# Caminhos e Tendências do uso de Banco de Dados em Bioinformática

**Emilio Mario Wieczorek, Eduardo Leal**

Curso de Sistemas de Informação – Centro Universitário Luterano de Palmas (CEULP)
Palmas, TO, Brasil

{wieczorek, eduardo}@ulbra-to.br

***Resumo.*** *Este artigo descreve os caminhos e tendências adotadas por empresas e institutos de pesquisa para a utilização de banco de dados em bioinformática, demonstrando algumas formas existentes para o armazenamento e acesso (busca) de dados (cadeias de DNA) provenientes de projetos de bioinformática, como o Projeto Genoma Humano.*

## 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.

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.

No campo de informática, a evolução dos sistemas computacionais seguem 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 cientificas, 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 área de 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).

Nosso estudo tem como objetivo o levantamento dos principais esforços computacionais realizados na área de bioinformática, enfocando o uso de banco de dados, a fim de identificar os caminhos e tendências adotados no uso dos bancos de dados em bioinformática (dados genômicos, moleculares, por exemplo), 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 genômicos e moleculares, fazendo com que novas descobertas a cerca 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 genômicos 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.

## 2 Revisão de Literatura

A aproximadamente vinte anos atrás, havia uma quantidade de dados genômicos (de genes) relativamente pequena, pois os laboratórios biológicos gastavam meses realizando experiências com pequenos fragmentos de DNA ou de proteínas. Devido a esta pequena quantidade de dados gerados os próprios biólogos administravam os resultados obtidos nestas pesquisas, geralmente armazenando estes resultados em arquivos de texto, codificados no padrão adotado pelo laboratório onde trabalhavam. Porém, com o Projeto Genoma Humano, a tecnologia avançou e a quantidade de dados gerada pelos laboratórios aumentou drasticamente, e as informações coletadas pelos experimentos utilizando genes do DNA não puderam mais ser armazenadas desta maneira "arcaica" [CRITCHLOW; MUSICK; SLEZAK, 2000], pois segundo a revista Ciência Hoje [COSTA, 2000], a molécula do DNA Humano pode conter Três bilhões de caracteres e entre eles, os 100 mil genes estimados para a espécie *Homo Sapiens*.

Para a solução destes problemas, surge a bioinformática, um campo interdisciplinar, que age como interface entre os campos científico e tecnológico. A bioinformática se caracteriza por tentar prover métodos computadorizados para interpretar os dados referentes ao sequenciamento do genoma, que gera grandes volumes de dados espalhados em várias partes do mundo, de forma a trazer novos avanços para a biologia molecular (cura de doenças).

A bioinformática representa, hoje, um dos grandes desafios para se tentar decifrar o genoma, pois ao mesmo tempo que é uma forma de se conseguir informações imediatas para os dados do genoma que vem sendo descobertos, também é a base para um sucesso científico futuro. (LENGAUER, 2001).

Com o grande volume de informação gerado pelos projetos de análise de transcriptomas (fragmentos de DNA), tem se tornado cada vez mais complexo o armazenamento, acesso e a análise dos dados. Para contornar tal dificuldade, devem ser implementados, bancos de dados, que disponibilizem, de modo confiável, os dados e

ferramentas de análise. Em muitos casos, esses bancos são abertos, o que aumenta ainda mais a aplicabilidade da pesquisa. [FÉLIX, 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.

Para se tentar solucionar os problemas relacionados com armazenamento, acesso e busca de dados, vários autores e institutos formulam propostas de como estes gigabytes de dados serão armazenados e como serão feitos o acesso e tratamento destes, para que se encontre algo de útil para a descoberta de "segredos" ainda guardados pelo corpo humano.

BANERJEE, S. [BANERJEE 2000], afirma que todos os genes humanos serão achados eventualmente, sendo desenvolvidos diagnósticos para todas as doenças hereditárias, modelos animais para pesquisa de doenças humanas serão mais facilmente desenvolvidos, e curas serão desenvolvidas para muitas doenças. Muitos destes desenvolvimentos acontecerão, não dentro de laboratórios biológicos, mas sim em plataformas computacionais de alto desempenho, com sistemas gigantescos para armazenar dados do genoma, bancos de dados para procurar pelos dados, semelhanças identificando padrões, como também integração de software para unificar as fatias de conhecimento desenvolvidas em instituições globalmente distribuídas.

Para BANERJEE, S. [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.

Um problema a ser superado quando se fala em banco de dados para bioinformática é que bancos de dados têm, de longe, 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. A figura 1 demonstra as etapas para o armazenamento de segmentos (sucessões de genes) de DNA em um banco de dados.

