CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS
COMUNIDADE EVANGÉLICA LUTERANA "SÃO PAULO"
Credenciado pelo Decreto de 06/07/2000 - D.O.U. nº 130 de 07/07/2000

Emilio Mario Wieczorek

Caminhos e Tendências do Uso de Bancos de Dados em Bioinformática

Palmas

2002

ULBRA
CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS
COMUNIDADE EVANGÉLICA LUTERANA "SÃO PAULO"
Credenciado pelo Decreto de 06/07/2000 - D.O.U. nº 130 de 07/07/2000

Emilio Mario Wieczorek

Caminhos e Tendências do Uso de Bancos de Dados em Bioinformática

"Relatório apresentado como requisito parcial da disciplina Prática em Sistemas de Informação I do Curso de Sistemas de Informação, coordenado pelo Prof. Eduardo Leal."

Palmas

2002

EMILIO MARIO WIECZOREK

CAMINHOS E TENDÊNCIAS DO USO DE BANCOS DE DADOS EM BIOINFORMÁTICA

"Relatório apresentado como requisito parcial da disciplina Prática em Sistemas de Informação I do Curso de Sistemas de Informação, coordenado pelo Prof. Eduardo Leal."

Aprovada em 03/12/2002

BANCA EXAMINADORA

_____

Prof. Eduardo Leal
Centro Universitário Luterano de Palmas

_____

Profª. Deise de Brum Saccol
Centro Universitário Luterano de Palmas

_____

Profª. CRISTINA DORNELLAS FILIPAKIS
Centro Universitário Luterano de Palmas

Palmas
2002

AGRADECIMENTOS

Agradeço a Deus, por me dar apoio quando necessito; a meus pais e minha namorada, por sempre me incentivarem na realização de meus sonhos e a meu professor orientador que muito contribuiu para a realização deste relatório.

DEDICATÓRIA

Dedico este relatório de estágio a Deus e a meus familiares, que sempre estiveram comigo para me auxiliar nos momentos mais difíceis de minha vida.

RESUMO

Este relatório de estágio traça os caminhos e tendências adotadas por empresas e institutos de pesquisa na utilização de bancos de dados na área de bioinformática, descrevendo as tecnologias de Banco de Dados, Data Warehouse e XML que são utilizadas para o armazenamento, transformação e acesso a dados biológicos provenientes de projetos de pesquisa, como o Projeto Genoma Humano.

Serão demonstrados os problemas existentes na utilização destas tecnologias, descrevendo também propostas formuladas por alguns autores para solucionar tais problemas. Serão abordadas, principalmente, técnicas envolvendo bancos de dados, pois este é o principal elemento deste estudo.

Palavras-chave: caminhos, tendências, bioinformática, banco de dados.

LISTA DE ABREVIATURAS

CAD	*Computer Aided Design*
CAM	*Computer Aided Manufacturing*
SQL	*Structured Query Language – Linguagem de Consulta Estruturada*
DNA	*Ácido Desoxirribonucléico*
XML	*eXtensible Markup Language – Linguagem de marcação de dados*
RNA	*Ácido Ribonucléico*
LOB	*Large Objects*
SGBD	*Sistema de Gerenciamento de Banco de Dados*
3D	*Três dimensões*
HTML	*HiperText Markup Language*
CGI	*Common Gateway Interface*
DTD	*Document Type Definition*
OID	*Object Identifier*
XSL	*Extensible Stylesheet Language*
XSLT	*Extensible Stylesheet Language Transformations*
DOM	*Document Object Model*
API	*Aplications Programming Interface*
SAX	*Simple API for XML*
W3C	*World Wide Web Conso*
PDF	*Portable Data Format*

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

1. 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. A bioinformática é um campo interdisciplinar que une biologia e informática, e tem como objetivo desenvolver e aplicar técnicas computacionais no estudo da genética, da biologia molecular e da bioquímica (Lengauer, 2001).

A bioinformática torna-se essencial para a construção de bases de dados contendo informações sobre os genes e proteínas dos organismos vivos, para a descoberta de novos genes, e de novos medicamentos, pois é através da bioinformática que novas técnicas para o mapeamento e armazenamento das informações extraídas dos genes vem sendo estudadas e estruturadas (Banerjee, 2000).

No campo da informática, a evolução dos sistemas computacionais segue a evolução das necessidades que as aplicações por ele tratadas devem atender. Por exemplo, nos anos 60 a preocupação era o tratamento de dados que envolviam aplicações tipicamente científicas, evoluindo para as aplicações comerciais (folhas de pagamento, etc) e hoje atente a diversas áreas como CAD, CAM, aplicações médicas, e outros. Para atender essas necessidades computacionais, os bancos de dados estão em constante evolução, uma vez que devem suportar os diferentes tipos de dados que essas aplicações requerem.

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).

Outro fator que merece atenção é a expansão que vem acontecendo no setor de biotecnologia médica, fazendo com que um grande número de institutos de pesquisa públicos e privados se voltem para as áreas de biotecnologia, mais precisamente para a bioinformática, uma área relativamente nova (últimos 10 anos), tornando assim esta área necessária para a descoberta de futuras curas para doenças, como o câncer (Lengauer, 2001).

Esse estudo tem como objetivo o levantamento do uso de banco de dados no domínio da bioinformática, a fim de identificar os caminhos e tendências adotados, para que num futuro breve, consigamos elaborar um padrão a ser utilizado por estes bancos de dados, facilitando assim, a integração de vários institutos de pesquisa que trabalhem com dados biológicos e moleculares. Esse estudo permite que novas descobertas acerca do genoma, principalmente do genoma humano, sejam realizadas mais rapidamente e com maior eficácia, pois a falta de um padrão tanto para a elaboração e construção quanto para o armazenamento e acesso aos dados biológicos e moleculares dificulta o tratamento dos dados provenientes de pesquisas envolvendo o DNA, além de não se conseguir uma integração maior entre os vários institutos de pesquisa que trabalham com estes dados.

Este trabalho está organizado da seguinte forma: o capítulo 2 apresenta a motivação para a realização deste relatório de estágio, o capítulo 3 apresenta a revisão bibliográfica, o capítulo 4 apresenta os materiais e os métodos utilizados no desenvolvimento deste relatório de estágio, o capítulo 5 apresenta os resultados e discussões, o capítulo 6 apresenta a conclusão e o capítulo 7 é destinado às referências bibliográficas.

