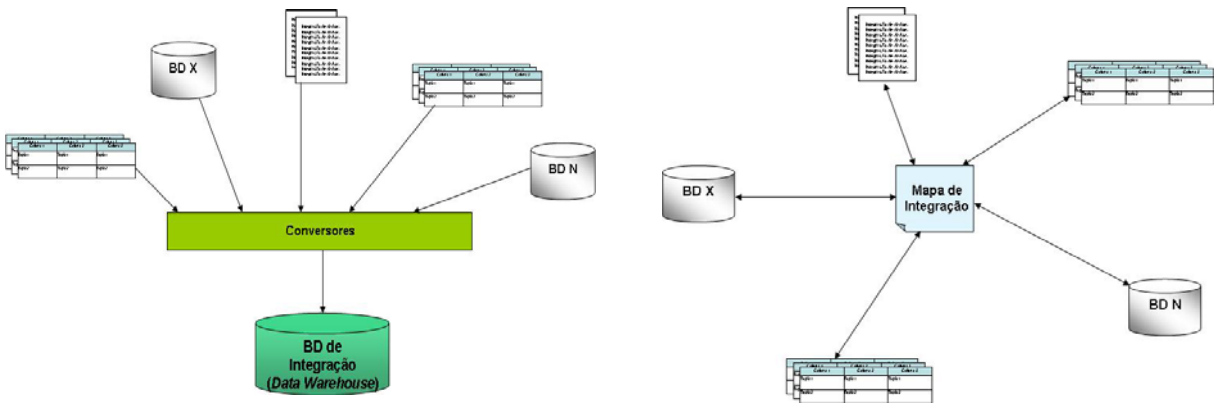
Lemos (LEMOS *et al.*, 2003b) observa que várias aplicações de Bioinformática armazenam seus dados em formatos não padronizados. Além disso, considerando-se a integração de fontes de dados públicas, observa-se também que tais fontes não podem ser controladas, mas que, no entanto, ao se conhecer suas estruturas, podem ter seus dados mapeados para um formato comum. Geralmente, os dados destas aplicações apresentam estrutura irregular, dos quais alguns objetos apresentam atributos omitidos, outros podem ter várias ocorrências de um mesmo atributo, um mesmo atributo pode ser de diferentes tipos em diferentes objetos e informações semanticamente relacionadas podem estar representadas diferentemente em vários objetos. Os dados com estas características são chamados de semi-estruturados.

O modelo semi-estruturado apresenta flexibilidade para a evolução de esquemas, uma vez que não necessita modificar dados já armazenados, além de facilitar a transferência e integração de dados entre laboratórios e a compreensão dos usuários. Assim, o modelo semi-estruturado pode se tornar uma opção muito conveniente para BDBMs. Contudo, esse não deve ser o único ponto para sua adoção em um BDBM. Questões como consultas semânticas e mais eficientes também devem ser consideradas, sendo esse um tema bastante discutido atualmente, e que esse modelo também se propõe a resolver com o apoio de outras abordagens como, por exemplo, o uso de ontologias.

Essas são algumas das razões que motivaram a escolha pelo modelo de dados semi-estruturado neste trabalho, sendo, para tanto, utilizado um banco de dados XML nativo para o armazenamento de biosseqüências e suas anotações.

4.4.2 Integração de Dados de Projetos Genoma

A integração de dados de diversas pesquisas que ocorrem simultaneamente é o maior desafio aos pesquisadores, principalmente pelo fato de que cada pesquisa atua em um determinado domínio de conhecimento da biologia e geralmente as armazenam em modelos de dados diferentes. Como a integração de BDBMs tem sido de grande relevância para novas descobertas biológicas, uma série de trabalhos vêm sendo realizados com esse objetivo. Existem trabalhos que realizam a integração materializada de diferentes fontes de dados (LEMOS *et al.*, 2003a; OLIVEIRA, 2005), convertendo-os para um formato de estrutura comum e armazenando-os em um único banco de dados (geralmente um DW) (LEMOS *et al.*, 2003b). Essa forma de integração é representada na Figura 8.a. Outra forma é a integração virtual, utilizada por (GOPALACHARYULU *et al.*, 2005) e também pelo banco de dados InterPro (APWEILER *et al.*, 2001), para os quais foram criados mapeamentos entre diferentes fontes de dados, onde cada entrada referente a uma fonte de dados indica os seus equivalentes em outras fontes. Esse tipo de mapeamento é representado na Figura 8.b.



(a) (b)

Figura 8 Modelos de integração de dados pós-esquema. Em (a), integração materializada por meio de um Data Warehouse. Em (b), integração virtual por meio de mapeamento.

Essas duas formas de integração citadas são abordagens utilizadas quando as fontes de dados são heterogêneas, com modelos e esquemas de dados diferentes. Uma terceira abordagem em BDBMs, chamada de esquema-unificado, considera o compartilhamento de um mesmo esquema entre um conjunto de bancos de dados homogêneos. Neste caso, os bancos de dados são projetados sobre um mesmo esquema, como, por exemplo, o esquema de

banco de dados modular CHADO (GENERIC MODEL ORGANISM DATABASE, 2008a), o qual é utilizado por diferentes projetos genoma. A Figura 9 apresenta esse tipo de integração.

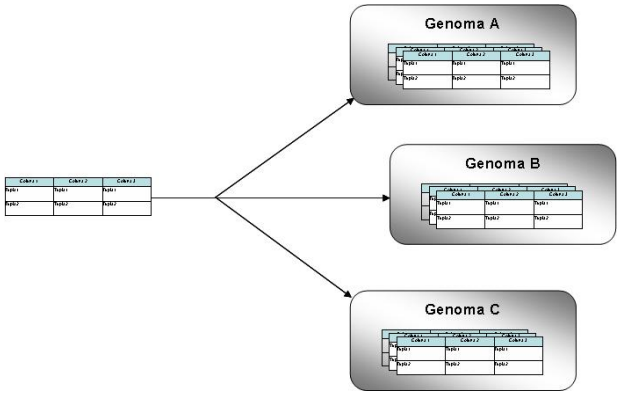


Figura 9 Modelo de integração esquema-compartilhado.

A abordagem esquema-unificado tem como vantagem a padronização de termos, os quais terão a mesma semântica para todo o conjunto de bancos de dados, tornando a integração e o compartilhamento de dados mais fácil. Contudo, os esquemas perdem em questão de autonomia, pois todos os bancos de dados devem ter exatamente o mesmo esquema, podendo levar a situações nas quais os esquemas não sejam tão adequados para suas aplicações. Para os BDBMs, essa característica não é bem aceita, pois projetos genoma de um mesmo domínio não necessariamente trabalham com o mesmo conjunto de anotações. Tais projetos tanto podem conter anotações semelhantes quanto complementares. Além disso, a evolução de esquemas também pode se tornar algo complexo, pois isso pode requisitar a reformulação do esquema base para manter a integração dos esquemas.

Assim como em qualquer banco de dados, também é importante que o esquema de um banco de dados XML seja projetado de acordo com os requisitos da aplicação, pois isso será determinante para que ele tenha um bom desempenho. Quando aplicado a um conjunto de BDBMs, os quais necessitem compartilhar seus dados e manter sua própria autonomia, se torna interessante criar esquemas XML estruturalmente diferentes, mas semanticamente integrados.

Com o intuito de integrar semanticamente fontes de dados é que ontologias têm sido usadas como esquemas conceituais em sistemas de informação, como descrito no capítulo 3. Dentre esses sistemas, há aqueles que utilizam uma abordagem com uma única ontologia global (YIGAL *et al.*, 1998), com múltiplas ontologias locais para cada fonte de dados (MENA *et al.*, 1996) e híbrida, com múltiplas ontologias locais e uma ontologia global à qual

são mapeadas todas as locais (VISSER *et al.*, 2000; BERGAMASCHI *et al.*, 2001). As formas de representação da ontologia também variam nesses sistemas, podendo ser encontradas em forma de linguagens de representação próprias com regras e restrições (GENESERETH *et al.*, 1997), *thesaurus* (BERGAMASCHI *et al.*, 2001), por meio de linguagens como a OWL, entre outros. Esses sistemas, no entanto, são empregados na integração de fontes de dados totalmente independentes e heterogêneas estruturalmente e semanticamente, quando então é necessário estudar suas estruturas e mapeá-las para um esquema comum.

Considerando-se todas essas questões abordadas, este trabalho propõe um modelo de integração de dados semi-estruturados que considera a geração de esquemas XML a partir de uma ontologia. Esse modelo é chamado *conceito-compartilhado* e visa facilitar a integração de esquemas de dados semi-estruturados, permitindo a construção de esquemas com diferentes estruturas, mas que mantenham a mesma semântica. O modelo *conceito-compartilhado* é apresentado no capítulo 6.

4.5 Considerações Finais

Neste capítulo foram descritos os aspectos principais do modelo de dados semi-estruturado e seu uso em bancos de dados. Assim, foi feita uma breve introdução à XML, linguagem de representação de dados semi-estruturados, e à XML Schema, linguagem para a definição de esquemas de documentos XML. Também foram contextualizados os modelos de dados utilizados por BDBMs, levando à motivação para o uso do modelo de dados semi-estruturado neste trabalho. Além disso, foram descritas algumas abordagens para a integração de dados biológicos, inclusive algumas que fazem uso de ontologias.

O capítulo 5 detalha alguns ambientes de anotação encontrados na literatura, os quais são comparados segundo critérios bem definidos. Dentre esses critérios, foram levantados qual o tipo de formalização de conceitos e qual modelo de dados são utilizados para a representação e armazenamento das anotações.

5 AMBIENTES DE ANOTAÇÃO

5.1 Considerações Iniciais

Ambientes de anotação são sistemas de *software* que agregam ferramentas de Bioinformática para o processamento, o armazenamento, a recuperação e a análise de anotações de projetos genoma, disponibilizando, assim, um conjunto de funcionalidades para que os pesquisadores realizem suas tarefas. As anotações de projetos genoma podem ter origem em bancos de dados externos (anotação importada), na execução de ferramentas automáticas (anotação automática) e na observação de pesquisadores (anotação manual).

O capítulo 5 descreve brevemente alguns ambientes de anotação estudados: suas arquiteturas e funcionalidades. Ao todo são 14 ambientes de anotação (seções 5.2.1 a 5.2.12). A seção 5.3 apresenta uma tabela comparativa entre eles, onde são verificados o modelo de dados, os tipos de anotação, se realizam ou não integração de dados e se usam vocabulário controlado ou ontologias para a representação de seus termos.

5.2 Ambientes de Anotação de Projetos Genoma

5.2.1 Bio-TIM

O Bio-TIM (*Bioinformatics – Transparent Information Management*) é um ambiente de anotação desenvolvido pelo Grupo de Banco de Dados (GBD) do Departamento de Computação (DC) da Universidade Federal de São Carlos (UFSCar). Sua arquitetura é composta por um *framework* de componentes de Bioinformática, chamado FrameEST (LOMBARDO, 2006), e por um DW para a integração de fontes de dados heterogêneas (OLIVEIRA, 2005).

O FrameEST é baseado no padrão MVC (Model-View-Control) e foi desenvolvido com o objetivo de construir componentes reutilizáveis por diferentes aplicações do domínio de biologia molecular. Dessa forma, diversas ferramentas de Bioinformática foram

componentizadas dentro desse *framework*, como o Phrap, o Phred e o BLAST. Esses componentes são instanciados para a composição de um *pipeline* para processamento e análise de cromatogramas.

O DW utilizado pelo Bio-TIM foi desenvolvido sob o modelo relacional. As anotações são importadas de fontes de dados externas e materializadas no DW, após passarem por tradutores responsáveis por sua normalização. Além disso, ele também pode armazenar as anotações automáticas produzidas pelas ferramentas de Bioinformática e as anotações manuais produzidas pelos biólogos.

Entretanto, o ambiente Bio-TIM ainda apresenta uma série de restrições, as quais tornam a funcionalidade de anotação bastante incompleta e parcial. Dentre elas está o fato de não apoiar a anotação manual com vocabulários controlados ou ontologias do domínio de biologia molecular. Outras restrições estão relacionadas ao modelo de dados utilizado, o relacional, e o tipo de dados tratados em projetos genoma, como anteriormente descrito na seção 4 do capítulo 4. Além disso, não há nenhum componente que possa prover padronização e semântica aos seus esquemas, dificultando a integração de seus dados com outros sistemas.

5.2.2 Apollo

Apollo (LEWIS *et al.*, 2002) é um sistema de anotações genômicas que permite aos pesquisadores explorarem anotações sob diversos aspectos e criar suas próprias anotações. Seu principal objetivo é permitir que especialistas conectem e comparem anotações (automáticas e manuais) com dados biológicos de forma completa para então validá-las ou não. Esses especialistas são chamados curadores.

Os curadores podem criar e modificar anotações de genes. Toda modificação realizada gera uma versão diferente da anotação, mantendo-se um histórico sobre elas. Para a anotação são utilizados campos com vocabulário controlado e com descrição livre.

O modelo de dados do Apollo é relacional, separado em duas visões: seqüências e características de seqüências. Aceita tanto anotações automáticas, importadas e manuais. O projeto *Generic Model Organism Database* (2008b) adotou o modelo do Apollo para o módulo de anotação.

Além disso, o Apollo oferece diversas ferramentas de Bioinformática para anotação e

também para visualização e navegação das anotações criadas. Diferentes níveis de detalhes são oferecidos ao pesquisador, um recurso que permite explorar profundamente uma seqüência.

As anotações importadas de fontes de dados públicas podem estar nos formatos GAME XML, GFF, GenBank e *flat-files*.

5.2.3 ASAP

O ASAP (*Alternative Splicing Annotation Project*) (GLASNER *et al.*, 2003) foi desenvolvido para armazenar anotações de todo o processo de seqüenciamento, portanto anotações importadas, automáticas e manuais são aceitas. Para tanto, um banco de dados relacional baseado no SGBD MySQL e uma interface *Web* foram desenvolvidos, os quais têm como objetivo armazenar, atualizar e distribuir os dados de seqüências de genoma e caracterizações funcionais. Suas anotações são organizadas em função da caracterização genética e bioquímica.

Este sistema provê suporte para três tipos de usuários:

- Usuários públicos, que podem somente navegar pelas anotações via interface *Web*; não lhes é permitido criar novas anotações;
- Usuários anotadores, que estão associados a projetos genoma e por isso têm a permissão para anotar novas observações sobre os dados genômicos; e
- Usuários curadores, que fazem a validação de todas as anotações realizadas para garantir a qualidade dos dados disponibilizados publicamente.

As anotações podem tanto conter campos com descrições livres como outros que seguem um vocabulário controlado. Entre as anotações estão indicações do autor da anotação e se a mesma foi ou não curada. Também é mantido um histórico das anotações, promovendo assim o versionamento das anotações realizadas.

5.2.4 BASys

BASys (*Bacterial Annotation System*) (DOMSELAAR *et al.*, 2005) é um servidor *Web* com apoio a anotações detalhadas e automáticas de seqüências de genomas de bactérias. Ele aceita dados de seqüências de nucleotídeos (DNA) e uma lista opcional informando a identificação do gene, e provê saídas de anotações textuais e imagens com *links*. BASys utiliza mais de 30 programas para determinar aproximadamente 60 subcampos de anotação para cada gene, incluindo o nome do gene/proteína, função dentro da *Gene Ontology*, possíveis parálogos e ortólogos, peso molecular, dentre outros.

Este sistema é composto por três módulos: (i) uma interface *Web* para submissão de dados genômicos, agendamento de anotações e monitoramento ou reportagem do progresso da anotação; (ii) um módulo de anotação para análise de dados cromossômicos e geração de anotações, as quais são armazenadas em um banco de dados relacional; (iii) um sistema de reportagem e apresentação de diversos gráficos, HTML e saídas textuais produzidas pelo BASys (GLASNER, LISS *et al.*, 2003).

Ele realiza uma busca em diversas fontes de dados com o intuito de determinar o máximo possível de anotações para cada gene. Sempre que possível, provê informações da fonte de anotação, a evidência utilizada para apoiar a anotação e uma indicação de sua qualidade. Essas informações são úteis para rastrear possíveis erros.

No entanto, o sistema BASys não fornece apoio à anotação manual, restringindo-se somente às anotações importadas e automáticas.

5.2.5 BioNotes

O BioNotes (LEMOS *et al.*, 2003a) é um sistema de anotação de biosseqüências dirigido por ontologias que permite a um pesquisador criar, recuperar e analisar anotações de biosseqüências, com o objetivo de seqüenciar genomas completos de DNA ou ESTs. Uma ontologia é utilizada para auxiliar o pesquisador a montar seu *workflow* de atividades, guiando-o na escolha da composição de ferramentas de anotação automáticas a ser utilizada dentro do sistema.

Seus desenvolvedores pesquisaram todas as necessidades às quais um sistema de anotação deveria se adequar para melhor oferecer suporte aos biólogos, e criaram o BioNotes com a intenção de oferecer todas as funcionalidades desejadas. Sendo assim, o BioNotes é composto por ferramentas de pesquisa, navegação por *hyperlinks*, para localização e tabulação de anotações. Além disso, ele possui recursos para versionamento de anotações, acesso seguro distribuído e facilidades para controlar programas de análise e para especificar anotações a armazenar. Isso faz dele um sistema bem completo.