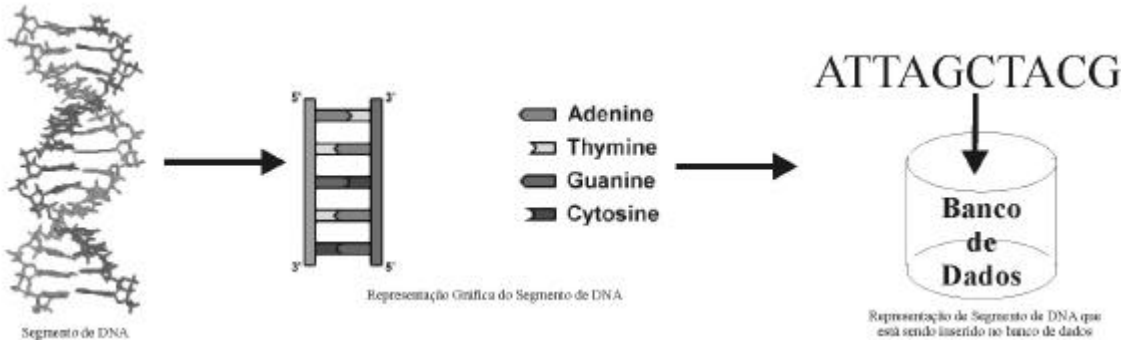


**Figura 1. Etapas realizadas para armazenar um segmento (sucessão de genes) de DNA em um banco de dados.**

Segundo BANERJEE, S. [BANERJEE 2000], para o caso específico de dados genômicos (de genoma), 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; 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.

A tabela abaixo mostra alguns bancos de dados que permitem trabalhar com sucessões genômicas, sendo a maioria de institutos e universidades, que vêm trabalhando na elaboração de bancos de dados específicos para trabalhar com expressões gênicas.

**Tabela 1. Bancos de Dados com capacidade de armazenar e buscar dados genômicos.**

| Banco de Dados | 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 de levedura | http://www.transcriptome.ens.fr/ymgv/ |
| Oracle 8*i*/9*i* – Banco de dados comercial | Oracle Corporation |

A Oracle [ORACLE CORP., 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 genômicos e otimizar a estensibilidade fazendo assim

uma ordenação inteligente dos predicados em questão, envolvendo tipos de dados 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 tenha operadores definidos pelo usuário, a Oracle [ORACLE CORP., 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 deixado 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 abaixo:

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

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

Para o acesso aos dados gerados por projetos de bioinformática (genômicos), o Lawrence Livermore National Laboratory [CRITCHLOW; MUSICK; SLEZAK, 2000] possui um projeto para a criação de um Data Warehouse (chamado de DataFoundry) para o ambiente de bioinformática (dados genômicos). 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 2, 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.

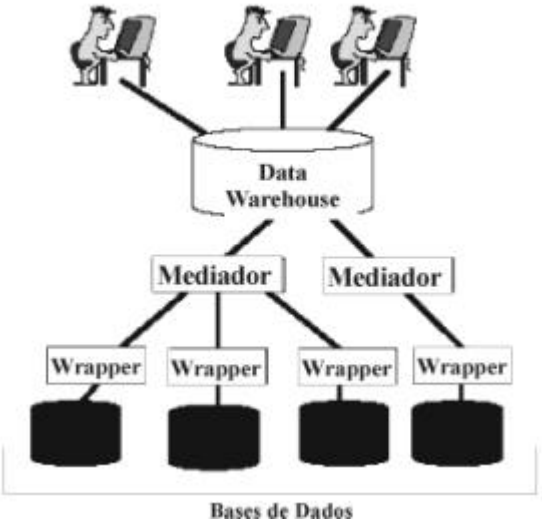


**Figura 2. Estrutura de um Data Warehouse**

Segundo [CRITCHLOW; MUSICK; 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; MUSICK; 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 podem 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.

Segundo BANERJEE, S. [BANERJEE 2000], os dados gerados pela bioinformática não serão armazenados em um mesmo banco de dados. É provável que dados de sucessão pertencentes ao genoma humano sejam colocados em banco de dados públicos ou privados. Também serão armazenados, por centenas de instituições, dados de sucessão para outros tipos de organismos. Novas pesquisas para estes dados farão com que novos dados sejam “criados” continuamente. Companhias farmacêuticas terão seus próprios dados privados, o que faz do mercado de banco de dados para bioinformática um dos melhores mercados para investimento nos próximos anos.

## 3 Trabalhos Futuros

Encontrar um banco de dados que suporte tudo o que a é gerado em projetos de pesquisa com genes através da bioinformática é sem sombra de dúvida, complexo, pois o banco de dados deverá se adequar ao domínio da aplicação. Muitas empresas e institutos vêm pesquisando a área de banco de dados para bioinformática, mas sem conseguir chegar a um padrão a ser adotado para todos os banco de dados utilizados em bioinformática, 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, não havendo 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, além de não haver um esforço mútuo para padronizar o esquema de pesquisa SQL para os dados genômicos (de genes).

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 genômicos, 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 genômicos.

