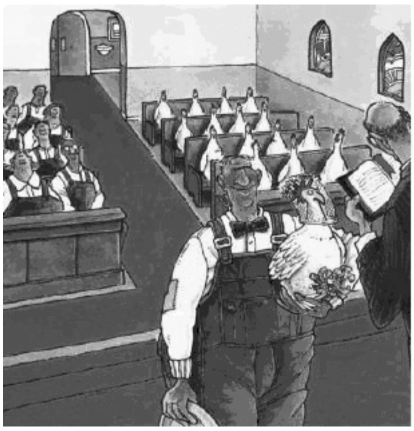
Introduction to BIOINFORMATICS

What is Bioinformatics?

The marriage between biology and informatics



Understanding nature's mute elegant language of living cells is the question of modern molecular biology.

From an alphabet of only four letters representing the chemical subunits of DNA, emerges a syntax of life processes whose most complex expression is man.

From the National Centre for Biotechnology Information (NCBI), http:://www.ncbi.nlm.nih.gov

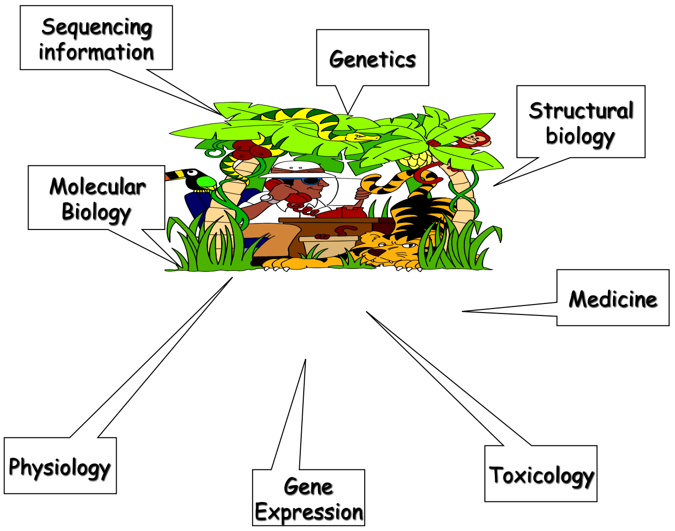
The challenge is in finding new approaches to deal with the volume and complexity of data, and in providing researchers with better access to analysis and computing tools in order to advance understanding of our genetic legacy and its role in health and diseases.

From the National Centre for Biotechnology Information (NCBI), http:://www.ncbi.nlm.nih.gov

Essentially, Bioinformatics has three components

- The creation of databases allowing the storage and management of large biological data sets.
- The development of algorithms and statistics to determine relationships among members of large data sets.
- The use of these tools for the analysis and interpretation of various types of biological data, including DNA, RNA and protein sequences, protein structures, gene expression profiles, and biochemical pathways

Data Jungle



Sequencing information

Genetics

Structural biology

Molecular Biology

Medicine

Physiology

Gene Expression

Toxicology

From: EBI SRS tutorial

Factors that made bioinformatics so important

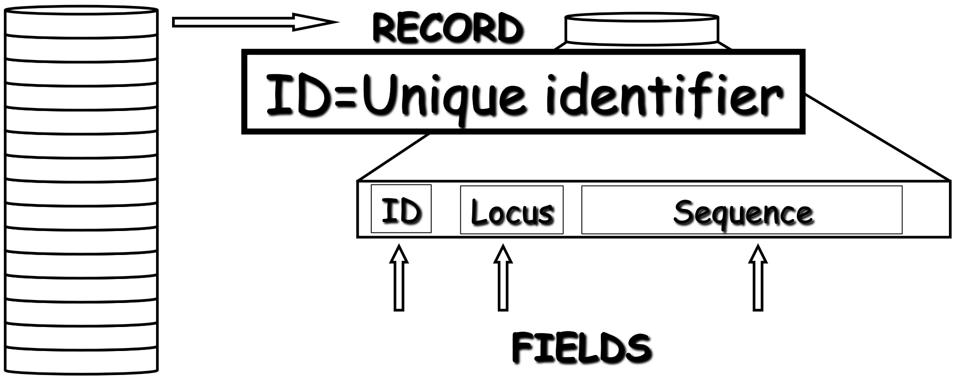
- **Rapid and cheap techniques for DNA sequencing**
- **The development of powerful computers**
- **Internet and the Wide World Web**

DATA SOURCES FOR DATABASES

- **Direct scientific submission**
- **Genome sequencing labs and groups**
- **Scientific literature**
- **Patent applications**

DATABASE: a collection of data that has a regular structure and that is organized in such a way that a computer can easily find and retrieve information.

A database is generally a collection of RECORDS, available through specific entries, each of which contains one or more FIELD.



