3
O Problema

3.1
Introdução

Neste capítulo nós argumentamos que é necessário um modelo de dados conceitual especialmente concebido para a biologia molecular. Esta afirmação é suportada pelo argumento que, exceto aspectos específicos do Universo do Discurso, as linguagens de modelagem conceitual tradicionais principalmente ER[1], EER [8], ORM[52] e UML[94]) não são adequadas para representar as informações biológicas.

3.2
Representação de Dados Biológicos

No mínimo dois aspectos tornam a modelagem conceitual de dados biológicos diferente dos modelos de dados conceituais padrões que merecem atenção especial. Esses aspectos estão relacionados com a natureza dos sistemas biológicos e dados biológicos. Embora alguns desses aspectos possam aparecer também em outros domínios eles são fundamentais para o domínio da biologia.

Sistemas biológicos parecem serem bem diferentes dos sistemas de engenharia onde é sabido exatamente "o que" é esperado do esquema de dados. Por exemplo, quando o projetista está construindo um esquema de dados para um sistema de informações geográficas, ele sabe exatamente quais são os conceitos e os relacionamentos que devem ser projetados. Além disso, o projetista do banco de dados conhece as consultas que ele deseja responder usando o esquema de dados.

Isto não significa que os sistemas biológicos não podem ser projetados. Uma combinação dos construtores das linguagens de modelagem tradicionais e as observações dos especialistas irão provavelmente levar o projetista do banco de dados a um esquema de dados adequado. No entanto, sempre que o projetista está pronto para caracterizar totalmente o modelo a partir dos conceitos base, o modelo por si só pode expressar conceitos emergentes e relacionamentos que não são evidentes para um esquema de dados. Essa necessidade pode dificultar a utilização do modelo de dados visto que ele deve evoluir constantemente e dependendo das evoluções exigidas, esta tarefa pode ser extremamente complexa.

Dados biológicos são extensos e diversos, cobrindo vários domínios de conhecimento como por exemplo: biologia molecular e celular, genética, biologia estrutural, farmacologia, fisiologia, etc. As propriedades dos dados biológicos identificadas abaixo fazem este campo particularmente desafiador:

O domínio da biologia possui uma grande diversidade e variabilidade de conceitos herdados da complexidade dos sistemas biológicos. Como ilustrado na figura 3.1, um conceito pode ser representado em diferentes níveis de abstração, tipo: átomos, moléculas, macromoléculas, células, organismos e ecossistemas. Similarmente, para cada elemento em um nível existe uma grande variabilidade dependendo de diversos fatores do tipo: organismo, idade, sexo, condições específicas, etc. Além disso, cada variabilidade de conceito pode ser representada em uma vasta escala, por exemplo, tempo do evento de evolução pode variar de milisegundos até séculos. Então, conceitos biológicos podem possuir um número combinatorial de possibilidades de representações.

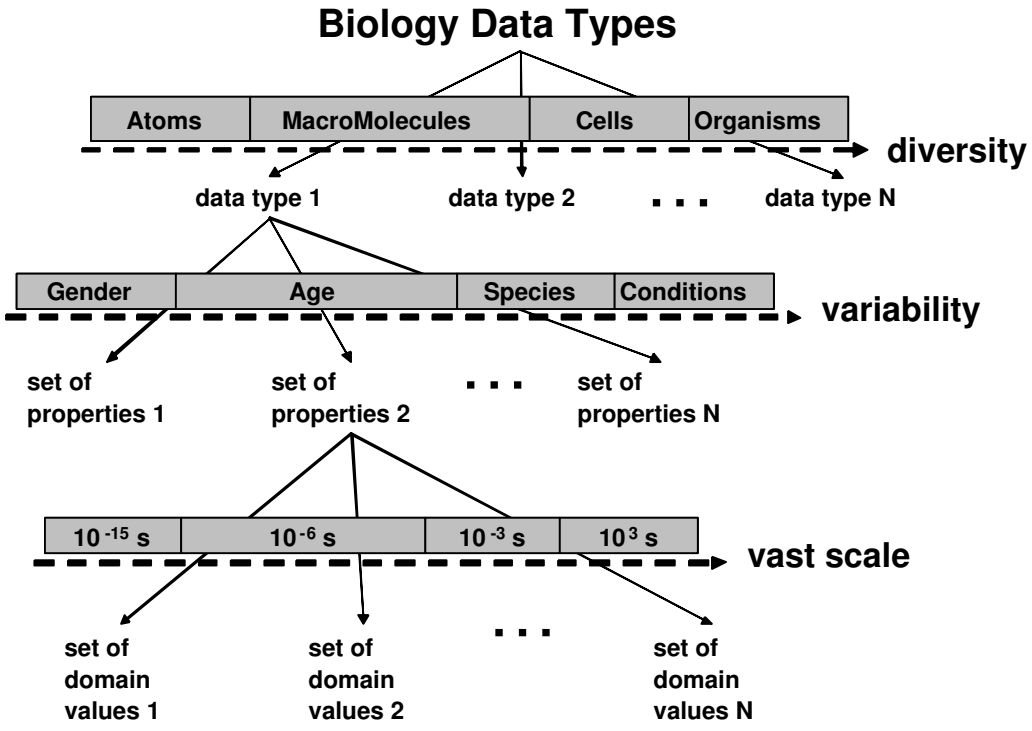

Figura 3.1: A diversidade e a variabilidade do tipos de dados biológicos

Existem também diversos conceitos similares que são difíceis de modelar. Por exemplo, em química, um aminoácido[1] é qualquer molécula que contém tanto os grupos funcionais amino e ácido carboxílico. Em bioquímica, este termo menor e mais geral é frequentemente usado para referir aos aminoácidos alfa: aqueles aminoácidos nos quais as funcionalidades do amino e carboxilate são anexadas ao mesmo carbono. Consequentemente, um esquema de dados que inclua o conceito aminoácido, dependendo em qual contexto ele seja usado, pode incluir uma interpretação incorreta deste conceito.

Dados biológicos são repletos de exceções por causa de duas razões: evolução dos sistemas biológicos e progresso tecnológico. As formas nas quais os sistema biológicos são organizados vem sendo alterada ao longo do tempo, eles evoluem. Suas estruturas casuais

[1] Aminoácidos são unidades estruturais básicas das proteínas.

não somente poderiam ter sido diferentes de fato, foram diferentes em diversos períodos na evolução da vida no planeta e em regiões distintas da terra e provavelmente serão diferentes no futuro. O fato da biologia ser uma ciência da descoberta e ainda muito conhecimento será apropriado pelos cientistas, é recorrente que os modelos, dados ou sistemas usados para representar esses conhecimentos sejam alterados com frequência. Além disso, avanços nas novas técnicas que investigam fenômeno biológicos e coletam dados biológicos têm sido responsáveis pela definição de novas restrições que devem ser garantidas. A questão é como acomodar todas as exceções e manter a checagem semântica. Por exemplo, dados biológicos que são derivados de experimentos de micro-array[91] devem ser validados diferentemente do dados derivados de observações manuais porque cada forma de coletar dados herda algum tipo de erro. Sendo assim, restrições devem ser especificadas de acordo com diferentes técnicas que se tornem disponíveis.