Também foram pesquisados os possíveis paradigmas de bancos de dados a serem usados em um projeto Genoma. Os autores fizeram uma análise comparativa entre esses paradigmas, seus prós e contras. Assim, escolheram o modelo semi-estruturado, baseado em documentos XML, para o seu sistema. Esse modelo está implementado sobre o modelo relacional no SGBD Oracle 9*i*, com colunas que armazenam dados do tipo XML.

Todos os tipos de anotações (automáticas, importadas e manuais) são armazenados em um DW XML, chamado Bio-AXS (SEIBEL, 2002). Para cada uma das fontes de dados públicas é criado um esquema XML (chamado de esquema local) para definir seus dados e, assim, serem armazenados no DW. Em seguida, com base em uma ontologia de aplicação, esses esquemas locais são integrados em um esquema global, com o mapeamento entre conceitos aplicados nas diferentes fontes. Algumas dessas fontes de dados já disponibilizam seus próprios dados em XML, num formato próprio, o qual é aproveitado. Somente é criado um esquema XML local para aqueles que não disponibilizam seus dados em XML. Da mesma forma, para todos os programas de análise que compõem o BioNotes é criado um esquema XML para armazenar os dados gerados por eles. Quanto às anotações manuais, elas também são validadas por um esquema XML, onde algumas das anotações seguem um vocabulário controlado ou ontologias, como a GO, enquanto outras podem ser descrições livres do pesquisador.

5.2.6 ERGO

O ERGO (OVERBEEK *et al.*, 2003) é um sistema para análise de genoma que integra dados biológicos de diversas fontes para alcançar uma análise compreensiva de genes e genomas.

As suas anotações podem ser importadas, automáticas ou manuais. As ferramentas de

anotação automática são utilizadas para verificar similaridades entre seqüências e predizer funções de genes. Elas são disponibilizadas *on-line*. Em seguida a essa análise, especialistas validam esses dados e criam suas anotações manuais.

Uma de suas principais características é um ambiente de anotações comparativas, no qual é possível verificar a qualidade tanto das anotações automáticas quanto das manuais. Assim, um usuário pode requisitar a comparação de todas as diferentes anotações disponíveis para um gene de um determinado genoma. Essas anotações podem tanto ser importadas como internas, ou seja, de usuários do sistema. Isso tudo permite uma verificação mais acurada da predição funcional de um gene.

As anotações manuais são criadas dentro de uma interface *Web* e são baseadas em uma ontologia própria do sistema, chamada ERGO-Ontology.

Todos os dados gerados pela execução do ERGO são armazenados no SGBD relacional PostgreSQL. Por haver um processo de avaliação de especialistas, pode-se considerar seu banco curado, ou seja, com informações consistentes.

5.2.7 GenDB

O GenDB (MEYER *et al.*, 2003) é um sistema de anotações de genomas de organismos procarióticos. Trata-se de um sistema modular, desenvolvido sob o paradigma de orientação a objetos. Isso o torna um sistema flexível e extensível. GenDB aceita tanto anotações manuais quanto automáticas e importadas.

Seu modelo de dados é relacional e divide-se em três conceitos: região, observação e anotação. Uma região é uma seqüência (ou subseqüência); observações são dados produzidos pela execução de programas de análise (anotações automáticas de ferramentas como BLAST e InterPro) e anotações são interpretações feitas por pesquisadores (anotações manuais).

A arquitetura do sistema GenDB é composta basicamente por um banco de dados relacional, um módulo (O2DBI) que controla o acesso ao banco e uma interface *Web*. O módulo O2DBI é responsável pelo mapeamento de objetos GenDB em tabelas e provê acesso aos dados por meio de interfaces (Perl ou C++). A partir dessas interfaces são implementadas interfaces de usuário, no lado cliente, as quais utilizam suas funcionalidades. No lado servidor, bancos de dados de seqüências podem ser acessados com o sistema SRS ou via interfaces criadas pelo projeto BioPerl (BIOPERL, 2008). Ferramentas intensivas

computacionais como BLAST ou InterPro podem ser gerenciadas e escalonadas por meio de um BioGrid (BIOGRID, 2008).

Para projetos de anotação, o sistema provê acesso à *Gene Ontology* e navegação em dados genômicos através de suas categorias.

5.2.8 GeneQuiz

O GeneQuiz (ANDRADE *et al.*, 1999) é um sistema automático para anotações funcionais de seqüências de aminoácidos, tendo como principal propósito definir uma funcionalidade específica e confiável a uma certa proteína. O sistema é composto por quatro módulos: GQupdate, GQsearch, GQreason e GQbrowse.

O primeiro, GQupdate, é responsável por manter o sistema integrado a fontes de dados públicos, de proteínas ou seqüências, e sempre atualizado. GQsearch aplica diversas ferramentas de análise à seqüência, convertendo e armazenando os resultados em um banco de dados relacional. GQreason utiliza esses resultados, juntamente com anotações encontradas nas fontes de dados públicos, na forma de palavras-chaves para concluir a respeito da funcionalidade celular geral e funcionalidade específica da seqüência consultada, partindo da análise de conjuntos homólogos. O último, GQbrowse, permite ao usuário, através de um navegador *Web*, examinar os resultados e imagens obtidos. Não provê meios para a anotação manual.

5.2.9 Ambiente de Anotação de Gopalacharyulu

Este trabalho (GOPALACHARYULU *et al.*, 2005) realiza a integração e a mineração de dados provindos de diversas fontes públicas. Sua integração é baseada na premissa de que relacionamentos entre entidades biológicas podem ser representados como uma rede complexa. A dependência contextual é alcançada por um uso prudente de medidas de distância nessas redes. A implementação do sistema é baseada em uma arquitetura multi-camadas usando um banco de dados XML nativo e uma ferramenta de *software* para consultar e visualizar redes biológicas complexas.

Por se tratar de um trabalho que objetiva integrar dados diversos, onde cada fonte pode ter seu modelo de dados próprio, optou-se pelo uso de ontologias para organizar o conhecimento biológico. São elas que formam essa rede complexa e as distâncias entre entidades biológicas são definidas através de uma abordagem que utiliza ontologias. Nesse trabalho, a *Gene Ontology* é a principal ontologia utilizada.

A sua arquitetura é dividida em três camadas. A primeira camada é a de dados, chamada de *back-end*. Uma camada intermediária provê as definições de ontologias, mapeamentos de esquemas, implementações de aprendizagem conceitual, além de conjuntos de algoritmos e módulos que processam e exibem resultados de consultas. A última camada é a de interface com o usuário, ou *front-end*. Os mapeamentos de esquemas fazem a associação entre entidades de múltiplas fontes de dados. Assim, é possível navegar em dados de uma fonte e diretamente relacioná-los a outros dados semelhantes descritos em outras fontes de dados. A camada de dados, por sua vez, é separada em um modelo relacional (SGBD Oracle 10g) para armazenamento de dados mais volumosos (como dados de expressão gênica) e um modelo semi-estruturado XML (Tamino XML Server) para representar dados de anotações.

A integração dos dados de diferentes fontes segue um procedimento para criação de esquemas XML, com a definição da estrutura dos dados e propriedades físicas, como índices e elementos raiz (também chamados *doctypes*) de instâncias XML. Esse procedimento requer o desenvolvimento de tradutores (*parsers*) para converter os dados ao esquema definido. Em seguida os documentos XML gerados são armazenados no Tamino XML Server.

5.2.10 MiGenes

O MiGenes (BASU *et al.*, 2006) é um sistema de anotação voltado para pesquisas de mitocôndrias, o qual utiliza organismos modelos para suas comparações de biosseqüências. Ele mantém um banco de dados relacional, o qual se conecta a bancos de dados públicos para manter-se sempre atualizado em relação a genomas de proteomas mitocondriais. Além dessas anotações importadas, o MiGenes também armazena anotações automáticas e manuais.

Além disso, é curado com base em termos da *Gene Ontology*, assim, asseguram que as anotações realizadas sejam de alta qualidade. As anotações manuais e a navegação sobre elas são realizadas por meio de uma interface *Web*.

Seu banco de dados relacional é composto por 14 tabelas e dividido em duas seções: gene e proteína. Esses dados são todos importados de bancos de dados públicos e têm os mesmos identificadores do GenBank (para genes) e do UniProt (para proteínas). Toda entrada no banco relaciona um gene a uma ou mais proteínas, formando um mapeamento entre anotações de diferentes fontes. Essa integração torna-se complicada, uma vez que há diversos bancos de dados especializados que não mapeiam anotações do GenBank ou do UniProt, e mesmo entre esses dois bancos não há indexação cruzada ou mapeamentos.

Por se tratar de anotações referentes a mitocôndrias, somente um subconjunto relevante de termos da *Gene Ontology* é considerado para as anotações. Alguns termos são mais abrangentes, outros mais específicos. Esse subconjunto de termos é mapeado em tabelas do banco para acesso e anotação.

5.2.11 PEDANT

O PEDANT (*Protein Extraction, Description and ANalysis Tool*) (FRISHMAN *et al.*, 2003) é um ambiente de anotação que provê ferramentas para análise automática de proteínas, determinando sua função e estrutura. É baseado em um esquema de banco de dados relacional e um sistema, BioRS, para mineração de dados.

O BioRS é um sistema com funções de integração e recuperação de dados. Ele é capaz de integrar e procurar fontes de dados baseadas em *flat-files* assim como bancos de dados relacionais. Além disso, o sistema indexa dados para melhorar o desempenho no processamento de consultas e a fonte de dados original somente é acessada quando o usuário requisita o dado completo ou quando a indexação é realizada.

Por fim, o PEDANT disponibiliza ferramentas para clusterização de seqüências, comparação entre genomas e interação proteína-proteína, o qual permite avaliar e visualizar um catálogo contendo interações entre proteínas.

5.2.12 Outros Ambientes de Anotação: Genotator, Artemis e GARSA

O Artemis (RUTHERFORD *et al.*, 2000) é um sistema simples que permite a

visualização de dados extraídos de documentos no formato GenBank, EMBL ou GFF. Também permite a execução de programas externos, como o FASTA e o BLAST. O usuário pode fazer anotações sobre os dados obtidos. Tem sido utilizado em projetos genoma de bactérias e pequenos organismos eucarióticos.

O GARSA (DÁVILA *et al.*, 2005) é um ambiente de anotação que tem como objetivo auxiliar pesquisadores na execução de ferramentas e visualização de resultados. Ele provê acesso a uma gama de ferramentas de Bioinformática (incluindo BLAST, Phrep/Phrap, etc.) e uma interface *Web*. Utiliza um banco de dados relacional para armazenar todos os seus dados. Faz *download* de seqüências de fontes de dados públicas como o GenBank (anotações importadas), armazena dados das ferramentas executadas (anotações automáticas) e permite algumas anotações de pesquisadores (anotações manuais).

O Genotator (HARRIS, 1997) é um sistema para anotação automática de seqüências, desenvolvido em 1997. É formado por um conjunto de programas de análise, um banco de dados e um navegador gráfico que permite ao usuário não somente visualizar os resultados como criar suas próprias anotações. O banco de dados é composto por *flat files* no formato ACE (o mesmo utilizado pelo AceDB), sendo que cada um contém um tipo de anotação.

5.3 Análise Comparativa entre os Ambientes de Anotação

Atualmente existem diversos ambientes de anotação já desenvolvidos para o processamento de dados gerados por projetos genoma. Na seção 5.2 foram apresentados apenas alguns dos ambientes encontrados na literatura. Muitas vezes esses ambientes são projetados com objetivos específicos de pesquisa, algumas vezes restringindo-se a um domínio também específico. Em (LEMOS *et al.*, 2004), é feito um levantamento dos requisitos funcionais necessários para um ambiente de anotação completo.

Uma análise comparativa entre os ambientes estudados é apresentada na Tabela 2. As características analisadas estão dentro do escopo do objetivo deste trabalho. Na última linha são apresentadas as características do ambiente BioFOX, proposto neste trabalho, o qual estende o Bio-TIM. Baseada no trabalho realizado em (LEMOS *et al.*, 2004), segue-se uma comparação entre esses ambientes de anotação, com respeito aos seguintes aspectos:

- como as anotações são modeladas;
- como as anotações são armazenadas.

Tabela 2 Tabela comparativa entre os ambientes de anotação apresentados neste capítulo. O campo 'Tipos de anotação' se refere às formas de anotação realizadas pelo ambiente, sendo (1) para anotação importada; (2) para anotação automática; e (3) para anotação manual.

* Somente foram consideradas as ontologias utilizadas pelo ambiente para organizar as anotações armazenadas.

Ambientes de anotação	Armazenamento de dados	Tipos de anotação	Integração de dados	Vocabulário controlado	Ontologia*
Apollo	SGBD Relacional	1, 2, 3	-	X	-
Ártemis	Formatos EMBL, GenBank e GFF	1, 2, 3	-	-	-
ASAP	SGBD Relacional	1, 2, 3	X	X	-
BASys	SGBD Relacional	1, 2	-	-	GO
BioNotes	SGBD Relacional Estendido (DW Relacional)	1, 2, 3	X	X	GO e ontologia de aplicação
Bio-TIM	SGBD Relacional	1, 2, 3	X	-	-
ERGO	SGBD Relacional	1, 2, 3	X	-	ERGO
GARSA	SGBD Relacional	1, 2, 3	-	X	GO
GenDB	SGBD Relacional	1, 2, 3	-	-	GO
GeneQuiz	SGBD Relacional	1, 2	X	X	-
Genotator	Flat files (ACE)	2	-	-	-
Gopalacharyulu	SGBDs Relacional e Semi-estruturado	1, 2, 3	X	-	GO
MiGenes	SGBD Relacional	1, 2, 3	X	-	GO
PEDANT	SGBD Relacional	1, 2, 3	X	-	-
BioFOX	**SGBD Semi-estruturado e DW Relacional**	**1, 2, 3**	**X**	**X**	**GO, SO e ontologia de aplicação**

5.3.1 Como as anotações são modeladas

Com respeito ao modelo de dados utilizado por esses ambientes de anotação, é possível verificar que a grande maioria utiliza o modelo relacional, como nos casos de ASAP, BASys, ERGO, GenDB, GeneQuiz, MiGenes, PEDANT, GARSA, Apollo e também o ambiente Bio-TIM.

Esses ambientes têm algumas desvantagens por adotarem um modelo estruturado relacional. Entre elas estão os fatos de que se torna difícil modelar todos os esquemas de

anotação implementados por fontes de dados externas e programas automáticos, e que o banco de dados pode gerar uma quantidade muito grande de tabelas (LEMOS *et al.*, 2004).

Ainda há aqueles ambientes que organizam seus dados em *flat files*, como é o caso do Artemis e do Genotator. Esses ambientes provêm muito menos recursos de armazenamento e acesso aos seus dados do que aqueles que utilizam SGBDs para tais tarefas, tornando-se inapropriados ao uso por parte de pesquisadores.

Como descrito no capítulo 4, o uso do modelo semi-estruturado propõe solucionar alguns problemas verificados por modelos estruturados, dentro do domínio de Bioinformática. O principal aspecto se refere ao requisito de evolução de esquemas de bancos de dados de biologia molecular, uma vez que estes são constantes. Apenas dois desses ambientes utiliza esse modelo de dados: Gopalacharyulu utiliza um banco de dados XML nativo para armazenar suas anotações e um banco relacional apenas para armazenar as seqüências biológicas, por serem dados volumosos; e o ambiente BioNotes implementa um banco de dados XML sobre relacional (modelo relacional estendido).

Uma das características apresentadas pelo BioFOX para estender o Bio-TIM é o uso de um banco de dados semi-estruturado, o qual se soma ao DW relacional. Para isso, foi utilizado um SGBD XML nativo, o Tamino XML Server (SOFTWARE AG, 2008), da empresa Software AG. Essa implementação objetiva melhor adequar o modelo de dados utilizado às características dos dados de anotações de projetos genoma. Portanto, nesse aspecto o BioFOX, assim como os ambientes de Gopalacharyulu e BioNotes é um ambiente que usa um modelo de dados semi-estruturado.

5.3.2 Como as anotações são armazenadas

Como pode ser verificado na Tabela 2, os únicos ambientes que não fornecem a opção de anotação manual são o BASys e o Genotator. Todos os demais trabalham com todas as formas de anotação (importada, manual e automática). Dentre os ambientes que fornecem a opção de anotação manual, alguns trabalham com vocabulário controlado, outros com ontologias e ainda outros com as duas formas.

Alguns desses ambientes utilizam um DW para armazenar localmente anotações provindas de fontes de dados externas, geralmente públicas. Essa abordagem permite maior agilidade para o acesso aos dados, mas também requer que sistemas de *software* sejam re-

executados regularmente e a atualização constante em relação a fontes de dados externas. Entre esses ambientes estão o BioNotes, o PEDANT e o Bio-TIM.

Dentre todos os ambientes analisados, apenas o BioNotes utiliza uma ontologia para organizar semanticamente seus esquemas de dados. Ele utiliza ontologias locais para representar cada uma das fontes de dados externas que ele integra e uma ontologia global, a qual relaciona e integra todas as locais. Essa abordagem, portanto, considera o uso de ontologias para a integração semântica de fontes de dados diversas.

