



José Antonio Fernandes de Macêdo

**Um modelo conceitual para biologia
molecular**

Tese de Doutorado

Tese apresentada ao Programa de Pós-graduação em Informática do Departamento de Informática da PUC-Rio como parte dos requisitos parciais para obtenção do título de Doutor em Informática

Orientador: Prof. Edward Hermann Haeusler

Rio de Janeiro
Agosto de 2005



José Antonio Fernandes de Macêdo

**Um modelo conceitual para biologia
molecular**

Tese apresentada ao Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC-Rio como parte dos requisitos parciais para obtenção do título de Doutor em Informática. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Edward Hermann Haeusler

Orientador

Departamento de Informática — PUC-Rio

Prof. Rubens Nascimento Melo

Departamento de Informática - PUC-Rio

Prof. Daniel Schwabe

Departamento de Informática - PUC-Rio

Prof. Marta L. Queirós Mattoso

Departamento de Informática - UFRJ

Prof. Antonio Basílio Miranda

Departamento de Bioquímica - Fiocruz

Prof. José Eugênio Leal

Coordenador Setorial do Centro Técnico Científico —
PUC-Rio

José Antonio Fernandes de Macêdo

Cursou Tecnólogo em Processamento de Dados na PUC-Rio em 1998. Atuou em empresas como Gerente de Informática, Analista de Sistemas, Administrador de Bancos de Dados, Programador de 1986 a 1991. Retornou ao mundo acadêmico para realizar Mestrado em Banco de Dados na PUC-Rio, encerrado em 2000, com publicações na área de Banco de Dados. Ministrou disciplinas relacionadas à Modelagem de Dados, Banco de dados, Linguagens de Programação em diversos cursos, inclusive no Bacharelado em Informática da PUC-Rio. Lecionou no curso de Administração e Tuning de Banco de Dados do CCE PUC-Rio. Possui interesse acadêmico e profissional nas áreas de Banco de Dados, Desenvolvimento de Aplicações baseadas em objetos e Engenharia de Software.(Rio de Janeiro, Brasil)

Ficha Catalográfica

Macêdo, José Antonio Fernandes de

Um modelo conceitual para biologia molecular/
José Antonio Fernandes de Macêdo; orientador: Edward
Hermann Haeusler . — Rio de Janeiro : PUC-Rio,
Departamento de Informática, 2005.

v., 93 f: il. ; 29,7 cm

1. Tese (doutorado) - Pontifícia Universidade
Católica do Rio de Janeiro, Departamento de Infor-
mática.

Inclui referências bibliográficas.

1. Informática – Teses. 2. Banco de Dados. 3. Mod-
elagem Conceitual. 4. Biologia Molecular. 5. Bioinfor-
mática. I. Haeusler, Edward Hermann. II. Pontifícia Uni-
versidade Católica do Rio de Janeiro. Departamento de
Informática. III. Título.

Aos meus pais, Lásaro e Maria Lúcia, que nunca mediram esforços para me oferecer a melhor educação. À minha esposa Janaína, pela seu amor, carinho, confiança e paciência, sem os quais teria sido muito difícil esta jornada. E ao meu filho Antonio que em breve nascerá e a quem aguardamos com muita ansiedade e amor.

Agradecimentos

Ao Professor Sérgio Lifschitz, orientador deste trabalho, pela motivação, pela disponibilidade e pela amizade que demonstrou ao longo de todo nosso convívio.

Aos Professores Fábio Porto e Philippe Picouet, pelas boas discussões e acolhida quando estive longe do Brasil.

Ao professor Edward Hermann Haeusler, pela ajuda na busca de soluções para os problemas encontrados neste trabalho.

Aos funcionários da secretaria, da biblioteca e do Lab-DI, pela atenção e presteza com que sempre atenderam às minhas solicitações.

Aos amigos da PUC, que tive a felicidade de conhecer durante esta minha jornada. Eles ajudaram a abrandar os momentos mais difíceis.

