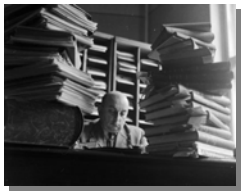
Banco de Dados aplicado a Sistemas Biológicos

Saulo França Amui
(sauloamui@gmail.com)

Departamento de Genética – FMRP/USP
Grupo de Bioinformática

Dados Biológicos

- A internet mudou a maneira como os cientistas compartilham os dados e possibilitou que um depósito central de informações atendesse totalmente a uma comunidade de pesquisa.



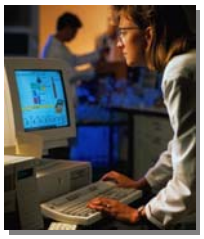
Dados Biológicos

- A tendência é armazenar dados biológicos brutos de todos os tipos em bancos de dados públicos, com acesso aberto pela comunidade de pesquisa.

Dados Biológicos

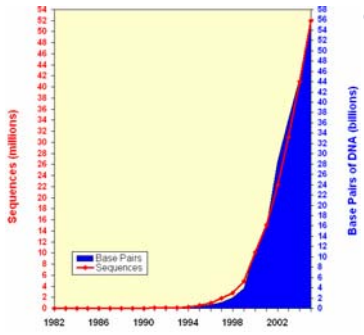
- Em vez de fazer pesquisas preliminar no laboratório, os cientistas vão primeiro aos bancos de dados

  → economia de tempo e recursos

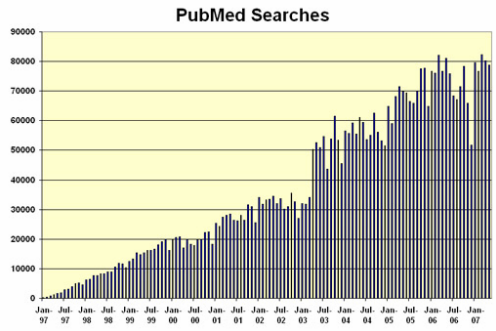


Dados Biológicos

- Com o avanço da tecnologia, existem cada vez mais sequências e anotações 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.



Crescimento do Gen-Bank (1982 a 2005)



Banco de Dados

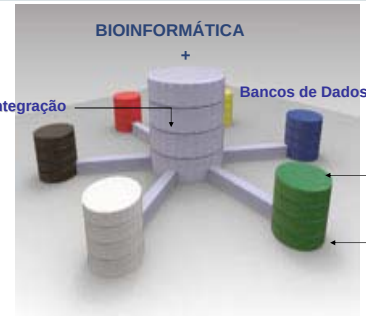
- Conjunto de informações relacionadas entre si, referentes ao mesmo assunto, organizadas prática e racionalmente, para que o usuário levante e **recupere informações**, tire conclusões e tome decisões.



Bioinformática e Bancos de Dados

- Representação, armazenamento e a distribuição de dados → aspecto funcional da bioinformática.
- O desenvolvimento de ferramentas analíticas para revelar o conhecimento contido nos dados é o segundo, e mais científico, aspecto da bioinformática.

Bioinformática e Bancos de Dados



Bancos de Dados Biológicos

- Todos os dados resultantes das análises de um projeto genoma são armazenados nos chamados bancos de dados biológicos.

Bancos de Dados Biológicos

- Constitui um grande conjunto de dados persistentes, geralmente associado a um software projetado para atualizar, consultar e recuperar componentes dos dados armazenados no sistema.
  (Bioinformatics FactSheet 2004)

- **Eficácia:** Fácil acesso às informações.
- **Objetivo:** Métodos para extrair somente as informações necessárias para responder uma específica pergunta biológica.

