Bancos de Dados de Genoma

Luiz Fernando Bessa Seibel, Melissa Lemos e Sérgio Lifschitz

Departamento de Informática

Pontifícia Universidade Católica do Rio de Janeiro

{seibel, melissa, lifschitz}@inf.puc-rio.br

Resumo: Os bancos de dados de genoma representam hoje uma das principais ferramentas de suporte para os biólogos moleculares e geneticistas. Para que estes bancos de dados possam ser realmente utilizados na prática é necessário tratar de vários pontos importantes, incluindo a definição do modelo de dados mais adequado, as necessidades de processamento, as análises e controles semântico dos dados e os meios de acesso e o problema da integração das bases de dados. Neste trabalho pretendemos apresentar os principais bancos de dados de genoma e os algoritmos envolvidos nas análises de sequências. Serão discutidos em particular os aspectos da integração destas bases de dados e alguns outros tópicos de pesquisa na área de banco de dados.

1 Introdução

Muitos projetos de análise de genoma estão sendo desenvolvidos atualmente. O Projeto Genoma Humano (PGH) é um dos maiores. Formalmente iniciado em Outubro de 1990, o PGH tem como objetivo principal descobrir todos os genes humanos e torná-los acessíveis para estudos biológicos posteriores, além de determinar a sequência completa das aproximadamente 3 bilhões de bases do DNA. Todos os organismos são focos deste projeto porque todos têm seu próprio genoma e estão relacionados através de similaridades de sequências de DNA. Assim, mesmo os genomas não humanos podem trazer novos conhecimentos sobre a biologia humana. Vários países têm estabelecido programas de pesquisas do genoma humano, entre os quais o Brasil [DOE00a] [DOE00b].

A informação detalhada do DNA será chave para o entendimento da estrutura, organização e função do DNA nos cromossomos. Mapas de genoma de outros organismos proverão a base para estudos comparativos que serão essenciais para o entendimento de sistemas biológicos mais complexos. Genes envolvidos em várias doenças genéticas serão encontrados, e estudos poderão ser feitos para se descobrir como tais genes contribuem para as doenças genéticas. Práticas médicas serão radicalmente alteradas quando novas tecnologias clínicas baseadas no diagnóstico de DNA forem combinadas com informações de mapas genéticos. A ênfase aos tratamentos de doenças será dada a prevenção. Pesquisadores serão capazes de prever indivíduos com tendência a doenças particulares e novas terapias poderão ser feitas baseadas em novas drogas, em técnicas de imunoterapia, em evitar condições ambientais que possam disparar a doença, e possivelmente, em substituição dos genes problemáticos [DOE00a][DOE00b].

Como parte do PGH, estudos paralelos têm sido feitos aqui no Brasil como o do organismo *Xylella fastidiosa* financiado pela FAPESP e do *Trypanosoma cruzi* realizado pelo grupo de pesquisa do Departamento de Bioquímica e Biologia Molecular (DBBM) da Fundação Oswaldo Cruz (FioCruz)[Fio00].

Entre os diversos assuntos pesquisados até agora destacam-se o armazenamento e o acesso aos dados de biologia molecular em bancos de dados, em particular as sequências de ácidos nucléicos e aminoácidos e suas respectivas anotações, e os algoritmos para análises destes dados.

Com o avanço da tecnologia, existem cada vez mais sequências e anotações [Doo90] e não é possível determinar a quantidade de informações que ainda será obtida de diversos organismos com o andamento do projeto genoma. Isso torna fundamental o uso de um banco de dados bem estruturado que permita o armazenamento, o acesso e o processamento destas informações de forma simples e eficiente.

Os bancos de dados de genoma representam hoje uma das principais ferramentas de suporte para os biólogos moleculares e geneticistas. É de fundamental importância para a pesquisa nesta área realizar cadastros de sequências e de algumas anotações relacionadas, e realizar consultas nestes bancos a fim de levantar dados para análises biológicas. Entre estas análises é possível destacar a comparação de sequências e o descobrimento de novos genes, funções e características de uma nova sequência. Para que estes bancos de dados possam ser realmente utilizados é necessário tratar de vários pontos importantes. Entre eles é possível destacar:

- Utilização de um modelo de dados apropriado;
- Adoção de algoritmos que permitam análises complexas nas sequências cadastradas no banco;
- Controle do cadastramento de sequências de forma a evitar múltiplas inserções do mesmo dado na base. Isto pode ser realizado através de algoritmos especialmente construídos para verificar a pré-existência de tais sequências no banco.

Existe ainda o problema da integração das bases de dados de genoma Atualmente, diversos centros de pesquisa têm feito esforços para cadastrar sequências de diferentes organismos. Assim, existem diversos bancos de dados, cada um com um modelo de dados distinto e utilizando diferentes tecnologias, sobre os quais os usuários têm necessidade de interagir.

Além disso, há vários estudos para a obtenção de algoritmos que façam análises eficientes em todo este volume de dados. Um dos problemas mais importantes para análises destes dados é o de comparação de sequências, pois ela é a base para várias outras manipulações mais elaboradas [MS94]. É possível citar duas principais famílias de algoritmos que realizam comparações de sequências armazenadas em bancos de dados, a FAST [Pea91] e a BLAST [AGM+90].

Este trabalho tem por objetivo apresentar os principais bancos de dados de genoma, as características de cada um e os principais algoritmos envolvidos nas análises de sequências em uma dada base. Em particular, será estudada a integração destas bases heterogêneas de forma a ser possível responder a determinadas consultas distribuídas.

O texto está organizado da seguinte forma: na seção 2 são apresentados conceitos de biologia celular e molecular considerados importantes para o entendimento deste trabalho. A seção 3 em seguida descreve as principais aplicações de informática na área de biologia hoje em dia, a saber, os bancos de dados e os algoritmos utilizados. É apresentada na seção 4 uma

classificação das implementações que visam a integração dos bancos de dados aplicados à biologia. Já a quinta seção apresenta as características e funcionalidades de alguns dos principais bancos de dados existentes e também de esforços de integração. Finalmente encerra-se com uma seção com comentários finais e trabalhos em andamento e futuros.

2 Conceitos de Biologia Celular e Molecular

Esta seção tem por objetivo apresentar alguns conceitos básicos da área de biologia celular e molecular, visando facilitar a compreensão do texto como um todo e foi baseado em [Rob85].

2.1 A Célula: Organização Estrutural

O estudo do mundo vivo mostra que a evolução produziu uma imensa variedade de formas. Existem em torno de quatro milhões de espécies diferentes de bactérias, protozoários, vegetais e animais, que diferem em sua morfologia, função e comportamento. Entretanto sabe-se agora que, quando os organismos vivos são estudados a nível celular e molecular, observa-se um plano único principal de organização. O objetivo da biologia celular e molecular é precisamente este plano unificado de organização – isto é, a análise das células e moléculas que constituem as unidades estruturais de todas as formas de vida. A célula é a unidade estrutural e funcional básica dos organismos vivos.

Células Procarióticas e Eucarióticas

As células são identificadas como pertencentes a dois grupos: procarióticas e eucarióticas. A principal diferença entre estes dois tipos celulares é a ausência de um envoltório nuclear nas células procarióticas. O cromossomo desta célula ocupa um espaço denominado nucleóide, estando em contato direto com o protoplasma. As células eucarióticas possuem um núcleo verdadeiro com um envoltório nuclear elaborado, através do qual ocorrem trocas entre o núcleo e o citoplasma.

2.2 A Célula: Organização Molecular

A estrutura celular visível aos microscópios óptico e eletrônico é resultante de um arranjo de moléculas numa ordem bastante precisa. Apesar de haver muito ainda a ser aprendido, começaram a surgir os princípios gerais da organização molecular de algumas estruturas celulares, como membranas, ribossomos, cromossomos, mitocôndrias e cloroplastos.

Numerosas estruturas celulares são constituídas por moléculas bastante grandes denominadas polímeros. Existem dois exemplos importantes de polímeros nos organismos vivos. São eles:

- Ácidos nucléicos, que resultam da repetição de quatro diferentes unidades denominadas nucleotídeos. A sequência linear de quatro nucleotídeos na molécula de DNA é a fonte básica da informação genética.

- Proteínas ou polipeptídeos são compostos por aproximadamente 20 aminoácidos, presentes em diversas proporções, unidos por ligações peptídicas. A ordem em que estes 20 monômeros podem se unir dá origem a um número astronômico de combinações em diferentes moléculas protéicas, determinando não só sua especificidade, mas também sua atividade biológica.

Ácidos Nucléicos

Todos os organismos vivos contêm ácidos nucléicos na forma de ácido desoxirribonucléico (DNA) e ácido ribonucléico (RNA).

O DNA é o principal armazenador da informação genética. Esta informação é copiada ou transcrita para moléculas de RNA, cujas as sequências de nucleotídeos contém o "código" para a ordenação específica de aminoácidos. As proteínas são então sintetizadas num processo que envolve a tradução do RNA. Refere-se frequentemente à série de eventos acima relacionada como o dogma central da biologia molecular; ela pode ser resumida na forma esquematizada na Figura 1:

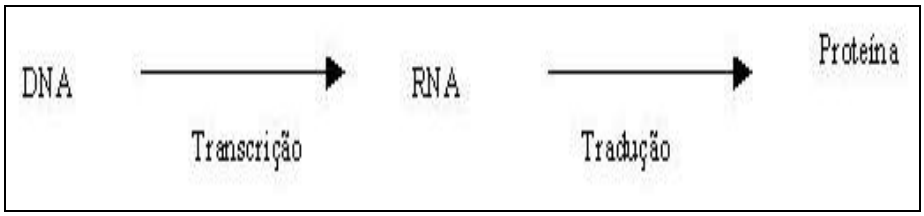

Figura 1.Processos transcrição e tradução.

Em células superiores, o DNA localiza-se principalmente no núcleo, dentro dos cromossomos. Uma pequena quantidade de DNA fica no citoplasma, contida nas mitocôndrias e cloroplastos. O RNA é encontrado tanto no núcleo, onde é sintetizado, quanto no citoplasma, onde tem lugar a síntese protéica.

Ácidos Nucléicos: uma Pentose, um Fosfato e quatro Bases

Os ácidos nucléicos são compostos por uma molécula de açúcar (pentose), bases nitrogenadas (purinas e piridiminas) e ácido fosfórico. Veja a Figura 2.

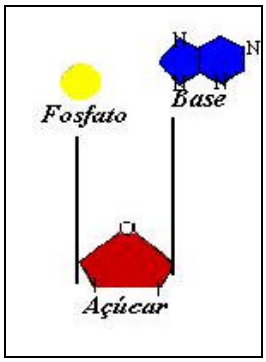

Figura 2. Ácido Nucléico

As pentoses são de dois tipos: ribose no RNA e desoxirribose no DNA.

As bases encontradas nos ácidos nucléicos são também de dois tipos: piridiminas e purinas. No DNA as piridiminas são timina (T) e citosina (C); as purinas são adenina (A) e guanina(G). O RNA contém uracila (U) no lugar de timina.

Toda a informação genética de um organismo vivo está armazenada em sua sequência linear das quatro bases. Portanto, um alfabeto de quatro letras (A, T, C, G) deve codificar a estrutura primária (i.é., o número e a sequência dos 20 aminoácidos) de todas as proteínas.

O DNA é uma Hélice Dupla

A estrutura do DNA é mostrada na Figura 3. Ele é composto por duas cadeias helicoidais de polinucleotídeos com giro para a direita, formando uma hélice dupla em torno de um mesmo eixo central. As duas fitas são antiparalelas, unidas por pontes de hidrogênio estabelecidas entre os pares de bases. Desde que existam uma distância fixa entre as duas moléculas de açúcar nas fitas opostas, somente certos pares de bases podem se encaixar na estrutura. Os únicos pares possíveis são o AT e o CG.

A sequência axial de bases ao longo de uma cadeia de polinucleotídeo pode variar consideravelmente, porém na outra cadeia a sequência deve ser complementar. Devido a esta propriedade, dada uma ordem de bases em uma cadeia, a outra é exatamente complementar.

Figura 3. A dupla hélice do DNA.

Estrutura do RNA: classes e conformação

A estrutura primária do RNA é semelhante à do DNA, exceto pela substituição da ribose pela desoxirribose e da uracila pela timina. A composição de bases do RNA não é similar a do DNA, pois as moléculas de RNA são compostas por uma única cadeia.

Existem três principais classes de ácido ribonucléico: o RNA mensageiro (mRNA), o RNA de transferência (tRNA) e o ribossômico (rRNA). Todos estão envolvidos na síntese protéica. O mRNA contém a informação genética para a sequência de aminoácidos, o tRNA identifica e transporta as moléculas de aminoácidos até o ribossomo, e o rRNA representa 50% da massa dos ribossomos, organelas que fornecem um suporte molecular para as reações químicas da montagem de um polipeptídeo.

Proteínas

As unidades constituintes das proteínas são os aminoácidos. Existem vinte tipos de aminoácidos, representados pelos caracteres A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y.

2.3 Biologia Molecular do Gene

O DNA transporta a informação genética de maneira codificada de célula a célula e dos pais para a progênie. Toda a informação necessária para a formação de um novo organismo está contida na sequência linear das quatro bases, e a replicação fiel desta informação é assegurada pela estrutura de dupla cadeia do DNA onde o A pareia-se somente com o T e o G com o C.

