

2

Contexto da Pesquisa

2.1

Introdução

Neste capítulo apresentaremos as principais questões relacionadas com contexto da pesquisa desta tese. Primeiramente discutiremos sobre modelos conceituais e conceitos sobre modelagem, incluindo ontologias, modelos semânticos e lógicos, todos relacionados com o desenvolvimento de sistemas de informação. Em seguida, detalharemos o domínio estudado nesta tese, em particular, a biologia molecular, e algumas das necessidades principais que devem ser consideradas. Estas serão a base para a discussão sobre a motivação para todo o trabalho, a qual deixaremos para o próximo capítulo.

2.2

Ciclo de Vida de um Sistema de Informação

O projeto de um sistema de informação é um processo para solução de um problema. Neste caso, problema significa um contexto, ambiente ou situação onde um sistema de informação poderia ser desenvolvido ou operado. Para o desenvolvimento de um sistema de informação com sucesso é necessário compreender todos componentes do problema, manipulação de dados, relações, regras, restrições, etc. Tal compreensão é processo difícil e demorado, a qual requer ferramentas especializadas para serem realizadas.

O projeto de um sistema de informação deve ser orientado por um processo sistemático chamado de metodologia de projeto. Várias metodologias de projeto foram propostas para suportar a construção de sistemas de informação. Em geral, essas metodologias englobam inteiramente o ciclo de vida de um sistema de informação que tipicamente inclui as seguintes fases:

1. Coleta e Análise de Requisitos
2. Projeto do Sistema (Especificação Lógica)
3. Implementação (Codificação)
4. Teste de aceitação
5. Implantação, Operação e Manutenção

Atualmente, um sistema de banco de dados é tipicamente parte de um sistema de informação, o qual inclui todos os componentes que são envolvidos na organização dos recursos. Esses recursos incluem os próprios dados, o software do SGBD, o hardware do sistema, a mídia de armazenamento e os aplicativos que acessam e atualizam os dados. Desta forma, o sistema de banco de dados é uma parte muito importante de toda a organização de um sistema de informação. Por causa disto, o ciclo de um sistema de informação deve também incluir o ciclo de vida de um banco de dados.

2.2.1

Banco de Dados e Sistema Gerenciador de Banco de Dados

Um banco de dados é uma coleção compartilhada de dados logicamente relacionados, projetado para atender as necessidades de informação de múltiplos usuários em uma organização. Os banco de dados armazenam dados relativos a um domínio particular e eles possuem as seguintes propriedades implícitas[41]:

- representa algum aspecto do mundo real, chamado de Universo de Discurso (UoD), o qual deve ser mantido atualizado dentro do banco de dados;
- é projetado, construído, e populado com dados de um domínio particular. Ele possui um grupo de usuários e algumas aplicações que interessam a esses usuários;
- é uma coleção consistente de dados logicamente inter-relacionados;

Um sistema gerenciador de banco de dados (SGBD) é uma coleção de componentes de software para criar, gerenciar e consultar um banco de dados. Um SGDB pode conter diversos bancos de dados, e cada um deles pode ter um administrador de banco de dados diferente (DBA).

É importante distinguir entre a descrição de um banco de dados, ou esquema de banco de dados, do próprio banco de dados. Um esquema de banco de dados pode ser usado para gerar um ou mais banco de dados, mas um banco de dados pode ser gerado apenas por um esquema de banco de dados. Além disso, banco de dados baseados em um mesmo esquema podem armazenar dados diferentes.

Os dados de um banco de dados em determinado momento do tempo é chamado de estado do banco de dados. Ele pode também ser chamado de conjunto corrente de ocorrências ou instâncias de um banco de dados. Por exemplo, a Figura 2.1 ilustra que podemos ter um esquema de dados relacionado a diversos bancos de dados. Um esquema de dados armazena informação sobre a estrutura do banco de dados, em particular, nome do cliente, idade e país. Cada banco de dados armazena dados de acordo com a estrutura do banco de dados, no nosso exemplo, (John, 33, USA) é uma instância do banco de dados #1.

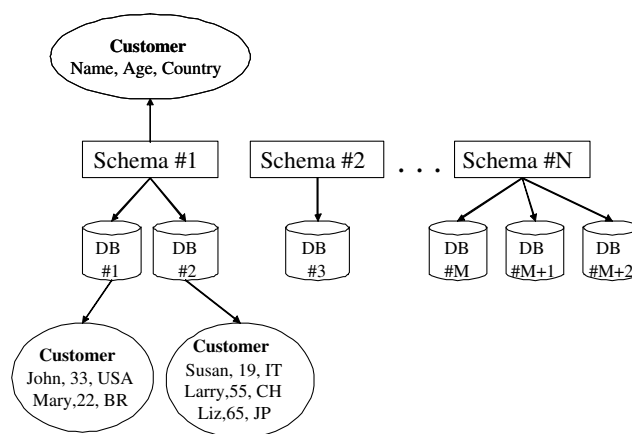


Figura 2.1: Banco de Dados e Esquemas de Dados

2.2.2

Projeto de Banco de Dados

O processo de criação de um banco de dados inicia com a análise de quais informações o banco de dados deve lidar. A seguir, a estrutura do banco de dados deve ser especificada através da declaração de tipos de dados, relacionamentos entre as informações e restrições que devem ser mantidas sobre os dados.

Após a criação da estrutura do banco de dados, precisamos popular o banco de dados em alguma mídia para armazenamento que é controlada pelo SGBD. Então podemos manipular o banco de dados usando operações, tais como: consultar o banco dados para recuperar dados específicos, atualizar o banco de dados para refletir mudanças do mundo real, e gerar relatórios a partir do dados.

Navathe et al [41] dividem o ciclo de uma aplicação de banco de dados em oito fases, tais como: projeto de banco de dados, implementação de banco de dados, carga e conversão de dados, conversão de aplicação, teste e validação, operação e monitoramento, e manutenção.