2. MOTIVAÇÃO

A motivação para a realização deste relatório de estágio é a crescente expansão do mercado farmacêutico em escala mundial, além do crescente interesse de instituições brasileiras em estudar e desenvolver soluções em bancos de dados para o domínio da bioinformática.

Outro fator importante é que devido este assunto ser muito recente, principalmente no Brasil, ele se torna um desafio a mais para se tentar aplicar técnicas de banco de dados que se adequem ao domínio da bioinformática, além de poder possibilitar a descoberta de informações hoje ocultas no genoma humano e de outros seres vivos.

3. REVISÃO DE LITERATURA

Neste capítulo serão descritos os principais esforços que vem sendo feitos para encontrar o melhor meio de prover o armazenamento de dados biológicos, como DNA, proteínas e genoma. Atualmente, não se encontra na literatura material disponível que trate os diversos aspectos relacionados a bioinformática. Em (Human Genome Program, 1992) são demonstrados os requisitos iniciais dos projetos de genoma; em (Alander, 1995) têm-se um índice bibliográfico com algumas informações de projetos envolvendo genética; em (Langdom, 1996) são abordados estruturas de dados para projetos de genoma; em (Biotech, 1998) tem-se informações sobre bioinformática, projetos envolvendo dados biológicos desenvolvidos por instituições espalhadas pelo mundo e algumas informações sobre tecnologias usuais em bioinformática; em (Bomtempi, 1999) o autor apresenta uma série de contribuições da informática para as áreas biológicas em geral; em (Júnior, Denipote, 1999) é abordado o projeto genoma de maneira geral; em (Leser, 1999) é abordado um projeto global para armazenar informações biológicas; em (Banerjee, 2000) têm-se a proposta apresentada pelo autor, além de algumas informações sobre o banco de dados 8*i* da Oracle; em (Basan, 2000) têm-se informações sobre ferramentas para sequenciamento e anotação de genes, incluindo algumas informações sobre bancos de dados biológicos; em (Costa, 2000) é abordado o trabalho desenvolvido por institutos de pesquisa e empresas no mapeamento do genoma humano; em (Critchlow; Musik; Slezak, 2000) é abordado a integração de bases de dados através de data warehouses; em (Oracle Corporation, 2000) têm-se a infra-estrutura do banco de dados Oracle 8*i* referente a dados biológicos; em (Pessini, 2000) são abordados os progressos de tecnologias de informática no âmbito da saúde mundial; em (Schroeder, 2000) são mostrados os recursos existentes no banco de dados desenvolvido pelo Centro Nacional de Biotecnologia; em (Tachinardi, 2000) são mostradas tendências que vêem a surgir em tecnologias voltada para a área de saúde; em (Lengauer, 2001) é feito um comparativo entre a era pós-genômica e a era pré-genômica, estabelecendo quais são as tecnologias envolvidas na área de bioinformática que merecerão

destaque nos próximos anos; em (Shui, 2001) é abordada a utilização de tecnologias de XML para o armazenamento e integração de dados biológicos; em (Ministério da Ciência e Tecnologia; Centro de Referência em Informação Ambiental, 2001) têm-se a tentativa de estabelecimento de padrões para o armazenamento de dados biológicos no banco Postgree SQL; em (Félix, 2002) são demonstradas algumas técnicas para a utilização de dados provenientes de projetos de genoma; em (Fugita, 2002) são demonstrados os passos realizados para efetuar a Anotação de Genes Associados com o Controle da Proliferação Celular e Origem de Tumores e em (Genoma e Genética, 2002) é feita uma análise dos desenvolvimentos ocorridos através do projeto genoma.

3.1 Bioinformática

A bioinformática, como demonstrado pela figura 1, é um campo interdisciplinar que une biologia e informática, e tem como objetivo desenvolver e aplicar técnicas computacionais no estudo da genética, da biologia molecular e da bioquímica (Lengauer, 2001).

Este novo campo apresenta um dos principais desafios deste século, pois representa uma grande área que está aberta para o desenvolvimento de novas pesquisas e de novas tecnologias, visto que tende a atender pesquisas envolvendo sistemas biológicos, organismos e células. A bioinformática é, ao mesmo tempo, uma solução para o desenvolvimento de aplicações imediatas e uma base para um sucesso científico e econômico futuro. (Lengauer, 2001).

A bioinformática surge ao término da era pré-genômica, era que foi caracterizada pelo esforço em mapear o genoma humano. A era pós-genômica se concentra em descobrir novas informações “escondidas” dentro deste “mapa” do genoma humano (lengauer, 2001).

A bioinformática, como mostrado pela tabela 1, se relaciona com várias áreas da informática, utilizando o melhor de cada área para solucionar os desafios apresentados no tratamento de dados provenientes de projetos biológicos, como o Projeto Genoma Humano (Biotech, 1996).

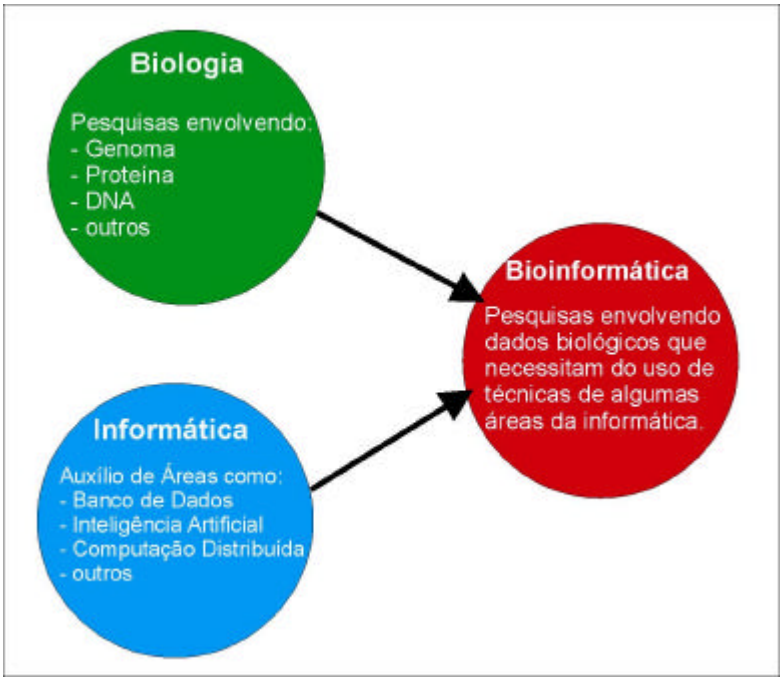

Figura 1- Representação Gráfica da área de Bioinformática.