O DNA não está livre dentro da célula, mas forma complexos com proteínas na estrutura denominada cromatina. No momento da divisão celular, a cromatina condensa-se na forma de cromossomos. Veja Figura 4.

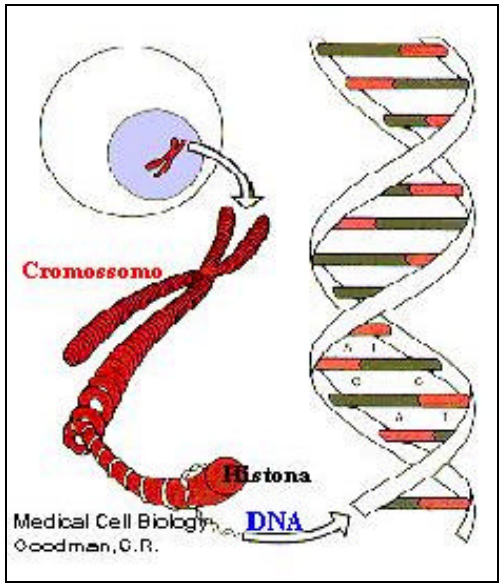

Figura 4. A célula e o cromossomo.

Os cromossomos são filamentos encontrados no interior do núcleo das células. Eles ocorrem normalmente em pares, têm diferentes tamanhos e formas e seu número é constante em cada espécie de ser vivo.

O gene é uma unidade hereditária que consiste numa sequência particular de bases no DNA e que especifica a produção de uma certa proteína (por exemplo, uma enzima).

Três Nucleotídeos codificam um Aminoácido

Os códons, ou unidades hereditárias que contém o código de informação para um aminoácido, são compostos por três nucleotídeos (um trio). Esta informação encontra-se no DNA, de onde é transcrita para o RNA mensageiro; assim, o mRNA possui a sequência de bases complementar à do DNA do qual foi copiado. O DNA e o mRNA possuem somente quatro

bases diferentes, enquanto que as proteínas contêm 20 diferentes aminoácidos. Dessa maneira, o código é lido em grupos de três bases, sendo três o número mínimo necessário para a codificação de 20 aminoácidos. Veja na Figura 5 a ilustração do código genético.

Por volta de 1964 todos os 64 códons possíveis haviam sido decifrados. 61 códons correspondem a aminoácidos e 3 representam sinais para a terminação das cadeias polipeptídicas. Sabendo que existem somente 20 aminoácidos, fica evidente que vários trios podem codificar para o mesmo aminoácido; isto é, alguns dos trios são sinônimos. A prolina, por exemplo, é codificada por CCU, CCA, CCG e CCC.

Mutação

Outro conceito importante da biologia é o de mutação, que é uma mudança no conteúdo do DNA. Os tipos de mudanças podem ser de substituição de base, inserção de base, remoção de base, e rearranjo ou troca na ordem de segmentos de base. Estas mudanças podem ser divididas em classes dependendo da escala com que elas ocorrem. Algumas mudanças são fenômenos localizados, enquanto outras ocorrem um milhão de vezes seguidas.

Genoma

O genoma é o conteúdo de todo DNA presente em uma célula, incluindo todos os genes e todas as regiões intergênicas.

First Position	Second Position of Codon: T	C	A	G	Third Position
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

Figura 5. O código genético.

Sequência e Biossequência

O termo **sequência finita de caracteres**, ou simplesmente **sequência** ou **cadeia**, será usado no sentido restrito de uma sequência finita de caracteres de um dado alfabeto S. Assim, se S = {A,C,T,G}, então ATTCCG e CCGA são sequências. Uma **biossequência** [MS94] é uma sequência onde o alfabeto S = {A,C,G,T} (DNA) ou S = {A,C,G,U} (RNA) ou S é formado pelos 20 aminoácidos citados anteriormente.

3 Bancos de Dados e Algoritmos de Biologia Molecular

As biossequências podem ser tratadas como cadeias de texto. Por este motivo, os biólogos moleculares podem coletá-las e guardá-las em arquivos texto. Foi isso o que foi feito no início dos processos de sequenciamento [Doo90]. No entanto, com o avanço da tecnologia, a produção de biossequências aumentou e, consequentemente, os dados armazenados em arquivos textos cresceram muito, tornando sua manutenção e a dos programas de aplicação relacionados muito trabalhosa. Diante disto os biólogos moleculares começaram a usar Sistemas Gerenciadores de Bancos de Dados (SGBD), mais apropriados para gerenciar grandes volumes de dados.

Quando se começou a armazenar, os dados eram obtidos através de publicações em artigos científicos. Com o avanço da tecnologia e, consequentemente, com o crescimento exponencial do volume de biossequências, tais dados passaram a ser submetidos aos bancos de dados através da Internet [Doo90]. Isto possibilitou uma grande facilidade na submissão de biossequências aos bancos de dados, o que é muito importante para que os biólogos possam acessar e fazer suas análises em novos dados mais rapidamente.

Atualmente os bancos de dados de biologia molecular (BDBM) utilizam sistemas de banco de dados relacional, sistemas orientados a objetos e ainda existem alguns que nem propriamente banco de dados são, utilizando apenas *flat files* [NK99].

É difícil estimar o número de BDBM existentes. Hoje em dia existem não somente os bancos de dados de sequências de nucleotídeos (DNA) e de aminoácidos (proteínas), mas também inúmeros outros com informações bem específicas, como organismos especiais (ex.: *Eukariotic* [PPJ+00], *Escherichia Coli* [NK99] e *Drosophilia* [Fly99]), biossequências específicas (ex.: tRNA e rRNA), enzimas, mutações, famílias de biossequências (filogenia), etc. Além disso, já existem bancos que guardam estruturas tridimensionais das biossequências, como por exemplo o PDB [BWF+00].

É possível destacar os seguintes BDBM como os maiores atualmente: GenBank Sequence Database [BML+00], EMBL Nucleotide Sequence Database [BBC+00], Genome Sequence Database (GSDB) [HCF+00], Genome Database (GDB) [LCP+98], PIR (Protein Identification Resource) - International Protein Sequence Database [BGH+00], e *A. Caenorhabditis elegans* DataBase (ACeDB). Nestes bancos de dados estão armazenadas anotações relevantes, além das próprias biossequências.

Apresentaremos aqui alguns exemplos de bancos de dados de biologia molecular (BDBM) , suas características mais importantes e alguns algoritmos para análises destes dados.

3.1 Exemplos de BDBMs

Existem inúmeros BDBM, a seguir está uma lista com alguns dos mais importantes deles.

GenBank Sequence Database [BML+00]

Organização responsável: National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), National Institutes of Healthy (NIH)

Informações principais: Sequências de nucleotídeos

URL: http://www.ncbi.nlm.nih.gov/Genbank/index.html

EMBL Nucleotide Sequence Database [BBC+00]

Organização responsável: EMBL OutStation - The European Bioinformatics Institute

Informações principais: Seqüências de nucleotídeos

URL: http://www.ebi.ac.uk/embl/index.html

Genome Sequence Database (GSDB) [HCF+00]

Organização responsável: Department of Energy (DOE) Federated Information Infrastructure -National Center for Genome Resources

Informações principais: Sequências de nucleotídeos

URL: http://www.ncgr.org/gsdb/gsdb.html

Genome Database (GDB) [LCP+98]

Organização responsável: U.S Department of Energy, com apoio adicional de U.S. National Institutes of Health, Japanese Science and Technology Agency, the British Medical Research Council, INSERM of France, e European Union.

Informações principais: Sequências de nucleotídeos

URL: http://www.gdb.org/

PIR (Protein Identification Resource)-International Protein Sequence Database [BGH+00]

Organização responsável: National Biomedical Research Foundation (NBRF), Munich Information Center for Protein Sequences (MIPS), e Japan International Protein Information Database (JIPID)

Informações principais: Seqüências de aminoácidos

URL: http://www-nbrf.georgetown.edu/

Swiss-Prot Protein Sequence Data Bank

Organização responsável: EMBL Outstation - The European Bioinformatics Institute (EBI) e Swiss Institute of Bioinformatics (SIB)

Informações principais: Seqüências de aminoácidos

URL: http://www.expasy.ch/sprot, e http://www.ebi.ac.uk/swissprot.

Protein Data Bank (PDB) [BWF+00]

Organização responsável: Federal Government Agency

Informações principais: Estruturas terciárias da proteína

URL: http://www.rcsb.org/pdb

A Caenorhabditis elegans DataBase (ACeDB)

Organização responsável: NIH National Center for Research Resources

Informações principais: *C. elegans, Human Chromosome 21, Human Chromosome X, Drosophila melanogaster, mycobacteria, Arabidopsis, soybeans, rice, maize, grains, forest trees, Solanaceae, Aspergillus nidulans, Bos taurus, Gossypium hirsutum, Neurospora crassa, Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Sorghum bicolor*.

URL: http://probe.nalusda.gov:8000/acedocs

FlyBase [Fly99]

Organização responsável: U.S. National Institutes of Health e British Medical Research Council.

Informações principais: *Drosophila*

URL: http://fly.ebi.ac.uk:7081/docs

Eukariotic Promoter Database (EPD) [PPJ+00]

Organização responsável: ISREC em Epalinges s/Lausanne (Switzerland)

Informações principais: *Eukariotic promoter*

URL: http://www.epd.isb-sib.ch

DNA Data Bank of Japan (DDBJ) [TMO+00]

Organização responsável: Center for Information Biology, National Institute of Genetics, Yata, Mishima, Japan

Informações principais: Sequências de Nucleotídeos.

URL: http://www.ddbj.nig.ac.jp

3.2 Características de BDBMs

3.2.1 Volume Grande de Dados

Um dos pontos mais importantes a considerar no contexto de BDBM é o volume de dados, que vem aumentando com o passar do tempo devido ao avanço da tecnologia e do grande interesse no Projeto Genoma. Como exemplo, é possível citar o Projeto Genoma Humano [HG00] [DOE00a] [DOE00b]. Nele existem aproximadamente 3 bilhões de bases arranjadas ao longo dos cromossomos, em uma ordem particular para cada indivíduo. Além do comprimento de um único genoma ser consideravelmente grande, há ainda a necessidade de armazenar genomas de vários seres e muitas informações relacionadas a eles.

O armazenamento, e posterior acesso e processamento a toda esta informação, é um grande desafio para profissionais de computação e especialistas em biologia e informática. Um

milhão de bases (chamada de megabase) de dados de sequência de DNA é equivalente a 1 megabyte de espaço de armazenamento de dados em um computador. Como o genoma humano tem aproximadamente 3 bilhões de pares de bases, um genoma precisaria de 3 gigabytes de espaço de armazenamento de dados em um computador [Cas92]. Isto somente para os dados da sequência de nucleotídeos, não incluindo anotações e outras informações que podem estar associadas aos dados da sequência.

A cada dia que passa mais anotações estão sendo associadas aos dados da sequência, o que não é uma surpresa porque a sequência é meramente um ponto de partida para entendimentos biológicos mais profundos. Além disso, vale ressaltar que estes dados (sequência e suas anotações) são informações de um único ser ou organismo. É necessário considerar informações de um número indeterminado de organismos e seres, o que torna fundamental o uso de um banco de dados bem estruturado que permita o armazenamento, o acesso e o processamento destas informações de forma simples e eficiente.

O GenBank, por exemplo, tem atualmente mais de 7GB de dados, sendo que tem aumentado de volume a taxas consideráveis, tendo dobrado de dezembro de 1999 a abril de 2000 [Gen00].

3.2.2 Informações Armazenadas

Os bancos de dados aplicados à biologia molecular podem se classificados de acordo com as informações biológicas que armazenam [AG97], que são, principalmente, de:

- sequências (de nucleotídeos ou de proteínas) e anotações sobre as mesmas,
- proteínas e informações sobre as respectivas funcões,
- estruturas de moléculas de proteínas (secundárias, representadas em um plano, ou terciárias, representadas em três dimensões),
- taxonomia (classificações dos organismos vivos),
- bibliografia na área de biologia molecular (artigos, jornais, periódicos, etc).

Sequências de nucleotídeos

Os bancos de sequências de nucleotídeos armazenam, além da própria sequência, anotações contendo dados de características biológicas relevantes sobre elas, que são: organismo a que pertence, sites das sequências que codificam moléculas de proteínas, função, fenótipo (características aparentes), e *links* para outros bancos de dados contendo informações biológicas sobre a sequência.

Embora exista um controle sobre êrros comuns detectados na submissão de sequências ao banco, a qualidade da informação é do pesquisador que submeteu a sequência. Os laboratórios que submetem sequências ao banco tem diferentes critérios sobre a qualidade da sequência que está sendo enviada. Além disso, alguns tem a preocupação de retirar da sequência os dados de clones vindos do sequenciamento, outros não agem desta forma, poluindo a sequência com informações desnecessárias. Assim, redundâncias e inconsistências são inevitáveis. Os bancos de dados de nucleotídeos apresentam, portanto, diversos êrros. As sequências existentes nestes bancos estão incompletas, contaminadas e com êrros oriundos do próprio sequenciamento. Os administradores de algumas dessas bases de dados resolveram

atacar o problema da redundância onde sequências similares foram agrupadas, desde que fosse possível inferir que uma delas era a origem das outras.

