# PADRÕES DE TIPOS E MÉTODOS PARA BANCO DE DADOS EM BIOINFORMÁTICA[1]

E. M. WIECZOREK[2], E. LEAL[3]

III Congresso Científico do CEULP/ULBRA

**RESUMO:** Com o início do Projeto Genomas, em 1990, vários SGBD (Sistemas Gerenciadores de Banco de Dados) vêm sendo adaptados para suportar os dados gênicos (de genoma), a fim de se conseguir armazenar e buscar estas informações (dados) da maneira mais excelente possível. A maioria dos bancos de dados para a bioinformática (biológicos) consiste em longas cadeias de caracteres para representar as bases do DNA G (Guanina), A (Adenina), T (Timina) e C (Citosina). Um problema a ser superado quando se fala em banco de dados para bioinformática é que bancos de dados têm sido em grande parte usados para administrar dados empresariais, números simples, caractere ou datas. Poucos bancos de dados tiveram uma habilidade nativa para lidar com dados complexos, como dados multimídia, dados espaciais, ou dados genéticos (sucessão de genes). Para a resolução de tal problema, deve ser elaborada uma padronização de tipos de dados e métodos que os banco de dados devem suportar, relacionando-os e descrevendo suas funcionalidades.

**PALAVRAS CHAVE**: Bioinformática, Banco de Dados, Padrões de Tipos e Métodos.

**ARVORE DO CONHECIMENTO:** Engenharias, Ciência da Computação, Sist. de Computação.

## PATTERNS OF TYPES AND METHODS FOR DATABASE IN BIOINFORMATIC

**ABSTRACT:** With the beginning of the Genomas Project, in 1990, some DBMS (Systems Manager of Data base) come being suitable to support the biological data (of genoma), in order to obtain itself to store and to search these information (given) in the possible way most excellent. The majority of the data bases for the bioinformática (biological) consists of long chains of characters to represent the bases of DNA G (Guanina), (Adenine), T (Timina) and C (Citosina). A to be surpassed problem when if it speaks in data base for bioinformática is that data bases have been to a large extent used to manage given enterprise, simple numbers, character or dates. Few data bases had had a native ability to deal with complex data, as multimedia data, space data, or genetic data (succession of genes). For the resolution of such problem, it must be elaborated a standardization of types of data and methods that the data base must support, relating them and describing its functionalities.

**KEYWORDS**: Bioinformatic, database, Patterns of Types and Methods.

**INTRODUÇÃO:** O mapeamento do genoma humano e de outros organismos gera diariamente um elevado volume de informações que são sistematicamente armazenadas em bancos de dados computacionais, sendo estas informações fontes de estudo para a biologia e medicina através da bioinformática. O desafio apresentado pela bioinformática é encontrar a melhor forma de armazenamento e de pesquisa (SQL) para os dados gerados por projetos de pesquisa na área da bioinformática, como o projeto genoma humano, que possui centenas de gigabytes de dados a espera para serem armazenados e tratados. Para tanto, surge a necessidade de se possuir formas de armazenamento, acesso e pesquisa sobre tais dados, para que se consiga trazer a informação da melhor maneira desejada possível, devendo existir assim, técnicas diferenciadas para o tratamento destes dados, que são nada mais do que grandes cadeias de DNA (em banco de dados, grandes cadeias de caracteres). Para que se possa atingir tal meta, é necessário identificar padrões de tipos de dados e de métodos (de acesso, de armazenamento, etc), para que se possa implementar um banco de dados que

1 Parte do Trabalho de Conclusão de Curso do Primeiro Autor
2 Aluno de Sistemas de Informação no CEULP/ULBRA.
3 Professor orientador no curso de Sistemas de Informação no CEULP/ULBRA.

tenha suporte aos dados gênicos (de genoma), para que se possa trabalhar de uma maneira melhor os dados que já existem e os que estão sendo gerados diariamente.