O domínio da biologia é altamente integrado. A informação tende a formar redes e hierarquias com vários fatos aceitando como verdade outros fatos. Se o banco de dados pode alterar a si mesmo como resultado da aplicação de uma regra semântica, então o resultado da mudança deve ser checado. O desafio com respeito a modelagem de dados é como conceitualmente definir regras semânticas de acordo com múltiplos fatos biológicos.

Dados biológicos são com frequência incompletos[2]. Este ocorre porque alguns objetos biológicos não têm uma definição clara (pode ainda estar sob investigação) e descrições completas levam tempo para serem obtidas. Isto também pode ocorrer por causa dos recursos limitados e das tentativas para colecionar os dados falharam. Por exemplo, a maior parte dos genomas no GenBank são incompletos (a maioria pequenos fragmentos); pedaços das sequências para proteínas do GenBank estão com frequência perdidas das suas estruturas armazenadas no PDB;

Vários conceitos biológicos são compreendidos parcialmente e provavelmente se tornaram mais sofisticados, evoluindo com o tempo. Por exemplo, é originalmente aceito que um único gene leva (no máximo) a um único produto. É sabido atualmente que o mesmo gene pode ser traduzido em múltiplas formas produzindo múltiplos produtos. Um mapeamento um-para-um entre gene e produto é por isso uma simplificação excessiva. Este tipo de problema ocorre não somente no domínio da biologia mas o fato da biologia ser um disciplina da descoberta, este problema se torna relevante.

Um problema mais frequente é que os itens de dados são modificado depois de terem sido armazenados no banco de dados. Se os dados mudam, é levantada a questão de como a checagem devem ser propagadas sobre objetos relacionados. Por exemplo, a sequência de proteína está em um banco de dados e é parte de um alinhamento múltiplo. Alterações na sequência pode alterar o alinhamento. Propagação de dados é também um problema conhecido em outros domínios mas a alta distribuição e integração no domínio da biologia transforma isto num problema fundamental.

A biologia é repleta de definições ambíguas e confusão conceitual. Diferente ramos da biologia (ex. biologia estrutural, genética, etc) usam o mesmo termo para conceitos diferentes. Vários aspectos dos sistemas vivos são parcialmente entendidos. Este problema

[2]Usamos a palavra "incompleto" para denotar dados perdidos

somente torna-se crítico quando é necessário integrar modelos de dados diferentes. Neste contexto, dois tipos problemas de integração podem ocorrer:

- termos têm o mesmo nome mas diferentes semânticas: por exemplo, a palavra colônia é usada em zoologia para significar um grupo de animais da mesma espécie que vive junto e depende um do outro. Em microbiologia a colônia é um grupo de microorganismos que foram desenvolvidos de uma única célula;
- termos têm diferentes nomes com a mesma semântica: por exemplo, um esquema de dados usa o termo gene para nomear o conceito gene enquanto outro esquema usa a palavra GeneHumano para definir o mesmo conceito de gene.

Na primeira situação, quando formos integrar os dois esquemas contendo o conceito colônia é necessário manter os dois conceitos separadamente no novo esquema integrado. As duas possibilidades são: (1) prefixar cada conceito de colônia com zoologia e microbiologia ou (2) permitir a repetição dos nomes dos conceitos e usar um descrição diferenciada para eles. Na segunda situação, um único nome de conceito deve ser escolhido para unificar os diferentes nomes com a mesma semântica.

Neste trabalho, focaremos nos problemas de representação de dados que ocorrem na fase de projeto de esquema de dados. Por causa disto, iremos orientar nossa discussão para os problemas sobre representação de dados biológicos usando linguagens de modelagem conceitual de dados tradicionais. Adicionalmente, outros problemas não diretamente relacionados com a representação de dados poderão ser considerados.

Neste capítulo faremos uso da notação da UML[94] para exemplificar alguns problemas detectados nas linguagens de modelagem conceitual tradicionais. Observamos que embora a UML seja uma linguagem para projeto de aplicações, estaremos interessados no subconjunto da UML usado para representar modelagem de dados.

3.3 Conceitos Atômicos - Uma visão Reducionista

A sequência de ADN é o bloco básico de construção da informação biológica a qual permite os cientistas descreverem os objetos biológicos através de diversas perspectivas. Em termos de modelagem conceitual, isto significa que quase todo conceito na biologia molecular pode ser modelado usando a sequência de ADN ou derivado dele. Esta abordagem é seguida por Paton et al [34] para representar informação do genoma ilustrada na Figura 3.2. Neste modelo, notamos que cada classe é uma sequência de ADN e o diagrama de classe é usado para representar e relacionar diferentes classificações da partes da sequência do ADN. Por exemplo, a classes (a) ChromosomeFragment, (b) PrimaryTranscript and (c) Spliced Transcript são partes da sequência de ADN.