O objetivo do projeto de um banco de dados é especificar a estrutura lógica e física de um ou mais banco de dados. Esta estrutura deve acomodar a informação necessária para usuários de uma organização para um conjunto definido de aplicações. Esses objetivos não são fáceis de se alcançar porque enquanto o processo de projeto inicia com os requisitos definidos informalmente, o seu resultado deve ser um esquema de banco de dados bem definido. Além disso, o processo de projeto envolve um balanceamento entre a compreensão do modelo e sua performance.

Embora existam diferentes propostas de metodologia para o projeto de banco de dados, o projeto de banco de dados pode ser dividido em quatro passos distintos[50, 51, 42]:

1. Durante a análise dos requisitos, parte do mundo real que precisa ser modelada deve levar em conta as necessidades de informação (i.e. consultas principalmente) dos potenciais usuários do banco de dados;
2. Usando as informações levantadas no passo anterior, a estrutura e comportamento do banco de dados deve ser especificada formalmente no esquema conceitual. Adicionalmente, as restrições devem ser definidas de maneira a definir restrições que devem

ser garantidas na checagem da consistência do banco de dados durante a inserção dos dados.

3. O esquema conceitual resultante da fase anterior é mapeado para um esquema de um modelo de dados lógico, por exemplo, do tipo relacional. Este passo, chamado de projeto lógico, é necessário para mapear as ricas estruturas conceituais, usadas na modelagem conceitual, nas estruturas processáveis usadas na implementação efetiva em SGBDs.
4. Finalmente, questões relacionadas com a implementação do banco de dados, tais como tipos de dados, índices, visões, estruturas de armazenamento, são tratadas no projeto físico do banco de dados.

Entre esses passos, o projeto conceitual possui um papel muito importante porque ele enriquece a comunicação entre o projetista e o especialista do domínio, facilitando futuras mudanças na aplicação ou na implementação do banco de dados. Supomos que os modelos conceituais podem ser a ferramenta ideal para facilitar a compreensão do problema a ser solucionado durante o projeto do banco de dados.

2.3

Modelos Conceituais de Dados

O resultado do processo do projeto conceitual de banco de dados é um esquema conceitual de dados. O escopo do termo Modelo Conceitual estendeu-se gradualmente desde a aprovação do padrão ISO para modelagem conceitual[3], onde seu objetivo era representar o domínio do discurso. O domínio do discurso é um conjunto de dados envolvidos no problema a ser resolvido e as operações que afetam os dados.

São dois os objetivos dos modelos conceituais. Primeiro, os modelos conceituais são usados para descrever a informação a ser manipulada por um sistema de informação. Segundo, os esquemas conceituais são traduzidos em esquemas de dados lógicos que serão usados para implementar um banco de dados. Isto significa que o termo modelo conceitual é usado duas vezes durante a construção do banco de dados, para dois diferentes propósitos:

- Descritivo: o modelo conceitual descreve uma parte do mundo, e é a saída do processo de modelagem conceitual. Desta forma, o modelo conceitual é derivado da realidade e seu propósito é ganhar compreensão sobre tal realidade
- Prescritivo: o modelo conceitual é usado para descrever o modelo lógico que é um modelo computacional, resultante da análise de requisitos e das atividades de projeto.

A Figura 2.2 fornece uma visão clara desta distinção, indicando que o desenvolvimento do banco de dados pode ser visto como dois momentos sucessivos e distintos. A caixa apresentada nesta figura simboliza todo o processo do projeto do banco de dados, e as setas o fluxo de atividades.

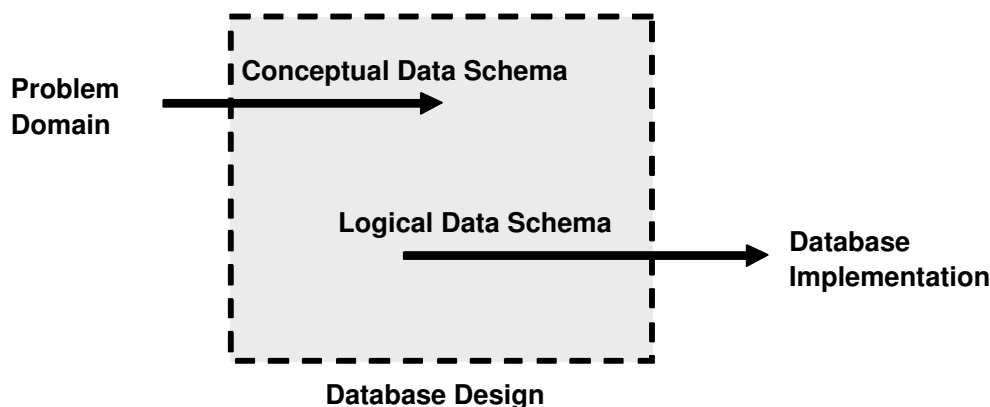


Figura 2.2: Um esquema do processo de projeto de um banco de dados [16]

2.3.1

Modelagem Conceitual de Dados

O processo de criação de esquemas conceituais durante o projeto de banco de dados é geralmente referenciado como modelagem conceitual, embora possam ser dados outros nomes dependendo da disciplina na qual isto se aplique, por exemplo, análise do problema, em Engenharia de Software[12] ou conceitualização, em Engenharia do Conhecimento[10]. Na área de banco de dados utiliza-se o termo modelagem conceitual de dados o qual possui os seguintes objetivos[17]:

- **Abordagem orientada ao projeto:** permite aos projetistas focarem no seu principal problema com a mínima interferência das restrições técnicas, e seu resultado é mais estável no tempo do que modelos orientados à implementação que se tornam obsoletos com a mudança da plataforma de implementação destino;
- **Aderência a linguagem humana:** provê um melhor suporte para interfaces visuais. Modelos entidade-relacionamento (ER)[1], por exemplo, permitem usuários visualizar facilmente os esquemas de banco de dados e assim compreender a conteúdo de informação do banco de dados. Tais modelos também suportam técnicas de manipulação através ferramentas especializadas (ex., ponto, clicar e arrastar elementos na tela do computador) para navegar pelo banco de dados, e expressar consultas e atualizações sem a complicação das linguagens textuais mais elaboradas;
- **Facilidades para troca de informação:** facilita a troca de informação entre banco de dados heterogêneos (distribuídos ou federados). Em tais contextos, um modelo conceitual provê o melhor veículo para a compreensão entre parceiros com diferentes bases técnicas. Abstrair a parte técnica pode facilitar a descrição global de toda informação relevante[7, 18].

