1
Introdução

1.1
Bioinformática e Modelagem Conceitual

A transição dos métodos tradicionais de descoberta *in-vivo* para os métodos de descoberta *in-silico* revolucionaram as ciências relacionadas com o estudo da vida[1]. Este fato reduziu o tempo e custo associado com a descoberta do conhecimento biológico. Consequentemente, uma nova disciplina - chamada de Bioinformática - foi criada juntando em um única disciplina a biologia e a ciência da computação.

A Bioinformática objetiva o gerenciamento e análise dos dados biológicos usando técnicas avançadas de computação, especialmente importantes na análise dos dados sobre pesquisa do genoma[60]. Além disso, a Bioinformática ajudou os cientistas na criação e manutenção de banco de dados para armazenamento das informações biológicas sobre sequência de ADN, síntese de RNA e geração de proteínas[107]. Porém com o advento das tecnologias de alta-performance (ex., micro-array, sequencing)[91, 71] e o progresso dos projetos de genoma[48, 61], ocasionou um crescimento exponencial do volume de dados, impulsionado pela aquisição de dados de forma rápida e consistente.

Através do sequenciamento do ADN de diversos genomas, os cientistas possuem atualmente um grande volume de informações sobre sequências, genes, proteínas, etc. Todas essas informações servem para estudar a organização e controle das vias genéticas.

Esta nova etapa da revolução biológica, denominada de era pós-genômica, está altamente associada com as disciplinas genômica, transcriptômica, proteômica e metabolômica, as "ômicas" de modo abreviado[35, 36, 37, 38, 39, 40]. Essas disciplinas relacionam uma sequência de ADN com (a) a estrutura de um produto para qual esta sequência codifica para (normalmente uma proteína); (b) a atividade de uma proteína e sua função dentro da célula; (c) do tecido e (d) do organismo. Desta forma, o foco da biologia é movido da caracterização molecular para o entendimento da atividade funcional.

A Bioinformática é, no entanto, somente um dos passos iniciais na remodelagem das ciências da vida. Para progressos futuros, será necessário o estudo dos sistemas biológicos como um todo, a compreensão do funcionamento dos órgãos (ex., coração) e dos sistemas associados (ex. sistema cardiovascular). Este estudo é parte de uma disciplina emergente chamada de Biologia Sistêmica (tradução adotada para Systems Biology)[62, 59, 72, 63, 73] a qual representa a pesquisa na era pós-genômica. Esta pesquisa objetiva desenvolver o

[1] Basicamente os métodos de descoberta *in-silico* consistem em coletar dados através de uma variedade de tecnologias anotando e explorando os resultados dos conjuntos de dados digitais.

entendimento em nível sistêmico dos sistemas biológicos. A Biologia Sistêmica adota uma abordagem orientada a sistemas para descrever os processos dinâmicos dentro e entre as células biológicas. Ele é mais próximo da aplicação da teoria de sistemas na biologia do que da aplicação da física na biologia.

Porém, avanços metodológicos na análise de dados são necessários para transformar técnicas experimentais - microscopia moderna, nanotecnologia, entre outras - em informação e conhecimento. Além disso problemas na era pós-genômica não serão apenas experimentais ou técnicos, mas também conceituais[73].

O sucesso da pesquisa em biologia dependerá da correta representação e manipulação dos dados biológicos permitindo os cientistas criarem, gerenciarem, manipularem, integrarem e analisarem os dados de forma a gerar informação e conhecimento. Lederberg and McCray reportaram 44 diferentes tecnologias "omicas" baseadas na consulta ao PUBMED[92] conduzida em 2001[49]. Os autores de [28] mostraram que o mRNA e a proteína não se relacionam no mesmo nível. Desta forma, não é possível simplesmente observar uma única dimensão "ômica" , como por exemplo, a expressão do mRNA em dados de microarray.