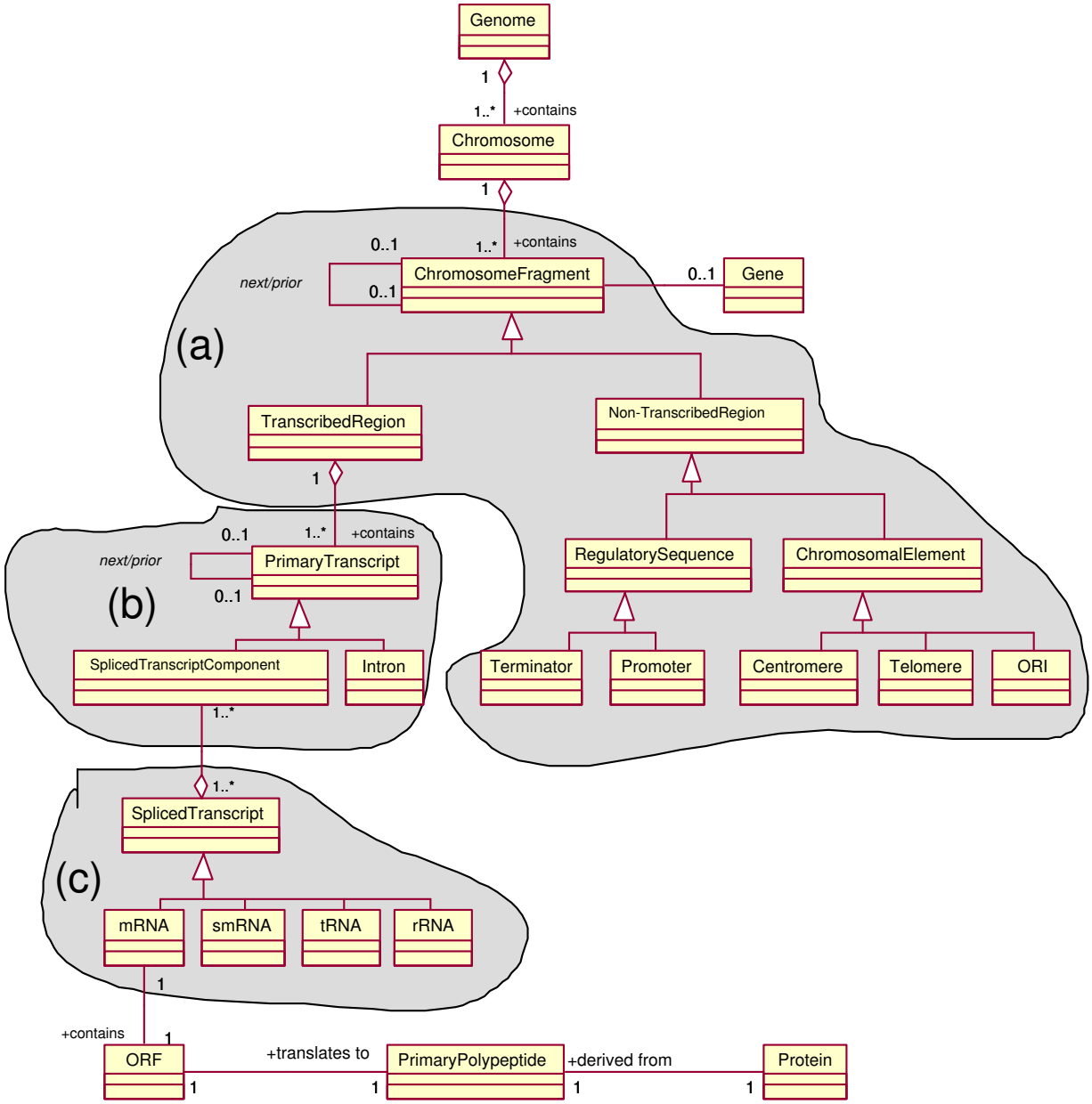

Figura 3.2: Diagrama de classe representando o Genoma[34]

Conceitos Pervasivos

Como partes de uma sequência de ADN eles são todos sequências também. Isto é um ponto interessante porque cada propriedade de uma sequência de ADN deve ser replicada dentro do esquema de dados. Sendo assim, o projetista do banco de dados é obrigado a criar informação redundante no modelo de dados e sincronizá-la a cada mudança que pode ocorrer sobre a definição do conceito ADN. Na Figura 3.2, podemos observar que informações redundantes aparecem no relacionamento "next/prior"descritos nas classes ChromosomeFragment e PrimaryTranscript. Além deste problema, a não aderência das linguagens tradicionais para representar conceitos pervasivos[3] como o ADN, podem induzir

[3]Conceitos que são representados em diversas partes do esquema de dados

alguns problemas na modelagem de dados. Isto pode ser visto como podemos ver na classe SplicedTranscript, que não possui o relacionamento "next/prior" embora ele seja uma sequência de ADN também.

Conceitos Ordenados e Padrões

Existem duas características importantes apresentadas nas sequências de ADN que devem ser levadas em conta durante a modelagem de dado:

A primeira característica é a ordem presente nas sequências de ADN. Como dito anteriormente uma sequência de ADN é uma longa fita de nucleotídeos. Para os cientistas é interessante reduzir a longa sequência em pequenas subsequências de forma a investigar suas funções biológicas, estilo estratégia divisão e conquista. Devido a isto, quando modelamos subsequências precisamos especificar a ordem entre elas. Tipicamente, em linguagens de modelagem tradicionais isto pode ser feito através da criação de um auto-relacionamento "next/prior" junto à classe representando a subsequência (ex. classe ChromosomeFragment) como ilustrado na Figura 3.2. No entanto, a ordem presente na sequência de ADN pode ser complexa de se representar caso existam dependências e restrições que governem esta ordem.

A segunda característica da sequência de ADN é a ocorrência de tipos especiais de padrões nas sequências, chamados de motivos[4]. Um padrão específico de subsequências são muito importantes para a pesquisa da biologia, tais como promotores, *stop codons*, etc. A Figura 3.3 ilustra uma expressão regular representando um motivo e uma lista de sequência de ADN que combina com este padrão.

Alguns tipos de motivos são muito complexos por causa das taxas de substituições e distância dos intervalos[102]. Desta forma, novas abordagens para especificar e detectar motivos foram desenvolvidas[46]. Tipicamente, motivos são representados por expressões regulares ou modelo probabilísticos[90, 33]. Embora, construtores de agregação e composição existentes em UML permitam especificar configuração de conceitos, eles não possuem poder de expressão para representar configurações complexas como as dos motivos.

Requisito	**#1** Representar relacionamentos com ordem complexa e padrões.

3.4 Sistemas Biológicos - Uma visão Holística

A integração de sistemas biológicos em diferentes níveis organizacionais demonstram que as funções celulares são distribuídas entre grupos de componentes heterogêneos que interagem dentro de grandes redes. Por exemplo, o proteoma se auto organiza em uma rede de interação de proteína e metabólitos são convertidos através de um intrincada teia metabólica.

[4]Motivo é um padrão associado com alguma função biológica

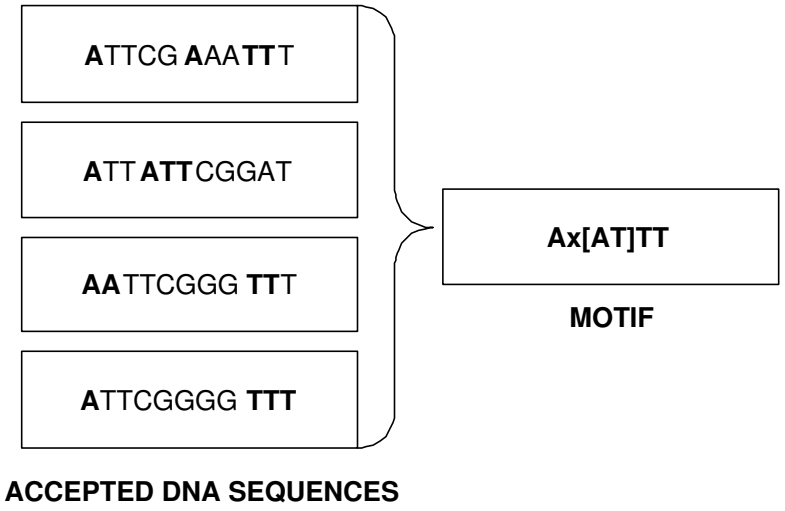

Figura 3.3: Sequências de ADN e especificação de motivos

A Figura 3.4 nos ajuda a entender a organização de um sistema da biologia molecular. A base da pirâmide apresenta a representação tradicional da organização funcional celular: genoma, transcriptoma, proteoma, e metaboloma (nível 1). No nível mais baixo, esses componentes formam motivos genéticos-regulatórios ou vias metabólicas (nível 2), os quais são o blocos de construção dos módulos funcionais (nível 3). Esses módulos são aninhados, gerando uma arquitetura hierárquica livre de escala (nível 4). A seta para cima sugere que os componentes de baixo nível compõem os sistemas de organização em larga escala que executam funções dos organismos. Nos sistemas de organização em larga escala, a percepção dos componentes de baixo nível é fraca. Reciprocamente, quando investigamos as especificidades do organismo a importância dos componentes de baixo nível é revelada. A Figura 3.4 traz para discussão a necessidade de levar em conta as visões holística e reducionista para o contexto da modelagem conceitual de dados.