A tabela 1 mostra algumas das áreas da informática que possuem um relacionamento com a bioinformática:

Tabela 1- Áreas de informática que possuem relacionamento com bioinformática (Biotech, 1996).

Área
Inteligência Artificial
Redes Neurais
Computação evolutiva, Algoritmos Genéticos e Programação Genética
Sistemas especialistas
Aprendizagem de máquina
Simulação de Sistemas
Estatísticas e Cálculos de Probabilidade

A bioinformática surge em um primeiro momento, devido à falta de mecanismos para o armazenamento de informações provenientes de projetos de pesquisa envolvendo o genoma, como o Projeto Genoma Humano, pois a grande quantidade de informações gerada deveria ser armazenado de forma cuidadosa, com organização, e possuindo indexação sobre as informações provenientes de sucessões genômicas (Critchlow; Musik; Slezak, 2000).

As tarefas mais efetuadas na bioinformática estão relacionadas com a criação e a manutenção de bancos de dados que contenham informações biológicas, envolvendo a análise de sucessões biológicas como (Shui, 2001):

- Encontrar genes nas sucessões de DNA pertencentes a vários organismos;
- Desenvolvimento de métodos capazes de predizer a estrutura e / ou a funcionalidade de proteínas descobertas em sucessões de RNA e DNA;
- Encontrar sucessões de proteína, agrupando-as em famílias de sucessões relacionadas, para que possam ser desenvolvidos modelos de proteínas; e
- Alinhamento similar de proteínas e elaboração de árvores filogenéticas geradas para examinar relações evolutivas.

Todas as características descritas acima foram de vital importância para o desenvolvimento de ferramentas que auxiliaram o desenvolvimento do Projeto Genoma.

3.1.1 Projeto Genoma

Iniciado a partir de 1990, o projeto Genoma Humano constitui-se em um esforço de 13 anos coordenado pelo Departamento Americano de Energia (Human Genome Program, 1992). O projeto originalmente foi planejado para ter uma duração de 15 anos, mas os rápidos progressos tecnológicos aceleraram as previsões para o ano de 2003. Os principais objetivos do projeto são: identificar todos os 100.000 genes humanos presentes no DNA, determinar as seqüências de 3 bilhões de pares de bases químicas, que constituem a base do DNA, armazenar estas informações em bancos de dados, desenvolver ferramentas para análise do material obtido, discutir e normatizar questões legais advindas do processo de pesquisa. (Genoma e Genética, 2002).

Segundo o Departamento Americano de Energia (Human Genome Program, 1992), a meta primária dos projetos de genoma públicos e privados é fazer uma série de mapas de diagramas descritivos de cada cromossomo humano a resoluções crescentemente melhores. Isto é feito dividindo os cromossomos em fragmentos menores que podem ser isolados, e ordenando estes fragmentos para corresponder aos locais respectivos dos cromossomos nos fragmentos. Depois que a ordenação é completada, o próximo passo é determinar a sucessão de bases A (Adenina),T (Timina), C (Citosina) e G (Guanina) em cada fragmento. Então, várias regiões dos cromossomos da seqüência serão “marcados” com sua respectiva

função. Finalmente podem ser catalogadas diferenças em sucessões entre indivíduos em um cenário global.

A figura abaixo mostra o processo pelo qual são usadas sucessões de DNA para modelar um modelo de proteína.

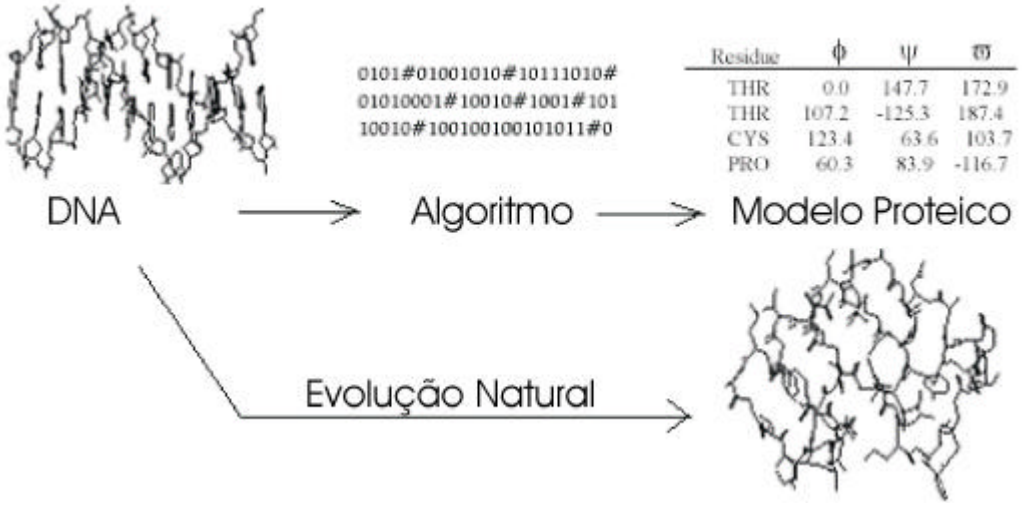

Figura 2- Processo pelo qual são usadas sucessões de DNA para modelar um modelo de proteína (Biotech, 1996).

A tabela 2 mostra alguns projetos que vem sendo desenvolvidos em Universidades e Centros de Pesquisa espalhados pelo mundo.

Tabela 2- Projetos desenvolvidos em Universidades e Centros de Pesquisas (Biotech, 1996).