2.4

Tipos de Modelos Conceituais

O termo Modelo Conceitual originalmente emergiu da área de banco de dados. Modelos conceituais foram usados para representar dados e relações, as quais eram

gerenciadas por um sistema de informação, independente de qualquer característica de implementação[2]. No entanto, como vimos anteriormente, o termo modelo conceitual estendeu-se por outras áreas, onde o seu objetivo era representar o domínio do discurso, o qual representa o conjunto de dados envolvidos no problema a ser resolvido e nas operações que afetam esses dados. Neste contexto, somente as operações que especificam restrições de integridade do problema a ser resolvido são representadas no modelo conceitual.

Uma grande variedade de modelos tem sido utilizadas na disciplina de projeto de sistemas de informação, a maioria destas pertencem ao conjunto denominado modelos conceituais. A melhor maneira de examinar os modelos pertencentes a este conjunto é agrupá-los de acordo com as suas orientações. Os grupos podem ser definidos como a seguir:

- Modelos Conceituais Funcionais, cujo paradigma é o DFD. Este tipo de modelo descreve as transformações de dados usadas no problema do domínio em análise. As transformações são descritas pelo conceito de processo, as quais recebem um conjunto de dados de entrada e gera um conjunto de dados de saída;
- Modelos Conceituais Orientados a Objeto, cujos representantes mais importantes são os diagramas de objetos e casos de uso. No entanto, casos de uso devem ser considerados como modelos conceituais funcionais. Modelos conceituais orientados a objetos descrevem os objetos ou classes de objetos e suas interrelações entre eles no domínio do problema[19];.
- Modelos Conceituais Orientados a Estado, incluem diagramas de transição de estado. Esses modelos conceituais descrevem a configuração de objetos, fatos, fenômenos, etc. dentro do domínio do problema, e as mudanças produzidas nesta configuração ao longo do tempo (ciclo de vida do objeto).

Embora esses modelos conceituais sejam usados no projeto de sistema de informação, todos eles usam algum tipo de representação de dados. Alguns deles separam representação de dados da transformação dos dados (manipulação) como os modelos funcionais. Modelos orientados a objetos juntam dados e suas manipulações em uma única entidade. Modelo conceitual orientado a estado expressam a transição dos valores dos dados das entidades ao longo do tempo. Sendo assim, modelos conceituais para sistemas de informação dependem fortemente da representação dos dados.

2.5

Classificação de Modelo de Dados

Na área de banco de dados, modelos são classificados de acordo com duas abordagens diferentes[50, 41, 42, 2]. A primeira abordagem é baseada nos níveis de abstração usados para representar dados[41]. A segunda abordagem é baseada nos níveis da arquitetura do SGBD[2].

Modelos de dados podem ser classificados de acordo com os níveis de abstração[41], por exemplo, de alto-nível (ex. modelos conceituais), modelos de implementação (ex.

modelos lógicos), e de baixo-nível (ex. modelos físicos). Modelos de baixo-nível provêm conceitos que descrevem detalhes de como os dados podem ser armazenados no computador e são geralmente entendidos por especialistas em computação. Modelos de dados de implementação também são chamados de modelo lógicos, os quais provêm conceitos que podem ser entendidos pelos usuários finais mas não estão longe da forma como eles serão armazenados computacionalmente.

Outra forma de classificar modelos de dados é relacioná-los com a arquitetura do SGBD. O relatório do comitê ANSI[2] propôs uma arquitetura em três esquemas para separar as aplicações dos usuários do banco de dados físico. Nesta arquitetura, esquemas podem ser definidos de acordo com os seguintes níveis: interno, conceitual e externo. Os níveis interno e conceitual possuem respectivamente dois esquemas associados. Ao nível externo podem ser associado vários esquemas externos, também chamados de visões dos usuários. Cada esquema pode ser gerado por um modelo de dados específico. O esquema conceitual e externo pode ser gerado por modelo de dados de alto-nível ou modelo de dados de implementação enquanto esquema de dados interno pode somente ser gerado por um modelo de dados físico (de baixo-nível).

A arquitetura em três níveis é uma ferramenta conveniente para o usuário visualizar os níveis de esquema no SGBD, diferentemente da classificação de projeto, que é focada no processo de descrever o banco de dados. A arquitetura em três níveis sugere o uso de modelo de dados de acordo com o tipo do nível do esquema (estrutura do SGBD).

2.5.1

Linguagem de Modelagem Conceitual

Modelos de dados conceituais são materializados através de linguagens ou notações que podem ser usadas por um projetista de banco de dados para criar um esquema de dados. Tipicamente, modelos conceituais de dados usam conceitos como entidades, atributos e relacionamentos. Uma entidade representa um objeto do mundo real ou um conceito, o qual é descrito no banco de dados. Um atributo representa alguma propriedade interessante que futuramente descreve uma entidade. Um relacionamento representa uma interação entre duas ou mais entidades.

Diversos modelos de dados conceituais foram propostos, tais como ER (entidade-relacionamento)[8], ORM¹ [52] ou OO (orientado a objetos) [43]. Embora modelos de dados orientados a objeto serem adequados para a implementação eles são próximos a modelos conceituais. A modelo de dados orientado a objetos mais importante foi proposto pela ODMG (Object Data Management Group)[93] que especificou um padrão para banco de dados de objetos. Modelos de dados objeto são frequentemente utilizados como modelos conceituais de alto-nível, particularmente no domínio de engenharia de software.

A maior parte do projeto de sistema de informação envolve o projeto de banco de dados. Por causa disso, algumas metodologias orientadas a objeto populares que foram inicialmente criadas para o projeto de sistemas de informação são também usadas

¹Do inglês, Object Role Model cuja tradução pode ser Modelo Objeto e Papel

para projeto de banco de dados. Este é o caso da metodologia OMT (Object Modeling Technique) [9] e UML (Unified Modeling Language)[94].