Bancos de Dados Biológicos

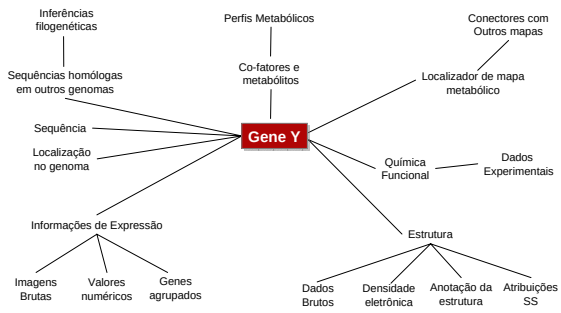
- Visão direcionada para a biologia molecular
- Base da biologia molecular: DNA - Nucleotídeos {A, C, G, T}
- Além do DNA, existem outros tipos de dados na linha de evolução (RNA e proteínas)
- Uma seqüência de DNA pode possuir milhares de pares de nucleotídeos.
- Cada seqüência possui uma identificação, funções biológicas e podem pertencer a vários organismos.

Organismos	Número de genes	Pares de base
Plantas	< 50.000	$< 10^{11}$
Mamíferos	100.000	3×10^{8}
Vermes	14.000	$\sim 10^{8}$
Bactérias	2-4.000	$< 10^{7}$
dsDNA vírus		
Vacina	< 300	187.000
ssRNA vírus		
Influenza	12	13.500

O que se pode descobrir sobre um gene por meio de uma busca a um Banco de Dados?

- **Informação evolutiva**: genes homólogos, freqüências dos alelos, ...
- **Informação genômica**: localização no cromossomo, intros, regiões reguladoras, ...
- **Informação estrutural**: estruturas da proteína correspondente, tipos de folds (grande similaridade estrutural), domínios estruturais,...
- **Informação de expressão**: expressão específica a um dado tecido, fenótipos, doenças, ...
- **Informação funcional**: função molecular/enzimática, papel em diferentes rotas, papel em doenças, ...

Gene
(Algumas informações associadas)

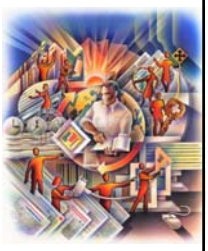


Bancos de Dados Biológicos

- Disponibilizar dados biológicos para os cientistas
- Dados publicados podem ser difíceis de encontrar ou acessar
- Coleta-los da literatura consume muito tempo
- Disponibilizar dados em formato que possa ser lido por um computador

Estruturação dos Dados X Acessibilidade

- O entendimento da diferença entre **dados estruturados** e desestruturados, e o projeto de um formato que se ajuste ao seu armazenamento de dados e às suas necessidades de acesso, é a chave para tornar os dados **úteis e acessíveis**.
- Formas de Armazenamentos
  - Diferentes bases de dados para armazenamento de dados biológicos



Formas de Armazenamento

- **arquivos em formato de texto**
- **arquivos estruturados**
- **bancos de dados relacionais**
  - *dados estão guardados em tabelas*
- **bancos de dados objeto-relacionais**
- **bancos de dados orientado a objetos**

Realidades e Problemas comuns

- Muitas bases de dados são construídas pelos próprios biólogos
- Não padronização da taxonomia
- Dificuldade na adoção de um vocabulário entre os grupos de pesquisa
- Termos diferentes para conceitos iguais
- Conceitos diferentes para termos iguais

Realidades e Problemas comuns

- **Qualidade dos dados disponíveis na Web**
  - Grupos de pesquisa submetem suas descobertas
  - Algumas bases aceitam de qualquer maneira
  - Muitas bases não verificam a qualidade dos dados
  - Outras bases preocupam-se com a qualidade dos dados, onde um comitê valida-os. Estas bases ganham destaque da comunidade científica.

Qualidade da Informação

- "Sua capacidade de julgar a qualidade da informação e do software encontrados na Web aumentará à medida que você continuar a atuar neste campo" (Gibas & Jambeck 2001)
  - Qual é a Fonte? Quem são os autores? Qual é o propósito da organização responsável pelas informações? É uma Organização acadêmica? Um Órgão Governamental? Uma empresa?
- Estima-se que qualquer sequência do GenBank provavelmente contenha pelo menos um erro.
  - O software oferecido por órgãos públicos como o NCBI e o PDB pode ainda estar em desenvolvimento. Grande parte dele é de boa qualidade.
- Transparência (Documentação) e Atualização

Realidades e Problemas comuns

- **Versionamento dos dados**
  - Dados não acurados podem ser melhorados
  - Novas versões sobre a anotação de uma seqüência são submetidas às bases de dados
  - Muitos pesquisadores "carregam" toda uma base de dados para seu ambiente local, logo, não possuem as últimas atualizações.
  - O custo de armazenamento e do tratamento das versões pode levar algumas bases de dados à simples atualização da "versão" corrente (a única)

Na prática ....