Os principais bancos de dados genéricos que armazenam sequências de nucleotídeos são aqueles que compõem o International Nucleic Acid Sequence Data Library, formado pelas bases de dados denominadas de Genbank, DDBJ e EMBL. Estes bancos armazenam também informações sobre partes das sequências que codificam moléculas de proteínas ou de RNA, além de anotações que contém outras informações biológicas relevantes. Tais informações são anotadas no campo *features*. A descrição completa do conteúdo de tal campo pode ser encontrada em http://ncbi.nlm.nih.gov/genbank/gbrel.txt.

Além destes, outros bancos de dados específicos de um dado organismo também armazenam informações sobre sequências, como por exemplo o AceDB e toda a família de bancos de dados que é baseada na sua arquitetura. A descrição completa da família de bancos de dados ACeDB pode ser encontrada em http://genome.cornell.edu/acedoc/index.html .

Outros bancos de dados especializados (em determinadas células ou componentes, em mutações, em funções gênicas, etc.) também armazenam informações de sequências, como por exemplo o Mitomap [KLB+97].

Sequências de proteínas

Os bancos de dados de sequências de proteínas armazenam além da própria sequência, informações sobre a função da proteína no organismo. Tais bancos de dados têm também como característica a redundância e inconsistência das informações sobre as sequências armazenadas.

O principal banco de sequências de proteínas é o Swiss-Prot. Existe maior cuidado com a qualidade da informação que é incluída neste banco, seu conteúdo é não redundante e inclui extensas anotações sobre as sequências. No entanto, este cuidado exige um intervalo de tempo entre a inclusão de uma sequência nos bancos de nucleotídeos e sua correspondente tradução para o Swiss-Prot. O banco que armazena esta tradução automática é o TrEMBL. Alguns bancos de nucleotídeos também armazenam sequências de proteínas, como por exemplo o Genbank.

Proteínas

Os bancos de dados de proteínas são especializados. O banco ENZYME e o banco LIGAND armazenam informações sobre enzimas. O banco PROSITE armazena documentações acerca de famílias de proteínas. Existem outros bancos de dados de grupamentos de proteínas segundo diferentes critérios / algoritmos, como por exemplo o banco BLOCKS.

Os bancos de dados de sequências de proteínas contém links para estes bancos de dados, que têm anotações mais completas sobre cada uma.

Estruturas de proteínas

Estes bancos de dados armazenam as representações da proteína em um plano ou em três dimensões. O principal banco de estruturas é o PDB, que armazena informações estruturais de moléculas de ácido nucléico. Estes bancos de dados não contém o mesmo volume de informações existente nos bancos de proteínas, devido ao difícil processo de obtenção de dados, feito via cristalografia.

Taxonomia

Os bancos de dados de taxonomia são bastante discutidos, uma vez que não existe consenso entre os especialistas sobre as classificações ali contidas. Os exemplos destes bancos de dados são: Species 2000, International Organization for Plant Information, Integrated Taxonomic Information System, The Tree of Life Project, entre outros. Cabe ressaltar que o Genbank mantém também informações de taxonomia, que são definidas e mantidas por um grupo de especialistas independente.

Publicações

Os bancos de dados de publicações armazenam e disseminam informações sobre a literatura científica de diversas áreas. Na área da biologia molecular, o mais importante repositório de tais informações é o MEDLINE, agora denominado PUBMED, que pode ser acessado via NCBI, através de uma interface denominada Entrez. O correspondente ao MEDLINE para a área agrícola é o AGRICOLA.

3.3 Modelo dos Dados

Diversos modelo de dados tem sido utilizados para representação das informações biológicas. Esta seção discute brevemente as implementações existentes e as vantagens e desvantagens de cada uma em termos de representação de fatos biológicos e de facilidades para os usuários.

Modelo Relacional

Diversas bases de dados de biologia molecular são implementadas em bancos de dados relacionais disponíveis no mercado. Tal tecnologia, no entanto, apresenta vantagens e desvantagens para esta aplicação [NK99], que serão resumidas a seguir.

O modelo relacional agrega a informação em tuplas, onde cada tupla (ou linha da tabela relacional) representa uma coleção de valores correlacionados, que não podem mais ser separados em relações mais simples. A normalização serve para eliminar problemas inerentes à duplicação de dados, que são: múltiplas atualizações e geração de tuplas espúrias na operação de junção.

Nos bancos de dados de biologia molecular é frequente a ausência de informações (atributos com valor NULL), fato que aumenta a decomposição dos dados em tabelas menores. Além disso, as frequentes exceções feitas às estruturas relacionais tendem a aumentar a decomposição, gerando novas tabelas. Assim, enquanto proliferam as tabelas do banco de dados, tornando os ítens de dados mais simples e de fácil entendimento de forma isolada, uma nova dificuldade aparece na compreensão e manutenção da estrutura dos dados, bem como no domínio completo do esquema. Em parte, isso deve-se ao fato de que o modelo relacional não representa relações existentes dentro de tuplas.

Por exemplo, na implementação relacional do Mitomap, a entidade *genetic locus* sofreu os seguintes desmembramentos ao longo do tempo:

Fase 1: Genetic locus (nome, start, stop, dados_mutação, etc.)

Fase 2: Genetic locus (nome, start, stop, id_mutação, etc.)

Mutação (id_mutação, tipo_mutação, dados_tipo_mutação, etc.)

Fase 3: Genetic locus (nome, start, stop, id_mutação, etc.)

Mutação (id_mutação, tipo_mutação, etc.)

Mutação_inserção (dados_mutação_inserção)

Mutação_exclusão (dados_mutação_exclusão)

Mutação_alteração (dados_mutação_alteração)

Ou seja, o objeto biológico se torna menos claro a cada decomposição.

Dado ao grande tamanho destas bases de dados e ao elevado número de tabelas, rapidamente estas bases de dados se tornam ingerenciáveis e mesmo incompreensíveis pelos próprios administradores.

A definição de relações no modelo E-R é ideal para representar relações (binárias) bem definidas entre as entidades. No entanto, os dados biológicos nem sempre se encaixam nessa categoria, devido às transformações existentes em virtude, por exemplo, de novas classificações ou de novas descobertas biológicas. Assim, é necessário que o modelo que represente tais dados seja mais flexível, de forma a facilitar a sua adequação ao mundo real. O modelo relacional não fornece tal flexibilidade.

A formulação de consultas ao modelo implementado exige o conhecimento da sua estrutura, limitando o tipo de consultas que poderiam ser feitas, desencorajando a exploração da base por usuários comuns. Ou seja, apenas especialistas em bancos de dados poderiam fazer tais consultas, fato que reforça a ênfase a ser dada na simplicidade do modelo de dados, para que possa ser compreendido pelos usuários.

Existem, no entanto, benefícios na adoção de um modelo relacional para os bancos de dados para biologia molecular. A teoria da normalização, baseada em dependências funcionais, garante a ausência de anomalias na base. A implementação relacional é responsável ainda pela obtenção de respostas rápidas às consultas, e por simplificar a tarefa de programação.

Em oposição a estes fatos, a validade da normalização se torna irrelevante se a tupla não pode representar o dado em questão, e a rapidez não pode ser avaliada se a consulta desejada não pode ser feita. Adicione-se a isso o fato de que o modelo relacional não se ajusta facilmente à natureza dos dados biológicos. Por exemplo, não é possível representar um atributo com múltiplos tipos de dados, mas isso pode acontecer na natureza.

Em resumo, o modelo relacional representa o mapeamento incompleto do mundo real para o conjunto de informações necessárias ao estudo da biologia molecular, tornando a compreensão e atualização dos dados bastante difíceis. Tais alterações só poderiam ser feitas com o completo conhecimento do esquema do banco, e não necessariamente com o completo conhecimento dos dados biológicos e de suas relações.

Uma dificuldade adicional é que o modelo relacional não provê facilidades de forma a compartilhar informações com outras bases de dados, sendo necessária a carga de tabelas nas várias bases e sua permanente atualização. Assim, as informações de uma base não podem apontar para ou serem apontadas por outras bases de dados, fato possível e de simples implementação em outros modelos.

Pelas razões descritas acima, diversas implementações usando outros modelos têm sido desenvolvidas.

Modelo Orientado a Objetos

Algumas bases de dados biológicas foram implementadas no modelo de dados orientado a objetos (OO). O modelo OO traz vantagens em relação ao modelo relacional, pois permite mapeamento direto de conceitos complexos do mundo real em estruturas de dados do modelo [NK99], [CM95], [Kro93].

O projeto dos objetos do modelo permite determinar o grau de normalização / simplificação de cada entidade / objeto envolvido (tal fato não está ligado às regras do modelo).

Com a adoção do modelo OO, o usuário final recebe o benefício do conhecimento do objeto de forma completa. Tal modelo também provê uma coleção de métodos e de estruturas para modelar, manter e consultar os dados.

Porém, o modelo OO também apresenta problemas. Objetos são representados em estruturas de dados fixas, têm métodos próprios e se relacionam através de ponteiros. Isso implica em que uma alteração no esquema do banco de dados pode acarretar na alteração da estrutura utilizada e mesmo na reprogramação dos métodos já implementados. Outro ponto problemático é a utilização de ponteiros para os objetos e do identificador único do objeto (OID), que, embora relevantes para o modelo, não são necessariamente relevantes em termos biológicos. Este fato pode dificultar a compreensão da referência aos dados por um usuário comum.

Para dificultar ainda mais o quadro, a herança biológica nada tem a ver com a herança advinda do modelo OO. As estruturas da biologia são representadas em uma enorme variedade de classes, que frequentemente não têm qualquer relação entre si. Assim, não existe nenhum benefício em herdar atributos de outras classes de objetos.

Embora o modelo orientado a objetos favoreça o mapeamento do mundo real, ainda existem inúmeras deficiências a serem resolvidas, que favoreceram o surgimento de novas implementações utilizando outros modelos [NK99].

Modelo Relacional-Objeto

O modelo relacional-objeto é o mais adequado para aplicações de biologia molecular porque são orientadas a consultas e requerem o uso de dados complexos.

Realmente os bancos de dados que utilizam o modelo de dados relacional-objeto tem sido recentemente utilizados para o armazenamento de dados de biologia molecular, uma vez que aliam a facilidade de consulta inerente ao modelo relacional com o tratamento de dados complexos.

Os bancos de dados que adotam este modelo permitem a formulação de consultas utilizando-se funções e operadores definidos pelos usuários. Tais requisitos não existem na definição da linguagem SQL-2, utilizada nos bancos que adotam o modelo relacional, porém são utilizados nas linguagens de consulta dos bancos de dados que adotam o modelo relacional-objeto (estes requisitos estão incluídos no padrão SQL-3).

O AatDB (banco de dados do genoma da Arabdopsis thaliana) pode ser citado como exemplo de implementação neste modelo.

Modelo de Dados Semi-Estruturados

Diversos bancos de dados biológicos implementam o arquivamento dos objetos utilizando dados semi-estruturados. É o caso do ACeDB e do GenBank. Outros bancos de dados se utilizam do código do AceDB e portanto utilizam o mesmo modelo. O AceDB optou por este tipo de implementação pelas facilidades inerentes à alteração dos objetos, sem necessáriamente exigir a alteração dos métodos já utilizados.

Para o AceDB, os objetos são definidos de acordo com uma linguagem cuja sintaxe é semelhante à XML [ABS00], onde a representação dos dados pode ser vista como uma árvore, cujos nodos podem estar presentes ou não e onde existem facilidades (inerentes da estrutura) no sentido de adicionar, excluir e alterar nós ou sub-árvores. Assim, o AceDB armazena os dados nessa estrutura (árvore), em formato binário.

Outras fontes de dados de biologia têm arquivos semi-estruturados, de forma a facilitar a troca de informações com outros bancos. É o caso do Genbank, que utiliza o padrão ASN.1 [IOS87].

Dados com formatos específicos

Os dados complexos podem ser também armazenados à parte em formatos específicos a fim de permitirem manipulação por algoritmos (programas) especiais. É o caso do formato T-FASTA, que facilita a execução dos algoritmos FASTA e BLAST, para verificação de pré-existência de uma dada sequência no banco. O próprio GenBank, além de outros bancos relacionais implementam este tipo de arquivamento.

3.4 Interface de Acesso

A interface para os usuários destes bancos precisa ser muito bem desenvolvida para que o acesso aos dados seja facilitado ao máximo. Ela pode prover mecanismos de consultas triviais, como buscas por palavras-chaves, autores, referências; mas também pode permitir consultas mais complexas, permitindo a utilização de operadores lógicos. Além disso, a interface deve permitir a execução de algoritmos necessários em biologia molecular, como os de comparação de sequências (por exemplo o FASTA [Pea91] e o BLAST [AGM+90]).

Interfaces para consultas são muito importantes para facilitar a interação dos cientistas com os bancos de dados. Os cientistas não estão preparados para manipular linguagens de consultas complexas e por isso preferem interfaces de usuário gráficas e mais intuitivas [MR95].

Não é simples construir uma interface que permita aos biólogos executar todas as operações em biossequências que desejam com seus respectivos parâmetros tendo em vista que a complexidade dos processamentos sobre as biossequências e as buscas eficientes sobre um grande volume de dados são problemas ainda não bem resolvidos.