Utilizamos a partir deste momento o termo "tradicional" para qualificar as linguagens de modelagem de dados que são populares no projeto de banco de dados. Seleccionamos as seguintes linguagens de modelagem como representantes dessas linguagens tradicionais para modelagem de dados conceituais: ER[1], ORM[52], EER [8] and UML³.

2.6

Modelo de dados Temáticos

Modelagem de dados conceitual é uma disciplina geral visto que pode ser usada para representar dados de domínios de aplicações diferentes. Por exemplo, nós podemos usar o modelo ER para criar um esquema de dados para aplicação financeira ou para uma aplicação médica. No entanto, a generalidade das linguagens de modelagem de dados limita a representação dos dados quando aplicada à algum domínio de aplicação específico, como por exemplo aplicações geográfica, Multimídia, Hipermídia, Científica, etc. Neste contexto, novos modelos de dados conceituais ou extensões aos modelos tradicionais foram propostas[53, 18, 20, 44, 13, 15] para suportar ou para facilitar a representação de dados específica. Alguns desses modelos são orientados para o projeto de aplicações incluindo modelagem de dados no nível conceitual[20, 44] enquanto outras são classificadas como modelo lógico de dados[53, 18, 15]. Esses novos tipos de modelos de dados orientados para um domínio específico são denominados de modelos de dados temáticos. Neste trabalho objetivamos investigar os requisitos necessários para um modelo de dados temático para biologia molecular e propor um novo modelo adequado para este problema.

2.7

Bases da Biologia Molecular

Biologia molecular é o estudo da biologia⁴ no nível molecular. A biologia molecular cobre outras áreas, particularmente a genética⁵ e a bioquímica⁶. A biologia molecular objetiva o conhecimento das interações entre os vários sistemas de uma célula, incluindo inter-relacionamento do ADN, RNA e síntese de proteínas. A biologia molecular também investiga como as interações são reguladas[95, 64].

A biologia molecular foca no processo de replicação, transcrição e tradução de material genético que é armazenado no ADN[6]. O dogma central da biologia molecular descreve como o material genético é transcrito em RNA e então traduzido para proteína. Embora sendo uma visão simplificada da biologia molecular, ele ainda provê um bom ponto de

²Os termos linguagem de modelagem de dados conceitual, linguagem de modelagem conceitual ou, simplesmente, linguagem conceitual serão usadas com o mesmo sentido neste documento.

³Embora UML tenha sido desenvolvida principalmente para o projeto de software, neste trabalho nos referimos somente a um subconjunto desta que é usado para modelar dados

⁴A ciência que estuda organismos vivos

⁵O ramo da biologia que estuda hereditariedade e variação nos organismos.

⁶A química orgânica dos componentes e processos que ocorrem nos organismos; o esforço para entender a biologia usando o contexto da química.

partida para entender esta área. Esta visão, no entanto, está sendo revisada sob a ótica dos novos papéis do RNA.

A maior parte do trabalho na biologia molecular é quantitativa, tendo sido realizada na interface entre a biologia molecular e a ciência da computação, denominada de bioinformática ou biologia computacional. O estudo da estrutura do gene e suas funções, tornou a genética molecular⁷ a mais proeminente sub-área da biologia molecular[95].

Vários outros campos da biologia focam em moléculas, tanto diretamente estudando suas interações, como no caso da biologia celular e biologia desenvolvimental, ou indiretamente, onde as técnicas da biologia molecular são usadas para inferir atributos históricos das populações ou espécies⁸, como nos campos da biologia evolucionária⁹ como é o caso da genética das populações e filogenética.

Bioinformática é, no entanto, somente o primeiro passo no sentido de remodelar as ciências da vida. Para o progresso futuro, o estudo completo dos sistemas biológicos são necessários para entender a biologia sob a ótica de um sistema. Este estudo é parte de um campo emergente denominado Biologia Sistêmica[62, 59, 72, 63, 73]. Biologia sistêmica é um novo campo da biologia que está tentando desenvolver a compreensão no nível dos sistemas biológicos.

A biologia sistêmica prove uma interface vital entre a biologia celular e aplicações biotecnológicas, podendo ser vista como uma arquitetura abstrata para gerenciamento de informações. Essas informações englobam parte significativa das ciências experimentais. Isto inclui descrições de experimentos, observações, dados experimentais, interpretações derivadas de modelos [89].

2.8

Sistemas de Informação Biológico

A biologia é uma ciência complexa pois precisa misturar o conhecimento de múltiplas disciplinas (ex. química, física, matemática, ciência da computação, etc) de forma a promover novas descobertas. Além disso, a biologia possui uma grande quantidade de conceitos que são difíceis de definir precisamente (ex. a definição de gene, proteína, classificações em geral, etc). Na biologia também encontramos relacionamentos entre esses conceitos apresentando uma semântica complexa.

Neste contexto, projetar sistemas de informação para biologia (denominado aqui de BIS¹⁰) é uma tarefa bastante complexa. De forma a entender e explicar as fontes dessas dificuldades, nós detalhamos o BIS usando uma arquitetura hipotética composta de quatro níveis lógicos baseados nos níveis identificados em [96]. Esses níveis são denominados por visão, modelo da aplicação, modelo do domínio e infra-estrutura. A Figura 2.3 ilustra essa arquitetura em quatro níveis e enumera alguns requisitos associados a cada um dos níveis.

⁷O ramo da genética trata da estrutura e atividade do material genético no nível molecular

⁸Grupos taxionômicos cujos membros podem herdar características de seus ancestrais

⁹a sequência de eventos envolvendo o desenvolvimento evolucionário de uma espécie ou taxonomia de um grupo de organismos