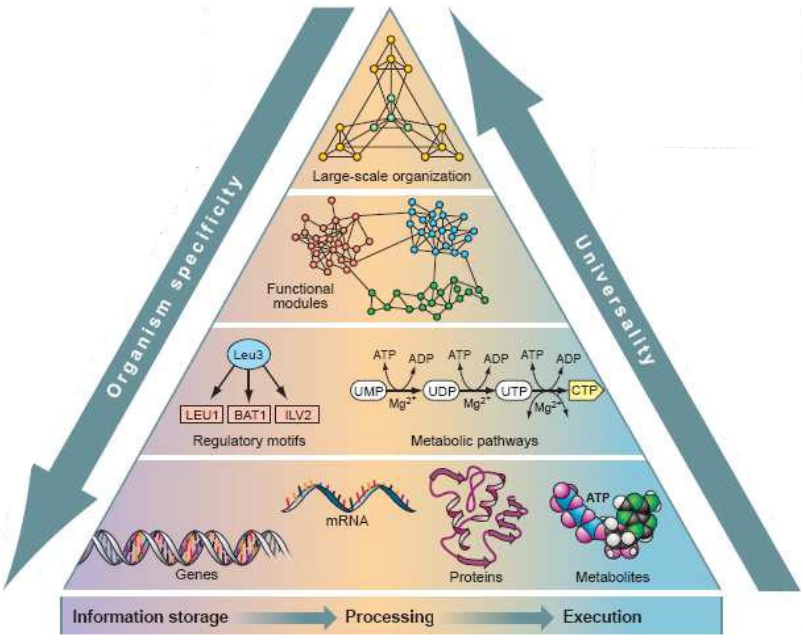

Figura 3.4: Complexidade de Pirâmide da Vida[59]

Uma abordagem Holística aplicada a modelagem conceitual para a biologia molecular significa projetar um efetivo esquema que envolvam aspectos da genômica, transcriptômica, proteômica e metabolomica ao mesmo tempo especificando, desta forma, uma rede de interações que executem um papel biológico específico.

Um aspecto fundamental em termos de modelagem conceitual é como modelar a função que emerge a partir das interações entre os elementos que compõem o sistema biológico sem a necessidade de modelar seus elementos subordinados. Funções emergentes aparecem em relacionamentos complexos, não somente olhando a rede estruturada de relacionamentos organismo-órgão-células-moléculas, mas também verificando os fatores ambientais.

As abordagens de visão e multi-dimensão adotadas na modelagem de dados tradicionais usam entidade dos esquemas conceituais de base para especificar novos conceitos de alto-nível (visões, fatos e dimensões). Em uma abordagem biológica Holística, novas funções emergentes ou propriedades podem aparecer independentemente dos elementos de base. Por exemplo, uma característica de uma doença causada por relacionamentos entre genes, proteínas e metabólitos podem especificar propriedades que não são diretamente associada com genes, proteínas e metabólitos. Esta associação pode ser desconhecida pelos cientistas ou pode não ser diretamente de derivada por causa da distância entre os níveis biológicos (ex. organismo e macromolécula).

Então, uma abordagem de modelagem Holística para biologia deve permitir a inclusão de novas propriedades nos conceitos de alto-nível que não são diretamente derivados dos conceitos de base, desconectando os conceitos de alto-nível dos conceitos de baixo-nível. Desta forma, abordagens tradicionais usada na modelagem conceitual como visões ou mecanismos multi-dimensionais não são adequadas para resolver este problema.

Requisito	**#2**	Incluir propriedades em conceitos denominados de alto-nível sem necessidade de alterar os conceitos de base ou de baixo-nível

3.5 As Ontologias e a Biologia

Ontologias são semanticamente mais ricas que esquemas conceituais de bancos de dados[22], e desta forma, mais perto do modelo de conceito de banco de dados. Ontologias podem ser úteis para evitar a construção esquemas de dados com pouco significado biológico e consultas[108]. Por exemplo, um gene pode ser representado diferentemente em diversos banco de dados, mas o conceito é único, no mínimo do ponto de vista da comunidade. Este ponto de vista é expressado na ontologia que esta comunidade especificou. Por exemplo, uma proteína é uma macromolécula, independente se ela é representada para fins de um sistema de informação,por uma string contendo letras A, C, T e G, um imagem de cromatograma, ou um modelo tridimensional. Um esquema conceitual que pretende capturar todas as peculiaridades de dados biológicos deve especificar diferentemente cada uma das três representações.

Desta foram, ontologias são ferramentas fundamentais para linguagens de modelagem de dados conceituais se enriqueçam com entendimento sobre o domínio dos conceitos. Ontologias pode dar suporte para a criação de modelos mais significativos. Adicionalmente, linguagens de modelagem conceitual não são mecanismos adequados para representar relações complexas entre conceitos da forma com as ontologias fazem. Tipicamente, se desejamos especificar a semântica do conceito gene em um esquema conceitual de dados precisamos usar uma observação textual para explicá-lo.

Requisito	**#3**	Enriquecer a semântica dos modelos conceituais usando ontologias

3.6 Relacionamentos Biológicos

Relacionamentos semânticos usados no domínio da biologia possuem mais restrições a serem mantidas do que as que são oferecidas por linguagens de modelagem tradicionais. Para evitar relacionamentos biológicos incorretos, por exemplo, o uso do relacionamento *is-a* deve levar em consideração o nível biológico que está sendo aplicado a cada classe (ex. organismo, célula, molécula, etc) e o tipo de inclusão que está sendo aplicada (ex. terminológica, funcional, papel, etc). Por exemplo, seria incorreto especializar ou generalizar dois conceitos com níveis biológicos não compatíveis como organismo e molécula.

Outro ponto importante são os tipos de relacionamento que carregam várias semânticas associadas. Por exemplo, os relacionamentos do tipo "parte-todo" apresentados no GeneOntology podem ser usados para descrever relacionamentos tanto de inclusão de vocabulário como de participação, porém em modelos orientados a objeto relacionamentos deste tipo podem representar ainda agregados compostos, ou agregados compartilhados. Além disso, Winston et al.[4] descreveram seis tipos de relacionamento "parte-todo", baseado nas seguintes propriedades:

1. configuração, se as partes tem um papel estrutural ou funcional com respeito a outra parte ou o todo que eles formam;

2. substância, se a parte é feita das mesmas coisas que o todo;

3. invariância, se a parte pode ser separada do todo.

[103] demonstrou que é fundamental usar todos os tipos de relacionamento "parte-todo" para representar os relacionamentos biológicos. Desta forma, a modelagem conceitual para biologia molecular deve incluir mecanismos para enriquecer a semântica dos relacionamentos tradicionais presentes nas linguagens existentes. Supomos que a melhor maneira seria permitir que o próprio projetista definisse as semânticas dos relacionamentos através de alguma linguagem lógica.

Requisito	**#4**	Permitir explicitar as semântica dos relacionamentos

3.7 Funções Biológicas

O sucesso da pesquisa da biologia depende da correta associação da função biológica com a estrutura dos elementos biológicos. A função biológica são objetos de primeira classe da pesquisa biológica porque eles ajudam a entender como os sistemas biológicos funcionam[104]. Funções biológicos são estudadas no sub-campo da biologia molecular, chamado de genômica funcional, a qual lida com a análise do fenótipo[104].

Função biológica é um conceito complexo e engloba dois diferentes conceitos na biologia. O Conceito da função biológica pode ser classificado com uma função local ou função integrada[47]. Função local trata das interações individuais que ocorrem entre a entidade biológica de interesse, e outras entidades. Por exemplo, a função local de uma enzima trata do substrato que ela age sobre, e sobre os ligantes que ativam ou inibem a enzima.