O BioFOX mantém as principais características do Bio-TIM, como trabalhar com todos os tipos de fontes de anotações e realizar a integração de dados provenientes de fontes externas. Contudo, ele introduz o uso de vocabulário controlado e ontologias para organizar e padronizar as anotações armazenadas. As ontologias GO e SO são utilizadas para auxiliar o processo de anotação manual. Além dessas duas ontologias de biologia molecular, o BioFOX utiliza também uma ontologia de aplicação para dar semântica a seus esquemas de dados.

O grande diferencial do BioFOX em relação ao BioNotes é a forma como a ontologia é utilizada para organizar e adicionar semântica aos seus respectivos esquemas de dados. Enquanto o BioNotes integra fontes de dados externas e, a partir delas, cria suas ontologias locais e a global, o BioFOX segue o caminho inverso. Neste, a ontologia deve guiar a criação dos esquemas de dados XML, ou seja, a partir de uma definição formal de um domínio é que são gerados esquemas de dados que compartilham uma mesma semântica. Dessa forma, pretende-se que diferentes projetos genoma partam de uma ontologia em comum para construir seus bancos de dados, os quais podem ser mais facilmente integrados em um passo posterior. Esse modelo é chamado *conceito-compartilhado* e é apresentado no capítulo 6.

5.4 Considerações Finais

Neste capítulo, foram descritos alguns dos principais ambientes de anotação encontrados na literatura, com o objetivo de apresentar e comparar as principais características desses ambientes, no que se refere à representação e armazenamento de anotações. Dentre as características analisadas estão o modelo de dados, os tipos de anotação, se realiza ou não integração de dados e se utiliza ou não vocabulários controlados e ontologias.

Também foram apresentadas as características que contemplam o ambiente BioFOX, o qual estende o ambiente Bio-TIM. Assim, relacionou-se suas novas características às motivações para seus usos. O modelo de dados semi-estruturado, introduzido no BioFOX, apresenta características que o torna mais apropriado para a representação de anotações de projetos genoma. Além disso, sendo o compartilhamento e integração de dados um fator de extrema importância para a evolução dessas pesquisas, o uso de ontologias para prover semântica aos dados e seus esquemas se torna essencial em tais projetos. Nesse sentido, o BioFOX utiliza uma ontologia que serve como modelo para a definição e criação de esquemas de dados semi-estruturados, os quais poderão, então, ser integrados. Assim, este trabalho segue um caminho inverso a outros que utilizam ontologias para descrever as fontes de dados já existentes e que devem ser integradas.

O capítulo 6 apresenta a arquitetura do ambiente BioFOX, proposto neste trabalho. Nele são descritos cada um dos módulos componentes do BioFOX, com destaques para o modelo de dados utilizado e para o auxílio à anotação manual.

6 ARQUITETURA DO AMBIENTE DE ANOTAÇÃO BIOFOX

6.1 Considerações Iniciais

Uma anotação consiste em uma informação associada a uma biosseqüência ou a seu produto com o propósito de relacioná-los a um conhecimento biológico. Como elucidado no capítulo 2, as anotações podem ser classificadas de acordo com sua fonte: importada, automática ou manual. Um grande problema encontrado por pesquisadores é de como essas informações devem ser organizadas, compartilhadas e integradas. Essa dificuldade surge devido ao fato de que geralmente não há padronização nos esquemas de dados entre diferentes bancos de dados ou mesmo de terminologias utilizadas por eles. A arquitetura proposta neste trabalho para o ambiente BioFOX utiliza-se de ontologias e bancos de dados XML para suprir essas necessidades.

A arquitetura proposta será integrada ao Bio-TIM, o qual recebe e processa cromatogramas até a análise da biosseqüência, produzindo diversos relatórios para os biólogos. Nesta arquitetura, ontologias são usadas para organizar semanticamente os esquemas de bancos de dados XML e suas anotações e também para guiar a inserção de dados e consultas. Essa característica visa propiciar ao ambiente de anotação flexibilidade e um modelo de dados compreensível. Assim, um *namespace* XML para vocabulários de anotação e uma ontologia de aplicação são peças importantes nesta arquitetura. Ambos trabalham em conjunto e permitem a conexão semântica entre esquemas de bancos de dados de mesmo domínio de projetos genoma (ex.: domínio de bactérias, protozoários ou de plantas). Além disso, para as anotações manuais também são utilizadas ontologias do domínio de biologia molecular, com o intuito de que os anotadores tenham maior apoio nessa tarefa.

As seções seguintes apresentam: (i) a arquitetura proposta para ambientes de anotação (seção 6.2), detalhando cada um de seus módulos e suas funções; (ii) o desenvolvimento dos *Namespace de Anotações Genômicas* (seção 6.3); (iii) o desenvolvimento da ontologia de aplicação (seção 6.4); e (iv) como é realizada a interação entre a arquitetura e seus usuários, ou seja, os pesquisadores (seção 6.5).

6.2 Arquitetura e seus Módulos

A arquitetura apresentada neste trabalho é composta por quatro módulos (Figura 10):

- Módulo Ferramentas de Bioinformática (MFB): agrega um conjunto de ferramentas computacionais para análise genômica.
- Módulo Administrador de Conhecimento (MAC): responsável pelo controle e manutenção das ontologias, de anotação e domínio, e do *Namespace* de Anotações Genômicas, os quais são essenciais para o funcionamento de todo o sistema.
- Módulo Repositório de Dados (MRD): composto por bancos de dados, onde serão armazenadas as anotações.
- Módulo Interface de Anotação (MIA): provê interfaces para a anotação de biosseqüências e, em especial, suporte ao processo de anotação manual.

As próximas seções detalham cada um dos módulos e a interação existente entre eles.

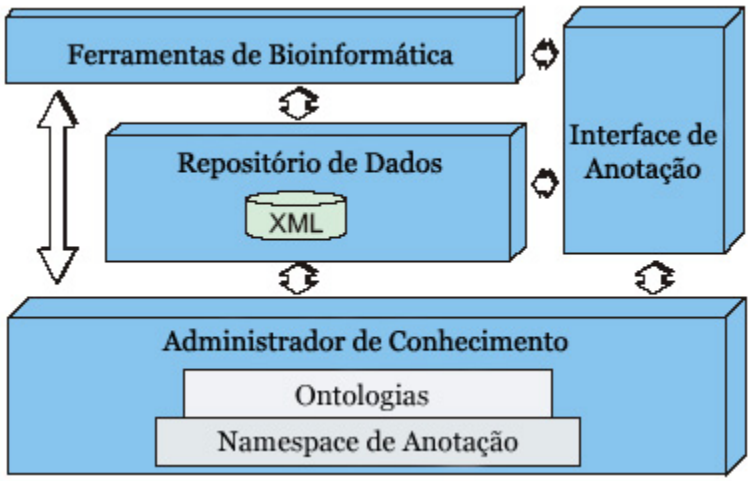


Figura 10 Arquitetura do ambiente BioFOX.

6.2.1 Módulo Ferramentas de Bioinformática (MFB)

Atualmente existe um grande número de ferramentas computacionais criadas para a análise de biosseqüências, sendo diversos os seus objetos de estudos, desde o reconhecimento da seqüência de nucleotídeos de uma região, buscas comparativas para reconhecimento de um

gene e visualização de uma região até determinação das funções de um gene. Essas ferramentas podem ser combinadas para gerar os mais diversos relatórios para os pesquisadores, sendo geralmente dispostas em um *pipeline*, no qual a entrada de dados para uma ferramenta é proveniente da saída gerada por outra.

Dentro desta arquitetura, o MFB é composto por um conjunto dessas ferramentas computacionais de Bioinformática, as quais podem ser executadas dentro do ambiente de anotação. Assim, o projetista do sistema escolhe aquelas ferramentas que serão úteis ao desenvolvimento do projeto e cria o *pipeline* de análises. Essas ferramentas são incorporadas ao sistema como componentes de programação por meio do *framework* FrameEST, desenvolvido em (LOMBARDO, 2006). Para cada uma dessas ferramentas, é preciso que as suas anotações sejam reconhecidas pelo MAC, pois só assim poderão ser consideradas para a composição dos esquemas de bancos de dados. A comunicação entre os demais módulos e o MFB está descrita em suas respectivas seções.

6.2.2 Módulo Administrador de Conhecimento (MAC)

Este módulo mantém todo o conhecimento necessário para o desenvolvimento e a manutenção do sistema e também para a integração semântica dos dados anotados. Dessa forma, o *Namespace* de Anotações Genômicas e as ontologias (isto é, a de aplicação e as de biologia molecular) são partes integrantes deste módulo. Assim, o MAC deve ser mantido sempre atualizado para poder refletir as mudanças em outros módulos, sempre que essas mudanças se referirem ao domínio de anotações.

A ontologia de aplicação representa todo o conjunto de anotações definido no *Namespace*, o que define um relacionamento estreito entre eles. Sempre que o *Namespace* é atualizado, suas alterações devem refletir diretamente na ontologia de aplicação. Essa relação pode ser vista sempre que uma nova ferramenta de Bioinformática for componentizada e adicionada ao MFB, quando, então, o MAC deve tomar conhecimento das anotações que podem ser realizadas por essa ferramenta. Assim, ambos, *Namespace* e ontologia de aplicação, devem integrar essas novas anotações.

Este módulo também provê suporte à modelagem do banco de dados. A ontologia de aplicação é responsável por auxiliar o projetista do banco de dados nessa tarefa, mostrando-lhe os conceitos existentes dentro do domínio de projetos genoma e permitindo-lhe manipulá-

los de forma a definir o esquema do banco de dados a ser criado. O uso da ontologia de aplicação também permite associar semântica ao esquema do banco de dados, o que pode contribuir substancialmente para a integração de dados.

Por fim, o MAC fornece ao MIA acesso a ontologias do domínio de biologia molecular que possam auxiliar os pesquisadores em suas anotações manuais. Atualmente as ontologias fornecidas são a GO e a SO. Outras ontologias podem ser futuramente incorporadas.

6.2.3 Módulo Interface de Anotação (MIA)

A anotação manual pode ser descrita como uma das fases mais delicadas e minuciosas de um projeto genoma. Isso ocorre porque biólogos e pesquisadores devem lapidar os dados e associá-los a processos biológicos conhecidos. Como grandes volumes de dados são produzidos, esse trabalho pode ser exaustivo para os pesquisadores, podendo levar a anotações incompletas, não padronizadas e de baixa qualidade. Uma razão para sua baixa qualidade é a falta de suporte aos pesquisadores, os quais geralmente produzem anotações muito subjetivas.

O ambiente de anotação deve, portanto, prover uma interface de anotação manual com a capacidade de oferecer suporte aos pesquisadores em suas descrições, diminuindo o esforço despendido nessa fase e permitindo criar anotações padronizadas e com maior qualidade.

O MIA deve contemplar interfaces para execução das ferramentas de Bioinformática, para a anotação manual e para acesso aos dados. Dentre elas, a interface de anotação manual é objeto de estudo e vem sendo desenvolvida para implementar as características mencionadas acima, com o suporte de ontologias de biologia molecular. Dentre os recursos já oferecidos por essa interface estão:

- **Auto-completar:** utiliza termos definidos na ontologia de domínio que são sugeridos à medida que o anotador for completando a anotação em um campo.
- **Sinônimos:** verifica na ontologia se há termos sinônimos ao que estiver sendo preenchido em um campo associado ao recurso; caso haja, esses termos são apresentados e podem substituir o termo original. Pode contribuir para o uso de termos mais técnicos na anotação.

- **Exemplos:** este é um recurso simples, mas que pode contribuir para que as anotações não sejam tão subjetivas. O anotador pode se espelhar em exemplos que aparecem em balões de texto para criar boas anotações, os quais podem ser formatados de acordo com protocolos de anotação definidos para o projeto, os quais indiquem, por exemplo, quais termos anotar e como anotá-los.

Além desses recursos, essa interface permite a criação de novas anotações, mesmo que elas não estejam explicitamente definidas no banco de dados. Isso quer dizer que elas poderão ser instanciadas em lugares pré-definidos no esquema de dados, nos chamados “pontos de extensão”. Essa flexibilidade é importante considerando-se que a biologia é uma área de estudos em constante evolução e que, por isso, há a possibilidade dos pesquisadores necessitarem realizar observações até então inesperadas. Essas novas anotações podem ser então recomendadas a integrarem o MAC e posteriormente participar de uma possível evolução do esquema do banco de dados, ou seja, serem explicitamente definidas no esquema.

6.2.4 Módulo Repositório de Dados (MRD)

O MRD é responsável pela manutenção das estruturas de armazenamento de dados de um projeto genoma. Ele é composto por um SGBD XML nativo, responsável por armazenar os esquemas XML e os dados de anotações, e por uma interface para a modelagem do banco de dados. As fontes de anotação podem ser importadas, automáticas ou manuais.

A interface para modelagem de bancos de dados XML, chamada de XML Database Design, visa guiar o projetista do banco de dados com o propósito de atingir os objetivos de criação de esquemas XML interoperáveis, flexíveis e ainda com semântica agregada. Para tanto, essa interface faz uso da ontologia de aplicação e do *Namespace* de Anotações Genômicas. A XML Database Design é apresentada no capítulo 7.

Agregar semântica ao esquema do banco de dados pode contribuir para o compartilhamento e a integração de dados entre um conjunto de bancos de dados que compartilhem a mesma semântica. Os BDBMs seguem, geralmente, os requisitos de seus projetos genoma sem se preocupar em como integrá-los a outros. Assim, a grande maioria dos projetos genoma conta com o seu próprio esquema de banco de dados, tornando-se heterogêneos e independentes. Com o uso de bancos de dados XML associados a uma ontologia de aplicação, é possível criar esquemas de dados adequados aos requisitos de cada

projeto e, simultaneamente, mantê-los semanticamente conectados. Assim, esta arquitetura apresenta um modelo de integração, ao qual chamamos **conceito-compartilhado** (Figura 11).

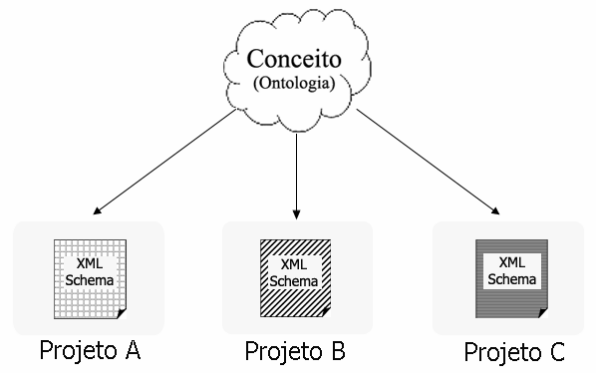


Figura 11 Modelo de integração proposto: conceito-compartilhado.

As vantagens desse modelo é que ele permite criar esquemas de dados diferentes sem, contudo, perder as características de compartilhamento de dados, uma vez que também padroniza suas terminologias e agrega semântica aos seus esquemas por meio de uma ontologia. Esse modelo ainda exige que os esquemas sejam convertidos para um formato comum. Porém, isso é mais simples de se alcançar com a XML.

6.3 Namespace de Anotações Genômicas

Uma das principais características da XML é o fato de ela ser um "*framework*" aberto para a definição de especificações padronizadas (ACHARD *et al.*, 2001). Isso permite à comunidade científica criar seus próprios modelos para troca de informações, ao compartilhar um esquema (composto por um vocabulário e sua estrutura) de documento padrão, atingindo-se interoperabilidade entre diferentes comunidades e grupos. Esse é um fator que tem sido explorado por comunidades de projetos genoma, que estabelecem vocabulários XML padrões para a troca de informações importantes para o desenvolvimento de suas pesquisas.

Cada especificação XML determinada por um XML Schema deve ser unicamente identificada, de forma a distinguí-la e permitir que um documento XML a referencie sem que haja ambigüidades. Esses identificadores são chamados *namespaces*, geralmente valorados com uma URL, indicando o local onde estão as definições do esquema da especificação.

O *Namespace* de Anotações Genômicas, desenvolvido neste trabalho, contém diversos esquemas XML voltados à representação de anotações genômicas em um banco de dados XML nativo. Esses esquemas definem vocabulários XML utilizados em aplicações do domínio de biologia molecular, os quais servem de base para a definição de esquemas de bancos de dados de anotações. A adoção de vocabulários XML tem como objetivo tornar bancos de dados de projetos genoma interoperáveis ao compartilharem os mesmos esquemas, ou ainda esquemas semelhantes, mas com a mesma semântica de dados. Considerando-se dois bancos de dados com o mesmo domínio de aplicação (ex.: projetos genoma de bactérias), as semelhanças entre eles podem ser classificadas como:

- **Esquemas idênticos:** as anotações armazenadas são de mesmo domínio, compostas pelo mesmo conjunto de dados e com esquemas de bancos de dados de mesma estrutura;
- **Conceitos idênticos:** as anotações armazenadas são de mesmo domínio, compostas pelo mesmo conjunto de dados, porém com esquemas de bancos de dados de estruturas diferentes;
- **Domínios semelhantes:** as anotações armazenadas são de mesmo domínio, mas não são compostas necessariamente pelo mesmo conjunto de dados;
- **Misto:** uma vez que um banco de dados XML pode ser composto por um ou mais esquemas, o tipo misto ocorre quando ambos os bancos de dados podem ter seus esquemas classificados em pelo menos dois dos tipos anteriores.