**MATERIAIS E MÉTODOS:** Através de pesquisas na biblioteca do Centro Universitário Luterano de Palmas, pesquisas na Internet através de sites de busca como http://www.google.com.br, http://www.altavista.com.br e outros, foi possível detectar a existência de várias abordagens para a utilização de bancos de dados no domínio da bioinformática. A primeira abordagem seria o uso de SGBD's que possuem suporte para a criação de novos tipos de dados e métodos (banco de dados extensível). Este banco de dados extensível dará apoio às necessidades do sistema para definir tipos de dados novos que sejam capazes de criar entidades de domínio como sucessão genotípica; uso de operadores definidos pelo usuário; indexação de domínio específico, fornecendo apoio para índices específicos de dados biológicos e otimizar a extensibilidade, fazendo assim uma ordenação inteligente dos predicados em questão, envolvendo tipos de dados definidos pelo usuário (ORACLE CORPORATION, 1999). Uma segunda abordagem pode ser através de sistemas envolvendo Data Warehouses (Armazéns de Dados), pois estes são utilizados pela indústria há muitos anos, e como demonstrado pela figura 1, constituídos tipicamente de 5 camadas: as fontes de dados, que contém os dados a serem integrados (adicionados) ao Data Warehouse através dos Wrapper's (analisadores gramaticais de dados), os mediadores (que traduzem os dados para a representação do Data Warehouse), o próprio Data Warehouse, que é um grande repositório de dados, geralmente um banco de dados relacional, que apresenta uma visão consistente dos dados provenientes das fontes de dados, e finalmente os usuários, que interagem com o sistema através de uma interface.

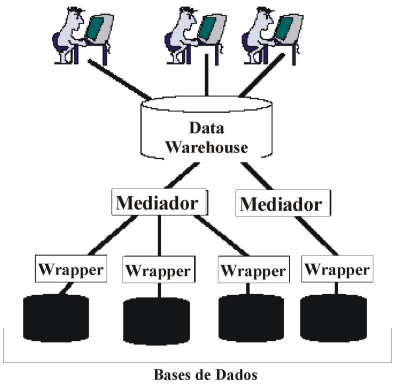


Figura 1- Estrutura de um Data Warehouse (Critchlow; Musik; Slezak, 2000).

Segundo (Critchlow; Musik; Slezak, 2000), o desafio para a criação de um Data Warehouse para o ambiente da bioinformática está no fato de que deve-se desenvolver uma infra estrutura flexível o bastante para controlar a natureza dinâmica do domínio, pois fontes de dados para aplicações científicas são extremamente dinâmicas. Sempre que uma fonte de dados muda seus dados, o Wrapper e o mediador devem ser atualizados para que estas atualizações sejam espelhadas no Data Warehouse. Isto se torna um grande desafio, pois deve-se manter um Data Warehouse extremamente funcional, mesmo integrando várias fontes de dados que sofram mudanças constantemente. Uma terceira abordagem seria a utilização de bases de dados XML, pois recentemente alguns esforços estão sendo dedicados para a construção de documentos de definição XML (DTD) que permitem conversões entre bancos de dados que se utilizam de diferentes tecnologias de XML (SHUI, 2002). Existem muitos projetos em andamento que provêem bibliotecas de repositório de dados em muitas linguagens, como Java e C/C++. Porém, muitos destes projetos estão preocupados em como analisar gramaticalmente os dados XML, ao invés de estabelecer um banco de dados XML bem formulado, capaz de integrar bancos de dados diferentes, criando assim um repositório de informação biológica. A grande preocupação neste caso é de como integrar estas diversas bases de dados XML, visto que os dados biológicos não possuem uma estrutura padrão, pois os dados podem variar de tipo de uma base para outra. É importante salientar que neste caso, por se tratar de um tema novo no Brasil, a maioria do material encontrado para a realização deste artigo foram artigos científicos escritos em inglês, com poucos materiais específicos do tema deste relatório em português.