3.5 Interação

No início da coleta e do armazenamento dos dados de biologia, os bancos de dados eram totalmente isolados, isto é, não existia nenhuma troca de informação entre eles. Com o passar do tempo, foi-se tendo a preocupação em integrar estes bancos de dados. Um biólogo, por

exemplo, além de pesquisar por informações de uma determinada sequência de nucleotídeos em um único banco de dados, gostaria de obter informações sobre a mesma sequência armazenadas em outros bancos de dados. Desta forma, os bancos de dados começaram a fazer referências a outros bancos de dados. O banco de dados GDB, por exemplo, possui referências ao GenBank, isto é consegue-se descobrir qual registro do GDB armazena dados de uma certa sequência que está no GenBank [SU94].

Além disso, diferentes bancos de dados podem possuir os mesmos dados. Por exemplo o GSDB [HCF+00], um banco de dados relacional implementado em Sybase, interage com os repositórios de sequências de DNA DDBJ, EMBL e GenBank [GSDB00]. Desta forma estes bancos mantêm seus dados replicados (parcial ou totalmente) em outros bancos.

3.5.1 Distribuição e Integração dos Dados

Cada banco de dados de biologia molecular consiste em um grande e variado montante de tipos de dados, que foram desenvolvidos independentemente, apesar de tais dados serem muito relacionados uns com os outros. Os cientistas que utilizam tais bancos precisam fazer consultas em vários destes bancos. Esta tarefa não é simples se eles não contarem com um sistema que os ajude. É por isso que é necessária a integração e o gerenciamento eficiente destes bancos de dados.

Já existem vários sistemas desenvolvidos que integram bancos de dados de biologia molecular [SU94],[MCK97],[KDG96]. Mas devido à grande dificuldade de se integrar tais bancos, existem muitos aspectos que ainda não atendem às necessidades dos biólogos.

Os bancos de dados de biologia molecular foram criados por diversos grupos internacionais. Ainda não existe um padrão em algum nível de abstração, muito menos em todos os níveis existentes de heterogeneidade, tais como o modelo conceitual, o modelo de dados, ou a linguagem de consulta. Isto faz com que tais bancos sejam completamente diferentes uns dos outros.

Uma infra-estrutura de informação federada precisa tratar da heterogeneidade como uma consideração primária e prover poderosas ferramentas que identifiquem a heterogeneidade imediatamente. Os métodos que não identificam a maioria destes níveis irão falhar mesmo quando confrontarem com um número moderado de banco de dados [Kar95].

3.5.2 Conceitos Diferentes

Além da heterogeneidade estrutural e de representação já mencionada, existe outra tão importante quanto e mais difícil de ser tratada: a heterogeneidade semântica. Os conceitos que foram usados na criação dos bancos de dados são muito diferentes. Como por exemplo, a palavra *gene* pode ter significados diferentes em bancos de dados distintos [Fre91].

Para que as informações em bancos de dados heterogêneos sejam comparadas, é preciso primeiro entender os diferentes conceitos em ambos. É preciso então escolher entre traduzir os significados e torná-los uniformes, ou deixá-los sozinhos e notar as diferenças. Isto é muito complicado de ser feito porque os conceitos não são claros e seu entendimento depende das pessoas que projetaram o banco. Além disso, encontrar documentação sobre estes bancos é uma tarefa bastante complicada pois há muito pouca informação disponível na literatura.

3.5.3 Gerenciamento de Memória

Outra característica a considerar sobre os BDBM é a estrutura de armazenamento físico (estrutura de dados em memória secundária) para a representação das biossequências. Em geral, os bancos de dados convencionais possuem estruturas de armazenamento e métodos de acesso como índices primários e secundários, que melhoram o tempo de acesso aos dados.

As aplicações não convencionais, como os bancos de dados temporais e espaciais, trouxeram inovações tanto de estrutura de armazenamento quanto nos métodos de acesso. E isto motiva um estudo com o objetivo de encontrar uma estrutura de armazenamento também para os BDBM, já que hoje em dia as biossequências são armazenadas como simples textos e seus acessos não levam em consideração nenhuma característica particular de alguma aplicação da biologia.

É possível supor que se o banco de dados e a memória principal que armazenam as biossequências para suas análises fossem estruturados de maneira mais *ad-hoc*, levando em consideração características particulares de determinadas aplicações da biologia molecular, estas aplicações poderiam vir a ter uma melhora significativa em suas performances.

3.6 Aplicações e Algoritmos

Existem diversas aplicações neste contexto de biologia computacional. Entre elas é possível destacar [MS94]:

- *Comparação de biossequências*

  Compara uma biossequência a outra a fim de encontrar trechos semelhantes entre elas;

- *Montagem de fragmentos de DNA*

  Dadas várias sequências de fragmentos de DNA, busca-se reconstituir (*fragment assembly*) o trecho de DNA do qual esses fragmentos provieram através de comparações entre elas;

- *Mapeamento Físico de Cromossomo ou Mapeamento Físico de DNA*

  Ao se iniciar o estudo de um cromossomo, uma das técnicas usadas é a de quebrá-lo em vários pedaços através de enzimas de restrição. Estes pedaços são então replicados através de um processo chamado *clonagem*, que cria cópias desses fragmentos. Essas cópias recebem o nome de *clones*. No processo de quebra, a informação de localização de cada clone no cromossomo é perdida e o problema consiste em recuperar esta informação;

- *Construção de Árvores Filogenéticas*

  objetivo principal é esclarecer histórias evolutivas dos organismos. Este esclarecimento é feito através da construção de árvores filogenéticas, que mostram como os organismos atualmente existentes se relacionam através de organismos ancestrais;

- *Predição de Estruturas*

  As biossequências que formam um ácido nucléico ou proteína são muito mais do que simples cadeias unidimensionais de nucleotídeos ou aminoácidos. Essas cadeias se dobram de diversas formas e apresentam diversas estruturas tridimensionais. Essas

estruturas estão intimamente relacionadas à função das moléculas e, portanto, sua determinação é fundamental para o estudo dos ácidos nucléicos e proteínas. Muitas estruturas ainda não foram desvendadas e, por esse motivo, muito esforço vem sendo feito na procura de métodos computacionais que auxiliem em suas predições.

3.6.1 Algoritmos de Comparação

Entre as aplicações apresentadas, a comparação de sequências é a operação primitiva mais importante na área de biologia computacional e serve de base para muitas outras manipulações mais elaboradas. A grosso modo, esta operação consiste em encontrar trechos semelhantes entre duas ou mais sequências. Contudo, por trás desta aparente simplicidade, esconde-se uma vasta gama de problemas distintos, com formalizações diversas, muitos deles exigindo algoritmos e estruturas de dados próprias para sua execução eficiente.

A seguir são dados alguns exemplos práticos de comparação de biossequências [MS94]:

1. Sejam duas sequências sobre o mesmo alfabeto com aproximadamente 10.000 caracteres. Suponha que elas possuem composições idênticas, exceto por divergências isoladas (inserções, remoções ou substituições de caracteres) que ocorrem a taxa de um erro a cada 100 caracteres. Deseja-se encontrar estes erros. Este problema aparece quando um gene é sequenciado por dois laboratórios diferentes e deseja-se comparar os resultados, ou quando a sequência foi digitada mais de uma vez e deseja-se tratar erros de digitação.
2. Sejam duas sequências de algumas centenas de caracteres sobre um mesmo alfabeto. Deseja-se decidir se existe um prefixo de uma delas que seja semelhante a um sufixo da outra. Em caso afirmativo um alinhamento entre as regiões semelhantes deve ser produzido. Suponha esta mesma situação, exceto que em vez de duas, existam 500 sequências que devem ser comparadas duas a duas. Estes problemas aparecem no contexto de montagem de fragmentos em programas de auxílio a sequenciamento de DNA em larga escala.
3. Sejam duas sequências de algumas centenas de caracteres sobre um mesmo alfabeto. Deseja-se decidir se há algum trecho de uma delas semelhante a um trecho de tamanho aproximadamente igual na outra. A semelhança não é medida em termos de porcentagem de caracteres idênticos, mas em termos de um esquema de pontuação que atribui uma nota fixa a cada par de caracteres do alfabeto. Dois trechos são considerados semelhantes se a soma das notas dadas a caracteres alinhados for superior a um dado valor. Suponha esta mesma situação, exceto que, em vez de duas, temos uma sequência fixa que deve ser comparada a várias outras. Estes problemas aparecem no contexto de buscas de semelhanças locais usando bases de dados de biossequências.

Famílias FAST e BLAST

Existem ainda os algoritmos de comparação que são utilizados especialmente em análises de biossequências armazenadas em bancos de dados. As famílias de algoritmos mais utilizadas atualmente são as FAST [Pea91] e BLAST [AGM+90].

Durante os anos 80, Lipman, Pearson e Wilbur descreveram em detalhes heurísticas usadas em seus programas para buscas em bases de biossequências [WL83][LP85] [PL88]. O primeiro programa a surgir foi o FASTP [LP85], que faz buscas com proteínas. A seguir apareceu uma versão para sequências de nucleotídeos, FASTN. Posteriormente ambos foram juntados num único programa chamado FASTA [PL88]. Estes programas efetuam

comparações locais e retornam apenas um alinhamento local - considerado o ótimo. Mais tarde, programas que também obtém vários alinhamentos locais (LFASTA, PLFASTA) foram incorporados à família de programas FAST. Um sumário destes programas encontra-se em [Pea90]. Um estudo extenso sobre a **sensibilidade** (capacidade de detectar homologias remotas) e **seletividade** (capacidade de detectar falsas homologias) de FASTA foi empreendido por Pearson [Pea91].

Na década de 90 surgiram os programas BLAST (*Basic Local Alignment Search Tool*) [AGM+90][AMS+97]. O algoritmo BLAST foi desenvolvido por Altschul, Gish, Miller, Myers e Lipman [AGM+90]. A motivação para o desenvolvimento de BLAST foi a necessidade de aumentar a velocidade do FASTA. Como na família FAST, o BLAST possui versões para proteínas (BLASTP) e ácidos nucléicos (BLASTN).

Comentários Finais

Atualmente existem diversos grupos de pesquisas em bioinformática nas áreas de algoritmos([MS97], [Sha99], [KRT96]), integração de BDBM ([MC95], [MCK97]), [Kar95], [BDO95], [BDH+95]) e construção de ferramentas para o funcionamento completo de um laboratório de biologia molecular incluindo interface com o usuário, banco de dados, entre outras [GRS94].

O nosso grupo de pesquisa, no Departamento de Informática da PUC-Rio, além de estudar estas áreas pesquisa estruturas de armazenamento em memória principal e secundária para as biossequências que sejam mais adequadas às aplicações de biologia computacional.

4 Distribuição e Integração de BDBMs

No capítulo anterior foi comentado o que são e porquê surgiram os BDBMs, a distribuição dos dados da biologia molecular e a necessidade da integração dos BDBMs. Neste capítulo serão apresentados requisitos que devem ser cumpridos e algumas suposições simplificadoras para a integração de BDBMs e alguns métodos que são utilizados para se integrar BDBMs.

4.1 Requisitos e Pressupostos de Integração

Procura-se nesta seção descrever o ambiente heterogêneo de fontes de informação de biologia molecular em termos de requisitos sobre as fontes de dados, as necessidades dos usuários e funcionalidades do ambiente de integração. Procura-se com esta descrição conhecer melhor o problema, que tem diversos aspectos, nem todos atendidos pela tecnologia atual de bancos de dados.

4.2 Características das Fontes de Dados

As fontes de dados de biologia molecular podem ser [DOB95]:

- arquivos com uma dada estrutura, que precisa ser conhecida para que se possa recuperar os dados (por exemplo, dados no formato ASN.1 e do GenBank);
- arquivos com dados em formato apropriado para troca de informações e que conta com interface gráfica para consulta (por exemplo, ACeDB);

- bancos de dados implementados via Sistemas Gerenciadores de Bancos de Dados (SGBD's), com modelos de dados relacional, orientado a objeto e relacional-objeto e interfaces de consulta bem definidas;
- arquivos com dados em formato apropriado para execução de determinadas aplicações (FASTA, BLAST).

Com o desenvolvimento de novas técnicas de experimentos na área da biologia molecular, novas leis e generalizações tem sido descobertas. Tal fato tem provocado mudanças radicais nos esquemas das fontes de dados. Mesmo que seja possível construir um esquema satisfatório que represente as necessidades da área, isto representa uma pequena parcela das informações biológicas. E mais, será também necessária a integração destas informações com outras não-biológicas, prevendo-se novas alterações de esquema advindas daí. Assim, o esquema das fontes de dados não é estático.

Fontes de dados são conectadas via Internet e devem ser capazes de atender a consultas complexas, embora algumas das existentes atualmente, não atendam a este requisito.

As atualizações feitas sobre uma fonte de dados local são restritas e controladas pelos seus mantenedores. É duvidoso supor que essa autonomia local seja abandonada para permitir maior flexibilidade nas transações, no sentido de suportar a implementação de atualizações globais. No entanto, os usuários priorizam o acesso aos dados mais recentes. Portanto, as atualizações são relevantes e devem ser feitas a tempo.

4.3 Necessidades dos Usuários

Os usuários tem necessidade de formular consultas complexas sobre a base de dados distribuída. Até recentemente, os usuários se satisfaziam em navegar através das fontes de dados e obter informações relacionadas a outras quase que por acaso. Muitos estão satisfeitos com os pacotes de *software* que utilizam, dotados de uma interface gráfica apropriada para a visualização de mapas do genoma em estudo. No entanto, a necessidade de análises avançadas sobre os dados exige facilidades de formulação de consultas complexas. Além disso, com os avanços tecnológicos na área de comunicação de dados, os usuários esperam que as respostas às consultas fiquem mais rápidas.