¹⁰BIS é o acrônimo de Biological Information System

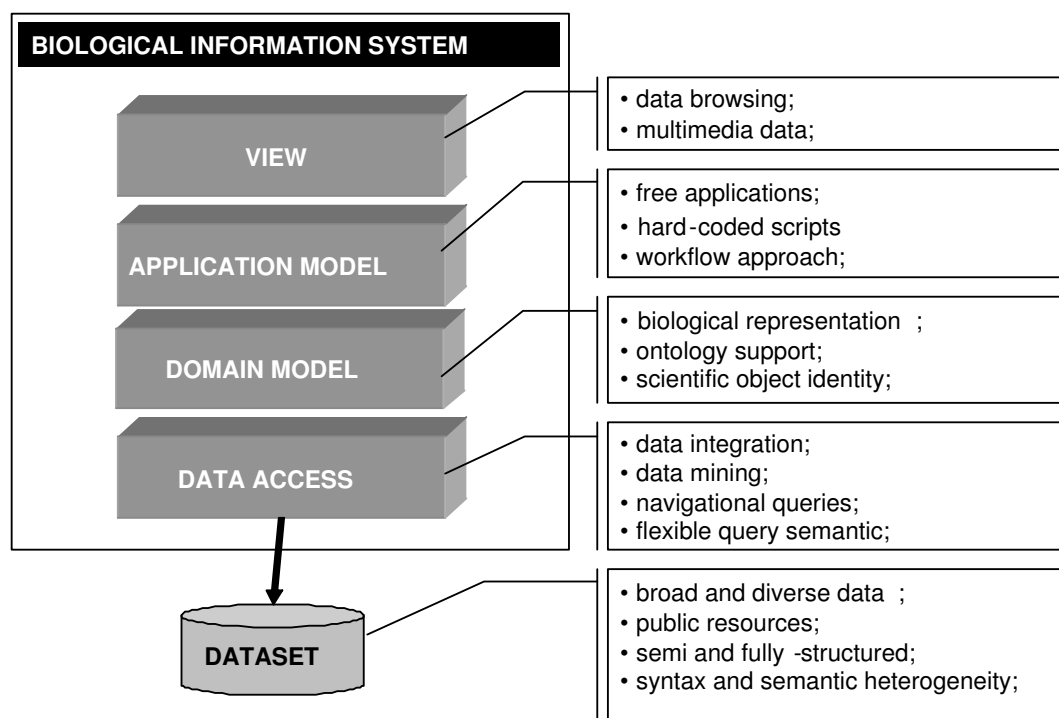


Figura 2.3: Camadas lógicas da Arquitetura de um Sistema de Informações Biológicas

A camada visão incorpora os componentes da interface com o usuário. Esta camada pode incluir componentes de interface gráfica (GUI). Um usuário interage com uma aplicação através de sua camada de apresentação. A Web tornou-se a abordagem preferida para disseminar dados científicos entre os pesquisadores no domínio da biologia[76]. Então um candidato natural para ser o componente de interação com usuário é o navegador Web (ex. Internet Explorer, Mozilla Firefox). Este fato pode ser decisivo para escolher uma interface gráfica visto que é difícil construir uma interface gráfica flexível usando linguagens de programação para Web.

A camada do modelo da aplicação (as vezes denominada como lógica da aplicação) armazena o conhecimento da aplicação, isto significa que qualquer componente específico para o domínio precisa residir nesta camada. O domínio da biologia colecionou uma variedade de ferramentas para assistirem os cientistas na busca, mineração e análise dos dados biológicos. Ferramentas para biologia incluem análise básica de biosequência como FAST, BLAST, Clustal, para citar algumas delas.

Atualmente o número de aplicações gratuitas disponíveis na Web é mais do que 300[77, 29]. Ferramentas legadas também incluem uma variedade de scripts codificados explicitamente em linguagens como Perl ou Python que implementam consultas específicas, links para repositórios de dados, e executam uma sequência de manipulações de dados pré-definidas. Desta forma, a integração entre essas ferramentas e linhas de consultas dentro de um BIS é uma requisito importante. A maioria das abordagens para integração de sistemas legados utilizaram uma abordagem orientada a workflow, a qual promove a colaboração entre cientistas, assim como a integração física de dados e ferramentas. Neste último caso, o problema da integração dos dados segue passo-a-passo uma única execução de consulta do usuário, incluindo tratamento de segurança e regras semânticas.

A camada do modelo de domínio define que tipos de entidades existem no domínio das

aplicações, e seus atributos e suas interconexões. As características mais importantes dos dados usados na biologia não estão relacionadas com o seu volume mas com sua diversidade e variabilidade. A natureza dos dados biológicos é intrinsicamente complexa, variando desde genes e proteínas, até interações entre proteínas, vias metabólicas e redes de regulação, para células e tecidos, organismos e populações, e finalmente os ecossistemas encontrados na terra.

Refletindo sobre a complexidade dos sistemas biológicos, os tipos de dados biológicos variam desde texto simples, sequências de ADN e proteína, estruturas tridimensionais de moléculas, imagens com diferentes resoluções, até saídas experimentais derivadas de tecnologias, tais como chip de microarray, géis, microscopia eletrônica, ressonância magnética nuclear e espectrometria de massa.

Associado aos tipos diversos de dados, existe uma vasta escala de complexidade, população, tempo e espaço, por exemplo, o tempo pode variar de 10^{-15} até 10^9 segundos que representa a variação entre o tempo de uma reação química até o tempo de evolução de uma espécie. Além disso, existem características que variam bastante entre diferentes organismos e espécies, como por exemplo, as estruturas e funções dos órgãos variam de acordo com a idade e sexo, em estados normais e de doença, ou ainda entre espécies. Por isso, toda a diversidade e variabilidade apresentam um grande desafio na modelagem do domínio da biologia e respectivamente no projeto de sistemas de informação biológicos[30].

A camada de acesso aos dados (ou camada de infra-estrutura) existe para permitir que a camada de domínio capture e mantenha os estados dos dados da aplicação. A camada de acesso aos dados pode ser formada por um ou mais banco de dados independentes ou por servidores de banco de dados distribuídos. Dados biológicos são disseminados em diferentes fontes de dados através da Web, disponíveis em vários formatos, anotados, e armazenados em arquivos planos, semi ou totalmente estruturados, relacionais ou orientados a objetos. Além disso, acessar dados heterogêneos é mandatório para um BIS, porque uma única consulta pode envolver arquivos planos (ex. Genbank), fontes Web(ex. GeneCards) e banco de dados específicos. As fontes de dados são na maioria das vezes textuais com dados semi (ex. HTML e XML)ou totalmente estruturados (ex. XML esquema). A proliferação de arquivos textuais na biologia é justificada pela necessidade dos cientistas entenderem os dados e seus esquemas permitindo um acesso fácil à informação sem interferência de aspectos tecnológicos. SGBDs tradicionais (relacional, orientado a objetos, relacional-objeto) são estruturados e seus esquemas são difíceis de alterar e entender porque demandam conhecimento sobre conceito do modelo de dados do banco de dados.