Função integrada trata do papel que uma entidade biológica executa em um grande sistema do qual ela faz parte. Por exemplo, uma enzima em uma via para biossíntese de lisina pode ter várias funções integradas, incluindo biossíntese de lisina e biossíntese de aminoácido. Esta enzima tem múltiplas funções integradas porque participa múltiplos sistemas biológicos com escopo hierárquicos e aninhados. Neste exemplo, o sistema de biossíntese de lisina é um subsistema do sistema biossíntese de aminoácido.

O fato de que a mesma proteína pode ter funções idênticas locais em dois diferentes organismos, mas diferentes funções integradas demonstra que existem diferenças entre esses conceitos. Por exemplo, considere uma enzima E que catalisa a mesma interação molecular em dois organismos mas opera-os no contexto de dois processos celulares distintos nos dois organismos.

Nas linguagens de modelagem tradicional, estrutura e função biológica pode somente ser representada por conceitos ou relacionamentos. Então, defendemos que modelos de dados conceituais para biologia molecular devem representar a diferença semântica entre estruturas e funções biológicas visto a importância que eles possuem para a pesquisa em biologia.

Requisito	**#5**	Representar funções e estruturas biológicas de forma distinta

3.8 Herança Não-Monotônica

Relacionamentos de herança na biologia são tipicamente não-monotônicos. As linguagens de modelagem conceitual tradicionais podem representar somente herança monotônica que é o relacionamento de generalização onde as propriedades das subclasses não podem redefinir as propriedades herdadas das superclasses. Por exemplo, a Figura 3.5 apresentam diagrama de classes UML com duas hierarquias com associações entre classes em diferentes níveis. O relacionamento entre os conceitos Biomolecule e StructuralComponent indica que é permitido falar de biomoléculas tendo componentes estruturais, mas sabemos que nem

todos os tipos de biomolécula podem ter qualquer tipo de componente estrutural. Seguindo o exemplo, Protein é um tipo de Biomolecule que pode ter somente uma estrutura do tipo AlphaHelix, a qual é um tipo de StructuralComponent. O relacionamento hasComponent entre os conceitos Protein e AlphaHelix explicitam este relacionamento mas não existe maneira para redefinir o relacionamento herdado das superclasse Biomolecule. Então existe uma contradição entre a herança estrutural oferecida pelas linguagens de modelagem estruturais e a semântica exigida para as heranças em biologia.

A herança não-monotônica é um problema já conhecido e discutido na literatura[66, 54]. Porém, as linguagens de modelagem tradicionais citadas neste trabalho não a implementam como ilustrado na Figura 3.5. Este problema pode ser superado usando alguns truques de modelagem. Por exemplo, poderíamos apagar o relacionamento Biomolecule-StructuralComponent da superclasse e incluir um relacionamento para cada subclasse (Figura 3.5).

Outra solução possível é criar duas subclasses para Biomolecule denominadas WithStructure e WithOutStructure. Essas classes representarão biomoléculas com e sem componente estrutural respectivamente. Em seguida, criamos um relação de associação entre a classe WithStructure e classe StructuralComponent. A primeira solução perdemos a semântica que a biomolécula possui componente estrutural e na última solução criamos classes somente para representar a existência ou não do relacionamento. Nas duas soluções estamos especificando em excesso ou em falta.

Por isso, uma linguagem para modelagem conceitual para biologia deve permitir a especificação de herança não-monotônica.

Requisito	**#6** Representar herança não-monotônica

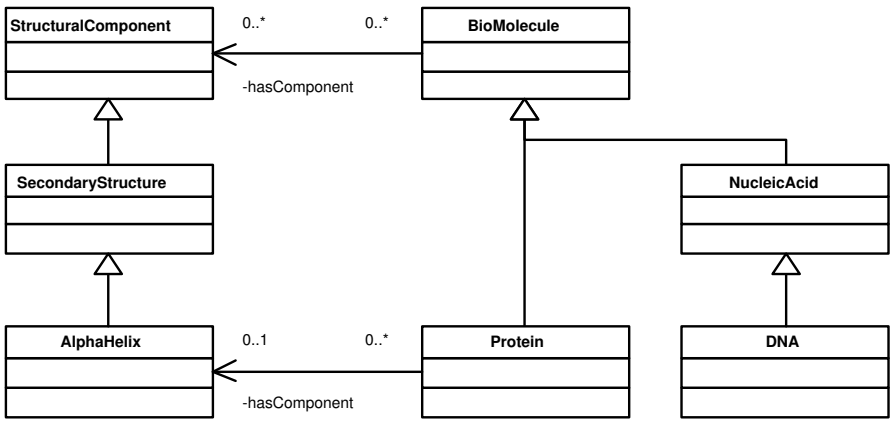

Figura 3.5: Exemplo de duas associações em níveis hierárquicos diferentes.

3.9 Fatores Externos

Em um organismo vivo a montagem de milhares de genes e seus produtos (ex. RNA e proteínas) trabalham de uma maneira intrincada para criar os sistemas biológicos. Então

a maioria dos relacionamentos biológicos entre dois elementos depende de um conjunto de fatores externos incluindo outros relacionamentos entre outros elementos. Por exemplo, a susceptibilidade dos alelos, os quais são variações diferente do mesmo gene podem ser influenciadas por três fatores. Infelizmente, linguagens de modelagem conceitual tradicionais não representam relacionamentos que podem variar de acordo com fatores independentes. Adicionalmente, o uso de especificações probabilísticas e empíricas (ex. depende muito, maioria de, etc) são também importantes para descrever relacionamentos biológicos[55].

Requisito	**#7**	Representar relacionamentos probabilísticos e empíricos

3.10 Classificações Biológicas

Técnicas de classificação exercem um papel importante no domínio da biologia porque elas permitem comparar diferentes organismos de diferentes domínios, reinos, fila, e classes, e contrastar esses organismos e entender seus relacionamentos evolucionários[67]. Na filogenia e evolução da biologia, a genômica comparativa tem dependido das comparações entre sequências no nível do gene e da proteína. E no futuro, isto dependerá mais e mais no rastreamento não somente das sequências de ADN mas como o genoma evolui com o tempo. O sistema de classificação biológico usa taxon, ou níveis de organização individual, para classificar todos os organismos.

Tipicamente essas classificações são feitas automaticamente. Porém, o perigo na geração de resultados incorretos é muito grande na filogenética computacional do que em vários outros campos da ciência porque a análise filogenética não possui bases empíricas[23]. Por exemplo, todos os métodos largamente usados assumem que as divergências evolucionárias são estritamente bifurcadas exceto em um dado conjunto de dados, este modelo pode ser violado por causa da transferência de material genético entre os organismos. Métodos filogenéticos tem dois requisitos fundamentais: a qualidade dos dados de entrada e a visão para os dados a partir de vários ângulos possíveis (distância, parsimônia máxima, etc)[24].

Em termos de modelagem conceitual, a qualidade dos dados de entrada pode ser garantida somente pela definição de restrições, enquanto a visão dos dados por vários ângulos possíveis podem ser especificadas usando visões. O poder de expressividade para representar restrições é limitado na maior parte das linguagens de modelagem conceitual exceto em UML que permite o uso de uma linguagem para definição de restrições denominada OCL (object constrained language)[82]. Desta forma, é um requisito que uma linguagem de modelagem conceitual para biologia molecular precisa especificar restrições complexas para evitar a entrada de dados incorretos. Além disso, esta linguagem conceitual deve incluir mecanismos para criar visões sobre o esquema conceitual.