É importante permitir que dois projetos genomas, mesmo que tenham o mesmo domínio de atuação, construam esquemas de dados diferentes para que possam ajustar a estrutura do banco de dados de acordo com suas necessidades. Isso pode afetar diretamente o desempenho de um banco de dados XML. No entanto, como citado acima, a semântica associada a eles deve ser a mesma. Portanto, qualquer anotação em comum deve conter a mesma semântica nos dois bancos de dados, facilitando a interoperabilidade entre eles.

O *Namespace* (Figura 12) pode ser subdividido em: esquemas de domínios específicos; esquemas de ontologias de biologia molecular; e esquemas de aplicativos. Os esquemas de domínios específicos contêm vocabulários XML referentes a sub-áreas da biologia molecular. Dessa forma, vocabulários específicos de anotações para os mais diversos projetos genomas podem ser incorporados ao *Namespace*. Foram utilizados vocabulários já definidos (como os citados anteriormente – GAME, BioML, etc.), os quais foram reestruturados para adequarem-se ao esquema geral da arquitetura.

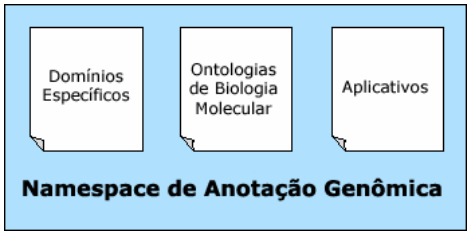


Figura 12 Namespace de Anotação Genômica subdividido em três categorias: domínios específicos, ontologias de biologia molecular e aplicativos.

Os esquemas de ontologias de biologia molecular devem servir de ponte entre as anotações genômicas e ontologias de biologia molecular. Assim, foram definidos esquemas contendo termos necessários para a associação de termos anotados no banco de dados e essas ontologias, e não para todo o conjunto de termos dessas ontologias. As ontologias de biologia molecular consideradas inicialmente foram a *Gene Ontology* e a *Sequence Ontology*. No caso da primeira, o consórcio que a criou propõe algumas observações, as quais devem ser realizadas para associar uma anotação à ontologia (apresentadas na seção 3.2, tabela 1). Portanto, tais observações é que foram mapeadas para um vocabulário XML, e não todos os termos referentes a produtos de genes que ela define. O mesmo acontece para a *Sequence Ontology*, que tem um conjunto de observações um pouco maior.

A terceira subdivisão contempla esquemas de aplicativos, ou seja, vocabulários utilizados por ferramentas de anotação automática. Esse trabalho de definição de vocabulários comuns a ferramentas de anotação automática já vem sendo realizado por pesquisadores do Projeto Hobit, na Alemanha (SEIBEL *et al.*, 2006). Entre os vocabulários já definidos neste projeto, há um que substitui o documento FASTA, chamado SequenceML. Esses vocabulários deverão ser integrados ao sistema em trabalhos futuros.

6.4 Ontologia de Aplicação

No contexto deste trabalho, a ontologia de aplicação deve ser capaz de, dentro do domínio de anotações de biologia molecular, representar o universo de anotações possíveis, e ser responsável por guiar a criação do esquema do banco de dados de um projeto genoma, partindo de seu conhecimento sobre as relações entre suas entidades. Para tanto, ela deve conter entidades que representem as anotações de domínio específico, de biologia molecular e a de aplicativos, e também estabelecer claramente os tipos de relacionamentos entre essas

entidades, de forma a guiar o projetista do banco de dados em sua função de modelagem do esquema do banco. A Figura 13 ilustra o uso da ontologia para o desenvolvimento do projeto.

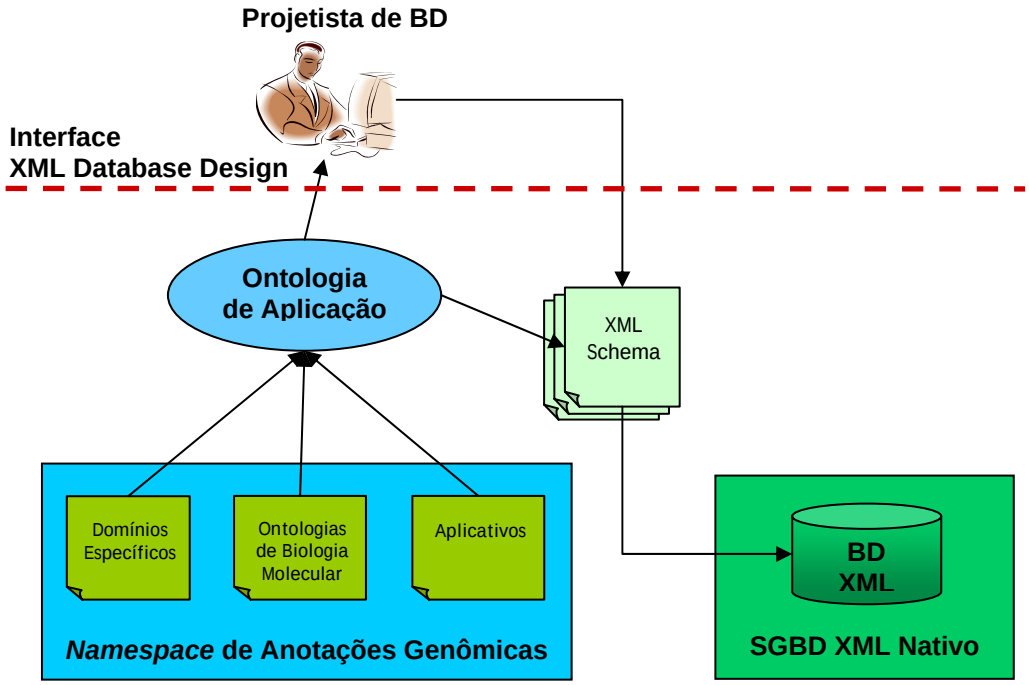


Figura 13 Desenvolvimento dos esquemas de bancos de dados de um projeto genoma.

A ontologia de aplicação é responsável por apresentar ao projetista os termos de conceitos de anotação contidos no *Namespace* por meio da interface *Web* XML Database Design. Através da interação do Projetista com a ontologia, será gerado um conjunto de documentos XML Schema que serão definidos no banco de dados XML do projeto. Essa ontologia é o componente principal para associar semântica às anotações e também para permitir interoperabilidade entre dois bancos de dados do mesmo domínio de projetos genoma, desde que façam uso da arquitetura proposta. Uma vez que as anotações de ambos os projetos estarão associados a uma ontologia em comum, sua integração poderá ser simplificada.

Para conseguir atingir um nível de interoperabilidade desejado, essa ontologia auxilia o processo de projeto do esquema do banco de dados. Enquanto o *Namespace* é responsável por todas as estruturas de esquema, a ontologia de aplicação é responsável por designar semântica às anotações e estabelecer um relacionamento entre elas. Assim, uma interface pode proporcionar interação entre o projetista do banco de dados e os conceitos da ontologia de aplicação, com o objetivo de propor esquemas de dados que estejam de acordo com os requisitos do projetista. A ontologia descreve relacionamentos estruturais que definem a

associação entre dois termos, caso eles sejam relacionados.

A ontologia de aplicação foi definida em OWL e, devido aos seus objetivos, seu projeto segue algumas especificações. Primeiro, os conceitos na ontologia são classificados como léxicos ou não-léxicos. Os conceitos léxicos são representados como propriedades de objetos da OWL. Os conceitos não-léxicos são representados como classes da OWL. Segundo, alguns relacionamentos entre os conceitos devem ser estabelecidos. Eles podem ser:

- **Associação:** relaciona dois conceitos não-léxicos. Esse relacionamento sugere que ambos os conceitos no relacionamento podem ser representados por meio do outro;
- **Agrupamento:** relaciona dois conceitos léxicos presentes em um mesmo conceito não-léxico, ou seja, relaciona duas propriedades de objetos pertencentes à mesma classe. Esse relacionamento sugere que ambos os conceitos devem ser sempre definidos juntos.
- **Parte de:** relaciona dois conceitos não-léxicos indicando que um é parte integrante do outro.

O relacionamento de associação é importante, pois conduz a diferentes visões de um mesmo dado, isto é, conduz a diferentes esquemas XML. Por exemplo, supondo-se duas classes, DNA e gene, caso elas estejam conectadas por esse relacionamento, elas podem ser estruturadas da forma mostrada na Figura 14.

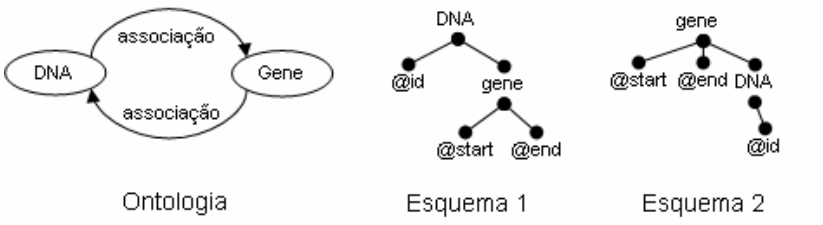


Figura 14 O relacionamento de associação e diferentes visões sobre o mesmo dado. O esquema 1 é uma visão centrada em DNA enquanto o esquema 2 é centrada em gene.

Assim, é possível definir um esquema centrado em DNA (esquema 1) ou um centrado em gene (esquema 2). Essa diferença estrutural permite que dois projetistas de bancos de dados modelem seus próprios esquemas de acordo com as necessidades e objetivos de seus respectivos projetos genoma, levando-se, assim, a dois esquemas diferentes representando o mesmo dado.

O relacionamento de agrupamento relaciona duas propriedades de objetos. Ele especifica que sempre que uma propriedade é escolhida pelo projetista do banco de dados

para compor o esquema final, a outra propriedade também deve ser escolhida. Assim, este relacionamento deve ser usado para definir restrições que agreguem semântica aos conceitos e evitem a composição de esquemas sem sentido. Para exemplificar, podemos citar os conceitos ‘*start*’ e ‘*end*’, ambos pertencentes a ‘*gene*’, os quais estão relacionados por meio de agrupamento. Para a anotação de um gene, não há sentido em apenas representar um ou outro conceito, o que poderia ser considerado um erro semântico. Ao contrário, eles devem sempre estar juntos. Ao agrupá-los, evita-se que o projetista possa cometer tal erro semântico.

6.5 Interação entre Pesquisadores e o Ambiente de Anotação

A interação entre pesquisadores de projetos genoma e o ambiente de anotação proposto pode ser visualizada na Figura 15. Uma interface *Web* é responsável pela comunicação entre o sistema e os pesquisadores.

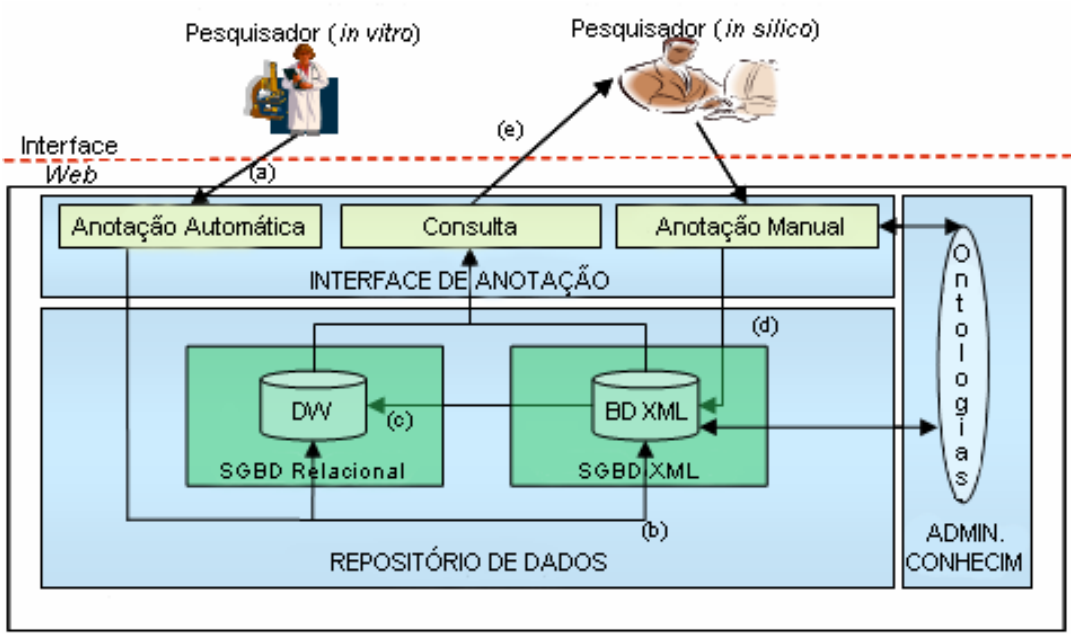


Figura 15 Interação dos pesquisadores de projetos genoma com o ambiente de anotação.

Os pesquisadores podem submeter cromatogramas ao *pipeline* de ferramentas de Bioinformática do sistema (Figura 15.a). A análise desses cromatogramas gera uma série de dados e relatórios que podem tanto ser armazenados no banco de dados XML provido pela arquitetura proposta, quanto ser armazenado diretamente no DW já provido pelo ambiente

Bio-TIM (Figura 15.b). No primeiro caso, os dados precisam antes passar pelo processamento de um conversor XML, sempre que a ferramenta não tiver a opção de gerar sua saída nesse formato. Caso ela gere a saída em XML, é necessário apenas armazená-los de acordo com o esquema definido no banco de dados. Os dados que forem armazenados somente no banco de dados XML podem ser carregados no DW em operações futuras (Figura 15.c). Para isso, um módulo de geração deve ser desenvolvido, de forma a obter os dados do banco de dados XML e convertê-los para as tabelas do DW.

Para as anotações manuais (Figura 15.d), o pesquisador utiliza a interface de anotação proposta pelo MIA. Neste momento, a interface de anotação manual faz acesso a ontologias do domínio de biologia molecular para que essas colaborem para que as anotações tenham maior qualidade.

As consultas aos dados do projeto genoma também devem ser realizadas por meio da interface *Web* (Figura 15.e). Elas podem tanto ser realizadas sobre o banco de dados XML quanto sobre o DW.

6.6 Considerações Finais

Este capítulo apresentou a proposta de uma arquitetura para o ambiente de anotação BioFOX, visando a melhoria do ambiente Bio-TIM. Foi dado enfoque ao módulo para armazenamento e gerenciamento de dados, o MRD, e ao módulo utilizado para administração de todo o conhecimento usado pelo sistema, o MAC. Também foram detalhados os principais componentes do MAC, a ontologia de aplicação e o *Namespace* de Anotações Genômicas e como eles atuam na comunicação entre os diferentes módulos da arquitetura.

O capítulo 7 apresenta a interface para modelagem de bancos de dados XML, desenvolvida como parte componente do MRD. São mostradas todas as suas telas e como ela pode ser manipulada.

7 IMPLEMENTAÇÃO, TESTES E DISCUSSÃO DOS RESULTADOS

7.1 Considerações Iniciais

Este trabalho apresenta uma arquitetura para o ambiente de anotação BioFOX, a qual será implementada no ambiente Bio-TIM. Como esclarecido previamente, o enfoque deste trabalho está no desenvolvimento de uma solução para a definição de bancos de dados dotados de aspectos como semântica e flexibilidade para a representação de um domínio complexo, tal qual o de biologia molecular. Para tanto, uma interface, chamada de XML Database Design, foi desenvolvida com o objetivo de auxiliar o projeto de bancos de dados XML. Ela pode ser um instrumento importante para a construção de esquemas de áreas não dominadas ou complexas, tendo sido dada ênfase, porém não limitada, a projetos de bancos de dados biológicos para suporte à anotação de genomas. A partir desta interface, foram realizados testes com usuários com experiência nas áreas de Bioinformática e Bancos de Dados, para verificar a utilidade dessa ferramenta e a sua contribuição para o domínio de projetos genoma. Ao fim dos testes, cada usuário teve que responder a questionários para que avaliassem a interface.

Assim, este capítulo apresenta a interface desenvolvida e os testes realizados, além de uma discussão sobre as contribuições e limitações da interface, com base nos resultados obtidos na aplicação dos questionários.

7.2 Interface Desenvolvida

A implementação da interface XML Database Design fez uso da tecnologia Java para *Web*, Java Enterprise Edition (JEE), e do *framework* Jena (CARROLL *et al.*, 2004) para a manipulação da ontologia. A sua arquitetura conecta-se ao MAC para ter conhecimento do domínio considerado e, então, poder guiar um projetista na elaboração de um esquema XML próprio para tal domínio. Ela permite criar diferentes esquemas, seja por meio de diferentes estruturas hierárquicas ou pela escolha dos tipos de dados a serem persistidos no banco de dados.

Para a modelagem dos esquemas XML, essa interface baseia-se em alguns conceitos advindos da AOM, como *asset* e Objeto de Negócio (ON). No entanto, devido às restrições de um ambiente *Web* para a representação gráfica de objetos, esses conceitos têm suas representações gráficas simplificadas neste trabalho, como pode ser notado nas seções subseqüentes. Por exemplo, as definições de propriedades de um *asset* não são feitas dentro de sua própria representação gráfica, mas em uma outra tela. Também não há arcos, os quais são substituídos pela ação de anexar os *assets* uns aos outros.