Os cientistas que mantêm suas próprias fontes de dados precisam de interfaces as quais ofereçam grande flexibilidade para adicionar e revisar os dados contidos dentro dessas fontes. Esta é a principal razão da baixa proliferação de banco de dados relacionais no domínio da biologia. Por outro lado, na abordagem que utiliza dados semi-estruturados, a organização dos dados permite mudanças no esquema de dados com mais facilidade, permitindo a adição ou eliminação de atributos sem muito esforço. Devido a isto, um sistema que lida com dados semi-estruturados não deve assumir uma representação de dados pré-definida. Tal sistema deveria oferecer uma grande facilidade sobre os SGBDs tradicionais. No entanto, essa flexibilidade pode afetar a qualidade dos dados por causa

das inúmeras fontes de dados que se sobrepõem, campos de dados que são preenchidos com diferentes objetivos, campos que são perdidos e pela falta de tipos e restrições.

Neste contexto, uma necessidade no projeto de uma camada de acesso a dados de um BIS é solucionar a integração de múltiplas fontes de dados. Visto que obter acesso a recursos biológicos, combinar dados de múltiplas fontes de dados, enquanto lida-se com a distribuição e heterogeneidade é uma tarefa bastante árdua[74].

As abordagens de integração desenvolvidas para integrar informações científicas (ex. KIND [78] , TAMBIS [108], Kleisli [79], K2 [79]), as quais podem ser usadas na camada de acesso aos dados, são classificadas como orientadas ao dado. O maior problema dessas abordagens é que elas são limitadas para oferecer mecanismos os quais permitam detectar discrepâncias na representação entre diferentes atributos de fontes de dados ou estender as capacidades das fontes para executar consultas mais ricas. Em contraste, outras abordagens focam na integração das aplicações com arquiteturas baseadas em agente (ex. CORBA), serviços Web e mecanismos de grade (ex. myGrid[80]). Essas abordagens são chamadas de orientadas a aplicação e seus arcabouços de integração se baseiam em APIs compartilhadas e descrições de recursos com pouco foco no gerenciamento de dados (avaliação de consulta, otimização e transações). A abordagem de integração de sistemas biológicos apresentada em [74] descreve uma arquitetura de mediação para prover gerenciamento de dados tradicional (via mediador LeSelect[97]) e mecanismos para acessar várias aplicações, web services e grades. Nesta abordagem, o conjunto de dados resultantes de uma execução da aplicação (ex. BLAST) é materializada e tratada por um wrapper derivado que estende as capacidades de consulta da aplicação remota. No entanto, uma necessidade mais crítica para a integração de dados na biologia é superar a heterogeneidade semântica pela falta de consenso sobre a terminologia biológica.

Além de todas as características apresentadas anteriormente, o processo de descoberta biológica pode impactar no projeto de todas as camadas da arquitetura de um BIS. O processo de descoberta biológica pode variar de uma abordagem reducionista até uma abordagem integrada, como foi apresentado no caso da biologia sistêmica. A abordagem reducionista tenta elucidar a estrutura e função dos elementos individuais (genes, proteínas, macromoléculas, etc) enquanto a abordagem integrada demanda o conhecimento das conexões entre as partes moleculares e como elas trabalham em conjunto. Em outras palavras, o projeto de um BIS deve seguir essas prerrogativas de forma a ser capaz de realizar previsões quantitativas e qualitativas. Ainda mais, a pesquisa para descoberta e síntese serão direcionadas pela informação complexa obtida pela própria biologia e pela diversidade e heterogeneidade das observações experimentais. Desta forma, um BIS deve oferecer mecanismos para incorporar facilmente novas fontes de dados (ex. GenBank), ferramentas computacionais (ex. BLAST) e resultados de experimentos (ex. microarray[91]) de forma a combiná-los resultando em uma infra-estrutura coesa de informação. Desta maneira, todas as camadas da aplicação devem ser afetadas pela incorporação desses componentes, i.e., fontes de dados, ferramentas computacionais ou resultados experimentais.

Para resumir a discussão acima, podemos deduzir que separando os diferentes aspectos em camadas, reduzimos a complexidade do projeto de um BIS permitindo associar

um conjunto de responsabilidades para cada camada. No entanto, a camada do modelo do domínio tem uma importante influência em todas as camadas porque ela especifica o modelo de dados da aplicação que direciona o projeto de todas as outras camadas da aplicação.

Por exemplo, se decidirmos inserir um novo conceito (ex. Ribossomo) na camada do modelo de domínio, devemos propagar esta mudança para a camada de acesso aos dados e para o modelo da aplicação de forma a especificar como os dados representados por este conceito serão tratados e acessados pela lógica da aplicação. Provavelmente essas mudanças podem refletir em outras mudanças na projeto da interface gráfica com o usuário, por exemplo mostrando a imagem do ribossomo.

Desta maneira, a camada de modelo do domínio pode ser vista como uma ponte entre a camada de acesso de dados e a camada do modelo da aplicação, responsável para fornecer uma visão integrada das fontes de dados para o modelo da aplicação. A camada do modelo do domínio se refere exclusivamente à modelagem de dados do domínio da aplicação.

2.8.1

Abordagem Reducionista versus Abordagem Holística

Basicamente, existem no mínimo duas abordagens para gerenciar a complexidade dos dados biológicos: reducionista e holística. Na biologia assim como em outras ciências metodológicas holismo representa o oposto do reducionismo metodológico.

Reduccionismo é uma abordagem para construir descrições de sistemas independente das descrições dos subsistemas que compõem este sistema, ignorando as relações existentes entre eles. Por exemplo, considerando um sistema biológico sendo composto por moléculas com certas estruturas, note que uma das posições dos átomos em uma molécula são especificadas, relacionamentos entre os átomos também são especificados. Igualmente, uma vez que as localizações relativas das moléculas sendo especificadas em um organismo, relacionamentos entre essas moléculas são também especificados. Levando em conta que tais relacionamentos estão além da abordagem puramente reducionista.