Macêdo, José Antonio Fernandes de; Haeusler, Edward Hermann.
Um modelo conceitual para biologia molecular. Rio de Janeiro, 2005. 93p. Tese de Doutorado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Projetos de genômica e biológica molecular estão gerando dados cujos volumes e complexidades jamais foram observados nesta área. Além disso, fontes de dados e de conhecimento são produzidas e utilizadas por grupos de pesquisa os quais utilizam terminologias diferentes (sinônimos, apelido e fórmulas), sintaxes diferentes (estrutura de arquivos e separadores) e semânticas diferentes (intra e interdisciplinares homônimos). Desta forma, 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. Neste trabalho, estudamos os problemas para representação de dados biológicos apresentados nas principais linguagens de modelagem tradicionais. Em seguida, levantamos os requisitos para um novo modelo de dados conceitual para biologia molecular. Finalmente, propomos um novo modelo conceitual contendo construtores específicos para solucionar alguns dos problemas estudados. Além disso, formalizamos o modelo proposto usando lógica de primeira ordem e a utilizamos para realizar inferências, as quais possam auxiliar o projetista de banco de dados na criação de um esquema conceitual de dados.

Palavras-chave

Informática – Teses; Banco de Dados, Modelagem Conceitual, Biologia Molecular, Bioinformática

Macêdo, José Antonio Fernandes de; Haeusler, Edward Hermann.
A conceptual model for molecular biology. Rio de Janeiro,
2005. 93p. PhD. Thesis — Departamento de Informática, Pontifícia
Universidade Católica do Rio de Janeiro.

Genomic and molecular biology projects are generating knowledge data whose volume and complexity are unparalleled in this research area. In addition, data and knowledge sources produced and used by research groups have terminological differences (synonyms, aliases and formulae), syntactic differences (file structure, separators and spelling) and semantic differences (intra- and interdisciplinary homonyms). In this context, data management techniques play a fundamental role for biological applications development because it offers adequate abstractions to design, implement, access and manage data, in order to generate knowledge. In this work, we study the representation problems presented in traditional languages. Following, we raise the main requirements for a new conceptual data model specially conceived for molecular biology. Finally, we propose a new conceptual data model with special types of constructor trying to solve some of the representation problems discussed before. In addition, we formalize our proposed model using first-order logic and we use this logical description to infer some properties that may help database designer during the elaboration of a conceptual database schema.

Keywords

Computer Science; Database Systems; Conceptual Modeling; Molecular Biology; Bioinformatics.

Conteúdo

1	Introdução	12
1.1	Bioinformática e Modelagem Conceitual	12
1.2	Objetivos da Tese	15
1.3	Estrutura da Tese	15
2	Contexto da Pesquisa	17
2.1	Introdução	17
2.2	Ciclo de Vida de um Sistema de Informação	17
2.3	Modelos Conceituais de Dados	20
2.4	Tipos de Modelos Conceituais	21
2.5	Classificação de Modelo de Dados	22
2.6	Modelo de dados Temáticos	24
2.7	Bases da Biologia Molecular	24
2.8	Sistemas de Informação Biológico	25
2.9	Classificações Biológicas e Ontologias	30
2.10	Conclusão	32
3	O Problema	34
3.1	Introdução	34
3.2	Representação de Dados Biológicos	34
3.3	Conceitos Atômicos - Uma visão Reducionista	37
3.4	Sistemas Biológicos - Uma visão Holística	39
3.5	As Ontologias e a Biologia	41
3.6	Relacionamentos Biológicos	42
3.7	Funções Biológicas	43
3.8	Herança Não-Monotônica	43
3.9	Fatores Externos	44
3.10	Classificações Biológicas	45
3.11	Redes de Vias da Biologia	48
3.12	Aspectos Temporais e Espaciais	49
3.13	Outros Aspectos Importantes	50
4	Descrição do Modelo de Dados	53
4.1	Introdução	53
4.2	Requisitos para um Modelo de Dados Biológico	53
4.3	BioConceptual e o Modelo de Dados ODMG	53
4.4	Tipo de Dado Objeto	54
4.5	Atributos de um Tipo de Dado Objeto	57
4.6	Tipo de Relacionamento	59
4.7	Relacionamentos de Associação	60
4.8	Ligações "É-UM" entre tipos de objetos	61
4.9	Restrições de Integridade	65
4.10	Múltiplas Percepções e Representações	66
5	Formalização do Modelo de Dados	73
5.1	Introdução	73
5.2	Uma visão geral do meta-modelo do BioConceptual	73
5.3	Abordagem para Formalização	74