- Para que estes bancos de dados possam ser realmente utilizados na prática é necessário tratar de vários pontos importantes:
  - a definição do modelo de dados mais adequado,
  - as necessidades de processamento,
  - as análises e controles semântico dos dados,
  - os meios de acesso e o problema da integração das bases de dados.

- Avanço da tecnologia → crescimento exponencial do volume de biossequências → dados submetidos aos bancos de dados através da **Internet**
- Grande facilidade na submissão de biossequências aos bancos de dados → muito importante para que os biólogos possam acessar e fazer suas análises em novos dados mais rapidamente.
- 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.

Bancos de Dados Públicos

- Bancos de Dados Públicos (mais de 348).
  - BD de seqüências de nucleotídeos
    - EMBL (http://www.ebi.ac.uk/embl)
    - GenBank (http://www.ncbi.nlm.nih.gov/GenBank)
    - DDBJ (http://www.ddbj.nig.ac.jp)
  - BD de seqüências de proteínas
    - SWISS-PROT, TrEMBL (http://www.expansy.ch/sprot)
    - PIR (http://pir.georgetown.edu)
  - BD de motivos
    - Pfam (http://www.sanger.ac.uk/Software/Pfam)
    - PROSITE (http://www.expansy.ch/prosite)
  - BD de estruturas macromoleculares 3D
    - PDB (http://www.rcsb.org/pdb)

O que se deseja saber sobre uma seqüência?

- Essa seqüência é similar a de algum gene conhecido? Quão semelhante? Qual é o seu significado?
- O que nós conhecemos sobre o gene 1234?
  - Genômica (localização, regiões regulatórias)
  - Estrutura (domínios?)
  - Funcional (doenças)
- Informações evolucionárias:
  - Esse gene pode ser encontrado em quais organismos?
  - Qual é a sua taxonomia?

Classificação dos bancos de dados

- Primários
  - Deposição direta de seqüências sem qualquer processamento ou análise (não curados).
  - GenBank, EMBL-Bank, DDBJ, etc.
- Secundários
  - Derivam dos primários porém com alguns tipos de análises (geralmente curados).
  - Swiss-prot, Uniprot, PROSITE, Blocks, PDB, etc.

Bancos de Dados Compostos

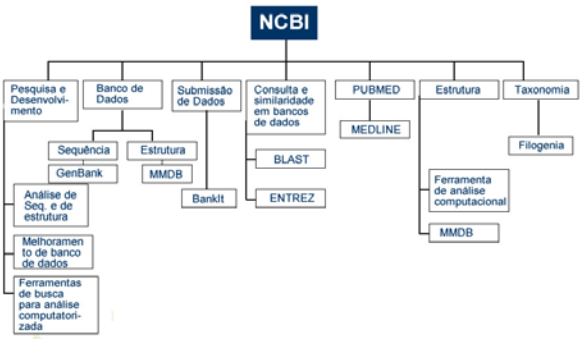
NRDB	OWL	MIPSX	SP + TrEMBL
PDB	SWISS-PROT	PIR1-4	SWISS-PROT
SWISS-PROT	PIR	MIPSOwn	TrEMBL
PIR	GenBank	MIPSTrn	
GenPept	NRL-3D	MIPSH	
SWISS-PROTupdate		PIRMOD	
GenPeptupdate		SWISS-PROT	
		EMTrans	
		GBTrans	
		Kabat	
		PseqIP	