Holismo sempre tem sido levado como tese que o todo é mais que a soma de suas partes[98]. Em uma interpretação, holismo é uma tese metodológica, para o efeito de que a melhor maneira para estudar o comportamento de um sistema complexo é tratar ele como um todo, e não meramente analisar a estrutura e o comportamento de suas partes componentes.

Tradicionalmente, a biologia é admitidamente reducionista[99] em vários sentidos. A investigação da estrutura de uma célula que vem sendo por três séculos, culminando com a descoberta da dupla hélice, é reducionista, analisando unidades cada vez menores e usando química e física para entender os processos da vida[5]. Isto não significa, porém, que existiu uma visão que toda a vida pudesse ser entendida inteiramente pela física ou química. Estudando organismos no nível molecular nunca evitou estudá-los em níveis mais elevados e integrando esses níveis em uma verdadeira compreensão Holística.

De acordo com a visão reducionista da biologia molecular, o ADN é o último repositório da complexidade biológica e é a informação básica para o processamento do sistema biológico. O ADN é abstraído por uma sucessão de letras (A,C,G e T)

representando a estrutura primária de uma real (ou hipotética) molécula de ADN (ou fita) por meio dos possíveis nucleotídeos. Com certeza, isto é geralmente aceito que o armazenamento de informação, o processamento da informação, e a execução de vários programas celulares residem em níveis distintos da organização biológica: o genoma da célula, transcriptoma, proteoma, e metaboloma. Transcriptoma é o conjunto de todas as moléculas mRNA (ou transcritos) em uma população de células para um dado conjunto de circunstâncias ambientais. Proteoma é o conjunto de todas as proteínas expressadas por uma célula ou órgão em um momento particular e sobre condições específicas. Metaboloma é o conjunto total de entidades metabólicas e pequenas vias¹¹ em uma célula, tecido, organ, organismo, e espécies.

Abordagens Holística ou reducionista são importantes para o completo entendimento do sistema biológico e tem um grande impacto em seu gerenciamento.

2.9

Classificações Biológicas e Ontologias

Classificação e taxonomia de termos exercem um papel fundamental no domínio da biologia, a qual tenta explicar os complexos conceitos da vida e seus relacionamentos. Além disso, o domínio da biologia apresenta algumas características que tornam abordagens terminológicas difíceis, tais como[76]:

1. Múltiplas definições para genes são a norma, mais do que exceções(banco de dados PUBMED[92]);
2. O mesmo nome refere para diferentes genes que não compartilham nenhum relacionamento entre si;
3. Pesquisadores trabalhando em diferentes organismos ou campos estabelecem seus próprios uso de nomes;

Algumas iniciativas para definir ontologias para aplicações de bioinformática foram criadas([31, 32, 45, 21]) de forma a criar uma nomenclatura comum para o domínio da biologia molecular e permitir o reuso e compartilhamento de suas classificações.

Para compartilhar e reutilizar uma informação é necessário haver algum tipo de modelo comum. É muito difícil para um conjunto independente de recursos de informações de diferentes domínios de aplicação usarem um único modelo de dados. Ontologias foram propostas como especificações em um nível de abstração acima dos modelos conceituais, uma forma de especificação explícita de uma conceitualização[14]. Ontologias são projetadas para enriquecer os conteúdos dos modelos de dados e capturar as afirmações e significados intencionais dos conceitos e declarações em um domínio particular[22].

Na prática, esquemas conceituais são limitado pelas representações disponíveis na tecnologia corrente dos computadores. Sendo assim, esquemas conceituais associam, para

¹¹Uma subvia conservada é frequentemente codificada por genes posicionalmente pareados no cromossomo e são especialmente úteis na predição das funções dos genes

cada um dos seus componentes, um ou mais representações computacionais adequadas, enquanto uma ontologia não requer tal acordo.

Os usos de ontologias na biologia recaem em três áreas distintas: definição de banco de dados (ex. EcoCyc[31] e RiboWeb[32], anotação e comunicação (ex. GO e OMB), e formulação de consultas e pesquisa baseada em ontologias (ex. TAMBIS[108]). Além disso, ontologias são expressivas e formais. O formalismo pode variar desde sistemas baseado em frames até terminologias baseadas em frases. As primeiras são relativamente ricas, expressivas e formais, enquanto as últimas são mais simples e menos expressivas.

A ontologia do TAMBIS foi a primeira ontologia para biologia a usar description logic[65] como sua representação e, como consequência, tem uma semântica mais bem definida. Em contraste com o uso e formalismo restritos das ontologias, o escopo e o detalhe do conteúdo das ontologias variam enormemente. Por exemplo, no RiboWeb[32] e no EcoCyc[31] as ontologias agem com um esquema para uma base de conhecimento. A ontologia do TAMBIS[108] funciona com um esquema global sobre diversas fontes de dados biológicas. Em ambas estratégias as ontologias guiam o usuário para criar consultas consistentes que permitem capturar o conhecimento do domínio com bastante fidelidade.

Um dos mais influentes desenvolvimentos com respeito ao GeneOntology (GO)[45], o qual está adquirindo o status de um padrão na tentativa de desenvolver vocabulários controlados para uso compartilhado através de diferentes domínios biológicos.

A versão do GO em maio de 2005 contém 18852 termos divididos em três sub-ontologias cujos nós principais são respectivamente: componente celular, função molecular e processo biológico. Esses três princípios organizacionais são usados para descrever produtos dos genes. Por exemplo, o produto do gene *cytochrome c* pode ser descrito pelo termo da função molecular *electron transporter activity*, pelos termos de processo biológico *oxidative phosphorylation* e *indução da morte da célula*, e pelos termos de componente celular *matrix mitochondrial* e *membrana interna da mitocondria*. Essas três redes são estruturadas pela relação de inclusão (*is_a*) e de inclusão (*part_of*).