7.3 Testes

Os testes foram realizados para avaliar e validar a aplicabilidade da interface XML Database Design, suas contribuições e limitações dentro do contexto de projetos genoma e da arquitetura de ambientes de anotação. Para fins de simplicidade e agilidade para a realização dos testes com a interface, foram considerados apenas conceitos retirados do projeto ShEST (SHEST, 2008), sem prejuízos ou impacto na análise. Dessa forma, apenas uma pequena parte do vocabulário definido no *Namespace* de Anotações Genômicas foi utilizada, simplificando-se também a criação de uma ontologia de aplicação própria para essa atividade.

Participaram dos testes 14 usuários, com experiência em Bioinformática ou em Banco de Dados ou em ambas áreas. Entre eles haviam alunos de pós-graduação, docentes e técnicos de bioinformática das universidades: Universidade de São Paulo (USP) *campi* São Carlos e Ribeirão Preto, Universidade de Passo Fundo (UPF), Universidade Federal do Rio de Janeiro (UFRJ), Centro Universitário Franciscano (UNIFRA) e Universidade Federal de São Carlos (UFSCar). Foram aplicados questionários para que os usuários avaliassem a interface quanto a questões relativas à experiência de cada um, à contribuição da interface, à aplicação da mesma no contexto de Bioinformática e também à sua usabilidade.

As próximas seções apresentam maiores detalhes de como a interface auxilia um projetista na modelagem de esquemas XML.

7.3.1 Recomendando Conceitos de um Subdomínio

A ontologia de aplicação atua como base de conhecimento do MAC. Além de definir os conceitos de um domínio, ela também define propriedades que relacionam esses conceitos, criando uma rede de conhecimento. Uma das propriedades de objetos definidas na ontologia de aplicação é a *part_of*, a qual permite relacionar os termos de um domínio a um subdomínio específico. Assim, considerando-se o domínio de projetos genomas, onde diversos conceitos são definidos, subdomínios podem ser encontrados (ex.: genoma de bactérias, de abelhas ou de camarões), cada um podendo ser composto por um subconjunto do total de conceitos definidos. Com isso, é possível sugerir ao projetista do banco de dados termos que ele possa utilizar em um determinado subdomínio, simplificando-lhe a tarefa de determinar quais termos são pertinentes ao projeto desenvolvido e diminuindo o tempo de esforço despendido.

Deve-se ressaltar, no entanto, que essa propriedade tem como objetivo indicar aqueles termos que aparecem mais freqüentemente dentro de um subdomínio, não sendo necessariamente os únicos pertencentes a ele. Dessa forma, é possível sempre atualizar esses relacionamentos e criar novos, bem como utilizar outros termos durante uma modelagem que se baseie nesse conhecimento definido na ontologia.

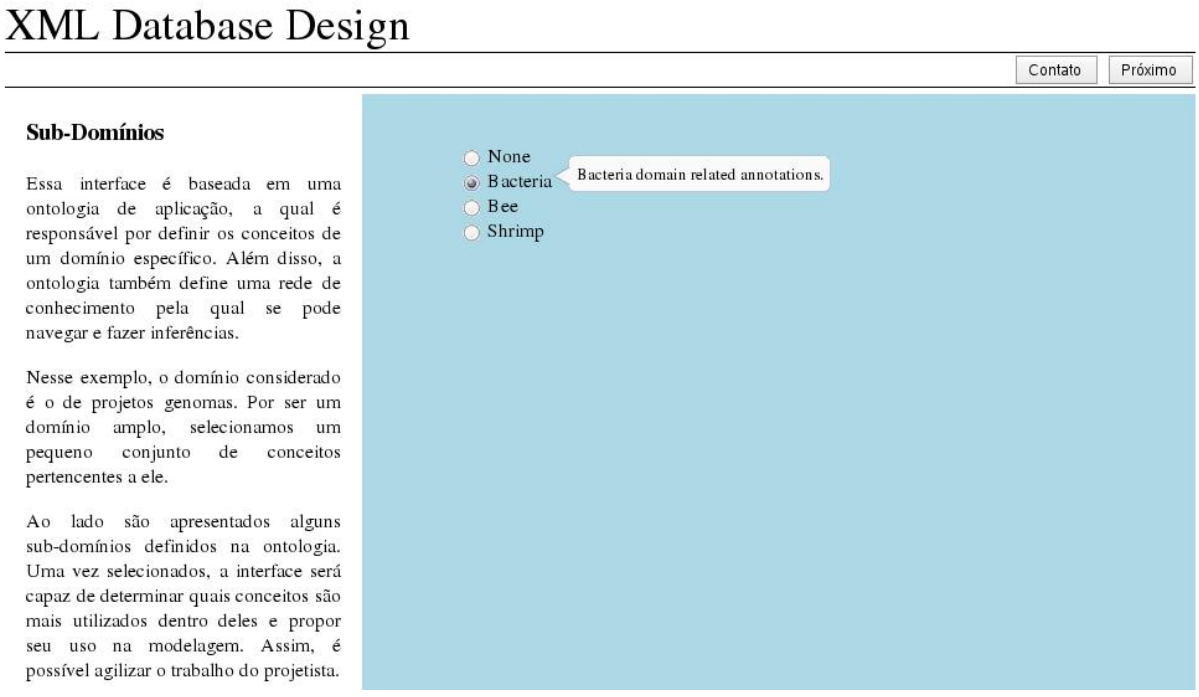


Figura 16 Escolha de um subdomínio para a recomendação de termos relacionados.

Dessa forma, o primeiro passo para o projetista é a escolha de um subdomínio (Figura 16), os quais são listados para que o projetista possa fazer uso dos mesmos. Por exemplo, se um banco de dados a ser projetado for para um projeto genoma de abelhas, o projetista pode selecionar a opção *"Bee"*, para que então lhe sejam sugeridos termos desse subdomínio. Em cada uma das opções é possível ler uma nota explicativa para o subdomínio posicionando-se o ponteiro do *mouse* sobre ela para que sua respectiva nota apareça. Caso nenhum desses subdomínios interesse ao projetista, este tem a opção *"None"*, a qual não considerará nenhum conhecimento previamente definido na ontologia e, assim, não irá sugerir nenhum termo para a modelagem, deixando essa seleção toda a cargo do projetista.

Uma vez escolhido um subdomínio com o qual irá trabalhar, o projetista pode se encaminhar ao passo seguinte através do botão "Próximo".

7.3.2 Definindo Assets e Objetos de Negócio

A tela para a definição de *assets* e ONs (Figura 17) é dividida em duas partes: à esquerda, um *frame* apresenta todos os conceitos não-léxicos de um domínio, também considerados conceitos chaves; à direita, uma área livre destinada à modelagem conceitual, por meio de figuras. Os conceitos não-léxicos de um domínio são definidos na ontologia e devem ser os mesmos que podem ser definidos como *assets* em uma modelagem AOM. Isso demonstra a forte ligação existente entre os vocabulários XML e a ontologia de aplicação, os quais devem estar em sintonia para representar os conceitos do domínio e manter a coerência da aplicação.

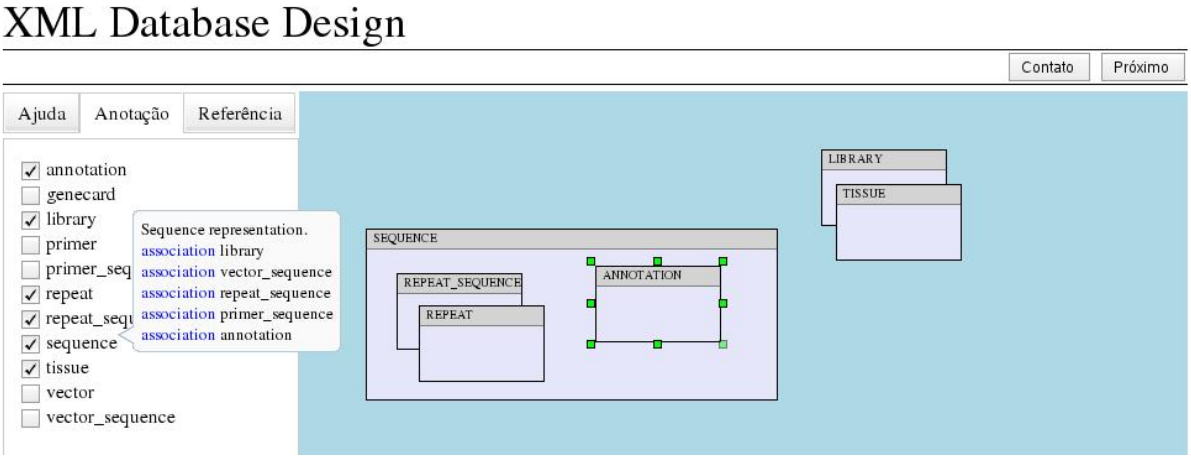


Figura 17 Modelagem conceitual do domínio, por meio de figuras.

Toda vez que um conceito chave é marcado, uma figura correspondente a ele é criada e desenhada na área de modelagem. As formas de manipulação das figuras são: (a) redimensioná-las, (b) arrastá-las ou (c) anexá-las a outra figura. Cada figura segue as regras de associação, definidas pela ontologia, correspondentes ao conceito que ela representa.

Ao anexar uma figura a outra, uma consulta à ontologia é realizada para verificar a propriedade de objetos de associação. Uma figura só será anexada a outra caso seus respectivos conceitos estejam relacionados por meio dessa propriedade. Para verificar quais são os relacionamentos de um conceito basta posicionar o ponteiro do *mouse* sobre ele no *frame* à esquerda. Uma nota é exibida com a definição do conceito e todos os seus relacionamentos de associação. Uma resposta visual é dada para facilitar a percepção de que uma figura pode ser anexada à outra (Figura 18): a figura que está prestes a anexar a outra tem a sua cor de fundo alterada, até que a ação se complete; caso contrário, sua cor de fundo permanecerá a mesma.

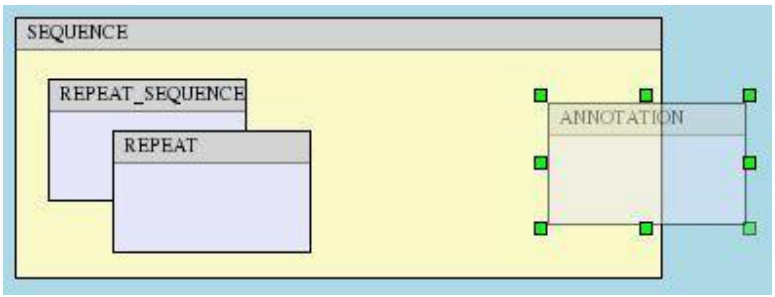


Figura 18 Verificação da associação entre os conceitos: a cor de fundo é alterada quando a figura a ser anexada é aceita.

Neste passo é possível notar algumas simplificações importantes em relação à representação da AOM:

- A representação gráfica de um *asset* contém apenas seu nome, sem qualquer definição de propriedades, chaves e restrições;
- Os arcos são substituídos pela ação de anexar uma figura a outra;
- Os ONs são definidos pelas figuras diretamente desenhadas na área livre, ou seja, aquelas que não estiverem anexas a outras figuras. Todas as outras que estiverem anexadas a elas farão parte do agrupamento que compõe o ON.

Não há qualquer restrição quanto ao número de figuras anexas a outra ou à profundidade do aninhamento criado entre elas, sendo apenas limitado às restrições de regras de associação.

Após o projetista modelar os conceitos chaves do seu domínio, ele deve seguir o botão

"Próximo", o qual o levará ao passo seguinte, no qual são apresentadas as propriedades de tipos de dados de cada um desses conceitos.

7.3.3 Definindo as Propriedades dos Assets

Para cada um dos conceitos-chaves de um domínio definidos na ontologia, existe uma lista de propriedades de tipos de dados que complementam a sua definição. Essas propriedades são apresentadas nessa quarta tela (Figura 19), para cada um dos conceitos que foram mapeados na tela anterior. Elas são apresentadas de forma hierárquica, como o desenho de uma árvore. Cada ON aparece como elemento raiz, suas propriedades são listadas em um nível abaixo e em seguida todos os outros *assets* que tenham sido mapeados dentro dele, e assim por diante, até que se encerre o aninhamento.

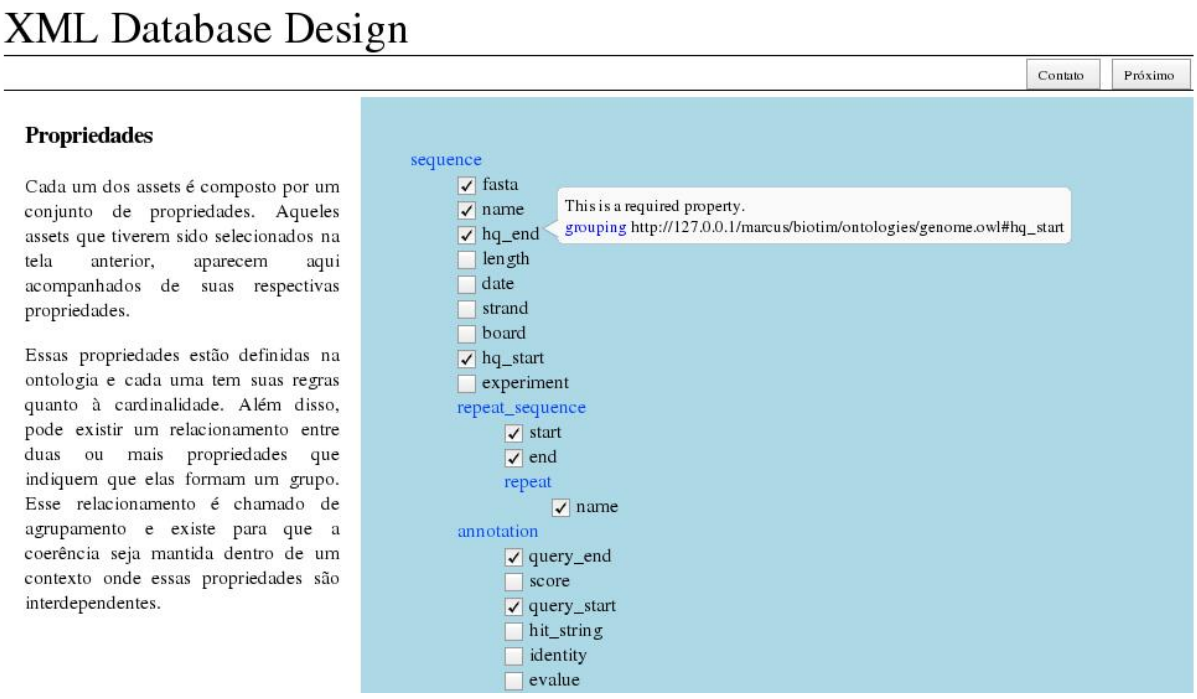


Figura 19 Seleção de propriedades relativas aos assets.

A ontologia de aplicação também define a cardinalidade dessas propriedades. Assim, é possível determinar se elas são requeridas (cardinalidade mínima é maior que zero) ou opcionais (cardinalidade mínima é igual zero). As propriedades requeridas são marcadas previamente e não podem ser desmarcadas pelo projetista. Dessa forma, apenas as

propriedades opcionais poderão ser escolhidas. Novamente há uma simplificação quanto à forma original de representação na AOM. Justamente porque essas propriedades já estão todas definidas no MAC, o projetista não tem liberdade para modificá-las neste passo, com o intuito de que sejam mantidas as mesmas estruturas definidas no *Namespace*. Modificações poderão ser feitas depois que o esquema for sugerido, como descrito na seção 7.7.

Outra propriedade de objetos definida pela ontologia é a de agrupamento, a qual permite agrupar propriedades de tipos de dados. Esses agrupamentos são criados com o intuito de indicar ao projetista que, nesse contexto, tais propriedades só existem se todas as outras do grupo também existirem. Sendo assim, os agrupamentos são importantes pois mantêm a coerência da modelagem. Esse é o caso das propriedades de tipos de dados *start* e *end*, definidas na classe *repeat_sequence*, as quais não fazem sentido se não estiverem juntas. Dessa forma, sempre que uma das propriedades listadas é (des)marcada, a interface verifica se há relacionamentos de agrupamento com outras. Caso haja, as demais propriedades do agrupamento também serão (des)marcadas.

Posicionar o ponteiro do *mouse* sobre uma das propriedades listadas permite ao projetista ler uma nota que indica a sua definição, se é requerida ou opcional, e ainda todas as demais propriedades com as quais ela esteja agrupada.

Esse é o último passo, dentro da interface, necessário para o projetista criar a sua modelagem. Em seguida, um esquema XML é proposto a ele, de acordo com a sua modelagem.

7.3.4 Apresentando o Esquema Proposto

A última etapa é a geração de um esquema XML a ser proposto ao projetista do banco de dados. O esquema é elaborado a partir da modelagem definida pelo projetista e vai buscar as definições XML no *Namespace* de Anotações Genômicas criado.

Por fim, um documento XML Schema é criado com todas as definições recuperadas do *Namespace* e apresentado ao projetista. Em poucos passos, um projetista é capaz de montar esquemas para bancos de dados XML, em domínios de aplicação complexos. O esquema proposto é apresentado na última tela (Figura 20) junto com o esquema original definido no *Namespace*, para que o projetista possa compará-los.