Requisito	**#8**	Representar restrições complexas as quais garantam a qualidade dos dados de entrada e mantenham a sua consistências

Requisito	**#9** Permitir a definição de visões sobre o modelo conceitual

O problema anterior trata das classificações que são realizadas "a posteriori" baseadas somente na disponibilidade dos dados. No entanto, existem classificações que precisam ser realizadas "a priori". Neste caso, existem três formas de representar essas classificações:

1. Especificar hierarquias "é um" no nível do esquema criando um tipo de conceito para cada elemento da hierarquia;

2. Representar esta classificação no nível dos dados e criar um tipo de conceito que represente os tipos de elemento da hierarquia;

3. Usar uma abordagem híbrida através da criação de um tipo de conceito para cada tipo de elemento relevante da hierarquia e um tipo de conceito geral para o resto dos elementos da hierarquia;

Nós utilizamos o seguinte exemplo para ilustrar essas abordagens. Suponha que precisamos criar um esquema conceitual para armazenar informação sobre microorganismos que ten um ou mais aspectos em comum, discutidos em [69]. Esses aspectos podem ser o grupo taxonômico, fonte isolante, condição de crescimento, bioquímico, morfologia ou outros determinantes. Além disso, microorganismos podem ser membros de mais de um grupo (ex. bioquímico e morfológico); alguns desses grupos pode ser subtipados posteriormente (streptococci, staphylococci etc.) e, existem sobreposições que definem um grupo.

A primeira solução é especificar uma hierarquia do tipo "é um" no nível do esquema criando um tipo de conceito para cada elemento da hierarquia. Esquematicamente, isto pode ser representado como conjuntos de microorganismos como descrito na Figura 3.6. Por esta razão, conjuntos possuem sobreposição assim como atributos identificadores distintos, os quais por si só podem ser um subtipo de outra hierarquia.

A Figura3.7 apresenta um esquema conceitual possível usando a notação ER. Neste esquema, nós criamos um tipo de entidade para cada microorganismo (OrganismA, OrganismB, OrganismC, OrganismD) e fizemos ele herdar, através de generalização, os atributos de entidades que representam um grupo de aspectos específicos. No entanto, cada aspecto pode ser colocado em mais de um grupo específico, por exemplo, o aspecto 'c' é apresentado em quatro grupos (GroupII, GroupIII, GroupIV and GroupV). Então precisamos criar uma entidade específica para cada aspecto específico (AspectA, AspectB, AspectC, AspectD, AspectE and AspectF) para evitar a duplicação da definição de atributos em diferentes entidades. Desta forma, nós podemos usar herança múltipla par representar cada entidade do grupo de aspectos. Embora todas as associações entre os aspectos, grupos de aspectos e organismos representados no diagrama ER sejam corretos, o diagrama torna-se denso e de difícil compreensão.

A segunda solução é representar a classificação do organismo no nível dos dados e criar um tipo de conceito que represente os tipos de elemento da hierarquia. Esta solução

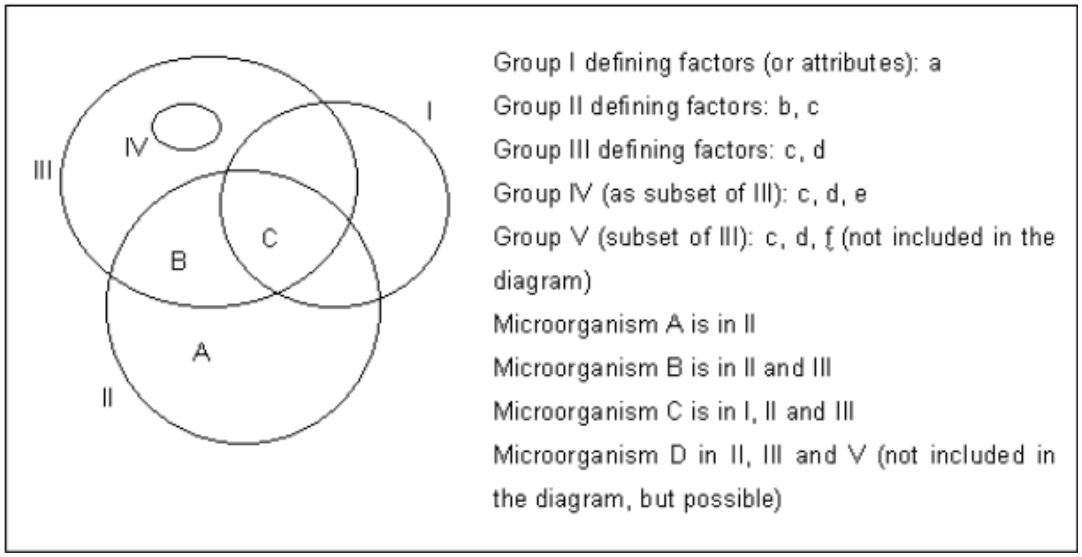

Figura 3.6: Representação esquemática dos grupos de organismos[69]

foi proposta em [69] através da criação de duas entidades (Microorganism and MOGroup) e um relacionamento m:n entre elas. Esta abordagem não mantém a semântica correta porque a classificação será representada pelos dados e não pelo nível do esquema de dados.

Em resumo, linguagens de modelagem tradicionais não são adequadas para representar hierarquias complexas. No caso de linguagens conceituais que não possuem herança múltipla isto é impossível.

A terceira solução usa uma abordagem híbrida. A linguagem UML na sua última versão[94] trata deste problema oferecendo um construtor denominado super-tipo. Este construtor é um meta-tipo cujas instâncias são subtipos de outro tipo. Por exemplo, TreeSpecies é um super-tipo do tipo Tree. Os subtipos de Tree (ex. Ash, Birchm Cherry) são todos instâncias da classe TreeSpecies. A UML oferece o relacionamento especial super-tipo que liga a classe super-tipo com a classe alvo. Desta forma, a classe alvo pode ter uma hierarquia de classes com somente os subtipos que são interessantes para o esquema conceitual. Cada subtipo deve indicar qual classe super-tipo ele pertence porque pode existir várias classes super-tipo associadas com a classe alvo.

Embora esta seja uma boa estratégia para selecionar somente elementos importantes para aparecerem na hierarquia, esta abordagem não resolve o problema de hierarquias grandes em modelos conceituais. Essas hierarquias rompem com o princípio da ortogonalidade onde precisamos balancear a simplicidade com o poder de expressividade da linguagem de modelagem[83]. Por isso, modelo de dados conceituais para biologia molecular devem lidar com o problema de grandes hierarquias.