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 pelos bancos de dados adotados na área de bioinformática, a fim de acabar com o problema causado pela falta de padronização, o que dificulta a troca de informações sobre os dados genômicos entre os mais variados institutos e organizações que pesquisam o genoma e que trabalham com biologia molecular.

## 4 Conclusões

Com o grande crescimento que a área de bioinformática vem apresentando, os esforços para se tentar adotar um padrão em bancos de dados voltados para a bioinformática vem aumentando consideravelmente, muitos institutos de pesquisa e empresas de soluções em banco de dados vêm desenvolvendo pesquisas na área para se tentar achar um padrão que seja utilizado em todos os bancos de dados para a bioinformática.

Esta padronização ajudará, não somente o armazenamento, acesso e busca dos dados "criados" por projetos de bioinformática, mas também irá auxiliar em muito a troca de informação sobre estes dados entre os mais diversos institutos de pesquisa espalhados pelo mundo que trabalham com este tipo de dado, facilitando assim a descoberta de curas para doenças e o tratamento preventivo das mesmas.

## 5 Referências Bibliográficas

[BANERJEE 2000], BANERJEE, S. "A Database Platform for Bioinformatics", Oracle Corporation, Redwood Shores, 2000.

[ALANDER 2000], ALANDER, J. T. “An Indexed Bibliografy of Genetic Programing”, University of Vaasa, Finlândia, 1995.

[CRITCHLOW; MUSICK; SLEZAK, 2000], CRITCHLOW, T.; MUSICK, R.; SLEZAK, T. “An Overview of Bioinformatics Research at Lawrence Livermore National Laboratory”, Department of Energy by University of California Lawrence Livermore National Laboratory, Califórnia U.S., 2000.

[LONGDON, 1996], LONGDON, W. D. “Data Structures and Genetic Programing”, University College London, Londres,1996.

[LENGAUER, 2001], LENGAUER, T. “Computational Biology at the Beginning of the Post-genomic Era”,University of Bonn, Berlin, 2001.

[LESER, 1999] LESER, U. “Designing a Global Information Resource for Molecular Biology (Short Paper)”, Technische Universität Berlin, Berlin, 1999.

[HUMAN GENOME PROGRAM, 1992], HUMAN GENOME PROGRAM, “Primer on MolecularGenetics”, Department of Energy, Washington D.C, U.S., 1992. Veja http://www.ornl.gov/hgmis/publicat/primer/intro.html acesso em 02/09/2002.

[ORACLE CORP., 1999], ORACLE CORP. “Oracle8i Data Cartridge Developer’s Guide: Release 8.1.5 (Part No. A68002-01)”, Oracle Corporation, Redwood Shores, 1999.

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

[BASAN, 2000], BASAN, A. L. C. “Ferramentas de Bioinformática para Sequenciamento e Anotação”. Universidade Federal do Rio Grande do Sul, Porto Alegre, 2000.

[MCT; CRIA, 2001], MCT, Ministério da Ciência e Tecnologia; CRIA, Centro de Referência em Informação Ambiental. “Sistemas de Informação: Estudos de Tecnologias e Padrões” Brasília, 2001.

[JÚNIOR; DENIPOTE, 1999], JÚNIOR, H. P. M.; DENIPOTE, J. G. “Projeto Genoma”, Universidade Estadual Paulista, São Paulo, 1999.

[SCHROEDER, 2000], SCHROEDER, L. F. “Recursos de Banco de Dados do Centro Nacional de Biotecnologia (NCBI)”, Centro Nacional de Biotecnologia, Brasília, 2000.

[FUGITA, 2000] FUGITA, A. “Anotação de Genes Associados com o Controle da Proliferação Celular e Origem de Tumores”, Universidade de São Paulo, 2002. Veja http://www.linux.ime.usp.br/~fugita/mac499/ acesso em 18/08/2002.

[CLICKZERO, 2002], CLICKZERO “Genoma e Genética”. Veja http://www.geocities.com/clickzero/genome.htm acesso em 15/08/2002.

[FÉLIX, 2002] FÉLIX, J. M. “Genoma Funcional”, Biotecnologia, Ciência & Desenvolvimento, Nº 24, janeiro a fevereiro, 2002.

[BOMTEMPI, 1999] BOMTEMPI, N. “Contribuições da Ciência Biológica no século XX e sua projeção para o século XXI”, O Mundo da Saúde, ano 23 nº 06, novembro a dezembro, 1999.

[PESSINI, 2000] PESSINI, L. “Tecnociência da Informação em Saúde”. O Mundo da Saúde, ano 24 nº 03, maio a junho, 2000.

[TACHINARDI, 2000] TACHINARDI, U. “Tendências da Tecnologia da Informação em Saúde”, O Mundo da Saúde, ano 24 nº 03, maio a junho, 2000.

[COSTA, 2000] COSTA, V. R. “Genoma Decifrado, Trabalho Dobrado”, Ciência Hoje. Vol. 28, nº 166, novembro, 2000.