Universidade / Centro de Pesquisa	Descrição
Centro Nacional Australiano de Pesquisas em Bioinformática.	Este centro contém uma lista de bases de dados e de softwares que trabalham com bioinformática, além de possuir um motor de busca interno de modo que se possa rapidamente encontrar o que se está procurando no Centro.
Departamento de Bioinformática da Universidade de Informática de Bergen, Noruega.	Desenvolvimento de bases de dados e softwares para bioinformática.
Instituto de Bioinformática da Universidade de Stanford.	Desenvolvimento de bases de dados e softwares para bioinformática, além de pesquisa envolvendo o Projeto Genoma.
Laboratório de Neuropsicológica e Bioinformática da Universidade de Tohoku, Japão.	Desenvolvimento de bases de dados e softwares para bioinformática, além de pesquisa envolvendo o Projeto Genoma.
Universidade de Campinas (UNICAMP)	Estudo de Arranjos (Microarrays) de DNA,

	construção de bases de dados e softwares para bioinformática.
Universidade Federal do Rio Grande do Sul (UFRGS)	Biologia Molecular.
Centro de Bioinformática da Universidade de Pune, Índia.	Desenvolvimento de Pesquisas na área de vírus em pestes agrícolas e em mapas do Genoma Humano.

3.1.2 Tecnologias existentes para a área de Bioinformática

Tecnologias computacionais proveram métodos capazes de armazenar e organizar informações sobre sucessão de genes em bancos de dados, permitindo assim uma análise mais rápida da sucessão de genes em questão. A evolução da computação e a alta capacidade de armazenamento têm causado o aumento de informações criadas sobre seqüências genotípicas, facilitando assim, a descoberta de novos genes. Cientistas desenvolvem novos e sofisticados algoritmos que permitem comparar sucessões usando teorias de probabilidade. Novas tecnologias como as de data warehouse permitem que estas informações sejam colocadas na Internet, facilitando a integração de mais pesquisadores, facilitando a construção de novas ferramentas que auxiliem o processo de bioinformática.

Para Banerjee (Banerjee, 2000), existem quatro tecnologias poderosas que se mostram promessas para resolver problemas intratáveis em bioinformática:

- a arquitetura de extensibilidade para armazenar uma sucessão de dados nativamente e executar estruturas de procura no banco de dados;
- tecnologias de warehousing para dados em padrões genéticos;
- tecnologias de integração de dados para habilitar questões heterogêneas por fontes biológicas distribuídas; e
- tecnologias de portal de Internet que permitem publicar informações de pesquisas na área da bioinformática, tanto para Intranets quanto para Internet.

A tabela 3 mostra alguns elementos que são utilizados para a busca de genes.

Tabela 3- Elementos utilizados para a busca de genes (Biotech, 1996).

Elementos	Descrição
Algoritmos para Reconhecimento de Padrão	São usados formulários de probabilidade para determinar se duas sucessões forem

	estatisticamente semelhantes.
Tabelas de Dados	Estas tabelas de dados contêm informações sobre sucessões iguais para vários elementos genéticos. Quanto mais informação se tem de um determinado fragmento de DNA, melhor será sua análise.
Diferenças de Taxonomia	Sucessões genotípicas de um indivíduo possuem taxonomia diferente em relação a outro indivíduo. A inclusão destas diferenças em um processo onde a velocidade de análise é alta minimizará erros.
Regras de Análise	Estas instruções de programa definem como os algoritmos são aplicados. Definem o grau de semelhança aceitado e se existem fragmentos inteiros de sucessões, considerando uma análise. Uma boa lógica no desenvolvimento do programa permite que os usuários possam ajustar estas variáveis.

3.2 Bancos de Dados

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). Cada sucessão de bases ou aminoácidos representa um gene particular ou proteína, respectivamente.

Enquanto que a maioria dos bancos de dados biológicos contém bases de DNA (nucleotídeos) e informações sobre sucessão de proteína, também há bancos de dados que incluem informação sobre taxonomia, como as características estruturais e bioquímicas de organismos (Basan, 2000).

A tabela 4 mostra alguns bancos de dados / SGBD´s que suportam dados biológicos.

Tabela 4- Bancos de Dados com capacidade de armazenar e buscar dados biológicos (Félix, 2002).

Banco de Dados/SGBD	Instituto/Empresa
NIH - Banco de dados de expressão gênica	Molecular Pharmacology of Cancer
SMD - Banco de Dados de Microarrays	Stanford University
YMGV - Visão global sobre Microarray	http://www.transcriptome.ens.fr/ymgv/

de levedura	
Oracle 8*i*/9*i* – Banco de dados comercial	Oracle Corporation

3.2.1 Problemas na Utilização de Banco de Dados

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, caráter ou datas. Poucos bancos de dados tiveram uma habilidade nativa para lidar com dados complexos, como dados multimídia, texto, dados espaciais, ou dados genéticos (sucessão de genes). A maioria destes dados fica difícil de ser controlado, como questões de achar a semelhança (em grandes cadeias de caracteres), questões sobre sucessões de gene e questões de localização de genes em cadeias de DNA (Oracle Corporation, 2000). A figura 3 demonstra as etapas para o armazenamento de segmentos (sucessões de genes) de DNA em um banco de dados.

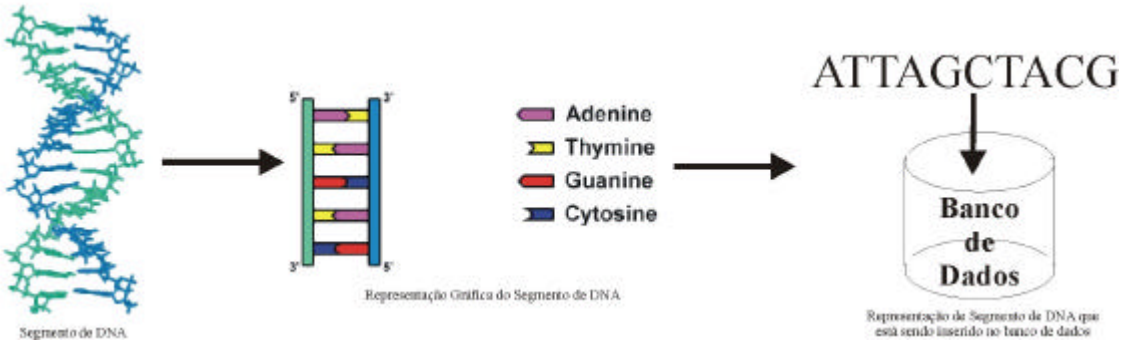

Figura 3- Etapas para o armazenamento de segmentos de DNA em um banco de dados.