Membros do Consórcio internacional

INSDC - International Nucleotide Sequence Database Collaboration

- Genbank (NCBI - National Center for Biotechnology Information)
- EMBL (European Molecular Biology Laboratory)
- DNA DataBank of Japan (DDBJ)

http://www.insdc.org



Genbank

Genbank (http://www.ncbi.nlm.nih.gov/)

- Responsável
  - National Center for Biotechnology Information (NCBI);
  - Desde 1982.
- Assunto
  - Dados sobre seqüências de ácidos nucléicos (DNA) e Proteínas derivadas;
  - Outros dados associados.
- Aproximadamente 61.132.599 seqüências

Genbank - Formato

- **HEADER**
  - Título
  - Taxonomia
  - Citação
- **FEATURES**
  - Gene
  - mRNA
  - AA
- **SEQÜÊNCIA**

NCBI - Vários

NCBI - Entrez

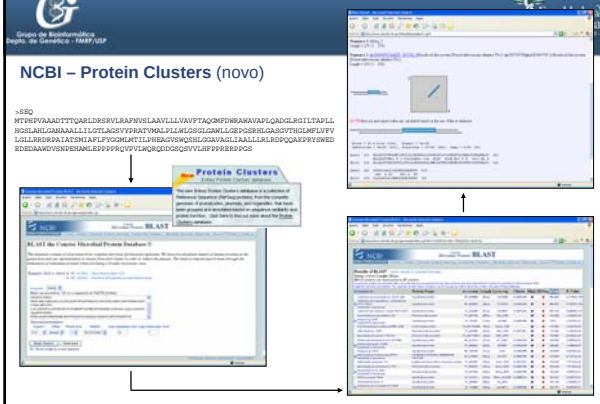
ENTREZ

- Interface de busca (keywords);
- Busca de genes, seqüências, proteínas, referências, etc.

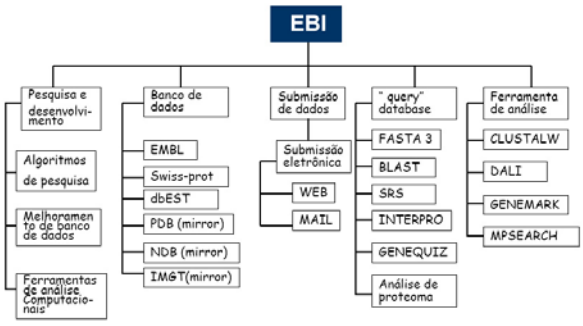
NCBI - PubMed

NCBI - Blast

NCBI – Protein Clusters (novo)



EBI

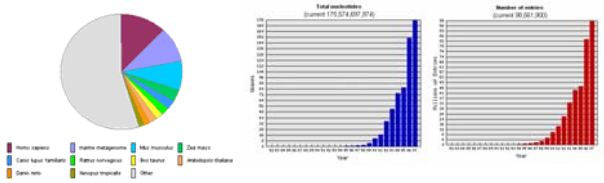


EMBL

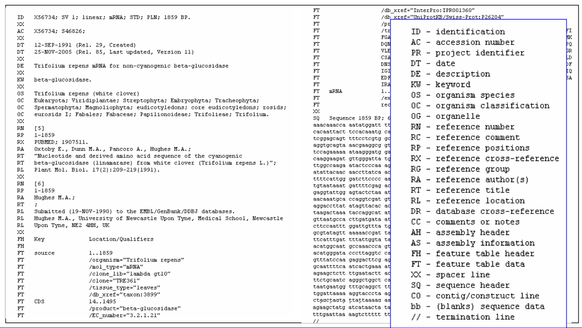
EMBL (http://www.ebi.ac.uk/)

- Responsável
  - European Bioinformatics Institute (EBI);
  - Desde 1982.
- Assunto
  - Dados sobre seqüências de ácidos nucléicos;
  - Outros dados associados.
- Aproximadamente 98.271.924 seqüências