A interface comumente adotada para consultas consiste na apresentação de um formulário onde os usuários preenchem lacunas e opções. Por trás deste formulário simples, devem, entretanto, existir camadas de *software* capazes de suportar consultas arbitrárias feitas à base distribuída e heterogênea, complementadas por otimizadores capazes de fornecer, de forma eficiente, respostas às consultas ad-hoc formuladas.

Atualmente existem aplicativos com interface web que possibilitam a formulação de consultas a um conjunto pré-definido e limitado de bancos de dados. No entanto, os usuários não devem ser "forçados" a restringir o número de bancos a serem acessados por uma consulta.

Para a formulação de consultas, os usuários também não devem conhecer locais físicos, esquemas ou mesmo mecanismos de acesso às fontes de dados.

4.4 Ambiente de Integração

Ferramentas especiais de alto nível devem capturar as mudanças de esquema porventura existentes em cada banco componente do ambiente heterogêneo e incorporar estas mudanças no esquema global (devem gerenciar a heterogeneidade). [Kar95]

Interfaces especiais sofisticadas devem ser elaboradas de forma a facilitar a formulação de consultas complexas pela comunidade científica em geral.

Em resumo, a meta da pesquisa na área de biologia molecular é a de permitir aos usuários a interação, com uma série de fontes de dados, como se estivessem interagindo com apenas uma. As fontes de dados envolvidas na interação são aquelas que contém informações relevantes para a mesma. Estas fontes de dados estão distribuídas, são heterogêneas e foram implementadas com modelos de dados distintos. A interatividade acima descrita significa acesso via Web, formulação de consultas a objetos biológicos específicos, formulação de consultas complexas e mesmo atualizações envolvendo um ou vários objetos e relações biológicas.

4.5 Métodos de Interoperabilidade de Bancos de Dados

O objetivo da pesquisa de interoperabilidade em bancos de dados é permitir que os usuários interajam com um conjunto de bancos de dados desconectados e heterogêneos como se estivessem interagindo com cada banco de dados individualmente. "Interação" possui vários significados, como, por exemplo, procurar informação sobre um objeto em particular, executar consultas complexas, executar atualizações. Será apresentado a seguir uma breve descrição dos métodos de interoperabilidade de BDBMs e uma avaliação deles com relação aos requisitos que foram expostos anteriormente.

4.5.1 Método 1: Referências Cruzadas

Neste método, um registro de um banco de dados pode possuir uma referência a um outro registro de um outro banco de dados. Com este tipo de referência, tornou-se possível que um usuário obtivesse informações que estão relacionadas umas com as outras. Por exemplo, o biólogo encontrou uma seqüência muito parecida com a de seu interesse em um determinado banco de dados A. Analisando as informações desta seqüência, ele descobre que mais informações sobre ela estão armazenadas em um outro banco de dados B. Logo para completar sua pesquisa, o biólogo deve se conectar com este outro banco. Neste método o usuário tem que fazer muitas tarefas que não estão automatizadas.

4.5.2 Método 2: Navegação em Hipertexto

Este método permite aos usuários navegar de um registro de um banco de dados para outro registro de outro banco de dados, através de links entre os dois. Geralmente somente uma operação é suportada: procurar dentro de um banco de dados para encontrar um ponto de partida (como por exemplo recuperar um registro do GenBank usando o nome de uma proteína), e então ir para outro banco de dados através de link. Por exemplo, um registro do GenBank possui link para o registro do Medline associado a ele, por isso o usuário através do GenBank pode ver o registros do Medline que o interessarem.

4.5.3 Método 3: Data Warehouse

Neste método, um conjunto de bancos de dados heterogêneos são traduzidos e carregados fisicamente dentro de um único banco de dados chamado data warehouse. Para cada banco de dados que é integrado no data warehouse, é preciso definir um tradutor do formato e do conceito do banco de dados, para o formato e o conceito do repositório central. Os conceitos do banco de dados data warehouse precisam conter todos os conceitos dos bancos de dados componentes que são incluídos no warehouse. Por exemplo, este método poderia ser utilizado para carregar o SwissProt, PDB, e o PIR dentro de um grande banco de dados Oracle. Traduções precisam ser definidas entre os diferentes conceitos do SwissProt, PIR e PDB para um conceito do warehouse. Uma vez que todos os bancos de dados estão presentes no warehouse do Oracle, consultas arbitrárias podem ser aplicadas aos dados. O processamento de consulta é mais rápido em sistemas warehouse porque os dados são locais.

4.5.4 Método 4: Bancos de dados Heterogêneos Fracamente Acoplados

Esta técnica permite aos usuários construir consultas complexas que são avaliadas entre vários bancos de dados fisicamente distintos e heterogêneos. Uma consulta identifica explicitamente todos os bancos de dados componentes, todas as tabelas e atributos (no caso de SGBD relacional) que são consultados em cada banco. Uma simples consulta pode incluir referências a vários bancos de dados.

4.5.5 Método 5: Bancos de Dados Heterogêneos com Acoplamento Forte

Sistema de bancos de dados heterogêneos com acoplamento forte é composto por um conjunto de sistemas de bancos de dados componentes, heterogêneos, cooperativos mas autônomos, integrados de tal forma na federação que consultas e atualizações podem ser realizadas, de forma transparente à localização dos dados e aos caminhos de acesso. Tal transparência é obtida pela tradução dos diferentes esquemas dos componentes para um modelo de dados comum e integrado, compondo um esquema global. Todas as transações que envolvem mais de um banco de dados são definidas em termos do esquema global [Uch94].

O acoplamento forte paga um preço alto na autonomia por ter integração de esquema. A fim de participar da integração, os usuários de bancos de dados individuais freqüentemente têm que comprometer seu jeito de entender e representar a semântica. Como resultado, eles freqüentemente têm que lidar com representações que não são naturais e nem tão adequadas para suas aplicações. A manutenção dos esquemas torna-se muito difícil pelo uso da integração. Qualquer mudança em um esquema individual, deverá estar de acordo com todos os esquemas participantes da integração, o que freqüentemente requisitará reprojetar o esquema integrado e recodificar todas as aplicações dependentes dele [Qia93]. O esquema de integração federado não tem sido utilizado em bioinformática devido possivelmente às constantes mudanças nos esquemas locais determinados pela evolução das pesquisas, ao uso de diferentes modelos de dados e tecnologias, além da complexidade inerente à sua implementação.

Comentários Finais

Nessa seção foram apresentados os requisitos e suposições acerca da integração de BDBMs e os métodos de integração de BDBMs passíveis de utilização. Grande parte da complexidade de implementação dos métodos de integração é devida à necessidade de se ter um conhecimento aprofundado em biologia.

As similaridades semânticas e as diferenças esquemáticas são assuntos muito importantes para qualquer método que trate da interoperabilidade de bancos de dados, assim, a pesquisa atual na área tem tratado de aplicar ontologias e de construir ferramentas de tradução de esquemas.

Outro aspecto importantíssimo e não tratado neste trabalho diz respeito às anotações biológicas nas diversas fontes de dados da pesquisa. Tais anotações requerem ainda uma observância cuidadosa com relação à qualidade da informação disponível e a integração dos bancos de dados que contém tais anotações irá facilitar sobremaneira a execução desta tarefa.

5 Bancos de Dados de Biologia Molecular

Nesta seção são apresentados alguns exemplos de bancos de dados de biologia molecular, considerados mais expressivos para o exemplificar o texto. Assim, são detalhados os seguintes bancos de dados: GenBank, que armazena os dados em *flat files* no formato ASN.1, o GSDB que constitui um exemplo de implementação relacional e o ACeDB, que é um exemplo de um banco de dados implementado especificamente para abrigar esta aplicação e que utiliza um esquema orientado a objetos, com dados armazenados em formato XML.

São ainda apresentados os esforços mais significativos de integração de bancos de dados aplicados à biologia, de acordo com os métodos apresentados no capítulo anterior. Desta forma, são apresentados os sistemas SRS (com método de integração via *links*), IGD (que utiliza como método a construção de um *data warehouse*) e CPL/Kleisli (que tem acoplamento fraco).

5.1 Exemplos de BDBMs

5.1.1 GenBank

O GenBank é hoje o mais importante repositório amplo de sequências de nucleotídeos. É usado como referência no sentido de verificar se uma dada sequência já está catalogada. O histórico do volume de sequências armazenadas no GenBank demonstra que, a cada ano, o número de sequências armazenadas, bem como o número de bases, cresce cerca de 70% por ano. A cada ano novas versões da base são disseminadas. Cada versão pode ter alteração na quantidade de informações armazenadas, bem como a inclusão ou alteração de atributos, ou mesmo a inclusão ou alteração de conceitos biológicos.

O GenBank mantém arquivos contendo estruturas ASN.1. Tais estruturas implementam um tipo de modelo de dados semi-estruturado, bastante útil para troca de informações com a comunidade científica. Segue-se um exemplo de descrição do formato ASN.1 para a entidade de dados “Publicações” do GenBank, utilizando a notação em CPL [BDH+95] .

```
Publications={[title: string,
  author: { || [name: string, initial: string] || },
  journal: < uncontrolled: string,
             controlled: < medline-jta: string,        % Medline journal title abreviation
                           iso-jta: string,            % ISO journal title abreviation
                           journal-title: string,      % Full journal title
                           issn: string > >            % ISSN number
  volume: string,
```

issue: string,

year: int,

pages: string,

abstract: string,

keywd: { string }] }

A notação utilizada no exemplo descrito anteriormente é apresentada a seguir.

Descrição dos tipos	**Notação**	**Terminologia ASN.1**
Lista	{ || *G*|| }	Sequência de
Conjunto	{ *G* }	Conjunto de
Registro	[*l*1: *G*1,…, *l*n: *G*n]	Sequência
	(campos rotulados)	
Variante	< *l*1: *G*1,…, *l*n: *G*n>	Escolha

(atributos de estruturas, do tipo union da linguagem C, rotulados)

Esquema e evolução

O Genbank armazena sequências de nucleotídeos e proteínas, além de informações biológicas relevantes sobre cada sequência, que são, por exemplo, o nome científico e a taxonomia do organismo de origem, um conjunto de anotações que especificam regiões codificantes na sequência e também outras regiões de relevância biológica. Nestas anotações estão incluídas ainda informações sobre as proteínas sintetizadas nas regiões codificantes que foram anotadas (função, estrutura, etc.). Um registro do GenBank é identificado pelo atributo número de acesso. A seguir é apresentado um exemplo de registro do GenBank . Cada registro possui rótulos que definem a informação que está armazenada.

LOCUS ABCRRAA 118 bp ss-rRNA RNA 15-SEP-1990

DEFINITION Acetobacter sp. (strain MB 58) 5S ribosomal RNA, complete sequence.

ACCESSION M34766

KEYWORDS 5S ribosomal RNA.

SOURCE Acetobacter sp. (strain MB 58) rRNA.

ORGANISM Acetobacter sp.

Prokaryotae; Gracilicutes; Scotobacteria; Aerobic rods and cocci;

Azotobacteraceae.

REFERENCE 1 (bases 1 to 118)

AUTHORS Bulygina,E.S., Galchenko,V.F., Govorukhina,N.I., Netrusov,A.I.,

Nikitin,D.I., Trotsenko,Y.A. and Chumakov,K.M.

TITLE Taxonomic studies of methylotrophic bacteria by 5S ribosomal RNA

sequencing

JOURNAL J. Gen. Microbiol. 136, 441-446 (1990)

FEATURES Location/Qualifiers

rRNA 1..118

/note="5S ribosomal RNA"

BASE COUNT 27 a 40 c 32 g 17 t 2 others

ORIGIN

1 gatctggtgg ccatggcggg agcaaatcag ccgatcccat cccgaactcg gccgtcaaat

61 gccccagcgc ccatgatact ctgcctcaag gcacggaaaa gtcggtcgcc gccagayy

Os rótulos referem-se às seguintes informações biológicas:

Locus: nome curto escolhido para sugerir a definição da sequência.

Definition: descrição concisa da sequência.

Accession number: número de acesso primário, um valor único e imutável atribuído para cada sequência.

Nid: identificador único da sequência ácido-nucléica que é atribuído pelo NCBI ao registro de sequência do GenBank. Enquanto o *accession number* é uma chave de recuperação única para um registro no banco de dados, mesmo que alguma modificação tenha sido feita, o *Nid* muda sempre que uma sequência é modificada.

Keywords: palavras-chave associadas ao gene ou a outras informações sobre o registro.

Segment: informações sobre a ordem em que este registro aparece na série de sequências descontínuas de uma mesma molécula.

Source/Organism: O campo *Source* consiste de duas partes. A primeira parte é encontrada depois do rótulo *Source* e contém o nome do organismo onde a sequência foi encontrada. A segunda parte consiste de informações encontradas depois do rótulo secundário *Organism*. Ela possui o nome científico formal do organismo (gênero e espécie, onde foi catalogado) seguido por sua taxonomia.

Reference: citações a todos os artigos que contêm dados sobre este registro. Ele é composto pelo número da referência e o local das bases na sequência citada e por mais cinco partes: *Authors*, *Title*, *Journal*, *Medline*, e *Remark*.

Authors: lista os autores na ordem em que eles aparecem no artigo citado.