5.4	Realizando Inferências ao Modelo	79
6	Conclusão	82
6.1	Revisão dos Objetivos e Resultados da Tese	82
6.2	Trabalhos Futuros	83

Lista de Figuras

1.1	Ciclo experimental do conhecimento da biologia	14
2.1	Banco de Dados e Esquemas de Dados	19
2.2	Um esquema do processo de projeto de um banco de dados [16]	21
2.3	Camadas lógicas da Arquitetura de um Sistema de Informações Biológicas	26
2.4	Complexidade das estruturas das Ontologias[87]	32
3.1	A diversidade e a variabilidade do tipos de dados biológicos	35
3.2	Diagrama de classe representando o Genoma[34]	38
3.3	Sequências de ADN e especificação de motivos	40
3.4	Complexidade de Pirâmide da Vida[59]	40
3.5	Exemplo de duas associações em níveis hierárquicos diferentes.	44
3.6	Representação esquemática dos grupos de organismos[69]	47
3.7	Diagrama ER relacionado com os microorganismos, seus grupos e aspectos	48
3.8	Estrutura de uma rede de vias metabólicas	49
3.9	Diagrama de classes UML modelando vias metabólicas	50
3.10	Um gene representado através de um lista encadeada de subsequências[70]	51
3.11	Um gene representado como um lista encadeada de fragmento de cromossomos[34]	52
4.1	instanciação de um Tipo de Dado Objeto do <i>BioConceptual</i>	55
4.2	Exemplo de População do Tipo de Dado Objeto Gene	56
4.3	Atributos Multivalorados	58
4.4	Coleção de Dados	58
4.5	Um diagrama mostrando um tipo de relacionamento ligando dois tipos de objetos	60
4.6	Um exemplo de ligação É-UM	62
4.7	Exemplos de Sobreposição e Disjunção	63
4.8	Neste exemplo é possível inferir sobreposição automaticamente	63
4.9	Refinando os tipos de relacionamentos	64
4.10	Representando Herança Não-Monotônica usando ligação "É-UM" entre tipo de relacionamento	64
4.11	Um diagrama exemplificando o refinamento de um papel através do uso de ligações É-UM	65
4.12	Exemplo de uso do construtor <i>Constraint</i>	66
4.13	Exemplo ilustrando as diferentes percepções possíveis do mesmo fenômeno do mundo real, e suas possíveis representações.	67
4.14	Usando percepções para associar a diferentes representações com o mesmo tipo de objeto Gene.	70
4.15	Três tipos de instâncias que variam de acordo com diferentes percepções	72
5.1	<i>BioConceptual</i> Tipos de Construtores	74
5.2	Representação gráfica do conceito Gene	76
5.3	Um diagrama mostrando um tipo de relacionamento ligando dois tipos de objetos	77
5.4	Um exemplo de ligação É-UM	78
5.5	Exemplo de uso do construtor <i>Constraint</i>	79
5.6	Esquema Conceitual apresentando Inconsistências	80

Lista de Tabelas

4.1	Resumo dos requisitos para um modelo de dados conceitual para biologia molecular	54
4.2	Exemplos de propriedades definidas pelo sistemas	71
4.3	Populando o banco de dados usando diferentes percepções	72