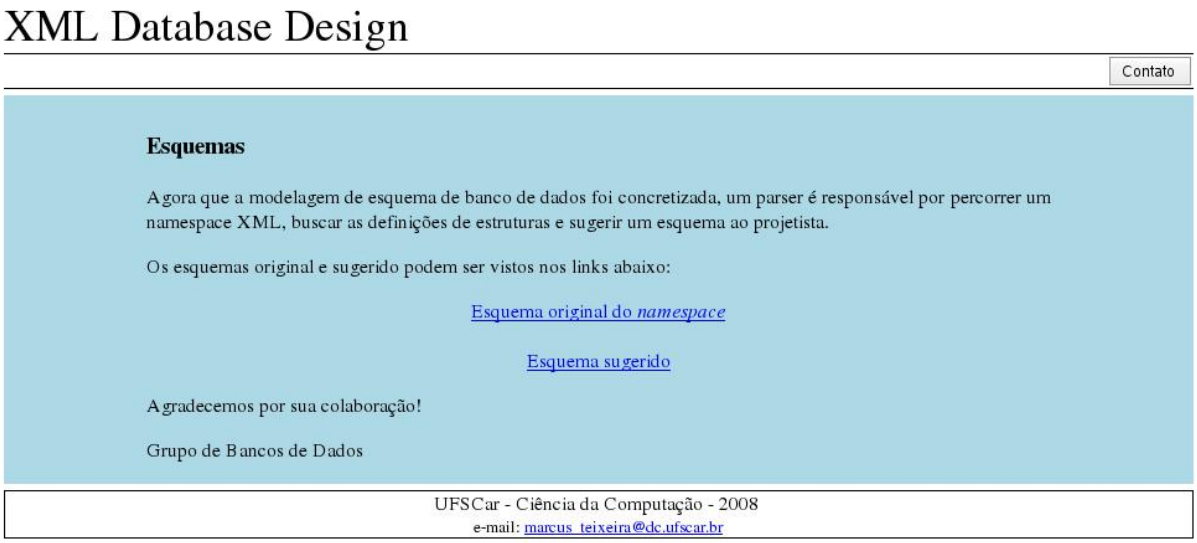


Figura 20 Apresentação do esquema original, definido no *Namespace*, e do esquema sugerido ao projetista.

7.3.5 Procedimentos Pós-Geração do Esquema

Uma vez que a interface desenvolvida está diretamente associada ao MAC, a modelagem e conseqüente geração de um esquema XML são restritas ao conhecimento definido na ontologia de aplicação. Porém, um fator importante da linguagem XML é o seu poder de evolução, que deve ser considerado, mas com cautela.

Após o sistema produzir e sugerir um esquema XML ao projetista, este pode modificá-lo. As mudanças podem ser relativas: (a) ao número de repetições de elementos; (b) à ordem de ocorrência de elementos; (c) à inclusão de novos elementos, os quais não foram identificados na ontologia; (d) à inserção de pontos de evolução por meio dos chamados *wild cards* (ou pontos de extensão) da linguagem XML Schema; ou mesmo (e) à definição de propriedades físicas que possam ser utilizadas por um SGBD, como indexação e gatilhos.

As únicas restrições que devem ser respeitadas se referem ao nome dos elementos e tipos de dados criados, os quais devem ser mantidos. Além disso, toda definição proveniente da interação com a interface deve sempre obedecer à ontologia de aplicação, como no caso dos relacionamentos de associação e agrupamento. Assim, nenhuma alteração feita pelo projetista deve afetar essa regra.

7.4 Discussão dos Resultados

A interface foi avaliada por meio de questões respondidas por cada usuário que realizou o teste. A primeira parte do questionário tinha como objetivo verificar a experiência de cada um em Bioinformática, em bancos de dados, em bancos de dados XML e o conhecimento de cada um a respeito do que são ontologias. O levantamento feito (Figura 21) indica que pouco mais de 70% deles têm experiência em Bioinformática, mais de 92% têm experiência em bancos de dados e pouco mais de 43% têm experiência com bancos de dados XML. Além disso, mais de 92% deles dizem ter conhecimento razoável ou bom sobre ontologias. Esses dados contribuíram para a qualidade da amostra, uma vez que os usuários demonstraram ter experiência em áreas na qual se insere este trabalho.

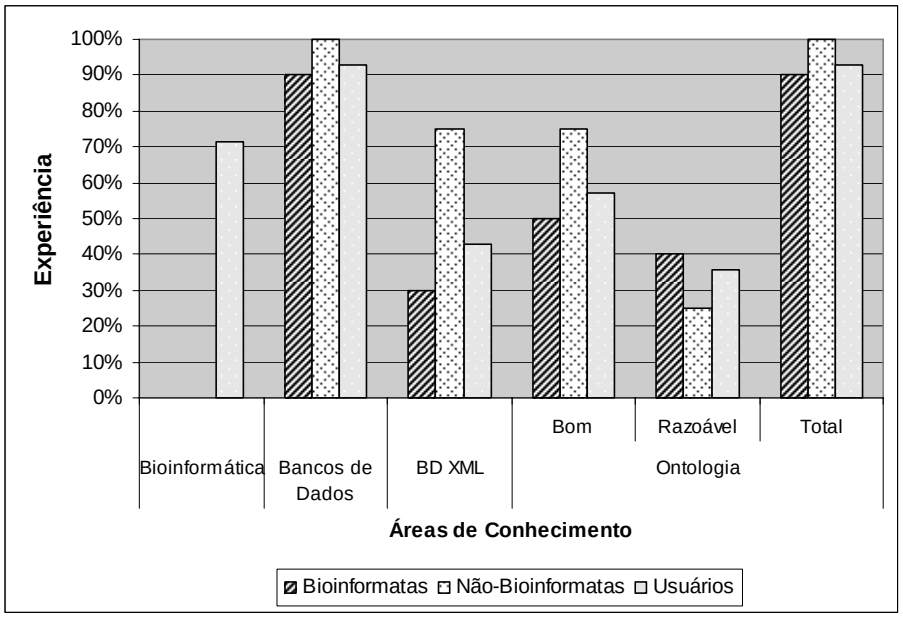


Figura 21 Gráfico indicando a experiência dos usuários que testaram a interface XML Database Design.

A segunda parte do questionário avalia alguns quesitos relativos às contribuições da interface XML Database Design. As análises são baseadas nos gráficos a seguir, sobre a avaliação de todos os usuários (Figura 22), dos usuários com perfil de bioinformata (Figura 23) e dos usuários com experiência em bancos de dados XML (Figura 24).

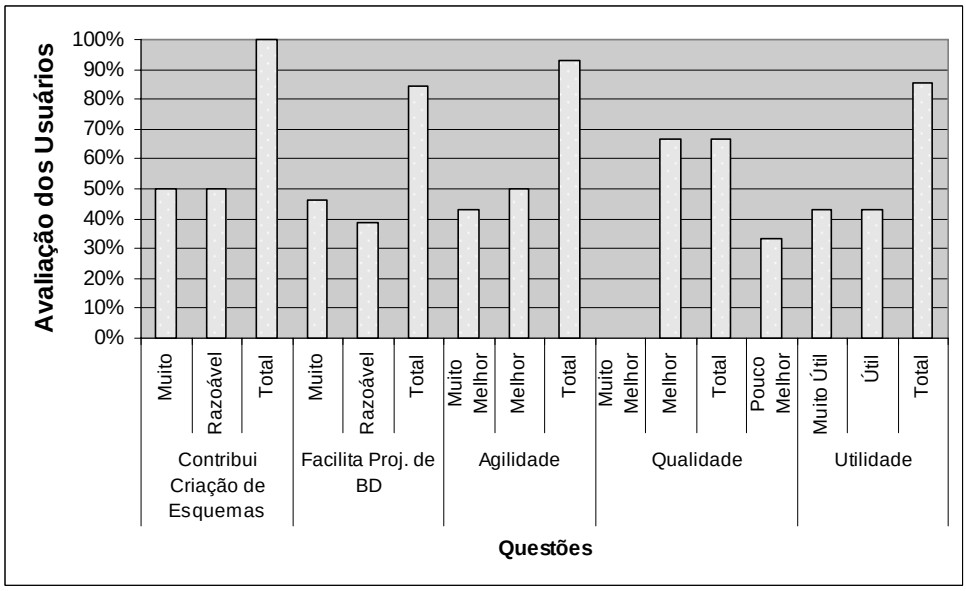


Figura 22 Gráfico de avaliação de todos os usuários.

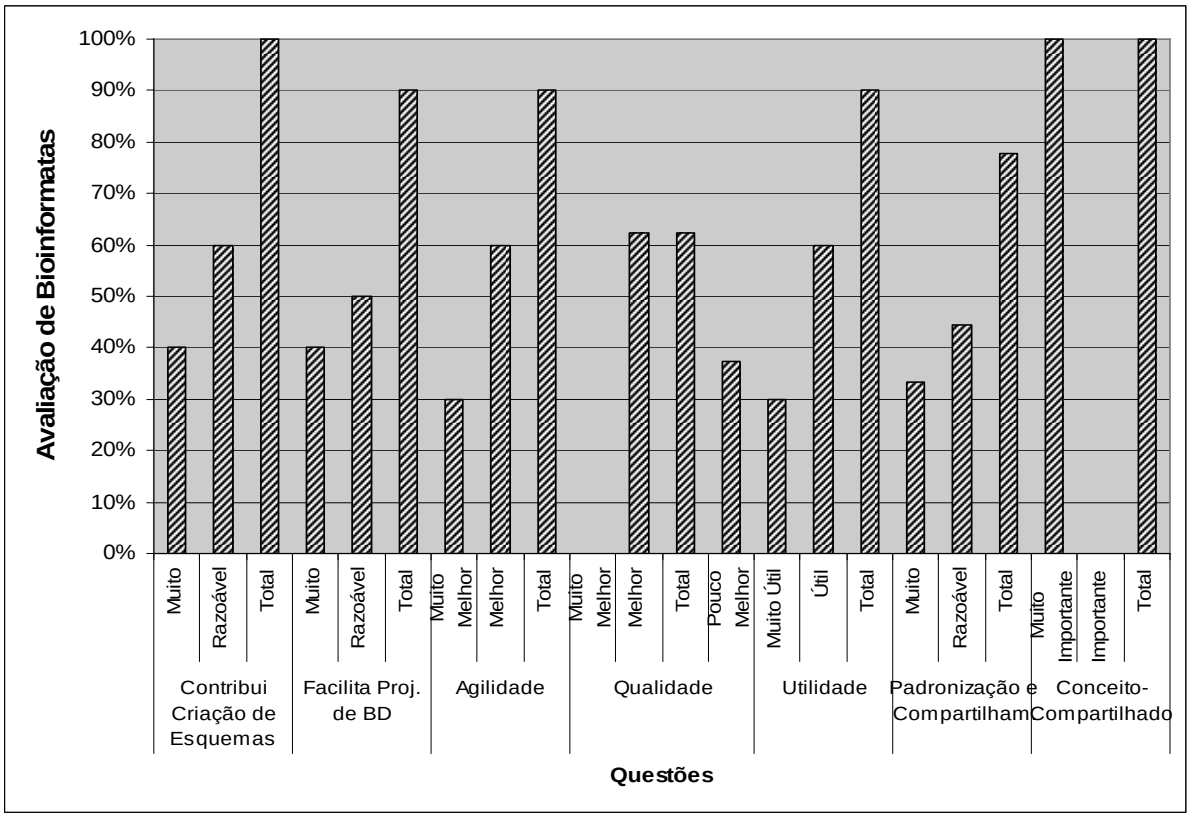


Figura 23 Gráfico de avaliação de usuários com perfil de bioinformata.

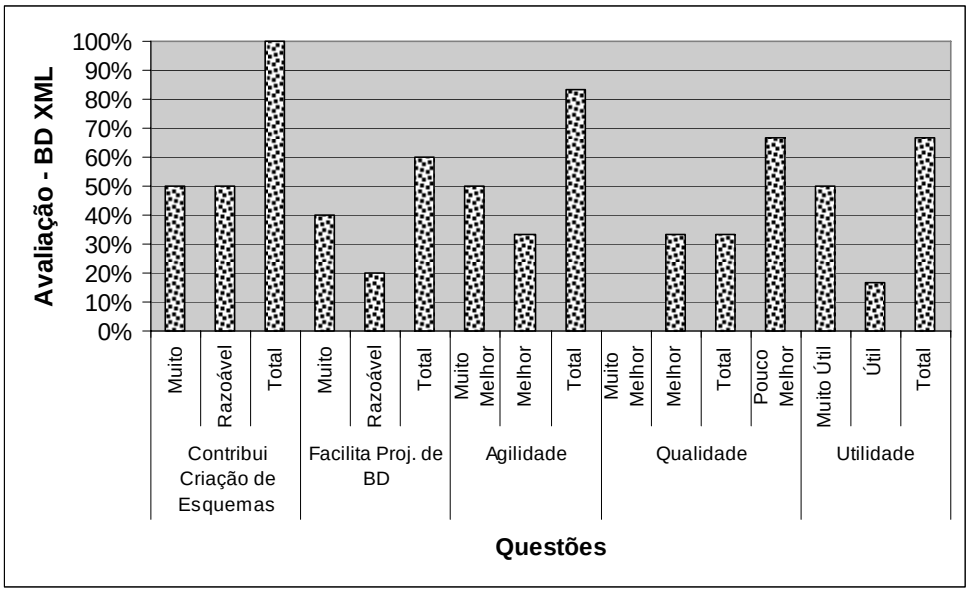


Figura 24 Gráfico de avaliação de usuários com experiência em bancos de dados XML.

Todos os usuários (100%) acreditam que a interface contribui para a criação de esquemas XML de forma razoável ou muito. Sob a perspectiva daqueles que têm conhecimento em bancos de dados XML, esse percentual é de 50% razoável e 50% muito, os quais são valores significativos. Para 84% dos usuários, essa interface facilita o projeto de banco de dados, chegando a 90% entre os bioinformatas e caindo para 60% entre aqueles que têm experiência com bancos de dados XML. Entre os aspectos considerados positivos pelos usuários para a avaliação desse quesito estão:

a. Visualização gráfica dos elementos em uma interface intuitiva e de fácil usabilidade;
b. Recomendação de um conjunto de termos e propriedades pertinentes ao domínio considerado, contribuindo para que o projetista não se esqueça de conceitos ou mesmo auxiliando-o em casos em que ele não conheça bem o domínio;
c. Uso de um vocabulário comum entre o projetista do banco de dados e biólogos especialistas no domínio;
d. Estabelecimento de regras do domínio, o que permite alertar para eventuais erros conceituais do projetista como, por exemplo, estabelecer algum relacionamento não permitido ou atribuir alguma propriedade inadequada no domínio considerado;
e. A representação de conhecimento pode ser compartilhada por mais de uma aplicação do domínio escolhido, diminuindo a heterogeneidade semântica

entre elas;

f. Diminuição no tempo despendido para desenvolvimento do projeto de banco de dados e também para criação de esquemas XML.

No entanto, críticas foram feitas em relação à dificuldade para se entender como fazer relacionamentos entre os conceitos de forma a criar hierarquias no esquema e também ao fato de essa interface não permitir a modelagem e criação de outros atributos e conceitos que não estiverem previstos na ontologia.

Com relação à maior agilidade para a criação dos esquemas quando do uso ou não uso dessa interface, a avaliação chega a 92%, sendo que 42% avaliam que ela agiliza muito o trabalho do projetista. Para aqueles usuários com experiência em bancos de dados XML, essa avaliação chega a 83%, sendo que 50% avaliam que ela agiliza muito esse trabalho. Muito dessa avaliação se deve ao fato de ser usada uma ontologia de aplicação que recomenda um vocabulário e ainda determina regras de domínio. Além disso, o fato do projetista não ter que escrever os esquemas XML também contribui. Outro aspecto levantado foi que o uso da ontologia para a modelagem tem potencial para diminuir a disparidade semântica do esquema em relação ao domínio em que a aplicação será empregada, reduzindo-se o custo de futuros esforços de integração. Um aspecto considerado como desafiante à agilidade adquirida com o uso da interface é o fato de que, uma vez que só é possível criar novos elementos e atributos manualmente, após o uso dessa interface, um projetista sem muito conhecimento em esquemas XML pode levar certo tempo para defini-los, principalmente se forem necessárias muitas modificações. Porém, vale ressaltar que outras ferramentas podem ser utilizadas para a modificação dos esquemas, não sendo necessário que o projetista edite diretamente os esquemas, sem qualquer *software* de apoio.

A respeito da qualidade dos esquemas propostos, dois terços (67%) dos usuários acredita que têm qualidade superior a esquemas gerado sem o uso da interface, sendo todas estas avaliações como melhor, ou seja, ninguém considerou que a qualidade é muito melhor. Por outro lado, 33% acreditam que a qualidade alcançada seja apenas um pouco melhor. Quando considerados apenas os usuários com experiência em bancos de dados XML, essa proporção se inverte. Novamente aqui, a qualidade é atribuída ao uso da ontologia, a qual asseguraria que os esquemas estejam em conformidade com um vocabulário padronizado e sem erros de semântica em relação ao domínio considerado. Deve-se ressaltar que a qualidade dos esquemas será tão boa quanto o vocabulário criado no *Namespace*, principalmente no que se refere à exploração dos recursos da linguagem XML Schema. Uma das críticas feitas cita

que não há como explorar os relacionamentos entre os *assets*, fator que deverá ser melhor explorado nas próximas versões dessa interface.