Title: título da publicação.

Journal: citação da literatura para o registro da sequência. A palavra ´*Unpublished*´ aparecerá depois do rótulo secundário *Journal* se os dados não aparecerem na literatura científica, mas foi diretamente depositado no banco de dados. Para as sequências publicadas a linha *Journal* contém a tese, a revista, ou o livro, incluindo o ano de publicação.

Medline: identificador único da *National Library of Medicine's Medline* para a citação (se conhecida).

Remark: comentário que especifica a relevância da citação do registro.

Comment: referências cruzadas para outras sequências, comparações com outras coleções, anotações de modificações no nome do *Locus* e outras observações.

Features: tabela que contém características encontradas em determinados sítios da sequência.

Base Count: sumário do número de ocorrências de cada código base na sequência.

Origin: especificação de como a primeira base da sequência relatada está localizada dentro do genoma. Isto possivelmente inclui sua localização dentro de um grande mapa genético.

Sequence: informa a sequência de nucleotídeos.

O Genbank passou por diversas alterações de esquema, cada uma delas para representar novas informações, tais como:

- representação de sequências de proteínas, a partir das de nucleotídeos que estão armazenadas no banco.
- dados de genes, observados nas sequências, que foram armazenados no formato EST (*Expressed Sequence Tags*).
- informações biológicas relevantes sobre uma sequência (e não apenas genes), que foram armazenadas no formato STS (*Sequence Tagged Site*).
- informações de sequências obtidas através de um processo de sequenciamento específico, que foram armazenadas no formato HTGS (*HighTthroughput Genomic Sequence*).
- informações de mutações de genes, no formato SNP (*Single Nucleotide Polimorphisms*).
- taxonomia.
- estrutura tridimensional de proteínas.
- *links* para a literatura (MEDLINE).

A cada alteração de esquema, os dados são atualizados, sendo que a sequência recebe um novo identificador (número de acesso). O identificador anterior é armazenado de forma a não se perder a referência anterior. Tal fato tem como objetivo permitir que os usuários acostumados a utilizar um conjunto de identificadores de sequências em suas pesquisas não necessitem atualizar tais identificadores a cada mudança de esquema.

Arquitetura do ambiente do GenBank

A submissão de sequências ao banco é feita através dos seguintes programas:

- BankIt, interface de submissão via Web.
- Sequin, software *stand alone* de interface de submissão via Mail.
- Existem também serviços *batch* para envio de sequências ao banco em formatos especiais, que são: EST (*expressed sequence tags*), STS (*sequence tagged site*) e HTGS (*high throughput genomic sequence*). Essas submissões geram o arquivamento das sequências em bancos de dados específicos.
- O GenBank tem, à parte, um banco de dados de mutações denominado SNP (*single nucleotide polymorphism*) onde é possível submeter sequências a esta base.

O GenBank troca dados com os bancos EMBL, DDBJ e GSDB de forma a manter o repositório de sequências o mais completo possível. Os dados do GenBank são

disponibilizados via WWW, rede local ou mesmo via execução local, cujo código é obtido por FTP, através das seguintes ferramentas:

- aplicativo Entrez, que consiste de uma interface de integração dos dados de sequências com dados de outros bancos contendo informações referentes à taxonomia, estrutura 3-D, população e *genome assembly*. Também são disponibilizados, através desta interface, dados de publicações relativas às sequências.
- similaridade de sequências, que é disponibilizada por um conjunto de programas que executam o algoritmo básico BLAST.
- buscas nos bancos de dados especializados dbEST, dbSTS e dbGSS (*Genome Survey Sequence*).

O mecanismo de consulta ao GenBank é dado através do aplicativo Entrez, que tem versão WWW. A consulta pode ser feita via atributos "palavra-chave", "sequência" e "UID". Não é permitido o acesso às estruturas do banco via *browse*. Um usuário comum do banco não acessa diretamente as suas estruturas, via SQL ou outras funções.

Existem mecanismos de exportação de dados que permitem aos usuários receber as sequências solicitadas em formato texto, ou mesmo a base completa em arquivo no formato ASN.1. O formato ASN.1 é usado para gerar estruturas de dados estáticas da linguagem C, a serem compiladas com as aplicações (por exemplo Entrez). Desta forma a interface é periodicamente modificada para acomodar mudanças no esquema do banco ou mesmo novos tipos de consultas. O mesmo pode ser feito com aplicações dos usuários.

Integração com outros bancos de dados

O GenBank conta com uma aplicação (Entrez) que implementa a integração entre diferentes bancos de dados, através de consultas baseadas em formulários. Ao se acessar o aplicativo, uma página www dinâmica é apresentada, onde é possível selecionar o banco a ser pesquisado, segundo critérios que são informados. Os dados resultantes da consulta podem ser utilizadas para uma consulta posterior.

Os bancos de dados que participam da integração são:

- Nucleotide - sequências derivadas do GenBank.
- Protein - proteínas derivadas de sequências do GenBank.
- Genome - montagens de código genético.
- Structure - estruturas 3-D de proteínas.
- PopSet - sequências de populações.
- PubMed - dados bibliográficos do MEDLINE e de outros bancos.

Cabe ressaltar que todos os bancos que participam da integração tem *links* entre si.

5.1.2 GSDB

O GSDB é um banco de dados relacional, implementado em Sybase, e se dedica a dar suporte à pesquisa científica através da criação, manutenção e distribuição de uma coleção de sequências de DNA e de informações correlatas. Em cooperação com os maiores repositórios de sequências de DNA (DDBJ, EMBL e GenBank), o GSDB permite o acesso e coleciona dados diretamente dos autores de diversas maneiras, incluindo as mais novas formas de acesso aos dados advindas das necessidades de sequenciamento em larga escala, a saber:

- direta atualização da base de dados. Centros de pesquisa que utilizam o SGBD Sybase podem implementar aplicações que atualizem diretamente a base de dados, utilizando um acesso cliente-servidor. Neste caso, o centro de pesquisa é responsável pela qualidade da informação armazenada.
- via World Wide Web. O servidor Web oferece diversos mecanismos de acesso, inclusive consultas ad-hoc em SQL. No caso de atualização da base, os dados submetidos passam por um processo de controle de qualidade do GSDB.
- cópia da base. Os centros de pesquisa que dispõem de uma licença do tipo cliente do Sybase podem acessar uma cópia *read-only* da base, utilizando tanto as ferramentas de acesso providas pelo SGBD como programas específicos para tal.

A evolução do GSDB teve os seguintes marcos:

- em 1979, início de operação no Los Alamos Sequence Library.
- de 1982 a 1992, operou como GenBank. A base de dados relacional foi implementada em 1989.
- em 1993, tornou-se Genome Sequence DataBase.
- em 1994, a base foi para o National Center for Genome Resources.
- em 1996, gerada uma nova versão da base, denominada 1.0.

Esquema e evolução

O GSDB armazena informações sobre sequências, publicações e membros da comunidade científica. Tais informações estão também disponíveis no GenBank. Estas bases de dados trocam informações diariamente no sentido de compatibilizar os respectivos conteúdos. Para modelar a base de dados, foi utilizado o modelo de entidades e relacionamentos [Che76]. Assim, por exemplo, a entidade sequência está relacionada à entidade gene e o tipo de relacionamento é de um-para-muitos. Da mesma forma, uma sequência pode constar de diversas publicações, cada uma elaborada por diversos autores. Um autor pode também participar de inúmeras publicações. O relacionamento entre as entidades sequência e publicações também é do tipo um-para-muitos, enquanto que o relacionamento entre as entidades publicações e autores é do tipo muitos-para-muitos.

A seguir, é apresentada uma breve descrição dos enfoques que influenciaram a evolução do modelo de dados do banco, a saber:

- o modelo tradicional de bancos de dados científicos (entrada de dados de sequências via citações em publicações científicas),
- o modelo de publicação eletrônica de dados (entrada de dados de sequências via submissão direta feita por laboratórios de sequenciamento ou por pesquisadores),
- anotações da comunidade científica (possibilidade de anotações de informações adicionais sobre as sequências feitas pela comunidade científica), e
- o modelo de banco de dados federados (divisão da base em três, uma contendo os dados das sequências, outra de publicações e a terceira de membros da comunidade científica). No caso, a dita federação é local e são mantidos *links* com outros bancos de dados.

A primeira implementação do GSDB foi baseada no modelo tradicional. Neste modelo, os dados de sequências, de artigos e de membros da comunidade científica eram coletados a

partir das publicações científicas e armazenados em arquivos do tipo texto. Estes arquivos eram então disponibilizados para a comunidade. Em 1986 a geração de sequências cresceu acima da capacidade administrativa do GSDB, que ficou impossibilitado de acompanhar tal crescimento. Além disso, as próprias editoras passaram a limitar a quantidade de novas sequências a serem publicadas. Desta forma, a informação contida no banco de dados ficaria incompleta se não ocorresse uma mudança no modelo de captura de informações.

O novo modelo foi denominado publicação eletrônica de dados. Neste modelo, os pesquisadores comunicam as suas descobertas diretamente ao banco de dados e são responsáveis por assegurar a qualidade da informação. Desta forma, a administração do GSDB trocou a função de coleta e garantia de qualidade dos dados por outras. Ficou responsável pela manutenção da estrutura do banco, pelo desenvolvimento de novas ferramentas de software, pelo projeto dos novos processos de obtenção dos dados e pelo suporte aos usuários. Assim, em 1987 o banco de dados passou a ser suportado por um SGBD relacional e a permitir a submissão de sequências via processo batch. O processo batch foi escolhido porque poucos membros da comunidade científica tinham acesso à Internet.

Em 1992 novas necessidades surgiram. Foi necessário reduzir a intervenção da equipe na base, no sentido de submissão manual de sequências e no suporte à comunidade para adição e correção de dados de sequências e de anotações biológicas. Assim, em 1994 houve novo re-projeto do banco de dados para suportar as seguintes necessidades:

- alteração das informações, de forma on-line (via Internet), pela comunidade científica, em substituição ao processo batch existente,
- facilidades de inclusão de novas anotações por pessoas da comunidade, que não aquelas que submeteram a sequência, de forma a se ter uma completa caracterização das mesmas,
- facilidades de modularização dos serviços e suporte a links com outros bancos de dados, de forma a se criar uma federação de serviços de genoma. Este suporte baseia-se na concepção e implementação de um banco de dados federativo que minimize o escopo de cada banco de dados participante da federação. Os mantenedores dos bancos de dados gastam recursos substanciais para armazenar informações adicionais sobre as sequências, tais como: taxonomia, genes e dados bibliográficos. O princípio básico da federação trata do armazenamento das sequências nos bancos de dados principais, com links para estas informações adicionais.

O novo esquema do banco foi criado em 1995 e aperfeiçoado em 1996, gerando a versão 1.0, que, de forma sucinta, contém as seguintes características:

- adoção de critérios de segurança e qualidade dos dados. O ponto central da segurança é o critério de propriedade. O usuário que inserir um dado no banco é o dono daquela informação e só ele pode modificá-la (os administradores do banco também podem fazê-lo). Outros usuários podem acessar o dado para leitura, desde que este seja um dado público. O dono da informação informa a privacidade do dado: público ou privado. Um software especial do SGBD verifica a qualidade dos dados públicos e os disponibiliza para a comunidade.
- inclusão de novos tipos de dados, como por exemplo a representação de alinhamentos de múltiplas sequências, sequências descontínuas (com informações sobre gaps), dados confidenciais de sequências e resultados de análises. O esquema do banco permite também que se represente coleções de elementos específicos do banco (grupos de

sequências, componentes de alinhamentos, coleções e unidades de publicação eletrônica).

Arquitetura do ambiente e interface com os usuários

O GSDB está implementado em um sistema de gerência de banco de dados comercial. Uma camada de *software* que permite a visualização dos objetos do banco está disponível aos usuários e esta camada faz acesso ao banco. Os usuários e desenvolvedores acessam o SGBD e suas aplicações diretamente (via SQL) ou através da camada de objetos.

5.1.3 AceDB

O ACeDB (A Caenorhabditis elegans Data Base) é um sistema de gerência de banco de dados que além de armazenar os resultados de projetos de sequenciamento e mapeamento de em larga escala, permite representar dados de experimentos genéticos de uma forma bastante flexível. O ACeDB foi criado por Richard Durbin (Sanger Centre, Cambridge - UK) e por Jean-Thierry-Mieg (CNRS, Montpellier – FR). O nome ACeDB além de indicar o software de gerenciamento de banco de dados, indica também a base de dados resultante do sequenciamento do DNA do nematóide C. elegans.

O *software* consiste de:

- módulo de gerenciamento de banco de dados central (*kernel*), com dados baseados em um modelo flexível, projetado especificamente para manipular informações biológicas,
- módulo de interface com os usuários, dotado de recursos gráficos e que tem telas específicas para representar tais informações,
- conjunto de ferramentas que lidam com informações biológicas (por exemplo, o software *gene finder*, desenvolvido por Phil Green, na *Washington University, St. Louis*).

O software ACeDB tem sido mantido pelos seus desenvolvedores. Como seu código fonte é distribuído gratuitamente, diversos pesquisadores tem feito implementações adicionais, que são tornadas públicas. Assim, a comunidade científica que utiliza tal ferramenta tem se beneficiado de constantes atualizações e de novas implementações.