Como podemos observar na figura 3, o grande problema no armazenamento de dados biológicos (de genoma) está no fato de que após ser feito o mapeamento das bases de DNA para o formato de caractere, este dado deverá ser armazenado, de forma que as pesquisas e busca de informações sobre estes dados não seja dispendiosa, e que estas retornem o que realmente se espera.

3.2.2 Propostas para a Utilização de Banco de Dados

Nas seções abaixo, abordaremos algumas propostas que se mostram interessantes para a utilização de bancos de dados no ambiente da bioinformática.

3.2.2.1 Banerjee

Segundo Banerjee (Banerjee, 2000), para o caso específico de dados biológicos (DNA, proteínas), deveria ser possível procurar por:

- Propriedades: Quais são as características (propriedades) de um segmento de DNA humano com tamanho igual ou superior a 10Kb e o que está associado a este segmento;
- Semelhança Estrutural: dado um segmento de genes qualquer (CGTAATGC), que outros segmentos existentes no banco de dados possuirão este mesmo segmento, tanto para este organismo quanto para outros organismos? A "operação de possuir" deve encontrar somente segmentos que possuem em algum ponto de sua extensão o segmento dado para a procura; e
- Local: dado um fragmento de DNA qualquer (CGTAATGC), qual é a seqüência de genes que o antecedem e o procedem.

A menos que bancos de dados possam tratar nativamente de dados complexos, aplicações especializadas têm que ser usadas como intermediárias para executar busca e localização de genes em fragmentos de DNA no banco de dados. Para a solução destes problemas, Banerjee (Banerjee, 2000) defende o uso de bancos de dados relacionais estendidos.

3.2.2.2 Oracle

A Oracle (Oracle Corporation, 1999) apresenta uma proposta interessante para a solução dos problemas de banco de dados em bioinformática: devem ser elaborados bancos de dados que sejam capazes de controlar tipos complexos, de modo a conseguir suprir as necessidades do domínio da aplicação, além de prover apoio a qualquer tipo de dado definido pelo usuário, ou seja, um 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 estensibilidade, fazendo assim uma ordenação inteligente dos predicados em questão, envolvendo tipos de dados definidos pelo usuário.

Criação de Tipos Definidos pelo Usuário

O sistema de tipos do ORACLE 8*i*/9*i* provê uma interface baseada em SQL para definir tipos. Estes tipos podem ser implementados em Java, C/C++ ou PL/SQL. O SGBD provê os serviços de infra-estrutura de baixo nível que são necessários para a criação automática destes novos tipos. Estes novos tipos podem ser objetos.

Um dado tipo objeto é diferente de tipos SQL nativos tais como tipos numéricos (NUMBER), literais (VARCHAR) ou data (DATE). Estes novos tipos geralmente são utilizados para estender as capacidades nativas do SGBD. Estes tipos podem ser utilizados para que possam ser feitos modelos que melhor representem o domínio do sistema, melhorando a "visualização" de dados do mundo real no SGDB. Além disso, ainda existe a possibilidade da utilização de outros dados pré-definidos pelo ORACLE 8i/9*i* para o armazenamento de dados "grandes", como o tipo LOB. Tipos objeto podem possuir métodos para acessar e manipular os atributos de objetos, e estes métodos podem ser invocados de dentro do SGBD.

Criação de Operadores Definidos pelo Usuário

Tipicamente, bancos de dados provêem um jogo de operadores pré-definidos para operar em tipos de dados embutidos. Podem ser relacionados os operadores matemáticos (+, -, *, /), de comparação (=,>, <), lógica booleana (NOT, AND, OR), comparação de strings (LIKE) e assim por diante. Para que se tenham operadores definidos pelo usuário, a Oracle (Oracle Corporation, 1999) acrescentou a seus bancos de dados (Oracle 8*i*/9*i*) a capacidade para definir operadores de domínios específicos, ou seja, se torna possível definir um operador para comparar sucessões genômicas. A implementação do operador é deixada ao usuário, este podendo escolher as funções, os tipos de métodos, pacotes, rotinas de bibliotecas externas e assim por diante. Pode-se ainda, serem invocados os operadores definidos pelo usuário em qualquer lugar, estes podendo ser usados como operadores embutidos, isto é, onde quer que aconteçam nas expressões. Os operadores definidos pelo usuário podem ainda ser usados em um comando SELECT, na condição de uma cláusula WHERE, na cláusula ORDER BY, e na cláusula GROUP BY. Depois que um usuário define um novo operador, este pode ser usado em comandos SQL juntamente com qualquer outro operador embutido.

Por exemplo: o usuário define um novo operador CONTEM () que possui um FRAGMENTO de DNA decodificado de uma sucessão particular, retornando TRUE se o fragmento contiver a sucessão especificada. Esta pesquisa poderá ser escrita da forma como mostra a figura abaixo:

```
SELECT ID FROM TABELADNA
WHERE CONTEM(FRAGMENTO'GCCATAGACTACA');
```

Figura 4- Pesquisa SQL representando operadores definidos pelo usuário.

Esta habilidade para aumentar a semântica dos operadores de domínio específico é um serviço oferecido pelo banco de dados.

Indexação Extensível

Bancos de dados apóiam-se em alguns métodos de acesso padrão a dados, como a utilização de árvores B+ e tabelas Hash. Como dados biológicos são dados complexos, surge então a necessidade de indexar tais dados utilizando técnicas de indexação específicas para o domínio em questão.

Para tipos de dados simples como tipos numéricos, literais e data, a indexação destes dados pode ser controlada facilmente pelo banco de dados. Para dados de sucessão de genes (dados biológicos/genômicos), são necessários índices especiais que possam executar comparações estruturais 3D, semelhança entre cadeias de DNA, e buscas sobre outros dados complexos.

O Framework para desenvolver novos tipos de índice está baseado no conceito de cooperação entre o usuário que irá desenvolver o novo tipo de índice e o banco de dados que irá dar suporte para a utilização deste novo índice criado para o controle de tipos complexos como dados genéticos ou espaciais. Neste caso, o usuário é responsável por definir a estrutura do índice, enquanto que o banco de dados o utiliza para realizar as transações com os dados genéticos. A estrutura de índice criada pode ser armazenada tanto no banco de dados como em um arquivo no sistema operacional, sendo que a melhor forma de armazenamento é no próprio banco de dados.