Requisito	#**10**Permitir representar hierarquias grandes

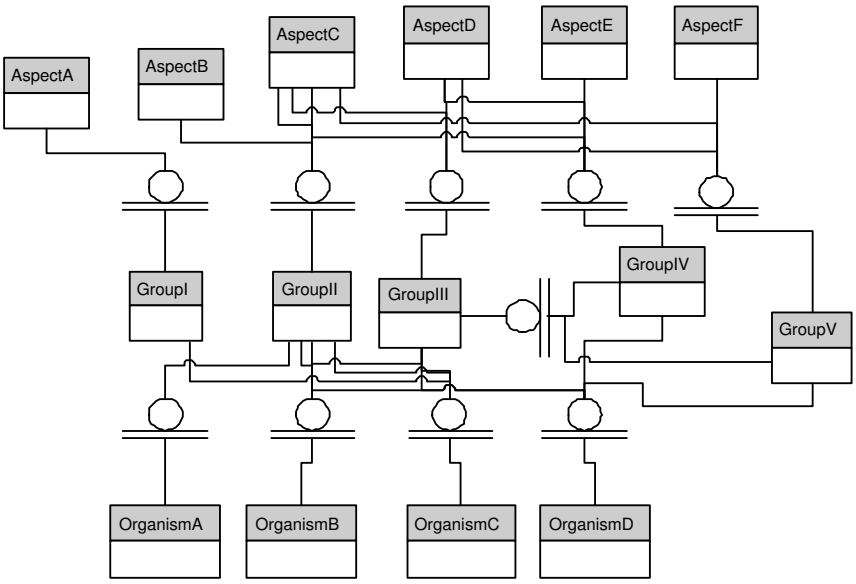

Figura 3.7: Diagrama ER relacionado com os microorganismos, seus grupos e aspectos

3.11 Redes de Vias da Biologia

Gerenciar e analisar informação genômica no contexto da sequência de ADN é apropriado para estudar questões da organização do genoma, evolução e mutação[48]. Porém, sequência de ADN não reflete o contexto nos quais a maioria dos genes age, i.e.m, genes relacionados funcionalmente não são normalmente agrupados no ADN, pelo contrário são distribuídos em sítios distantes. Por isso, vias biológicas[84][5] é uma alternativa para estudar a informação genômica.

Normalmente o desenvolvimento de bancos de dados de vias metabólicas tentam capturar mais informações complexas do que simples sequências de genes por causa do tipo de interações entre as moléculas (reações químicas): os objetos que formam os dados são nós das redes ligadas por arestas representando as reações químicas[56], e assim modelos especializados, tais como modelos baseados em grafos, deve ser projetados para facilitar a análise de redes bioquímicas[85]. Tais modelos geralmente contém poucos dados quantitativos e são, primariamente, obtidos pela análise qualitativa.

Desta forma, modelos conceituais de dados que objetivam representar redes bioquímicas devem permitir o projetista criar um esquema analítico baseado em elementos de baixo nível do tipo sequências, elementos químicos, enzimas e assim por diante.

Outro aspecto importante das vias metabólicas é seu aspecto dinâmico. Por exemplo, a Figura 3.8 exemplifica a estrutura de uma rede de vias metabólicas e a formulação matemática correspondente. A rede metabólica é representada com um grafo bipartido onde as letras A até D representam os componente como metabólitos e nucleotídeos. As letras u, v, p e q representam as taxas da reações da rede. A taxa da rede u representa a conversão de A+D para B. A taxa de reação u é ativada pelo componente C como indicado no gráfico. Formulações matemáticas para as taxas da reação u e v também são

[5]Vias biológicas são redes complexas de interações pelas quais as proteínas são expressas pelos genes executam a função biológica das células

apresentadas na figura. O sistema de equações diferenciais para a rede de vias metabólicas especifica a dinâmica do sistema.

A dinâmica das vias metabólicas deve ser capturada pelos modelos conceituais. Atualmente, as linguagens de modelagem conceituais tradicionais provêm diagramas de estados e diagramas de interação para modelar o comportamento das vias metabólicas, porém eles não fornecem mecanismos para especificar os aspectos espaciais e temporais, assim como a complexidade das reações e seus modelos matemáticos.

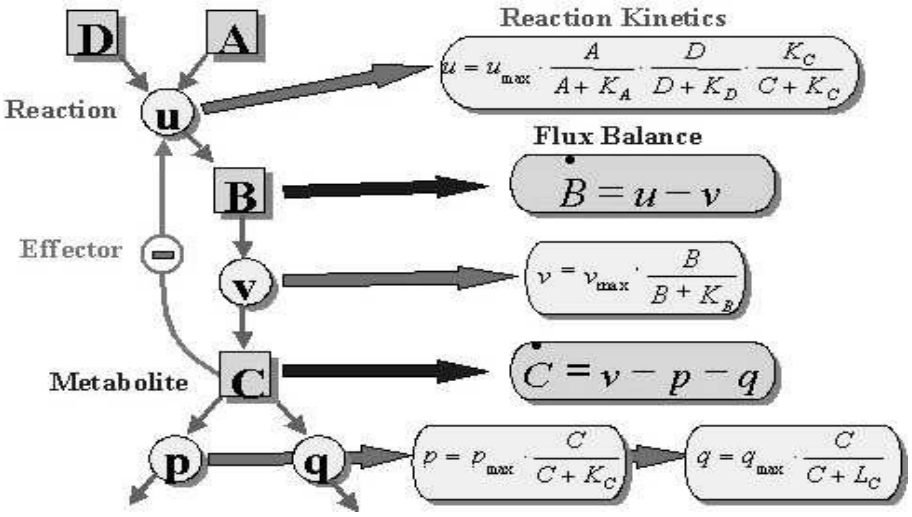

Figura 3.8: Estrutura de uma rede de vias metabólicas

3.12 Aspectos Temporais e Espaciais

Estudos recentes mostram que organização espacial e temporal dos processos intracelulares tem um papel crucial na troca de sinais entre células e no metabolismo celular[86][57]. Porém, este tipo de informação é difícil de representar em linguagens de modelagem conceitual tradicionais como mostrado em[68]. Usamos um exemplo, ilustrado na Figura 3.9 para explicar esses problemas. Suponha que desejamos representar uma via metabólica usando um diagrama UML. Este diagrama é ilustrado na Figura 3.9 onde é representado o grupo de interações entre os processos que ocorrem em uma via metabólica. Cada tipo de processo pode ser um ativação, inibição, transcrição, transição ou catálise. Um tipo de processo agrupa diversas interações entre tipos de moléculas. Um tipo de molécula pode ser uma proteína, uma enzima ou um aminoácido.

Observamos que neste diagrama (Figura 3.9) os aspectos temporais usam construtores de baixo nível (classes e relacionamentos) os quais são também usados para definir os elementos biológicos de base (moléculas). Desta foram, as dimensões espaciais e temporais são representadas usando os mesmos tipos de construtores. Isto pode ser um vantagem para domínios que não possuem semânticas e restrições complexas. No entanto, em domínios como o da biologia, construtores de alto nível poderia simplificar a construção de esquemas de dados e dar mais expressividade a eles.

O uso de construtores de baixo nível pode induzir alguns problemas do tipo: (1) aumentar a complexidade do esquema conceitual, (2) demandam projetistas com alta capacidade e (3) dão a mesma importância para elementos de diferente dimensões, por exemplo, o conceito proteína é conceitualmente mais relevante do que informações sobre localização celular. Desta forma, nós afirmamos que os modelos de dados para biologia molecular devem representar aspectos espaciais e temporais dos processos biológicos como fornecer construtores de alto-nível para aumentar a expressividade do modelo de dados.