O ACeDB é portanto uma ferramenta genérica, que é utilizada por diversos laboratórios, para armazenamento de resultados de sequenciamentos de diversos organismos: bactérias, fungos, plantas e mesmo de alguns cromossomos humanos.

Esquema e evolução

O ACeDB é um sistema orientado a objetos. Para os biólogos, esta forma de representação dos dados é mais intuitiva que a utilizada em tabelas relacionais. Cada objeto é representado por um único identificador, o seu nome, e contém diversos atributos organizados sob a forma de uma árvore. Os nós da árvore são também nomeados e apontam para outros objetos ou são folhas e contém valores, que podem ser numéricos ou cadeias de caracteres. Assim, o modelo é flexível porque permite, com facilidade, a adição de novos nós, em substituição às folhas da árvore. Cada objeto é alocado a uma classe e, através desta representação, é possível a construção de sub-classes de objetos. Comentários podem ser adicionados em qualquer ponto da árvore.

Cada classe tem portanto uma estrutura de dados em forma de árvore, onde está delimitada a altura de cada sub-árvore e os tipos de dados ou sub-classes que são permitidos em cada posição. A esta estrutura de dados é dado o nome de “modelo”. Objetos são instâncias das classes e, em geral, seus dados não contém todas as informações válidas e possíveis da estrutura. Esta representação traz as seguintes vantagens:

- objetos ainda pouco estudados podem ser representados pois ramos da árvore com informações ainda desconhecidas, podem estar ausentes. Mesmo que tais objetos sejam numerosos no banco de dados, ocupam pouco espaço em disco e em memória, aumentando a eficiência do sistema.
- se houver necessidade de extensão do esquema, fato que é bastante comum e frequente na área, basta alterar a estrutura com a extensão desejada. Cabe observar que todos os dados que existiam na base permanecem válidos. Apenas não contém informações sobre a extensão feita.
- é possível a inclusão de anotações biológicas relevantes sobre os dados (na forma de comentários), sem afetar os algoritmos de busca de informações.

Os desenvolvedores do ACeDB, de forma deliberada, evitaram a implementação da herança múltipla mas permitiram que dois objetos possam ter sub-árvores comuns. Por exemplo considere a representação de dois objetos do tipo Gene, um estudado através da genética clássica (não-clonado) e outro obtido por similaridade com uma proteína de outro organismo (clonado). Estas instâncias podem ser consideradas como arquétipos de duas sub-classes da classe Gene.

No ACeDB os objetos são representados em duas classes: a classe tipo B, que representa objetos na forma de árvore e a classe tipo A, que representa objetos como arrays de dados, forma esta de representação das sequências de DNA.

A razão do sucesso do ACeDB está nesta representação flexível do esquema do banco, que permitiu a sua adoção para armazenamento de dados do sequenciamento de diversos organismos, bastando adequar a estrutura (modelo) dos dados às informações que se deseja representar.

Para a definição do modelo de dados, O ACeDB conta com uma linguagem própria (*Data Definition Language*). Para exemplificar a linguagem é apresentada a seguir uma parte da definição da classe Gene e um exemplo de uma instância da classe.

// definição da classe Gene

```
?Gene  Reference_allele          ?Allele
       Molecular_information     Clone   ?Clone XREF Gene
                                 Sequence ?Sequence XREF Gene
       Map Physical pMap UNIQUE ?Contig XREF Gene UNIQUE Int
                 Autopos
           Genetic gMap ?Chromosome XREF Gene UNIQUE Float UNIQUE Float
                 Mapping_data    2point ?2point_data
                                 3point ?3point_data
```

```
        Location  ?Laboratory  #Lab_Location
?Lab_Location  Freezer   Text
                Liquid_N2 Text
```

// instância da classe Gene

```
ced-4   Reference_allele        n1162
        Molecular_information           Clone   MT#JAL1
        Map Genetic gMap III -2.7
                Mapping_data 2point "ced-4 unc-32/+ +"
        Location Cambridge Freezer A6
```

O ACeDB representa internamente os dados em forma de árvore, em formato binário. A entrada dos dados (e saída) é feita via arquivos ASCII denominados “ACE files”, onde as informações são representadas de acordo com uma sintaxe específica, semelhante à XML [XML98]. A seguir, é apresentado um exemplo de arquivo de entrada de dados do ACeDB. No exemplo dado, é definida a sequência de nome ACT3, com título, referência à base EMBL e o seu DNA. Em seguida, no mesmo arquivo é mostrada a forma de atualização dos dados armazenados através da troca do nome de uma sequência de zk643 para ZK643 (se a primeira existir).

```
// definição de uma sequência
Sequence ACT3
Title ``C. elegans actin gene (3)"
Library EMBL CEACT3 X16798
// DNA correspondente à sequência (classe A)
DNA ACT3
aagagagacatcctcccgctcccttcccacacccacttgctcttttctat
tgaccacacattatgaagataaccatgttactaatcaaattcgtgttctt
ttccaatttctttttc
// troca do nome de uma sequência
-R Sequence zk643 ZK643          // R significa “rename”
```

O software conta também com uma linguagem de consulta própria denominada AQL (ACeDB Query Language) que foi projetada de acordo com os conceitos utilizados nas linguagens OQL [Cat94] (proposta pelo ODMG para linguagem de consulta a bancos de dados orientados a objetos), Lorel (linguagem de consulta a dados semi-estruturados no sistema Lore, desenvolvido em *Stanford* [GMW+97]) e Boulder (http://stein.cshl.org/software/boulder/) sistema de acesso aos dados via valor de atributo, desenvolvido por Lincoln Stein para o *Whitehead Genome Center*).

Integração com outros bancos de dados

O ACeDB não tem integração com outros bancos de dados. No entanto nada impede que uma dada definição de um objeto (modelo do objeto) tenha ponteiros para objetos de outros bancos de dados. Além disso, como o formato dos arquivos de entrada e saída são bem definidos, isto é, contam com uma sintaxe própria, suas informações são, de certa forma, apropriadas para integração. Faltando aliar o componente semântico através, por exemplo, da adoção de uma ontologia.

Interface com os usuários

O ACeDB permite acesso às informações da base via interface textual e gráfica. No entanto, é bastante aceito na comunidade científica em virtude de sua interface gráfica, que apresenta, para os usuários as informações biológicas em um formato bastante apropriado. As telas gráficas disponíveis incluem a exibição do mapa genético, do mapa físico e da sequência. O ACeDB permite também a adição de imagens aos dados, assim, é possível apresentar por exemplo a imagem do gel, entre outras. O mapa genético informa os sítios geneticamente relevantes (por exemplo, posições de mutações ou de marcadores moleculares). O mapa físico fornece uma visão de superposições de sequências no código genético, via contigs, sequências e marcadores. A exibição de dados de sequências contém regiões codificantes (genes), regiões de similaridade e regiões promotoras, entre outras. O ACeDB também exibe o gel resultante do sequenciamento.

5.2 Integração de Bancos de Dados de Biologia Molecular

5.2.1 SRS - Sequence Retrieval System

O sistema SRS (http://expasy.cbr.nrc.ca/srs5) é um exemplo de integração utilizando *links*. Integra mais de trinta e cinco bases de dados com informações biológicas através da implementação de *links* entre objetos que compõem estas bases. A lista completa está em http://srs.ebi.ac.uk/srs5list.html.

O sistema permite a formulação de consultas através de uma linguagem própria. A linguagem foi inicialmente projetada no sentido de interpretar informações em bancos de dados que utilizam arquivos texto como forma de armazenamento (*flat files*). A sintaxe de cada um é descrita em camadas. Inicialmente são descritos os registros de um banco, em seguida as estruturas de dados de cada registro e finalmente os ítems de dados que compõem as estruturas.

Na linguagem pode-se especificar que bancos procurar e sobre que atributos efetuar as consultas. Por exemplo, a consulta: “Selecione o atributo DEFinition no banco de dados PIR, onde o valor do atributo = elastase”, seria expressa como:

[pir-def:elastase]

Os comandos da linguagem podem ser embutidos na linguagem C, através de uma API especialmente desenvolvida, tornando a ferramenta bastante útil.

Exemplo de uso da API:

```
#include <stdio.h>
#include "srs.h"

int main ()
{
  SrsEnv ();
  LibOpen ();

  if (Query ("[swissprot-def:elastase]", "Q1"))
    printf ("query Q1 found %d entries\n", SetSize ("Q1"));
}
```

Este método de integração é bastante popular entre os pesquisadores em biologia molecular e existem diversas implementações baseadas em *links*.

Em algumas implementações, os *links* são percorridos no sentido de se atender a uma dada consulta e, neste caso, existe perda de significado semântico no percurso. Por exemplo, o banco LinkDB utiliza os percursos via links para atender a consultas. Ao se questionar, neste banco, “quais as publicações que se referem a uma dada proteína ?”, o atendimento à consulta procura pela proteína no banco de dados Swiss-Prot, porém este banco não tem links para o banco Medline (de publicações), não permitindo o acesso de forma direta. Assim, por exemplo, pode-se caminhar via GenBank, que pode ser percorrido via Swiss-Prot (a proteína tem link com as sequências onde aparecem), e que tem links para o Medline. O problema aparece quando a sequência do GenBank tem mais de uma proteína anotada pois a resposta das publicações pode ser referente a uma outra proteína da sequência e não aquela que deu origem à consulta.

5.2.2 IGD

O Integrated Genomic Database (http://igd.rz-berlin.mpg.de/~www/lpi.html) [Rit94] é um exemplo de um data warehouse de biologia molecular [Mar95].

O IGD provê um esquema comum para os bancos de dados subjacentes, uma interface de usuário gráfica popular (AceDB) e facilidade de consulta. Como a maioria das atualizações aos bancos de dados ACeDB são feitas através de arquivos textos e não através de sistemas de gerenciamento de transações esperado na maioria dos SGBDs, atualizações diárias só são eficientes porque não é feita muita checagem de restrição [DOB95].

Neste método a integração reside fisicamente em um local, pode ser consultada sem acesso remoto a banco de dados e por isso permite acesso rápido aos dados. No entanto, é possível imaginar um cenário onde consultas em um esquema virtual IGD são traduzidas em consultas em dados originados dos bancos de dados bases do IGD [DOB95].

O custo de manutenção deste sistema é muito alto. Não está claro quais ferramentas foram construídas no projeto do IGD para tratar da evolução dos esquemas e dos dados. Atualmente, os BDBMs possuem tamanhos modestos e o recarregamento do banco de dados é praticável. Mas isto não será verdade no futuro onde atualizações incrementais serão inevitáveis. Nesta hora, questões como evolução de esquemas, manutenção do nível dos dados, e manutenção de tabelas de ligação serão predominantes [DOB95].

5.2.3 CPL/Kleisli

O sistema CPL/Kleisli tem como método de integração o acoplamento fraco. Foi desenvolvido por um grupo da University of Pennsylvania [BDH+95] [HWO+94] [Won94]. Sua implementação, chamada Kleisli, inclui uma poderosa linguagem de consulta chamada CPL que modela complexos tipos de dados de bancos de dados tais como listas, conjuntos, registros, e variações usadas em ASN.1 [IOS87]. CPL pode expressar consultas em tais tipos de dado, e pode codificar regras de tranformações entre tipos de dados, tais como projeções para simplificação de tipos complexos. Kleisli tem sido usada para responder com sucesso uma das consultas consideradas desafios pela DOE Informatics Summit [Rob94]: "Encontre informação nas seqüências de DNA conhecidas de um cromossomo humano 22, assim como as informações de seqüências homólogas de outros organismos". Kleisli responde tal consulta combinando informações de localização de cromossomo de um servidor Sybase GDB com seqüências e dados homólogos do servidor GenBank Entrez (ASN.1) [BDH+95] [Kar95].

O sistema CPL/Kleisli [BDH+95] suporta consultas ad hoc formuladas sobre bancos de dados distribuídos e heterogêneos. Hoje o sistema tem sido usado para integrar recursos autônomos, somente de leitura, através de visões de usuários(mediadores). Neste modo, CPL/Kleisli oferece as seguintes vantagens: uma interface uniforme para sistemas heterogêneos, construção barata, e manutenção relativamente barata de consultas complexas entre os múltiplos bancos de dados; tratamento uniforme dos recursos heterogêneos e de algoritmos de análises do banco de dados (ex. BLAST); otimização de consultas distribuídas incluindo paralelismo e *lazy evaluation*; um sistema de tipos necessário para a integração de recursos heterogêneos; e modularização dos drivers de dados para acesso aos recursos distribuídos [DOB95].

No entanto, existe uma desvantagem significativa no estilo do mediador desta integração. Experimentos com o sistema CPL mostraram que o sistema de rede existente é muito frágil e muito lento para permitir tempos de respostas adequados para muitas consultas distribuídas. É claro que isto depende fortemente do recurso em particular que está sendo acessado; consultas no servidor Entrez são intoleráveis, enquanto que as consultas nos sistemas de bancos de dados relacionais locais são rápidas, robustas e podem ser paralelizadas para obter significativas melhoras na performance. Além disso, enquanto atualizações em sistemas individuais subjacentes podem ser executadas dentro do sistema CPL-Morphase [DHK95], atualizações a nível global ainda não são suportadas [DOB95].