GO tem sido criticado por suas inconsistências e por sua falta de orientação clara de como as relações entre as suas três sub-ontologias. GO quebra a regra da unicidade: termos deveriam ter os mesmos significados, ou seja, apontar para as mesmas referências em todas as situações de uso. Por exemplo, a relação de parte-todo a qual está no centro de sua hierarquia organizacional e com a qual no mínimo três diferentes significados estão associados([100]:

1. A parte de B significa: A é algumas vezes parte de B no sentido de que existe para cada instância de A algum momento no qual ele é parte de uma instância de B. Exemplo: replicação bifurcada é em alguns momentos durante o ciclo da célula parte do nucleoplasmo;
2. A parte de B significa: A pode ser parte de B no sentido de ser uma relação de inclusão independente do tempo, a qual pode ser resumida como: classe A é parte da classe B se e somente se existe alguma subclasse C de B a qual é tal que todas as instâncias de A são incluídas como partes das instâncias de C e todas as instâncias

de C tem partes as quais são instâncias de A. Exemplo: flagelo faz parte da célula (alguns tipos de células tem flagelos como partes);

3. A parte de B significa: o vocabulário A está incluído dentro do vocabulário B. Exemplo: o termo componente celular é parte do geneontology.

O segundo aspecto relevante em ontologias é sua estrutura interna usada para representar os relacionamentos entre os conceitos. Uma estrutura da ontologia pode ser representada como um grafo onde os vértices (nós, folhas) e arestas (linhas conectando os nós) representam os termos e o relacionamento entre os termos da ontologia. As estruturas das ontologias podem variar desde simples hierarquias até grafos como ilustrado na Figura 2.4. Como citado em [87], a maioria das ontologias (ex. GeneOntology) possuem estruturas completas como grafos direcionados acíclicos e grafos não direcionados.

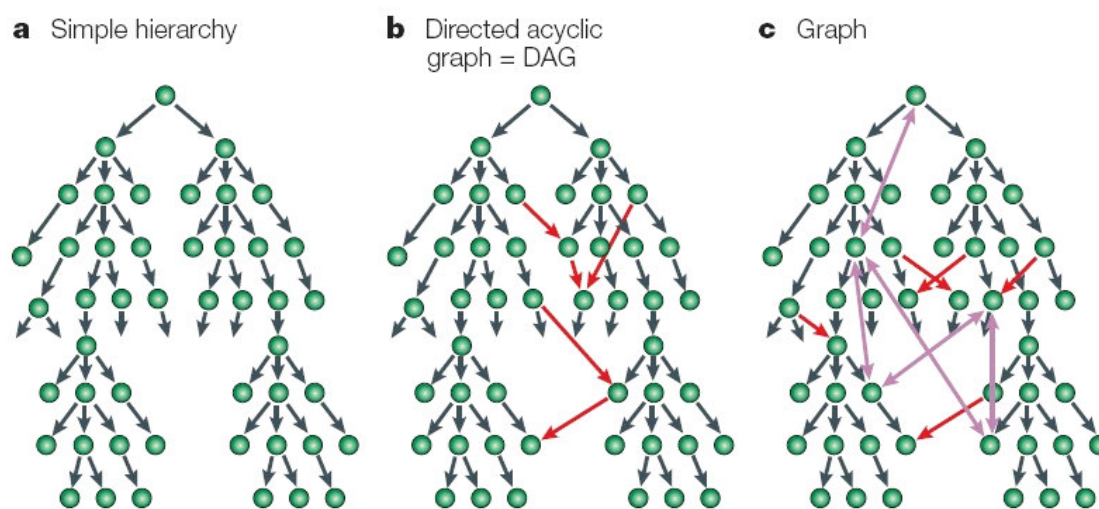


Figura 2.4: Complexidade das estruturas das Ontologias[87]

2.10

Conclusão

A maior parte dos bancos de dados são projetados usando linguagens de modelagem conceituais tradicionais. No entanto, a tarefa de projetar e implementar banco de dados a partir do zero é árdua e custosa[15]. Para lidar com esta dificuldade alguns esforços foram realizados objetivando reduzir o caminho da conceituação do banco de dados para sua implementação. Tipicamente, essas abordagens podem ser classificadas tais como: modelos de dados estendidos, framework de aplicação ou ferramentas especializadas.

Markowitz et al [15] propõem um modelo de dados orientado a objetos estendido denominado OPM com novos construtores para modelagem científica de experimentos (protocolos). Adicionalmente, o OPM provê ferramentas para geração automática das especificações de bancos de dados e consultas para implementação sobre um SGDB relacional comercial. Por isso, o OPM pode ser considerado uma implementação de um modelo de dados lógico mais do que um modelo de dados conceitual por causa dessas facilidades.

Outras abordagens usam técnicas de engenharia de software para prover frameworks de aplicação para o domínio da biologia molecular[81, 101]. [101] é um framework orientado a objetos para desenvolvimento rápido de aplicações para modelagem molecular e bioinformática estrutural. BALL provê um conjunto de estruturas de dados extensível assim como classes para mecânica molecular, métodos avançados para comparação e análise de estruturas de proteínas, importação de arquivos, e visualização. [81] é um framework orientado a objetos que objetiva a integração de bancos de dados de genoma assim como aplicações relacionadas.

Em geral, abordagens do estilo framework de aplicação incorporam esquemas pré-definidos dentro da arquitetura do framework e mecanismos limitados para estendê-los. Além disso, frameworks de aplicações demandam conhecimento especializado sobre técnicas de implementação e arquitetura do framework. Sendo assim, frameworks de aplicação são abordagens mais adequadas para o projeto de aplicações biológicas do que para modelagem de dados conceitual.

Nenhum trabalho diretamente relacionado com a proposta desta tese foi encontrado na literatura. Acreditamos que a falta de trabalhos nesse sentido são explicadas pela dificuldade encontrada na compreensão da complexidade do domínio da biologia. A necessidade de conhecimento profundo nesta área e na falta de consenso sobre os conceitos da biologia dificultam as pesquisa na definição de um modelo conceitual apropriado. Desta forma, tentaremos neste trabalho preencher esta lacuna, tentando levantar os requisitos para um modelo conceitual adequado e propondo um modelo que atenda a esses requisitos. O levantamento de requisitos é apresentado no próximo capítulo e a proposta é apresentada no capítulo subsequente.