Requisito	**#11** Representar aspectos espaciais e temporais do processos biológicos

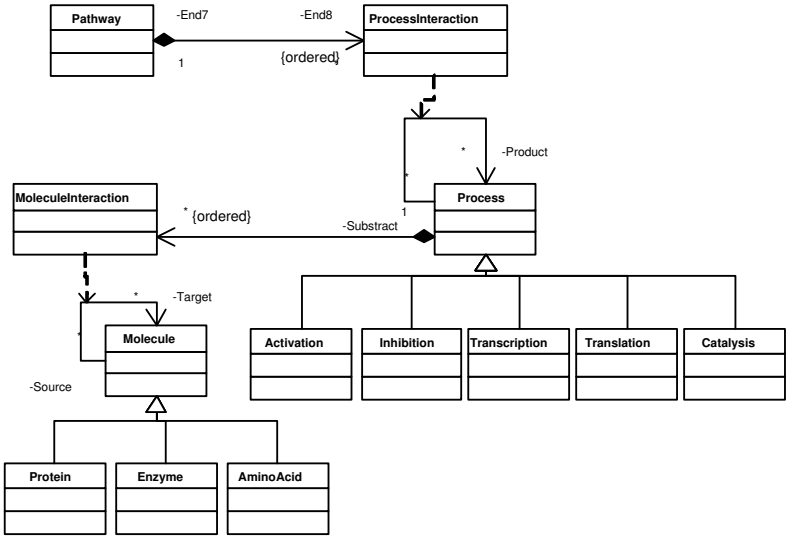

Figura 3.9: Diagrama de classes UML modelando vias metabólicas

3.13 Outros Aspectos Importantes

Observamos que existem algumas questões de caracter geral com respeito as linguagens de modelagem tradicionais. Normalmente esse problemas aparecem quando essas linguagens são utilizadas para modelar domínios complexos, como é o caso da biologia. Essas questões estão relacionadas com o tamanho do esquema de dados gerado e com a falta de nomenclatura comum.

Tamanho do Esquema Conceitual No domínio da biologia onde existe uma grande diversidade e variabilidade de tipos de dados, como ilustrado na Figura 3.1, este é um problema crucial. Cada uma das propriedades de um conceito (atributo e relacionamento) pode variar de acordo com diversas condições e cada atributo do domínio pode apresentar uma vasta escala de valores. Desta maneira, tentar modelar todos os tipos de conceitos com suas propriedades pode gerar um número exponencial de tipos de dados, atributos, relações e domínios. Neste contexto, o único modo para reduzir a complexidade do modelo conceitual de dados é oferecer construtores pré-definidos representando conceitos biológicos básicos como sequência, gene, proteína, etc. Além disso, uma linguagem de modelagem conceitual deve prover mecanismos para criar novos construtores baseados em outros construtores.

Falta de uma nomenclatura comum Representando conceitos biológicos com construtores de baixo nível pode aumentar a complexidade do esquema conceitual e esconder importantes semânticas dos conceitos modelados como ilustrado na seguinte citação:

"Diferente abordagens para modelar informação genética pode gerar diferentes interpretações sobre alguns termos biológicos, mesmo para noções familiares como Gene e ORF."[34]

Como um exemplo, é muito difícil escolher um nome de conceito apropriado durante a criação de um modelo porque existe uma falta de padronização no vocabulário usado no domínio da biologia[56, 25, 26, 27]. Existem alguns esforços nesta direção [58, 105] e o estabelecimento do consórcio GeneOntology[106] demonstra uma preocupação com a criação de uma terminologia unificada.

Mesmo termos usais como gene podem ter múltiplas interpretações que podem confundir a interpretação do esquema pelo leitor. Por exemplo, as figuras 3.10 e 3.11 apresentam dois esquemas diferentes contendo o conceito gene. Ambos definem que um gene pode ser representado como uma lista encadeada de outros conceitos (splicing unit and chromosome fragment). Porém, a segunda figura classifica o fragmento de cromossomo como uma região transcrita e não transcrita enquanto o primeiro define somente unidade particionada (spliced unit) a qual é uma região transcrita. Surpreendentemente o mesmo nome gene tem duas semânticas diferentes: uma sequência que tem somente informações codificantes na Figura 3.10 ou uma sequência que pode ser ou não ser informação codificanteFigura (cf., Figure 3.11).

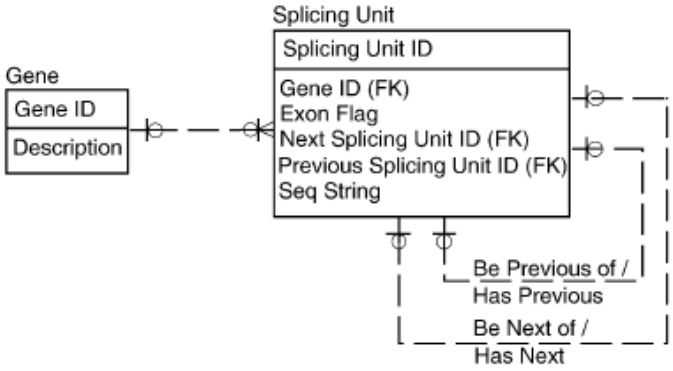

Figura 3.10: Um gene representado através de um lista encadeada de subsequências[70]

Em resumo, essas duas questões levam a necessidade de um último requisito para um modelo de dados conceitual para biologia molecular que está relacionada com a expressividade do modelo. Imaginamos que esta expressividade pode ser melhorada se novas linguagens de modelagem conceitual puderem ser estendidas para incorporarem novos construtores mais expressivos.

Requisito	#**12**Permitir a definição de construtores de alto-nível baseados nas definições de construtores de baixo nível

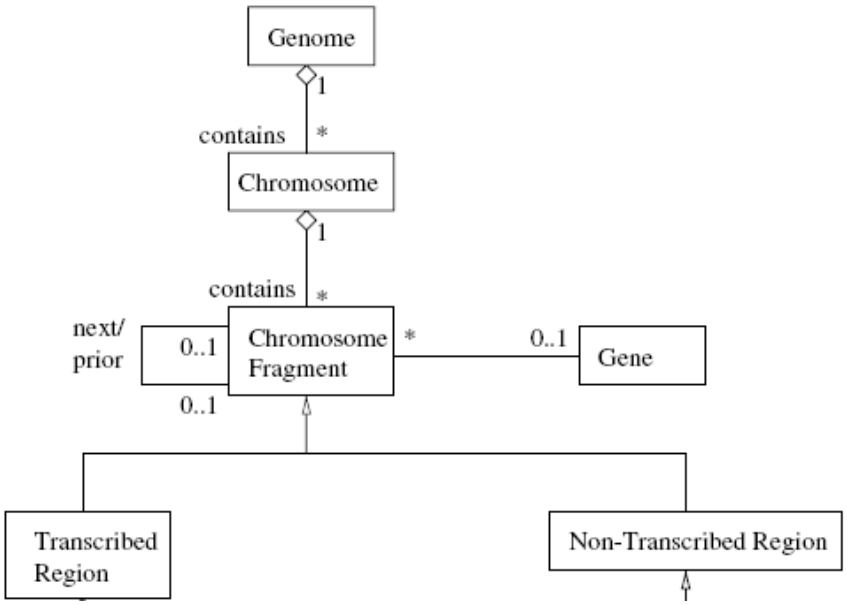

Figura 3.11: Um gene representado como um lista encadeada de fragmento de cromossomos[34]