EMBL - Estatísticas

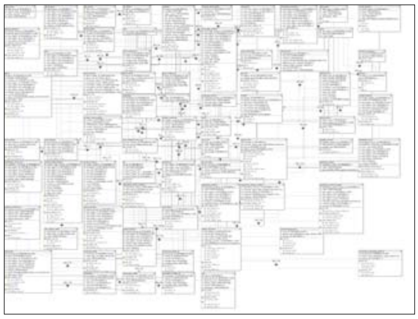


EMBL – Formato

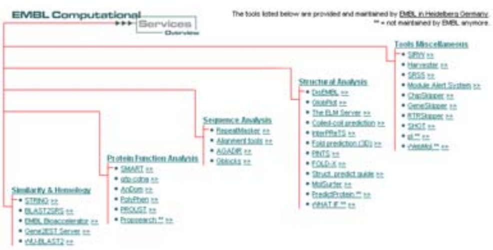

ID X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC X56734; S46826;
XX
DT 12-SEP-1991 (Rel. 29, Created)
DT 25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW beta-glucosidase.
XX
OS Trifolium repens (white clover)
OC Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids;
OC eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolieae; Trifolium.
XX
RN [5]
RP 1-1859
RX PUBMED; 1907511.
RA Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT "Nucleotide and derived amino acid sequence of the cyanogenic
RT beta-glucosidase (linamarase) from white clover (Trifolium repens L.)";
RL Plant Mol. Biol. 17(2):209-219(1991).
XX
RN [6]
RP 1-1859
RA Hughes M.A.;
RT ;
RL Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ databases.
RL Hughes M.A., University of Newcastle Upon Tyne, Medical School, Newcastle
RL Upon Tyne, NE2 4HH, UK
XX
FH Key Location/Qualifiers
FH
FT source 1..1859
FT /organism="Trifolium repens"
FT /mol_type="mRNA"
FT /clone_lib="lambda gt10"
FT /clone="TRE361"
FT /tissue_type="leaves"
FT /db_xref="taxon:3899"
FT CDS 14..1495
FT /product="beta-glucosidase"
FT /EC_number="3.2.1.21"

ID - identification
AC - accession number
PR - project identifier
DT - date
DE - description
KW - keyword
OS - organism species
OC - organism classification
OG - organelle
RN - reference number
RC - reference comment
RP - reference positions
RX - reference cross-reference
RG - reference group
RA - reference author(s)
RT - reference title
RL - reference location
DR - database cross-reference
CC - comments or notes
AH - assembly header
AS - assembly information
FH - feature table header
FT - feature table data
XX - spacer line
SQ - sequence header
CO - contig/construct line
bb - (blanks) sequence data
// - termination line

EMBL – Schema (SGBD – Oracle)



EMBL – Serviços



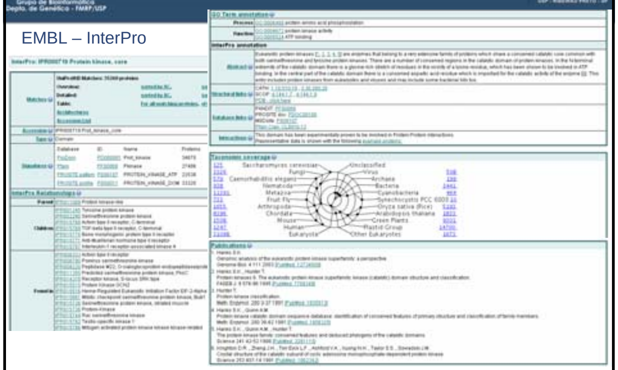
EMBL – InterPro

Fonte de recursos documentados que amalgama esforços dos bancos de dados PROSITE, PRINTS, Pfam e ProDom.

Cada entrada de InterPro inclui uma descrição funcional, anotação, referências bibliográficas e links para o banco de dados relevante.

EMBL

EMBL – InterPro



DDBJ

DDBJ (http://www.ddbj.nig.ac.jp/Welcome-j.html)

- Responsável
  - National Institute of Genetics – Japan (NIG);
  - Desde 1988.
- Assunto
  - Dados sobre seqüências de ácidos nucléicos;
  - Outros dados associados.