6 Comentários Finais

A área de bioinformática é hoje em dia uma das mais interessantes e importantes da computação, trazendo várias novas questões e problemas em aberto para os pesquisadores da área. Com a atual "corrida" de sequenciamento e a evolução tecnológica, o volume de dados é hoje bastante considerável e tende a crescer muito nos próximos anos. Assim, é natural que os SGBDs sejam sistemas pensados em prover suporte ao armazenamento e acesso eficientes aos dados.

Nesse trabalho foram descritos alguns aspectos relacionados ao uso de sistemas de bancos de dados, em particular as análises de sequências e respectivos algoritmos e os problemas de integração das bases. Além disso, foram apresentados alguns projetos e bancos de dados dentre os mais representativos discutidos na literatura.

Existem alguns projetos de pesquisa sendo realizados em todo mundo, no Brasil em particular, mas na área de banco de dados ainda há relativamente poucos resultados obtidos. O foco principal tem sido na integração das bases de dados e todos os aspectos relacionados. Porém, há vários outros problemas interessantes em aberto, entre eles, a própria definição do modelo de dados mais apropriado.

Outro assunto interessante que tem interessado nosso grupo de pesquisa diz respeito aos esquemas de armazenamento e gerência de memória para lidar com as biossequências. Como em outras áreas para os quais SGBDs foram especializados – como por exemplo, SGBD espaciais, temporais, etc. - é possível que se pense em estruturas de armazenamento melhor adaptadas ao contexto da aplicação. Entre outros temas de pesquisa há o estudo de possíveis índices (ou filtros) para uso nos algoritmos de análise e comparação de sequências. Quanto à gerência de memória, pode-se pensar em estruturas e métodos para disponibilizar o maior número de biossequências para utilização nos programas de comparação.

Independente de ser uma área relativamente nova e na qual poucos pesquisadores têm tido acesso à bibliografia e aos dados e processos, é fato que se trata de uma área das mais promissoras para a computação, em particular a área de banco de dados. Sem o devido suporte de SGBDs "científicos", o volume de dados esperado no banco de dados é tão grande e complexo que poderia vir a inviabilizar todo o esforço de sequenciamento feito até agora.

Referências

[ABS00] S. Abiteboul, P. Buneman e D. Suciu. *"Data on the Web - From Relations to Semistructured Data and XML"*. Morgan Kaufmann, 2000.

[AG97] M.Ashburner, N.Goodman *"Informatics – Genome and Genetics Databases"*, Current Opinion in Genetics & Development, 1997, 7:750-756.

[AGM+90] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, e D. J. Lipman. *"A basic local alignment search tool"*. J. of Molecular Biology 215, pp. 403-410, 1990.

[AMS+97] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, e D.J.Lipman.*"Gapped blast and psi-blast:a new generation of protein database search programs"*. Nucleic Acids Research, 25(17), pp.3389-3402, 1997.

[BBC+00] W. Baker, A. van den Broek, E. Camon, P. Hingamp, P. Sterk, G. Stoesser, M. Ann Tuli. *"The EMBL Nucleotide Sequence Database"*. Nucleic Acids Research 28(1), pp. 19-23, 2000.

[BDO95] P.Buneman, S.B. Davidson, C.Overton. *"Challenges in Integrating Biological Data Sources"*. Journal of Computational Biology 2 (4), pp.557-572, 1995.

[BDH+95] P.Buneman, S.B.Davidson, K.Hart, G.C.Overton, L.Wong. *"A Data Transformation System for Biological Data Sources."* Proceedings of 21th International Conference on Very Large Data Bases, pp 158-169, 1995.

[BGH+00] W. C. Barker, J. S. Garavelli, H. Huang, P. B. McGarvey, B. C. Orcutt, G. Y. Srinivasarao, C. Xiao, L. L. Yeh, R. S. Ledley, J. F. Janda, F. Pfeiffer, H. Mewes, A. Tsugita, C. Wu. *"The Protein Information Resource (PIR)"*. Nucleic Acids Research 28(1), pp. 41-44, 2000.

[BML+00] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, D. L.

Wheeler. *"GenBank"*. Nucleic Acids Research 28(1), pp. 15-18, 2000.

[BWF+00] H. M. Berman, J. Westbrook, Z. Feng, G. Gillil, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. *"The Protein Data Bank"*. Nucleic Acids Research 28(1), pp. 235-242, 2000.

[Cas92] Denise Casey. *"Primer on Molecular Genetics"*. HGP, U.S. Department of Energy, 1992. http://www.ornl.gov/hgmis/publicat/primer/intro.html.

[Cat94] *"The Object Database Standard: ODMG-93"*, Cattell R.G.G., San Francisco: Morgan Kaufmann, 1994.

[CM95] I.A.Chen, , V.M Markowitz. *"An Overview of the Object-Protocol Model (OPM) and the OPM Data Management Tools"*. Information Systems, 20(5):393-418.

[DHK95] S. Davidson, C. Hara, A. Kosky. *"Morphing sparsely Populated Data"*. Julho, 1995. Disponível em http://www.cis.upenn.edu/~kosky/mimbd95.html.

[DOB95] S.B.Davidson, C.Overton, P.Buneman. *"Challenges in Integrating Biological Data Sources"*. Julho, 1995. http://db.cis.upenn.edu/Publications/.

[DOE00a] U.S. Department of Energy. *"Human Genome Research"*. http://www.er.doe.gov/production/ober/hug_top.html, 2000.

[DOE00b] U.S. Department of Energy. *"Human Genome Project Information"*. http://www.ornl.gov/TechResources/Human_Genome/home.html, 2000.

[Doo90] Russel F. Doolittle, editor. *"Molecular Evolution: Computer Analisys of Protein and Nucleic Acid Sequences."* Methods in Enzymology. Academic Press 183, 1990.

[Fio00] FioCruz. http://www.dbbm.fiocruz.br/genome/tcruzi/tcruzi.html, 2000.

[Fly99] The FlyBase Consortium. *"The FlyBase Database of the Drosophila Genome Projects and community literature"*. Nucleic Acids Research 27 (1), pp. 85-88, 1999.

[Fre91] K. A. Frenkel. "*The Human Genome Project and Informatics*". Communications of the ACM 34(11), 1991.

[Gen00] GenBank, http://www.ncbi.nlm.nih.gov/GenBank/index.html., 2000

[GG95] N. Guarino, P. Giaretta. *"Ontologies and Knowledge Bases towards a Terminological Clarification"*. Towards Very Large Knowledge Bases, pág.25-32. IOS Press, Amsterdam.

[GMW+97] R. Goldman, J.McHugh, J. Widom e S. Abiteboul, *"Lore: A Database Management System for Semi-structured Data"*, SIGMOD Record, 26(3):54-66, September 1997 (http://www-db.stanford.edu/lore).

[GRS94] N.Goodman, S.Rozen, L.Stein. *"Managing Laboratory Workflow with LabBase"***.** Proceedings of the 1994 Conference on Computers in Medicine.

[GSDB00] The Genome Sequence DB. http://www.ncgr.org/research/sequence/, 2000.

[HCF+00] C. Harger, G. Chen, A. Farmer, W. Huang, J. Inman, D. Kiphart, F. Schilkey, M. P. Skupski1, J. Weller. *"The Genome Sequence DataBase"*. Nucleic Acids Research 28(1), pp 31-32., 2000.

[HG00] The Natl Human Genome Research Inst. http://www.nhgri.nih.gov/, 2000.

[HWO+94] K. Hart, L. Wong, C. Overton, P. Buneman. *"Using a Query Language to Integrate Biological Data "* http://www.cis.upenn.edu/~cbil/mimbd94/mimbd94CPL.html.

[IOS87] International Organization for Standardization(1987). *"Information processing systems - Open Systems Interconnection - Specification of Abstract Syntax Notation One (ASN.1)"*. Technical Report ISO-8824, International Organization for Standardization, Switzerland.

[Kar95] P.D.Karp. *"A Strategy for Database Interoperation"*. Journal of Computational Biology 2(4), pp. 573-586, 1995.

[KDG96] G. J.L.Kemp, J. Dupont, P. M.D.Gary *" Using the Functional Data Model to Integrate Distributes Biological Data Sources"*. *Proceedings: Eighth International Conference on Scientific and Statistical Database Systems*, IEEE Computer Society Press, pp. 176-185, 1996.

[KLB+97] A.Kogelnik, M.Lott, M.Brown, S.Navathe, D.Wallace *"MITOMAP: An Update on the Human Mitochondrial Genome Database"*, Nucleic Acid Research, 25(1), 1977.

[Kro93] P. Kroha. *"Objects and Databases"*. The McGRAW-HILL International Series in Software Engineering. The McGraw-Hill, 1993.

[KRT96] R. Karp, L. Ruzzo, M. Tompa. *"Algorithms in Molecular Biology"*. http://www.cs.washington.edu/education/courses/590bi/96wi/, 1996.

[LCP+98] S. I. Letovsky, R. W. Cottingham, C. J. Porter, P. W. D. Li. *"GDB: the Human Genome Database"*. Nucleic Acids Research 26(01), pp. 94-99, 1998.

[LP85] D. J. Lipman e W. R. Pearson. *"Rapid and sensitive protein similarity search."* Science 227, pp. 1435-1441, 1985.

[Mar95] V.M. Markowitz. *"Heterogeneous Molecular Biology Database Systems"*. Disponível em http://gizmo.lb.gov/.

[MC95] V. M. Markowitz, I.A. Chen. *"An Overview of the Object Protocol Model (OPM) and the OPM Data Management Tools"***.** Inform Systems 20 (5) 1995.

[MCK97] V.M.Markowitz, I.A.Chen, A.S.Kosky:*"Exploring Heterogeneous Molecular Biology Databases in the Context of the Object-Protocol Model"*. Theoretical and Computational Genome Research, pp. 161-176, Plenum Press, 1997.

[MR95] V.M.Markowitz, O.Ritter. "*Characterizing Heterogeneous Molecular Biology Database Systems*". Journal of Computational Biology, 2(4), 1995.

[MS94] J. Meidanis e J. C. Setúbal. *"Uma Introdução à Biologia Computacional"*. IX Escola de Computação. Recife, 1994.

[MS97] J. Meidanis e J. C. Setúbal. *"Introduction to Computacional Molecular Biology"*. PWS Publishing Company, 1997.

[NK99] S.B.Navathe, A.M.Kogelnik. *"The Challenges of Modeling Biological Information for Genome Databases"*. P.P.Chen et al. (Eds.): Conceptual Modeling, LNCS 1565, pp. 168-182, 1999.

[Pea90] W. R. Pearson. *"Rapid and sensitive sequence comparison with FASTP and FASTA."* Em [Doo90], pp. 63-98.

[Pea91] W. R. Pearson. *"Searching Protein Sequence Libraries: Comparison of the Sensitivity and Selectivity of the Smith-Waterman and FASTA algorithms."* Genomics 11, pp.635-650, 1991.

[PL88] W. R. Pearson e D. J. Lipman. *"Improved Tools for Biological Sequence Comparison."* Proceedings of the National Academy of Sciences of the U.S.A. 85, pp. 2444-2448, 1988.

[PPJ+00] R. C. Périer, V. Praz, T. Junier, C. Bonnard, P. Bucher. *"The Eukaryotic Promoter Database (EPD)"*. Nucleic Acid Research 28(01), p.302-303, 2000.

[Qia93] X. Qian. "*Semantic Interoperation via Intelligent Mediation*". In Proc 3rd Intl Workshop on Research Issues in Data Engineering: Interoperability in Multidatabase Systems, pp. 228-231. IEEE Computer Society Press, 1993.

[QR95] X. Qian, L. Raschid. "*Query Interoperation among Object-Oriented and Relational Databases*". In Proceedings of the Eleventh Conference on Data Engineering, págs. 271-278. IEEE Computer Society Press.

[Rit94] O. Ritter. *"The Integrated Genomic Database"*. Computational Methods in Genome Research (S.Suhai, ed.), 57-73, Plenum, New York, 1994.

[Rob85] E.M.F. De Roberts, Jr *"Bases da Biologia Celular e Molecular"*. Ed. Guanabara, 1985.

[Rob94] R. Robbins. *"Report of the invitational DOE Workshop on genome informatics"*. 26-27 April 1993; Genome Informatics I: Community databases. Journal of Computational Biology, 1(3): 173-190.

[Sha99] R.Shamir. *"Algorithms in Molecular Biology"*. http://www.math.tau.ac.il/~shamir/algmb/algmb98.html, 1999.

[SL90] A. Sheth, J. Larson. "*Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases*". ACM Computing Surveys, 22(3), Setembro, 1990.

[SU94] N. Sakamoto, K. Ushijima, "*Designing and Integrating Human Genome Databases with Object-Oriented Technology*". DEXA, pp.145-152, 1994.

[TMO+00] Y.Tateno, S. Miyazaki, M.Ota, H.Sugawara, T.Gojobori. *"DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams"*. Nucleic Acids Research 28(01), pp. 24-26, 2000.

[Uch94] E. Uchôa. "*HEROS – Um Sistema de Bancos de Dados Heterogêneos: Integrando Esquemas*". Departamento de Informática PUC-Rio, Dissertação de Mestrado em Informática: Ciência da Computação, 1994.

[WL83] W.J. Wilbur e D. J. Lipman. *"Rapid similarity searches of nucleic acid and protein data banks."* Proc Natl Academy of Sciences USA, pp.726-730, 1983.

[Won94] L. Wong. *"Querying Nested Collections"*. PhD thesis, Univ. of Pennsylvania.