**RESULTADOS E DISCUSSÃO:** Através da análise do tópico acima relacionado, podemos inferir que encontrar um banco de dados que suporte tudo o que é gerado em projetos de pesquisa com genes e outros dados biológicos através da bioinformática é sem sombra de dúvida, complexo, pois o banco de dados deverá se adequar ao domínio da aplicação. Tecnologias de Data Warehouse se mostram promissoras na tentativa de integrar bases de dados heterogêneas distribuídas geograficamente, mas somente isto não ajudará no desenvolvimento de um padrão específico para dados biológicos, pois é através desta padronização que poderá se trabalhar com várias bases de dados, sem haver nenhuma perda de performance. As tecnologias de XML (SGBD XML) para bioinformática se mostram promissoras, principalmente no que diz respeito à integração de dados biológicos provenientes de bases de dados heterogêneas distribuídas geograficamente que armazenem os dados biológicos como documentos XML, mas tais tecnologias ainda estão no início de seu desenvolvimento, o que faz com que tecnologias de XML entrem no mercado da bioinformática daqui a alguns anos (SHUI, 2001). Muitas empresas e institutos vêm pesquisando a área de bancos de dados para bioinformática, mas sem conseguir chegar a um padrão a ser adotado para todos os bancos de dados utilizados para o armazenamento e busca de dados biológicos, pois estas empresas e institutos tentam somente adequar o domínio de suas aplicações aos bancos de dados já existentes no mercado, tentando solucionar suas necessidades imediatas. Até o presente momento, não existe um esforço maior para se tentar encontrar um padrão para ser adotados na elaboração e construção de novos bancos de dados com objetivo específico de atender às necessidades da bioinformática, o que impossibilita de certa forma, a troca de informações sobre projetos que envolvam dados biológicos pelos mais diversos centros de pesquisa espalhados geograficamente.

**CONCLUSÃO**: Soluções como as da Oracle (ORACLE CORPORATION, 1999) oferecem um melhor "suporte" para a realização de tais pesquisas, mas tal solução é somente o início de pesquisas que devam surgir nos próximos anos, a fim de fazer com que dados biológicos possuam uma facilidade de tratamento e busca tal qual existe para dados comuns, como NUMBER, DATE e CHAR. A utilização de data warehouse é uma solução interessante quando falamos em interligar bases biológicas de várias entidades, mas esta solução não pode ser aplicada separadamente, sem utilizarmos formas de otimização de pesquisas e tratamento dos dados biológicos, pois se somente a integração destes bancos não nos garante que as buscas por informações referentes a dados biológicos vá se dar de uma forma eficaz. A utilização de tecnologias XML é muito interessante, mas esta tecnologia ainda não está bem formulada para o domínio de dados biológicos, sendo implementada e testada aos poucos, principalmente se apoiando nos conceitos oferecidos pela W3C.

## REFERÊNCIAS BIBLIOGRÁFICAS


BANERJEE, S. **A Database Platform for Bioinformatics**. Redwood Shores: Oracle Corporation, 2000.

CRITCHLOW, T.; MUSICK, R.; SLEZAK, T. **An Overview of Bioinformatics Research at Lawrence Livermore National Laboratory**. Califórnia: Department of Energy by University of California Lawrence Livermore National Laboratory, 2000.

LENGAUER, T. **Computational Biology at the Beginning of the Post-genomic Era**. Berlin: University of Bonn, 2001.

ORACLE CORPORATION. **Oracle8i Data Cartridge Developer's Guide: Release 8.1.5 (Part No. A68002-01)**. Redwood Shores: Oracle Corporation, 1999.

SHUI, W. M. **Utilizing Multiple Bioinformatic Information Sources: An XML Database pproach 2001 Bioinformatics Honours Thesis**. Sydney: University of New South Wales, 2001.

WIECZOREK, E. M. **Caminhos e Tendências do uso de Banco de Dados em Bioinformática**. Prática de Sistemas de Informação I - Estágio. Curso de Sistemas de Informação, Centro Universitário Luterano de Palmas - CEULP/ULBRA, 2002.