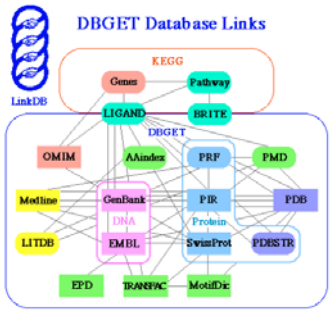
DDBJ - Formato

DDBJ

similar

Genbank

DBGET Database Links



Outros Bancos de Dados e Métodos

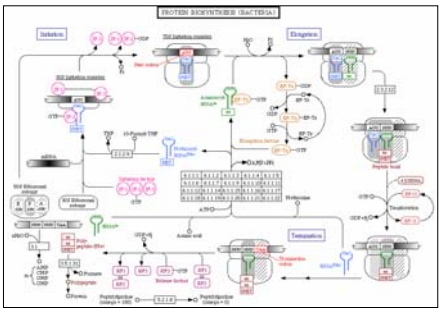
- **Swiss-prot**
  - Banco de dados curado com seqüências de proteínas;
  - Clara separação entre os dados anotados;
  - Redundância mínima;
  - Integração com outros bancos de dados;
  - http://expasy.org/sprot
- **Prosite**
  - Aponta domínios funcionais em uma proteína utilizando consensos armazenados em seu banco de dados;
  - http://expasy.org/prosite/
- **Blocks**
  - Múltiplos alinhamentos de regiões altamente conservadas de proteínas;
  - http://blocks.fhcrc.org/
- **PDB**
  - Banco de dados de estruturas de proteínas;
  - http://www.rcsb.org/pdb/home/home.do
- **KEGG**
  - Coleção de representações e mapas de vias metabólicas;
  - http://www.genome.jp/kegg/

Swiss-prot

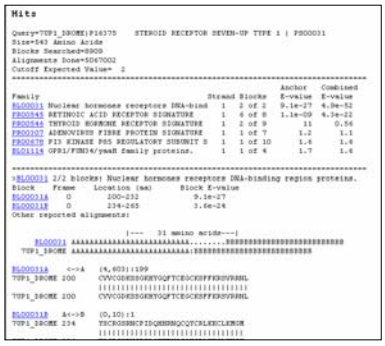
UniProtKB/Swiss-Prot: Q9ZZ38

KEGG

Kyoto Encyclopedia of Genes and Genomes → Informações por vias metabólicas

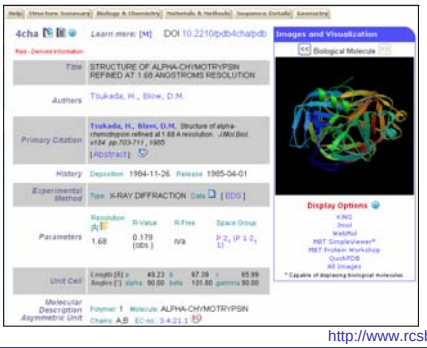


Blocks



Alinhamentos de várias sequências sem intervalos que correspondem às regiões de proteínas mais conservadas

PDB
(Protein Data Bank)



http://www.rcsb.org/pdb

- Diversidade de bancos de dados (Nucleic Acids Research)

http://nar.oxfordjournals.org/

Referências

B. BOECKMANN, A. BAIROCH, R. APWEILER, et al. "*The SWISSPROT protein knowledgebase and its supplement TrEMBL in 2003*". Nucleic Acids Research, pp. 365-370, 2003.

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.

Gibas, C.; Jambeck, P.; Desenvolvendo Bioinformática: ferramentas de software para aplicações e, biologia; Rio de Janeiro, Ed. Campus, 2001

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

L. FALQUET, M. PAGNI, P. BUCHER, et al. "*The PROSITE database, its status in 2002*". Nucleic Acids Research, pp. 235-238, 2002.

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.

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.

Y. TATENO, K. FUKAMI-KOBAYASHI, S. MIYAZAKI, et al. "*DNA Data Bank of Japan at work on genome sequence data*". Nucleic Acids Research, pp. 16-20, 1998.