Outro quesito avaliado se refere à utilidade da interface XML Database Design para o projeto de bancos de dados XML. Nesse quesito, a aprovação dessa interface chegou a 85% entre todos os usuários, chegando a atingir 90% entre os bioinformatas e 67% entre aqueles com experiência em bancos de dados XML. Considerando-se somente os que a avaliaram como muito útil, esses índices chegam a 42% entre todos os usuários, 30% entre os bioinformatas e 50% entre aqueles com conhecimento em bancos de dados XML.

O gráfico da Figura 25 apresenta uma comparação entre as avaliações de todos os usuários, dos usuários com perfil de bioinformata e dos usuários com experiência em bancos de dados XML.

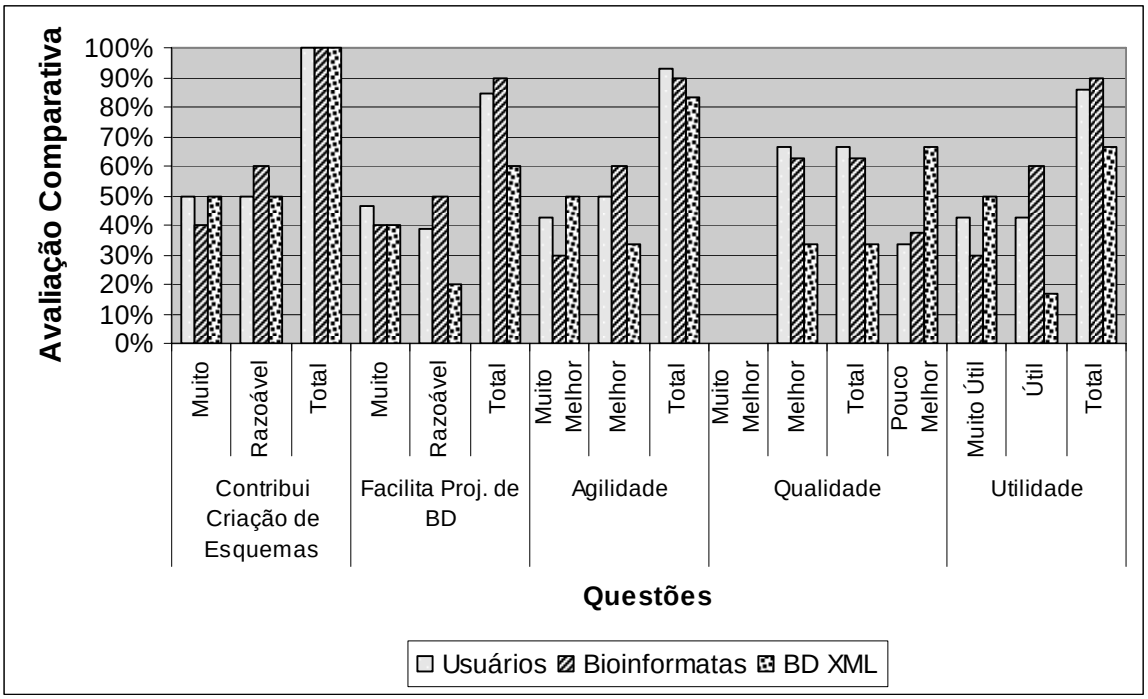


Figura 25 Gráfico comparativo entre todos os usuários e outros dois perfis analisados.

Alguns quesitos foram direcionados aos usuários com experiência em Bioinformática. O primeiro deles é em relação à contribuição que a interface desenvolvida pode dar para a padronização e compartilhamento de dados de projetos genoma. Para 33% dos bioinformatas essa interface pode contribuir muito e para outros 44% ela pode contribuir razoavelmente, chegando-se a um total de aprovação em torno de 77%. O segundo quesito se refere ao modelo conceito-compartilhado apresentado neste trabalho e explorado por essa interface. Houve grande abstenção na avaliação deste quesito, totalizando apenas 60% de respostas

entre os bioinformatas. Contudo, todos os que responderam a essa questão declararam ser este um fator muito importante. Entre os aspectos ressaltados por eles estão a flexibilidade para a criação de esquemas de dados que atendam às necessidades de um projeto e o uso de uma ontologia para controlar as correspondências semânticas entre diferentes esquemas. No entanto, diversos são os vocabulários e ontologias sendo desenvolvidas para essa área e, por isso, encontrar pontos de concordância nas definições e vocabulários em Bioinformática é muito difícil. Dessa forma, o desenvolvimento de uma ontologia de aplicação deve considerar esses fatores.

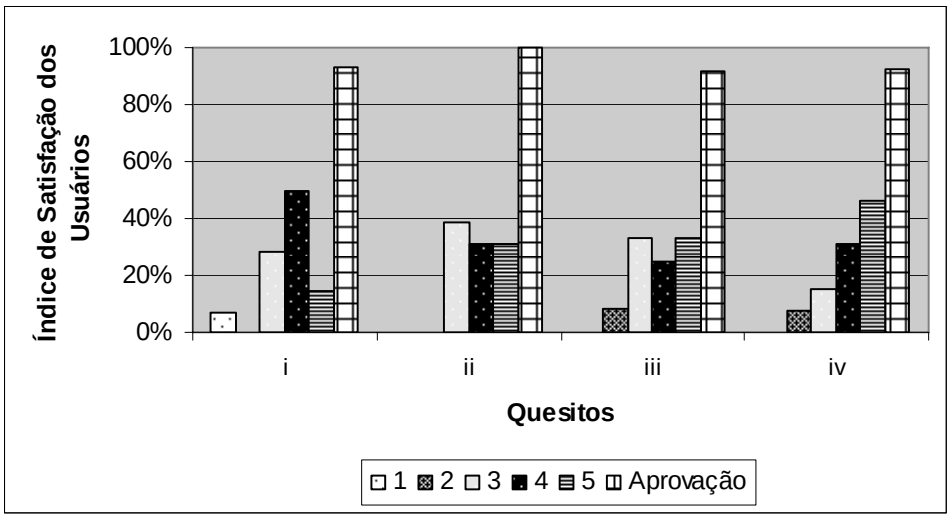


Figura 26 Índice de satisfação de todos os usuários.

Por fim, foram avaliados os índices de satisfação desses usuários (Figura 26) quanto à: (i) satisfação no uso da interface; (ii) adequação da interface à resolução do problema; (iii) instruções passadas ao usuário; e (iv) à conclusão da tarefa. A variação desses índices vai de 1 a 5, considerando-se satisfatórias as avaliações superiores ou iguais a 3, com os valores expressos em média ± desvio padrão (Média ± DP). Os resultados obtidos foram de:

- 93% para (i), sendo 50% avaliado em 4 e média de 3,64 ± 1,01;
- 100% para (ii), sendo 31% avaliados em 4 e mais 31% em 5, e média de 3,92 ± 0,86;
- 92% para (iii), sendo 33% avaliados em 5 e média de 3,83 ± 1,03;
- e 92% para (iv), sendo 46% avaliados em 5 e média de 4,15 ± 0,99.

Considerando-se somente os bioinformatas (Figura 27), os resultados obtidos foram de:

- 90% para (i), sendo 40% avaliados em 4 e média de 3,40 ± 1,07;

- 100% para (ii), sendo 50% avaliados entre 4 e 5, e média de 3,70 ± 0,82;
- 89% para (iii) e média 3,56 ± 1,01;
- e 90% para (iv), sendo 40% avaliados em 5 e média de 4,00 ± 1,05.

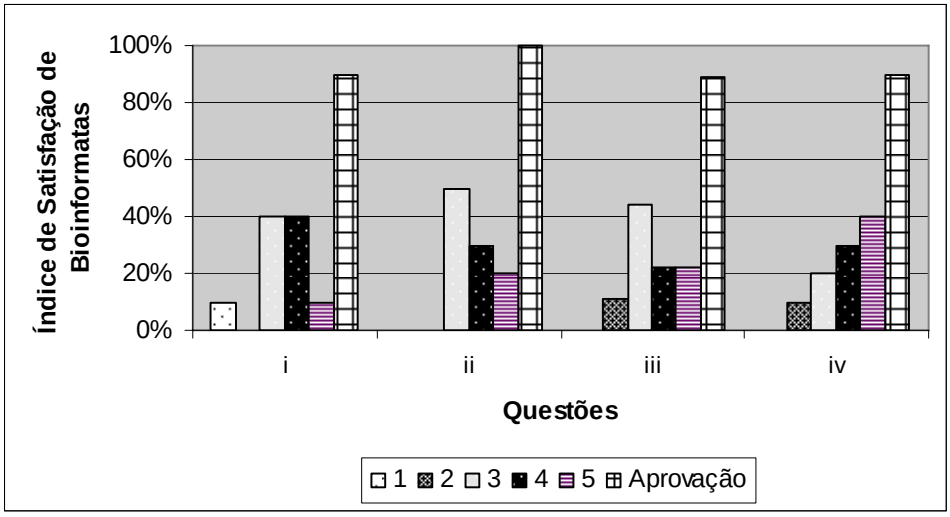


Figura 27 Índice de satisfação dos usuários com perfil de bioinformatas.

Dessa forma, podemos concluir que a interface XML Database Design alcançou um grau elevado de aprovação em relação aos quesitos analisados. Dentre todas as análises realizadas, nenhuma aparece com baixo índice de aprovação; ao contrário, a grande maioria apresentou aprovação acima de 80%.

7.5 Considerações Finais

Este capítulo apresentou o desenvolvimento da interface XML Database Design, para o projeto de bancos de dados XML baseados em uma ontologia de aplicação, e os testes realizados por usuários para a avaliação e validação da mesma. Foram apresentadas todas as interações da ontologia de aplicação com a interface, de modo a auxiliar o projetista do banco de dados na modelagem de um domínio. Os resultados dos testes foram analisados e expressos em termos estatísticos, considerando-se diferentes perfis de usuários, como bioinformatas, não-bioinformatas e aqueles com conhecimento em bancos de dados XML.

O capítulo 8 apresenta as conclusões deste trabalho, incluindo os resultados alcançados, principais contribuições e trabalhos futuros.

8 CONCLUSÃO

Este trabalho se insere no domínio de Bioinformática e propõe uma arquitetura para um ambiente de anotação de projetos genoma, chamado BioFOX. Uma vez que tais projetos produzem um grande volume de dados, é necessário a utilização de SGBDs com estruturas de dados adequadas para o armazenamento e o gerenciamento desses dados. Além disso, esses dados devem ser compartilhados entre os mais diversos grupos de pesquisa a fim de que as pesquisas em biologia molecular evoluam e se produzam novos conhecimentos, os quais poderão ser utilizados em áreas como medicina, agricultura e pecuária. Assim, a arquitetura do ambiente BioFOX propõe organizar dados genômicos por meio de ontologias e bancos de dados semi-estruturados. Ela foi desenvolvida com o objetivo de: (i) padronizar os conceitos estabelecidos para o domínio; (ii) agregar semântica aos dados; e (iii) criar bancos de dados flexíveis e apropriados para a evolução de esquemas, fatores relevantes em um domínio em constante evolução.

O ambiente BioFOX é composto por quatro módulos: (i) Módulo de Ferramentas de Bioinformática (MFB), o qual agrega um conjunto de ferramentas computacionais para análise genômica; (ii) Módulo Administrador de Conhecimento (MAC), responsável pelo controle e manutenção das ontologias, de anotação e domínio, e do Namespace de Anotações Genômicas; (iii) Módulo Repositório de Dados (MRD), composto por bancos de dados onde são armazenadas as anotações; e (iv) Módulo Interface de Anotação (MIA), o qual provê suporte ao processo de anotação manual. Neste trabalho, enfoque foi dado ao desenvolvimento do MRD, com o desenvolvimento da interface XML Database Design, e do MIA, com o desenvolvimento de um protótipo de interface para a anotação manual.

A interface XML Database Design foi desenvolvida para atender às necessidades de criação de bancos de dados com características apropriadas para o domínio de biologia molecular e projetos genoma. Assim, ela permite a geração de esquemas XML a partir da interação entre o projetista do banco de dados e uma ontologia de aplicação, a qual indica os conceitos existentes dentro desse domínio e aplica regras e restrições desse domínio. Os esquemas de dados gerados, dessa forma, compartilham de uma mesma semântica, mesmo que sejam caracterizados por estruturas e dados diferentes, seguindo a proposta do modelo de integração conceito-compartilhado. Esse fator é relevante para o bom desempenho de um banco de dados XML.

Os testes sobre a interface foram realizados por usuários com experiências em Bioinformática e bancos de dados. A análise dos resultados, obtidos a partir da aplicação de questionários, indica que a interface XML Database Design foi bem aceita, com bons índices de aprovação em diversos quesitos, inclusive quanto à contribuição para a criação de esquemas XML e quanto à facilidade para o projeto do banco de dados XML. Verificou-se que o uso de uma ontologia de aplicação contribui para que o projetista tenha um melhor entendimento sobre o domínio do projeto e para que ele não cometa erros semânticos que possam comprometer o seu esquema. Além disso, ela também contribui para o compartilhamento dos conceitos envolvidos no domínio considerado, o que implica na diminuição da heterogeneidade semântica entre bancos de dados de diferentes projetos.

As seções seguintes contemplam: as contribuições deste trabalho (seção 8.1) e uma lista de trabalhos futuros (seção 8.2).

8.1 Contribuições

As principais contribuições deste trabalho são listadas a seguir:

- Proposta de uma arquitetura para um ambiente de anotação, dividida em módulos bem caracterizados, segundo suas funcionalidades, para a manipulação e a organização de dados provindos de um projeto genoma;
- Desenvolvimento de uma interface para a modelagem conceitual de domínios complexos, como o de biologia molecular, e criação de esquemas de dados XML;
- Criação de esquemas de dados estruturalmente independentes, mas com semântica associada por meio de uma ontologia;
- Proposta do modelo de integração conceito-compartilhado, o qual explora as características de esquemas de dados XML associados a uma ontologia;
- Desenvolvimento de uma ontologia de aplicação contendo definições de conceitos e regras de domínio, com o objetivo de guiar e auxiliar a modelagem conceitual de dados;

Desenvolvimento de um *namespace* de vocabulários XML para a anotação de projetos genoma.

8.2 Trabalhos Futuros

Dentre as idéias formuladas para a continuidade deste trabalho, destacam-se:

- Consultas semânticas sobre os dados XML, a partir da ontologia de aplicação, incluindo-se a expansão semântica dessas consultas;
- Exploração da integração de dados a partir do modelo conceito-compartilhado apresentado neste trabalho;
- Componentização da interface de anotação manual;
- Expansão dos vocabulários de domínios já definidos no *Namespace* de Anotação Genômica e a inclusão de novos domínios;
- Aplicação da interface XML Database Design a diferentes domínios de conhecimento, além do domínio de biologia molecular;
- Melhoria da interface XML Database Design, com a implementação de funcionalidades como a representação de relacionamentos, a definição de novos atributos e entidades, visualização gráfica do esquema sugerido e também da ontologia de aplicação. Essas foram algumas das sugestões feitas pelos usuários que testaram essa interface.

REFERÊNCIAS

ACHARD, F.; VAYSSEIX, G.; e BARILLOT, E. XML, bioinformatics and data integration. Bioinformatics, v. 17, p. 115-125, 2001.

ALTSCHUL, S. F. et al. Basic local alignment search tool. Journal of Molecular Biology, v. 215, p. 403-410, 1990.

ALTSCHUL, S. F. e KOONIN, E. V. Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases. Trends Biochemistry Science, v. 23, p. 444-447, 1998.

ANDRADE, M. et al. Automated genome sequence analysis and annotation. Bioinformatics, v. 15, v. 5, May, p. 391-412, 1999.

APWEILER, R. et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Research, v. 29, p. 37-40, 2001.

ASHBURNER, M. et al. Gene Ontology: tool for the unification of biology. Nature Genetics, v. 25, May 2000

ASSET ORIENTED MODELING. Asset Oriented Modeling (AOM). Disponível em: <http://www.aomodeling.org/>. Acesso em: 01 abr. 2008.

ATTWOOD, T. K. et al. PRINTS-S: the database formerly known as PRINTS. Nucleic Acids Research, v. 28, p. 225-227, 2000.

BALL, C. A. et al. Integrating functional genomic information into the *Saccharomyces* Genome Database. Nucleic Acids Research, v. 28, p. 77-80, 2000.

BASU, S. et al. MiGenes: a searchable interspecies database of mitochondrial proteins curated using gene ontology annotation. Bioinformatics, v. 22, p. 485-492, 2006.

BENSON, D. A. et al. GenBank. Nucleic Acids Research, v. 33, Database issue, p. D34-D38, 2005.

BERKELEY DROSOPHILA GENOME PROJECT. Chaos-XML. Disponível em: <http://www.fruitfly.org/chaos-xml/>. Acesso em: 03 abr. 2008.

BERKELEY DROSOPHILA GENOME PROJECT. GAME - Genome Annotation Markup Language. Disponível em: <http://www.fruitfly.org/annot/gamexml.dtd.txt>. Acesso em: 03 abr. 2008.

BESEMER, J. e BORODOVSKY, M. Heuristic approach to deriving models for gene finding. Nucleic Acids Research, v. 27, p. 3911-3920, 1999.

BESEMER, M. G. et al. Genie - gene finding in *Drosophila melanogaster*. Genome Research, v. 10, p. 529-538, 2000.

