Pontifícia Universidade Católica
do Rio de Janeiro

Cristian Tristão

Uma Abordagem para Modelar, Armazenar e Acessar Sequências Biológicas

Tese de Doutorado

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Informática do Departamento de Informática da PUC-Rio.

Orientador: Prof. Edward Hermann Haeusler

Rio de Janeiro
Julho de 2012

Cristian Tristão

Uma Abordagem para Modelar, Armazenar e Acessar Sequências Biológicas

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Informática do Departamento de Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Edward Hermann Haeusler
Orientador
PUC-Rio

Prof. Sérgio Lifschitz
PUC-Rio

Prof. Marcus V. S. Poggi de Aragão
PUC-Rio

Prof. Duncan Dubugras Alcoba Ruiz
PUCRS

Prof. Antônio Basílio de Miranda
FIOCRUZ

Prof. José Eugênio Leal
Coordenador(a) Setorial do Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 09 de julho de 2012

Cristian Tristão

Possui graduação em Bacharelado em Ciência da Computação pela Pontifícia Universidade Católica do Rio Grande do Sul (2004) e mestrado em Ciência da Computação pela Pontifícia Universidade Católica do Rio Grande do Sul (2007). Tem experiência na área de Ciência da Computação, com ênfase em Banco de Dados, atuando principalmente nos seguintes temas: Biologia Computacional, Banco de Dados Biológicos e Gestão e Análise de Processos. Natural de Gravataí, Rio Grande do Sul.

Ficha Catalográfica

Tristão, Cristian

Uma abordagem para modelar, armazenar e acessar sequências biológicas / Cristian Tristão ; orientador: Edward Hermann Haeusler. – 2012.
107 f. : il. (color.) ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2012.
Inclui bibliografia.

1. Informática – Teses. 2. Base de dados biológicos. 3. Modelagem conceitual de dados. 4. Estrutura de acesso e manipulação de dados biológicos. I. Haeusler, Edward Hermann. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Aos meus pais pelo incentivo e apoio incondicional.
A minha esposa, Elizabeth, pela paciência e carinho constantes.
E ao meu pequeno anjo Gabriel que chegou para dar mais alegria e motivação.

"Pois te esquecerás dos teus sofrimentos e deles só terás lembrança como de águas que passaram. A tua vida será mais clara que o meio-dia; ainda que lhe haja trevas, serão como a manhã. Sentir-te-ás seguro, porque haverá esperança; olharás em derredor e dormirás tranquilo. Deitar-te-ás, e ninguém te espantará; e muitos procurarão obter o teu favor."
(Jó, 11:16-19)

Agradecimentos

À Deus por me amparar nos momentos difíceis, me dar força interior para superar as dificuldades, mostrar os caminho nas horas incertas e me suprir em todas as minhas necessidades.

À minha família, que apesar da distância sempre esteve perto e pronta para me acolher e incentivar. Esse doutorado é para você meu pai, que me ensinou que o maior legado é a educação.

Ao meu orientador, professor Sérgio Lifschitz, que acreditou no meu potencial e pela oportunidade oferecida. Em especial ao professor Hermann, que foi peça fundamental na reta final desta jornada, pela disposição e disponibilidade incondicional.

Aos professores e funcionários da Pós-Graduação, pelo excelente convívio nestes anos de trabalho. Aos amigos, professores e pesquisadores da FIOCRUZ, Antônio, Catanho e Thomas, que compartilharam grandes momentos de discussões e novas ideias.

Aos colegas, amigos e pesquisadores do Labbio, Daniel goiano, Zé Maria, Maíra, seu Paulo, Carlos Juliano, Ana Carolina, Luciana, Andreia, Márcia, Renato e Percy. Esta Tese carrega um pedaço de cada um de vocês. Peço desculpas caso tenha esquecido alguém. Também gostaria de deixar meus agradecimentos à instituição FAPERJ, pelo apoio financeiro.

À minha pastora Virgínia, que cuidou com grande carinho do meu lado espiritual.

Por fim, queria fazer um agradecimento especial a dois outros frutos desta Tese, minha esposa Elizabeth, pelo amor e apoio incondicional em todos os momentos, e ao meu pequeno anjo Gabriel, que trouxe alegria para nossas vidas e um incentivo todo especial.

Resumo

Tristão, Cristian; Haeusler, Edward Hermann. **Uma Abordagem para Modelar, Armazenar e Acessar Sequências Biológicas.** Rio de Janeiro, 2012. 107 p. Tese de Doutorado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

As pesquisas na área da biologia molecular vêm produzindo um grande volume de dados e estes precisam ser bem organizados, estruturados e persistidos. Na sua grande maioria os dados biológicos são armazenados em arquivos no formato texto. Para grandes volumes de dados, o caminho natural seria utilizar SGBDs para gerenciá-los. Contudo, estes sistemas não possuem estruturas adequadas para representar e manipular dados específicos ao domínio. Por exemplo, sequências biológicas normalmente são tratadas como simples cadeias de caracteres (tipo texto/varchar) ou BLOB, e desta forma perde-se todo um conjunto de informações composicionais, posicionais e de conteúdo. Esta tese argumenta que a gerência de dados (estrutura, armazenamento e acesso de dados) se transformou em um dos principais problemas para o domínio de pesquisas da bioinformática. Desta maneira propõe-se um modelo conceitual biológico para representar informações do dogma central da biologia molecular, bem como um tipo abstrato de dado (ADT – do inglês *Abstract Data Types*) específico para a manipulação de sequências biológicas e seus derivados.

Palavras-chave

Base de Dados Biológicos; Modelagem Conceitual de Dados; Estrutura de Acesso e Manipulação de Dados Biológicos.

Abstract

Tristão, Cristian; Haeusler, Edward Hermann (Advisor). **An Approach to Model, Store and Access Biological Sequences.** Rio de Janeiro, 2012. 107 p. DSc Thesis - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The researches in molecular biology have been producing a large amount of data and they need to be well organized, structured and persisted. Mostly biological data are stored on files in text format. For large volumes of data, the natural way would be to use DBMS to manage them. However, these systems do not have adequate structures to represent and manipulate data specific to the domain. For example, biological sequences are typically treated as simple strings (type text/varchar) or BLOB, and thus lost a whole set of compositional, positional and content information. This thesis argues that the management of data (structure, storage and data access) has become a major problem for researches in bioinformatics. Thus we propose a conceptual model for representing biological information of the central dogma of molecular biology, as well as an Abstract Data Types (ADT) specific for the manipulation of biological sequences and its derivatives.

Keywords

Biological Database; Conceptual Modeling of Data; Access Structure and Manipulation of Biological Data.

Sumário

Lista de figuras

Lista de tabelas