Para dar suporte a esta necessidade, o ORACLE 8*i*/9*i* apresenta o conceito de um IndexType, cujo propósito é habilitar procura e recuperação de dados em domínios

complexos como domínios de bioinformática de forma eficiente. Com tal funcionalidade, o usuário pode:

- Definir a estrutura de um índice que será utilizado em um determinado domínio como um Indextype novo;
- Armazenar os tipos de dados criados no próprio banco de dados ou em arquivo no sistema operacional; e
- Administrar e utilizar os dados de índice para avaliar consultas.

Na ausência de índices de domínio definidos pelo usuário, diversas aplicações mantêm em memória separada índices para dados complexos armazenados em arquivos. Para tal feito, uma quantia considerável de código e esforço é requerida, pois manter a consistência entre os índices externos e os dados do banco de dados que se relacionam com estes índices não é uma tarefa muito fácil.

Otimizador de Extensibilidade

Um otimizador típico gera um plano de execução para uma instrução SQL. Considerando uma instrução SELECT, temos que a execução planeja tal instrução, incluindo um método de acesso a cada tabela na cláusula FROM, ordenando estas tabelas de acordo com a melhor forma (forma menos dispendiosa, mais rápida) para realizar a execução da cláusula. Métodos de acesso definidos pelo sistema incluem índices, clusters hash e scaneamento de tabelas. O otimizador escolhe um plano gerado através de um jogo de ordenação e permutação, computando o custo de cada consulta, selecionando assim a consulta com mais baixo custo.

Para cada tabela existente na consulta, o otimizador calcula o custo de cada método de acesso. Bancos de dados colecionam e mantêm estatísticas sobre os dados em tabelas, como número de valores distintos, histogramas de distribuição e assim por diante, o que ajuda o otimizador a realizar seus cálculos para encontra a melhor instrução SQL.

Sempre que um índice de domínio é analisado, uma ligação é feita com uma coleção de estatísticas definidas pelo usuário. A representação e o significado destas estatísticas colecionadas pelo usuário não é de conhecimento do banco de dados, mas será utilizada pelo usuário calculando o custo ou a seletividade de uma operação de domínio. A seletividade de um predicado ou cláusula que utiliza uma tabela escolhida para formar a consulta SQL é usada para determinar a otimização desta consulta. Assim, se fôssemos

elaborar um índice de domínio para sucessões de genes e implementar um operador CONTEM () baseado neste índice, seria necessário também que fosse especificado a seletividade do operador. Depois disto, um usuário executa uma consulta da forma mostrada pela figura 5, fazendo com que um plano de execução seja gerado para determinar se o operador CONTEM deverá ser aplicado antes do operador > ou vice versa:

```
SELECT * FROM TABELADNA
WHERE CONTEM(FRAGMENTO'GCCATAGACTACA')
AND ID > 100;
```

Figura 5- Pesquisa SQL representando o otimizador de extensibilidade.

O otimizador também calcula o custo de vários caminhos de acesso escolhendo uma instrução SQL ótima. Este otimizador de extensibilidade já é implementado pelos bancos Oracle 8*i*/9*i*.

3.2.3 O Uso de Data Warehouse's para a Integração de Bases de Dados Biológicas

Para o acesso aos dados gerados por projetos de bioinformática (dados biológicos, como DNA, Proteínas), o Lawrence Livermore National Laboratory (Critchlow; Musik; Slezak, 2000) possui um projeto para a criação de um Data Warehouse (chamado de DataFoundry) para o ambiente de bioinformática (dados biológicos). O projeto começou a ser desenvolvido em outubro de 1996 e sua tarefa inicial era desenvolver uma infra-estrutura que permitiria criar e manter uma visão consistente de várias fontes de dados autônomas.

Uma outra 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 6, são 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.

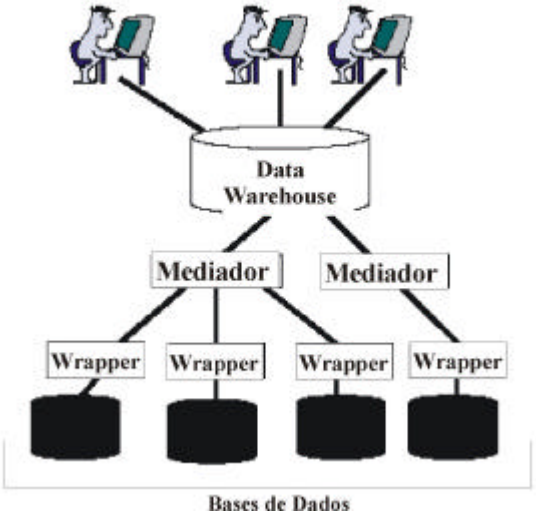

Figura 6- Estrutura de um Data Warehouse (Critchlow; Musik; Slezak, 2000).

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.

A infra-estrutura de meta dados do DataFoundry (Critchlow; Musik; Slezak, 2000) contém um gerador de mediador, um programa que automaticamente gera um mediador que usa uma coleção de meta dados declarativos, definindo uma biblioteca de classes que pode ser usada pelo wrapper para representar dados obtidos da fonte de dados. Isto simplifica a integração (adição) de novas fontes de dados, pois o administrador somente definirá o conjunto de meta dados apropriados e escreverá um wrapper que usará tais classes resultantes, ao invés de ter de escrever o wrapper e o mediador. Também irá simplificar a manutenção da Data Warehouse, pois é significamente mais fácil atualizar o conjunto de meta dados do que atualizar o mediador. O DataFoundry proverá acesso para os usuários através de interfaces desenvolvidas basicamente em HTML e Scripts CGI,

podendo esta interface ser desenvolvida também em uma linguagem de programação da escolha do laboratório/usuário, como PERL, C/C++, e outras.

3.4 Tecnologias de XML para a Bioinformática

Bancos de dados biológicos provaram ser úteis para o armazenamento de dados biológicos (genoma, DNA, proteínas), especialmente para a análise de dados não-trabalhados. Ferramentas computacionais para a identificação de sucessão, análise estrutural e visualização de cadeias de DNA foram elaboradas para acessar estes bancos de dados. Isto torna difícil a integração de dados de diferentes fontes. Para se tentar solucionar este problema, pode-se utilizar recursos que integram bancos de dados para bioinformática (biológicos) diferentes através da utilização de documentos XML (Shui, 2001).