BIOGRID. BioGRID. Disponível em: <http://www.thebiogrid.org/>. Acesso em: 06 jun. 2008.

BIOPERL. BioPerl. Disponível em: <http://www.bioperl.org/>. Acesso em: 06 jun. 2008.

BLAKE, J. A. et al. The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. Nucleic Acids Research, v. 28, p. 108-111, 2000.

BRAGANHOLO, V. d. P.; HEUSER, C. A. XML Schema, RDF(S) e UML: uma comparação. IDEAS'2001 (IV Workshop Iberoamericano de Ingeniería de Requisítos y Ambientes de Software). Santo Domingo, 2001. p. 78-90.

BURLIN, C.; KARLIN, S. Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology, v. 268, p. 78-94, 1997.

BUTEMAN, A. et al. The Pfam protein families database. Nucleic Acids Research, v. 28, p. 263-266, 2000.

CAENORHABDITIS elegan genetic and genomics. Disponível em: <http://elegans.swmed.edu/genome.shtml>. Acesso em: 03 abr. 2008.

CARROLL, J. J. et al. Jena: implementing the semantic web recommendations. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE ON ALTERNATE TRACK PAPERS & POSTERS, 13., New York, 2004. Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters. New York: ACM, 2004. p. 74-83.

CORPET, F.; GOUZY, J.; KAHN, D. Recent improvements of the ProDom database of protein domains families. Nucleic Acids Research, v. 27, p. 263-267, 1999.

DÁVILA, A. M. R. et al. GARSA: genomic analysis resources for sequence annotation. Bioinformatics, v. 21, n. 23, Oct., p. 4302-4303, 2005.

DNA DATA BANK OF JAPAN. DNA Data Bank of Japan. Disponível em: <http://helix.genes.nig.ac.jp/homology/ssearch-e_help.html>. Acesso em: 03 Abr. 2008.

DOMSELAAR, G. H. V. et al. BASys: a web server for automated bacterial genome annotation. Nucleic Acids Research, v. 33, May, p.W455-W459, 2005.

DORNELES, C. F. Extração de dados semi-estruturados com base em uma Ontologia. 2000. 49 f. (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2000.

EILBECK, K. et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biology, v. 6, n. 5, p. R44, 2005.

EUROPEAN BIOINFORMATICS INSTITUTE. The EMBL Nucleotide Sequence Database. Disponível em: <http://www.ebi.ac.uk/embl/>. Acesso em: 03 abr. 2008.

EUROPEAN BIOINFORMATICS INSTITUTE. SWISS-PROT and TrEMBL. Disponível em: <http://www.ebi.ac.uk/swissprot/>. Acesso em: 03 abr. 2008.

FRISHMAN, D. et al. The PEDANT genome database. Nucleic Acids Research, v. 31, p. 207-211, 2003.

GENERIC MODEL ORGANISM DATABASE. Chado Schema. Disponível em: <http://www.gmod.org/chado>. Acesso em: 03 abr. 2008.

GENERIC MODEL ORGANISM DATABASE. Generic Model Organism Database. Disponível em: <http://www.gmod.org/>. Acesso em: 03 abr. 2008.

GLASNER, J. D. et al. ASAP, a systematic annotation package for community analysis of genomes. Nucleic Acids Research, v. 31, p. 147-151, 2003.

GOPALACHARYULU, P. V. et al. Data integration and visualization system for enabling conceptual biology. Bioinformatics, v. 21, p.177-185, Jan. 2005.

GRAVES, M. Projeto de Banco de Dados com XML. Local de pulicação: Pearson Education do Brasil, 2003.

GRUBER, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human-Computer Studies, v. 43, p. 907-928, 1995.

GUARINO, N. Formal ontology in information systems. Netherlands: IOS Press., Ed. Amsterdam, 1998.

GUERRINI, V. H.; JACKSON, D. Bioinformatics and XML. On Line Journal of Bioinformatics, v. 1, p. 1-13, 2000.

HARRIS, N. L. Genotator: A Workbench for Sequence Annotation. Genome Research, v. 7, p. 754-762, 1997.

HENIKOFF, J. G. et al. Increased coverage of protein families with the BLOCKS database servers. Nucleic Acids Research, v. 8, p. 228-230, 2000.

HOFFMAN, K. et al. The PROSITE database, its status in 1999. Nucleic Acids Research, v. 27, p. 215-219. 1999.

INTERNATIONAL ORGANIZATION FOR STANDARTIZATION (ISO). ISO 8824:. Information processing systems - Open Systems Interconnection - Specification of Abstract Syntax Notation One (ASN.1). Switzerland. 1987.

JENSEN, R. A. Orthologs and paralogs - we need to get it right. Genome Biology, v. 2, n. 8, Aug. 2001.

KROGH, A. Two methods for improving performance of an HMM and their application for gene finding. In: INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS FOR MOLECULAR BIOLOGY. Halkidiki, Greece, 1997. ISBC, 1997. p. 179-186.

LEMOS, M. Workflow para bioinformática. 2004. 239 f. (Doutorado em Ciência da Computação) - Departamento de Informática, Pontífica Universidade Católica-Rio, Rio de Janeiro, 2004.

LEMOS, M.; SEIBEL, L. F. B.; CASANOVA, M. A. BioNotes: A System for Biosequence Annotation. In: INTERNATIONAL WORKSHOP ON DATABASE AND EXPERT SYSTEMS APPLICATIONS, 14., Prague, 2003. Prague, Czech Republic: IEEE Computer Society, 2003a. p. 16.

LEMOS, M.; SEIBEL, L. F. B.; CASANOVA, M. A. Sistemas de anotações em biosseqüências. Rio de Janeiro: 2003b.

LEMOS, M., SEIBEL, L. F. B. E CASANOVA, M. A. Functional Requirements of Biosequence Annotation Systems. Rio de Janeiro: 2004.

LEWIS, S. E. et al. Apollo - a sequence annotation editor. Genome Biology, v. 3, n. 12, Dec. 2002.

LOMBARDO, L. R. FrameEST: um framework de componentes, no padrão MVC, para o domínio de biologia molecular. 2006. 80 f. (Mestrado em Ciência da Computação) - Departamento de Computação, Universidade Federal de São Carlos, São Carlos, 2006.

MELLO, R. d. S. et al. Dados Semi-Estruturados. SBBD. João Pessoa, Paraíba, Brasil**:** p. 39 2000.

MEYER, F. et al. GenDB - an open source genome annotation system for prokaryote genomes. Nucleic Acids Research, v. 31, p. 2187-2195. 2003.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. XML at NCBI. Disponível em: <http://www.ncbi.nlm.nih.gov/IEB/ToolBox/XML/>. Acesso em: 03 abr. 2008.

NECASKY, M. XSEM - A Conceptual Model for XML Data. In: ASIA-PACIFIC CONFERENCE ON CONCEPTUAL MODELLING, 4. Australia: Australian Computer Society, 2007. p. 37-48.

OBJECT MANAGEMENT GROUP. OMG Unified Modeling Language Specification. Disponível em: <http://www.omg.org/technology/documents/formal/unified_modeling_language.htm>. Acesso em: 03 abr. 2008.

OKURA, V. K. Bioinformática de Projetos Genoma de Bactérias. 2002. 117 f. (Mestrado em Ciência da Computação) - Instituto de Computação, UNICAMP, Campinas, 2002.

OLIVEIRA, G. B. d. Bio-TIM - Ambiente para convergência de informações em Bioinformática. 2005. 98 f. (Mestrado em Ciência da Computação) - Departamento de Computação, Universidade Federal de São Carlos, São Carlos, 2005.

OVERBEEK, R. et al. The ERGO genome analysis and discovery system. Nucleic Acids Research, v. 31, p. 164-171, 2003

PELLEGRINI, M.; MARCOTTE, E. M.; YEATES, T. O. A fast algorithm for genome-wide analysis of proteins with repeated sequences. PUBMED, 1999.

PONTING, C. P. et al. SMART: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Research, v. 27, p. 229-232, 1999

PROTEOMETRICS. Biopolymer Markup Language. Disponível em: <http://www.proteometrics.com/BIOML/>. Acesso em: 03 abr. 2008.

PSAILA, G. ERX: a conceptual model for XML documents. In: ACM Symposium ON APLIED COMPUTING, 2000, Como, Italy, p. 898-903, 2000.

RUTHERFORD, K. et al. Artemis: sequence visualization and annotation. Bioinformatics, v. 16, p. 944-945, 2000.

SANGER INSTITUTE. AceDB. Disponível em: < http://www.acedb.org/>. Acesso em: 03 Abr. 2008.

SCHULZE-KREMER, S. Ontologies for molecular biology. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING, 3., 1998, Maui, Hawaii: AAAI Press, 1998. p. 693-704.

SEIBEL, L. F. B. Bio-AXS: uma arquitetura para integração de fontes de dados e aplicações de biologia molecular. 2002. 181 f. (Mestrado ou Doutorado em Área de Concentração) - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2002.

SEIBEL, L. F. B.; LEMOS, M.; LIFSCHITZ, S. Bancos de Dados de Genoma. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 15., João Pessoa: SBBD, 2000.

SEIBEL, P. N. et al. XML schemas for common bioinformatic data types and their application in workflow systems. Bioinformatics, v. 7, p. 490, Nov. 2006.

SENGUPTA, A.; MOHAN, S.; DOSHI, R. XER - Extensible Entity Relationship Modeling. In: XML CONFERENCE & EXPOSITION, 2003, Philadelphia. Philadelphia: IDEAlliance, 2003. p. 140-154.

SHEST. ShEST - Projeto Genoma EST do Camarão *Litopenaeus vannamei*. Disponível em: <http://shrimp.dc.ufscar.br/>. Acesso em: 03 abr. 2008.

SOFTWARE AG. Tamino XML Server. Disponível em: <http://www.softwareag.com/tamino/>. Acesso em: 03 abr. 2008.

SOLOVYEV, V.; SALAMOV, A.; LAWRENCE, C. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucleic Acids Research, v. 22, p. 5156-5163. 1994.

SOLOVYEV, V.; SALAMOV, A.; LAWRENCE, C. Identification of human gene structure using linear discriminant functions and dynamic programming. In: INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS FOR MOLECULAR BIOLOGY, 3., Cambridge, United Kingdom, 1995. p. 367-375.

STAMMA, S. et al. Function of alternative splicing. Gene, 344, p.1 – 20, 2005.

STEIN, L. Genome annotation: from sequence to biology. Genetics, v. 2, p.493-505, Jul. 2001.

STOFFEL, K.; TAYLOR, M.; HENDLER, J. Efficient Management of Very Large Ontologies. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 14., Providence: MIT–AAAI Press, 1997. p. 442-447.

THALHEIM, B. Entity-Relationship Modeling - Foundations of Database Technology. Berlim: Springer, 2000.

THE FLYBASE. The FlyBase database of the *Drosophila* genome projects and community literature. Nucleic Acids Research, v. 27, p. 85-88, 1999.

THE GENE ONTOLOGY. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, v. 32, Database issue, p. D258-D261, 2004.
THE GENE ONTOLOGY. Gene Ontology. Disponível em: <http://www.geneontology.org/>. Acesso em: 03 abr. 2008.

THE SEQUENCE ONTOLOGY. Generic Feature Format. Disponível em: <http://www.sequenceontology.org/gff3.shtml>. Acesso em: 03 abr. 2008.

UBERACHER, E.; MURAL, R. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proceedings of the National Academy of Sciences USA, v. 88, p. 11261-11265, 1991.
USCHOLD, M.; GRUNINGER, M. Ontologies: principles, methods and applications. Knowledge Engineering Review, v. 11, n. 2, p. 93-155, 1996.

W3C. Resource Description Framework (RDF). Disponível em: <http://www.w3.org/RDF/>. Acesso em: 03 abr. 2008.

WAUGH, A. et al. RNAML: a standard syntax for exchanging RNA information. RNA, v. 8, v. 6, Jun., p. 707-717, 2002.

ZHANG, M. Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proceedings of the National Academy of Sciences USA, v. 94, p. 565-568, 1997.

APÊNDICE A – QUESTIONÁRIO DE AVALIAÇÃO DA INTERFACE XML DATABASE DESIGN

Avaliação da Interface XML Database Design

Possui experiência em bioinformática?

( ) Sim

( ) Não

Possui experiência em projetar/modelar bancos de dados?

( ) Sim

( ) Não

Se sim, com que freqüência?

( ) Sempre

( ) Muito freqüentemente

( ) Freqüentemente

( ) Pouco freqüentemente

( ) Raramente

Possui experiência em projetar/modelar bancos de dados XML?

( ) Sim

( ) Não

Se sim, com que freqüência?

( ) Sempre

( ) Muito freqüentemente

( ) Freqüentemente

( ) Pouco freqüentemente

( ) Raramente

Indique o seu grau de conhecimento, anterior ao uso dessa ferramenta, a respeito do que são ontologias e como elas são ou podem ser empregadas no meio computacional?

( ) Bom
( ) Razoável
( ) Pouco
( ) Nenhum

Esta ferramenta contribui efetivamente para a criação de esquemas de bancos de dados XML?

( ) Muito
( ) Razoável
( ) Pouco
( ) Nada

O uso da ferramenta facilitou o projeto do banco de dados? Como?

( ) Muito
( ) Razoável
( ) Pouco
( ) Nada

_____
_____
_____
_____
_____

Como você avalia o uso dessa ferramenta, em termos de agilidade para a composição de esquemas, quando comparado à não utilização da mesma?

( ) Muito melhor
( ) Melhor
( ) Pouco melhor
( ) Nada
( ) Pior

Críticas e sugestões:

_____

_____

_____

_____

_____

Como você avalia o uso dessa ferramenta, em termos de qualidade dos esquemas produzidos, quando comparado à não utilização da mesma?

( ) Muito melhor

( ) Melhor

( ) Pouco melhor

( ) Nada

( ) Pior

Críticas e sugestões:

_____

_____

_____

_____

_____

Considera natural a seqüência de passos apresentada pela ferramenta para um projeto de bancos de dados?

( ) Sim

( ) Mais ou menos

( ) Não

Sugestões de modificações dos passos e/ou adição de passos extras:

_____

_____

_____

_____

Como você avalia a forma como esta ferramenta permite a criação de diferentes esquemas XML?

( ) Muito útil

( ) Útil

( ) Pouco útil

( ) Desnecessário

As questões 11 e 12 são mais específicas para o domínio de Bioinformática, devendo ser respondida preferencialmente por pessoas dessa área. Quem não for dessa área mas se julgar em condições de respondê-las, sinta-se à vontade.

Considerando-se o contexto de Bioinformática, caracterizada por um complexo conjunto de dados, você acredita que essa ferramenta pode contribuir para a padronização e compartilhamento de dados de projetos genoma?

( ) Muito

( ) Razoável

( ) Pouco

( ) Nada

Considerando-se a integração de dados em Bioinformática, você acredita que a definição de diferentes esquemas a partir de um mesmo vocabulário, proporcionada por essa ferramenta, pode contribuir para a independência estrutural dos bancos de dados sem, contudo, perder a semântica entre eles? Como?

_____

_____

_____

_____

_____

Qual a importância disso?

( ) Muito importante

( ) Importante

( ) Pouco importante

( ) Desnecessário

Sobre a Interface de Anotação, como avalia as funcionalidades de auto-completar e sugestão de palavras correlatas em determinados campos?

( ) Muito útil

( ) Útil

( ) Pouco útil

( ) Desnecessário

Críticas e sugestões de novas funcionalidades:

_____

_____

_____

_____

_____

APÊNDICE B – QUESTIONÁRIO DE AVALIAÇÃO DA SATISFAÇÃO DA INTERAÇÃO DO USUÁRIO

QUIS – Questionário da Satisfação da Interação do usuário

Por favor, em cada questão abaixo, indique com um "x" a alternativa que melhor define a sua impressão sobre o uso deste sistema. A satisfação do usuário deve variar de grau 1 (mais baixo) a 5 (mais alto) ou N/A (Não se aplica).

PARTE A – Reação do Sistema

		1	2	3	4	5		N/A
	Frustrante						Satisfatório	
	Tedioso						Estimulante	
	Difícil						Fácil	
	Inadequado						Adequado	

PARTE B – Terminologia e Informação do Sistema

		1	2	3	4	5		N/A
Termos usados pelo sistema	Confuso						Claro	
Terminologia relativa à tarefa	Nunca						Sempre	
Mensagens que aparecem na tela	Confuso						Claro	
Localização das mensagens na tela	Confusa						Clara	
Instruções para o usuário	Nunca						Sempre	
Sistema mantém você informado sobre o progresso da tarefa	Nunca						Sempre	
Mensagens de erro	Inútil						Útil	

PARTE C – Aprendizado

		1	2	3	4	5		N/A
Aprender a operar o sistema	Difícil						Fácil	
Explorar por tentativa e erro	Difícil						Fácil	
Tarefas podem ser executadas de uma forma rápida e/ou lógica	Nunca						Sempre	
Conclusão da tarefa	Confuso						Claro	

PARTE D – Capacidade do Sistema

		1	2	3	4	5		N/A
Velocidade do sistema	Lento						Rápido	
O sistema é confiável	Não Confiável						Confiável	
Corrigir seus erros	Difícil						Fácil	
Projetado para todos os níveis de usuários (iniciantes e experientes)	Nunca						Sempre	

PARTE E – Sugestões e Comentários