Neste contexto, técnicas para gerenciamento de dados possuem um papel fundamental para o desenvolvimento de aplicações biológicas porque elas fornecem abstrações adequadas para projetar, implementar, acessar e gerenciar os dados. Nós observamos que as técnicas de gerenciamento de dados tradicionais não são adequadas para lidar com dados biológicos por causa dos seguintes desafios, enumerados abaixo. Esses desafios também podem ser encontrados em outros domínios de aplicações, como por exemplo em aplicações geográficas:

1. Complexidade dos dados: dados biológicos são extensos e diversos, englobando vários domínios de conhecimento como biologia molecular e celular, genética, biologia estrutural, farmacologia e fisiologia. Cada um desses domínios contemplam tipos de entidades sobrepostas e complementares, e com suas próprias terminologias e necessidade de dados. Além disso, a variedade de procedimentos experimentais e analíticos resultam em dados relacionados mas não idênticos. Isto gera uma quantidade incomparável de tipos de dados, desde sequências até estruturas tridimensionais, imagens, estruturas de grafos, tabelas de dados, textos semi-estruturados e sem estruturas. Ainda mais, o progresso da ciência adiciona outra dimensão de instabilidade aos tipos de dados;

2. Acessar, integrar e compartilhar dados biológicos: a informação biológica é acessível através da Web e a maior parte desta informação aparece sob a forma de texto. Somente uma pequena parte reside em banco de dados ou em formatos específicos. Várias das fontes de dados não são padronizadas e não documentadas. Consequentemente, a integração e compartilhamento de dados biológicos se torna um grande desafio.

3. Desenvolvimento de banco de dados: para os cientistas, o desenvolvimento de banco de dados normalmente significa a produção de um conjunto de dados. No entanto, a separação da aplicação da representação dos dados não é reconhecida, assim como

a necessidade de um sistema gerenciador de banco de dados (SGBD). Mesmo assim, um SGDB é usado e princípios para projeto de banco de dados não são seguidos. A tecnologia de gerência de dados está distante e algumas vezes não é compreendida pelos cientistas, e igualmente, a biologia soa complicada para os especialistas em banco de dados.

A comunidade de banco de dados tem estudando esses problemas e um grande esforço tem sido feito neste sentido, tentando propor adequadas ao domínio. A maioria dos trabalhos em gerenciamento de dados para ciências da vida, apresentados na literatura, tem se focado na integração de dados biológicos[74, 88, 75]. Pouca atenção tem sido dada na representação e manipulação de dados biológicos de forma a suportar o ciclo experimental de conhecimento da biologia.

Nós propomos na Figura 1.1 um esquema típico do ciclo de pesquisa da biologia computacional. Primeiramente, (1) modelos qualitativos - tais como mapas de vias metabólicas - e modelos quantitativos baseados em dados *in vivo* e *in vitro* e hipóteses. Em seguida, simulações são executadas usando propriedades numéricas e discretas do modelo quantitativo, gerando predições sobre o comportamento do sistema. Os resultados das análises sugerem novas hipóteses (análise e interpretação), os quais subsequentemente são testados por experimentos em bancada, iniciando um novo ciclo.

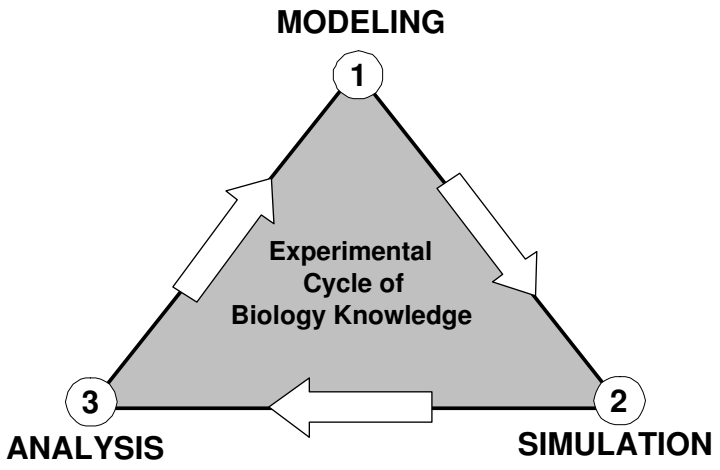

Figura 1.1: Ciclo experimental do conhecimento da biologia

A integração das técnicas de gerenciamento de dados com o ciclo experimental significa que a gerência de dados deve englobar atividades do tipo: modelar, simular e analisar dados experimentais biológicos. O início desta atividade de integração é a modelagem conceitual de dados devida as seguintes razões:

- a modelagem conceitual permite uma representação abstrata dos dados, a qual se assemelha com a forma que os usuários percebem realmente o mundo real, reduzindo (com respeito aos modelos de dados mais tradicionais) a distância semântica entre o domínio e sua representação;
- ele pode ser uma boa forma de comunicação entre o projetista de banco de dados e o usuário do banco de dados;

– uma linguagem de modelagem conceitual permite descrever de uma forma declarativa e reutilizável o domínio de aplicação, seu vocabulário relevante, e restrições para o uso de dados. Um bom modelo conceitual é necessário para permitir a integração semântica dos dados ;

1.2 Objetivos da Tese

As três maiores objetivos desta tese intitulada "BioConceptual: Um modelo conceitual para biologia molecular" são os seguintes:

Enumerar os requisitos para um novo modelo conceitual de dados para biologia molecular. O objetivo é identificar necessidades biológicas em termos de representação de dados. A partir dessas necessidades analisar quais são as limitações nas linguagens de modelagem tradicionais que impedem a correta representação dos dados. Para cada problema de representação detectado é levantado um requisito específico para o modelo de dados conceitual para biologia molecular.

Propor um modelo de dados conceitual para o domínio da biologia molecular. Usar todos os requisitos levantados pelo último objetivo apresentado para propor um modelo de dados conceitual adequado para o domínio da biologia molecular. Este objetivo inclui a seleção de quais requisitos levantados que serão implementados.

Formalizar o modelo proposto. A formalização de um modelo é essencial para demonstrar sua viabilidade. Faremos uso da teoria de conjuntos e lógica para formalizar o modelo proposto. O objetivo é mostrar que o modelo proposto gera esquemas de dados corretos e não-redundantes.

Essas três contribuições serão usadas para estruturar a discussão do que será estudado e desenvolvido. Ao final deste trabalho, será elaborado um resumo das conclusões sobre os resultados obtidos.

1.3 Estrutura da Tese

Esta tese é dividida em seis capítulos seguido pela bibliografia. Cada capítulo é resumido da seguinte forma:

Introdução: Capítulo 1, a introdução, apresenta a motivação que orienta a pesquisa deste trabalho, descreve o escopo da tese, e o que está além do tema discutido. Em seguida descreve a estrutura da tese e apresenta as linhas condutoras que orientam este trabalho.

Preliminares e contexto da pesquisa: No Capítulo 2, o contexto da pesquisa é introduzido. Primeiro, as áreas onde o contexto se aplica são brevemente apresentadas. Essas áreas incluem sistemas de informação, modelagem de dados, modelagem conceitual de dados, assim como uma breve introdução. Segundo, uma visão geral sobre biologia molecular, a qual engloba sistemas de informações biológicos e classificações biológicas (em especial ontologias).

Detalhamento do Problema: Capítulo 3 enumera os problemas relacionado com a representação dos dados biológicos usando linguagens de modelagem tradicionais. Esses problemas são apresentados de acordo com a sua relevância.

Descrição do modelo de dados: Capítulo 4, descreve o modelo de dados proposto detalhando cada construtor deste modelo. Características importantes são apresentadas em detalhe, tais como: herança não-monotônica, relacionamento ordenado, declaração de restrições e contextos ontológicos.

Formalização do modelo de dados: No Capítulo 5 um meta-modelo orientada a objetos é proposto no sentido de permitir a sua implementação. Foi escolhido o padrão para modelos de dados orientados a objetos da ODMG como modelo base para o modelo proposto. Finalmente, é apresentada uma formalização de cada um dos construtores do modelo, usando lógica de primeira ordem.

Conclusão: No Capítulo 6, os comentários conclusivos são apresentados. Um resumo breve da pesquisa realizada é provido, assim como uma enumeração da contribuição da tese. Finalmente, trabalhos futuros na área de modelagem conceitual para biologia molecular é descrito.