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. Vários modelos de dados propostos para dados semi-estruturais são semelhantes, com variações secundárias. Estes modelos modelam uma DTD como um gráfico rotulado, onde cada nó do elemento XML possui um identificador associado (OID), permitindo assim indexar mais rapidamente uma recuperação de dados do documento (Shui, 2001).

Cada nó do documento XML será formado por várias “folhas”, onde cada folhar possuirá um valor atômico. Foram propostos várias linguagens de consulta para estes dados semi-estruturados, onde a característica comum é a de descrever o conteúdo abordado dentro de um determinado nó do documento XML. Tecnologias como XSL, XSLT, DOM API para XML (SAX) foram desenvolvidos para auxiliar o controle de visualização de documentos XML. Estas tecnologias permitem que documentos XML

possam ser convertidos em documentos HTML, post-script, PDF e outros formatos de documentos, podendo estes serem utilizados para vários propósitos (Shui, 2001).

3.4.1 Propostas de Utilização de XML para Bancos de Dados Biológicos

Aqui é apresentada uma solução proposta por William M. Shui (SHUI, 2000), onde é abordado o desenvolvimento de um sistema de banco de dados XML para o tratamento de dados biológicos. O modelo proposto está baseado em um SGBD XML e aborda a utilização de fontes de dados biológicos diferentes, integrando funcionalidades variadas de bioinformática com um sistema de bancos de dados. Este modelo permite a provisão de funções de bioinformática através da interface do SGBD.

O sistema é projetado como um módulo separado do SGBD XML, possuindo sua própria API que permite que o sistema de banco de dados o utilize através de um plug-in. Esta API provê acesso a uma coleção de dados biológicos através de ferramentas de análise que filtram e procuram por dados durante a execução de uma pesquisa XML.

A API também traz algumas informações estatísticas como o tempo médio gasto para se efetuar uma pesquisa, o valor das variáveis utilizadas na pesquisa e outras informações que sejam necessárias através de funções bio-analíticas. Isto permite realizar a otimização de eventuais consultas que venham a ser feitas em cima de dados biológicos.

A figura abaixo representa a interpretação do modelo atual de dados biológicos para análise.

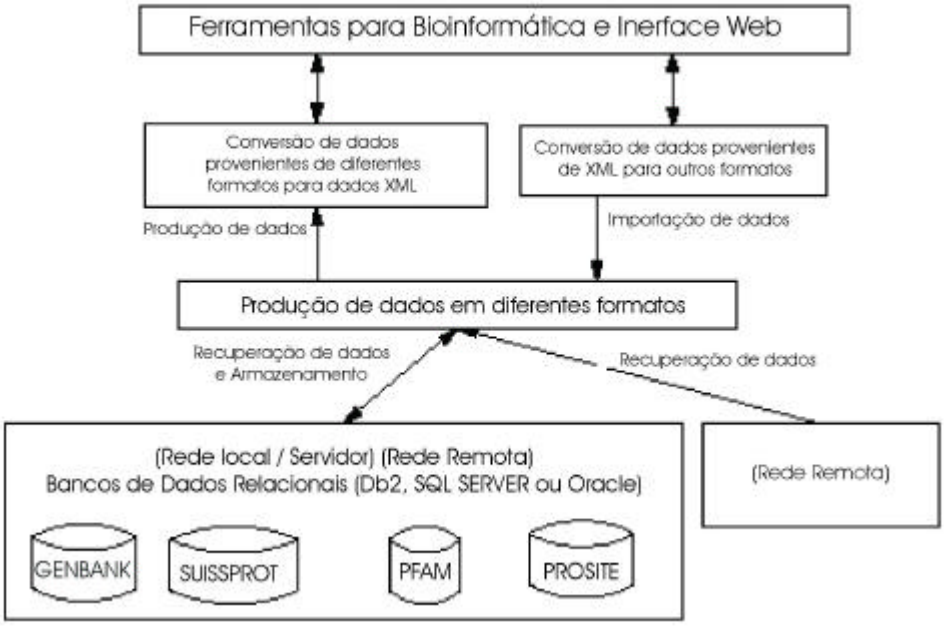

Figura 7- Representação da interpretação do modelo atual de dados biológicos para análise (Shui, 2001).

4. MATERIAIS E MÉTODOS

Para a realização deste relatório de estágio, foram feitas 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, discussões com o grupo de estudos da área de banco de dados do Centro Universitário Luterano de Palmas, além de reuniões semanais com o professor orientador.

É importante salientar que neste caso, por se tratar de um tema novo no Brasil, a maioria do material encontrado para a realização deste relatório de estágio foram artigos científicos escritos em inglês, com poucos materiais específicos do tema deste relatório em português.

Tentou-se também um estabelecimento de contato com outras universidades brasileiras como a PUC-Rio, a UNICAMP e a UFRS, que são grandes nomes em pesquisas envolvendo bioinformática, mas não houve um interesse destas universidades em ceder algum ou parte de algum material sobre pesquisas envolvendo bancos de dados para o armazenamento de dados biológicos (de bioinformática).

Apesar dos contratempos encontrados devido a falta de material adequado em nosso idioma, esse relatório apresenta-se com uma boa bibliografia, a fim de complementar certos aspectos teóricos discutidos neste relatório.

5. RESULTADOS E DISCUSSÕES

O início deste estudo serviu para a elaboração do artigo “Caminhos e Tendências do Uso de Banco de Dados em Bioinformática”, apresentado no IV Encontro de Estudantes de Informática, II Encontro de Informática do Tocantins e IV Escola de Informática Norte da Sociedade Brasileira da Computação, que aconteceu em outubro de 2002, no auditório do Centro Universitário Luterano de Palmas - TO.

Um fator relevante que foi verificado é a falta de material que trate do assunto aqui levantado (bancos de dados para bioinformática), o que foi, sem dúvida, o maior problema para a realização deste relatório, pois se trata de um relatório onde o principal objetivo é o levantamento bibliográfico do que está acontecendo na área de banco de dados para bioinformática.

Outro fator verificado é a falta de integração dos centros de pesquisa brasileiros, pois como o assunto é uma novidade para a maioria destes centros, os resultados obtidos com experiências com banco de dados biológicos estão sendo, de certa forma, mantidos em sigilo, o que é um ponto contra quando novas pesquisas que visam auxiliar este desenvolvimento começam a ser desenvolvidas, pois a falta de informação e principalmente a falta de integração são um dos grandes desafios a ser superados.

Através da análise dos tópicos relacionados, 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, facilitando então o andamento de projetos que envolvam dados biológicos distribuídos em diversos centros de pesquisa.

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.

Além de não existir um esforço para a padronização dos bancos de dados para bioinformática, também não existe uma tentativa significativa para que seja feita a padronização do esquema de pesquisa SQL para os dados biológicos (DNA, proteínas, etc).

Atualmente, a Oracle, uma das grandes empresas do ramo de soluções para banco de dados, iniciou suas pesquisas na área de bioinformática, objetivando atender as necessidades dos institutos e empresas particulares que trabalhem com dados biológicos, sendo assim, uma das primeiras empresas a tentar padronizar o “esquema” de banco de dados para bioinformática, pois até então os esforços para a área eram escassos e individuais, não possibilitando assim a construção de um padrão a ser adotado para o armazenamento e tratamento de dados biológicos.

Além da Oracle, outros grandes institutos e centros de pesquisa, tanto de computação quanto de biologia molecular, de universidades, órgãos do governo de vários países (inclusive o Brasil) e empresas privadas (principalmente farmacêuticas), espalhadas pelo mundo estão tentando entrar em comum acordo para elaborar um padrão que venha a

ser adotado por todos os bancos de dados que venham a trabalhar com dados biológicos, a fim de acabar com o problema causado pela falta de padronização.

Com a adoção de um padrão específico para os bancos de dados biológicos (para bioinformática), a troca de informações dentre os mais variados institutos de pesquisa será amplamente melhorada, facilitando assim a descoberta de novos medicamentos e novas curas para doenças consideráveis intratáveis como o câncer (Lengauer, 2001).

Com o recente interesse da indústria farmacêutica em estudar o genoma para desenvolver medicamentos melhores e mais eficazes, as tentativas para a padronização dos bancos de dados biológicos (de bioinformática) se darão de uma maneira mais rápida, pois os valores investidos na área de bancos de dados para bioinformática pelas por empresas farmacêuticas fará com que diversos centros de pesquisa tentem desenvolver um padrão a ser adotado (Banerjee, 2000).

6. CONCLUSÕES

Como foi mostrado neste relatório de estágio, muitas instituições e empresas privadas de vários países vêm tentando desenvolver soluções de bancos de dados que venham a auxiliar na pesquisa de dados genéticos (DNA, proteínas), a fim de que o desenvolvimento de novas pesquisas envolvendo dados biológicos se dê de uma maneira mais rápida e eficiente.

O interessante neste caso é que poucos ou nenhum dos bancos de dados existentes no mercado para a bioinformática consegue retornar de forma eficaz os resultados esperados nas consultas efetuadas sobre uma base de dados biológica. Soluções como as da Oracle 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.

Devido o que foi visto neste relatório, podemos concluir que o desenvolvimento de soluções para bancos de dados no domínio da bioinformática está em crescente expansão, o que fará com que pesquisas melhores no âmbito da bioinformática venham a acontecer, facilitando assim a descoberta de novos dados biológicos/genômicos que possibilitem a descoberta de novos medicamentos.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ALANDER, Jarmo T. **An Indexed Bibliografy of Genetic Programing**. Vaasa: University of, 1995.

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

BASAN, Ana Lúcia C. **Ferramentas de Bioinformática para Sequenciamento e Anotação**. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2000.

BIOTECH. **Bioinformatics**. Texas: jun. 1998. Disponível em <http://biotech.icmb.utexas.edu/pages/bioinform/BIintro.html>. Acesso em 22/08/2002.

BOMTEMPI, Nelson. Contribuições da Ciência Biológica no século XX e sua projeção para o século XXI. **O Mundo da Saúde**, São Paulo, ano 23, n. 6, p. 399-405, nov. a dez. 1999.

COSTA, Viviane Rita. Genoma Decifrado, Trabalho Dobrado. **Ciência Hoje**, São Paulo, v. 28, n. 166, p. 22-35, nov. 2000.

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

FÉLIX, Juliana M. Genoma Funcional. Biotecnologia, **Ciência & Desenvolvimento**, São Paulo, n. 24, p. 60-67, jan. a fev. 2002.

FUGITA, André. **Anotação de Genes Associados com o Controle da Proliferação Celular e Origem de Tumores**. São Paulo: Universidade de São Paulo, ago. 2002. Disponível em <http://www.linux.ime.usp.br/~fugita/mac499>. Acesso em 18/08/2002.

GENOMA E GENÉTICA. São Paulo: ClickZero, ago. 2002. Disponível em <http://www.geocities.com/clickzero/genome.htm>. Acesso em 15/08/2002.

HUMAN GENOME PROGRAM. **Primer on MolecularGenetics**. Washington: Department of Energy, ago.1992. Disponível em <http://www.ornl.gov/hgmis/publicat/primer>. Acesso em 02/08/2002.

JÚNIOR, Hermes P. Morais; DENIPOTE, Juliana Gouveia. **Projeto Genoma**. São Paulo: Universidade Estadual Paulista, 1999.

LANGDON, W. B. **Data Structures and Genetic Programing**. Londres: University College London, 1996.

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

LESER, Ulf. **Designing a Global Information Resource for Molecular Biology (Short Paper)**. Berlin: Technische Universität Berlin, 1999.

MIT, Ministério da Ciência e Tecnologia; CRIA, Centro de Referência em Informação Ambiental. Sistemas de Informação: Estudos de Tecnologias e Padrões. Brasília, DF, 2001.

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

PESSINI, Léo. Tecnociência da Informação em Saúde. **O Mundo da Saúde**, São Paulo, ano 24, n. 3, p. 163-164, mai. a jun. 2000.

SCHROEDER, L. F. **Recursos de Banco de Dados do Centro Nacional de Biotecnologia (NCBI)**. Brasília: Centro Nacional de Biotecnologia, 2000.

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

TACHINARDI, Umberto. Tendências da Tecnologia da Informação em Saúde. **O Mundo da Saúde**, São Paulo, ano 24, n. 3, p. 165-172, mai. a jun. 2000.