Data Resources at NCBI

Databases: Primary and Derivative

Primary Databases

- Archival submissions of experimental results
- Database staff organize but don’t add additional information

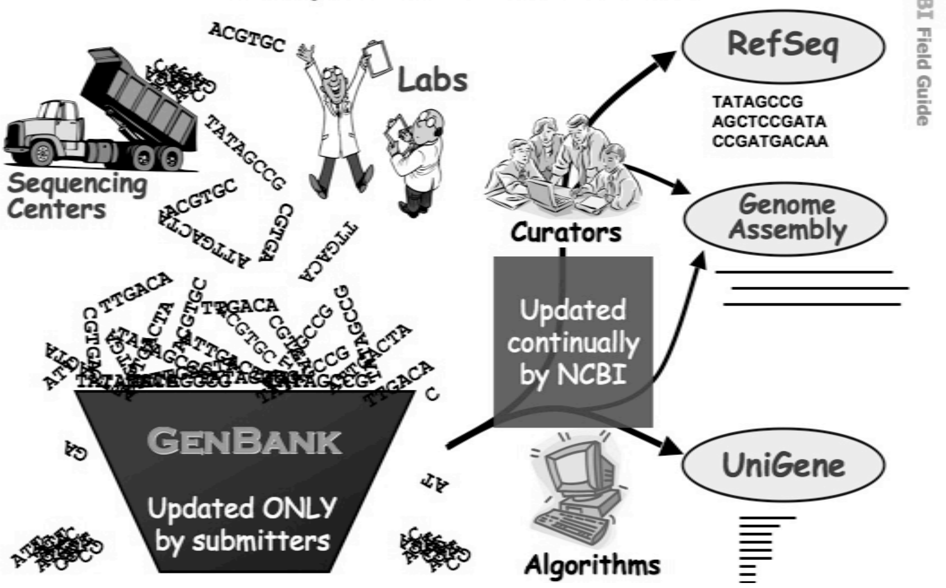
Genbank dbEST dbSNP Probe

Derivative Databases

- Curated/expert review
- Computationally derived
- Combinations

Refseq Genomes UniGene UniSTS Homologene

Primary vs. Derivative Sequence Databases



- **<u>Bioinformatics Developers</u>**

  They develop tools for bioinformatics

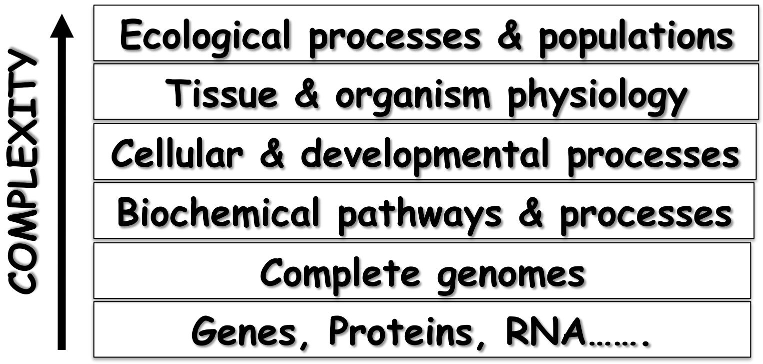
  - **Experts in Mathematics, Statistics and Informatics**
  - **Computational biologists**

- **<u>Bioinformatics Users</u>**

  They use the tools of bioinformatics

  - **Researchers (Biologists, Biotechnologist,…)**

Biological complexity



COMPLEXITY

Ecological processes & populations

Tissue & organism physiology

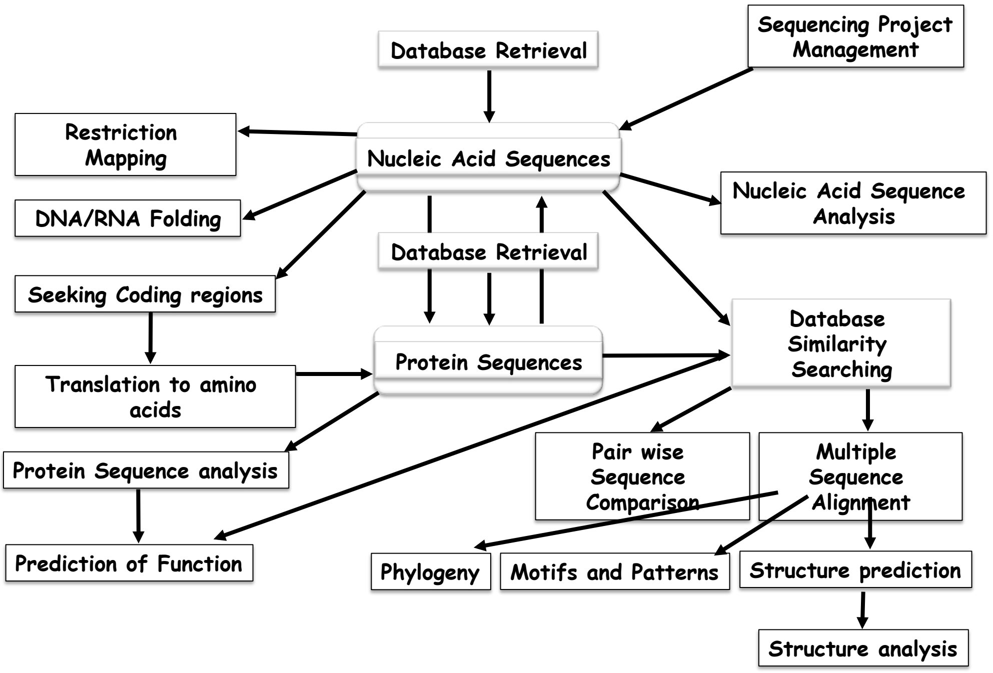
Cellular & developmental processes

Biochemical pathways & processes

Complete genomes

Genes, Proteins, RNA…….

Introduction to Bioinformatics: www.bioinformatics.com/courses.com/bioinfom



Database Retrieval

Sequencing Project Management

Restriction Mapping

Nucleic Acid Sequences

DNA/RNA Folding

Nucleic Acid Sequence Analysis

Database Retrieval

Seeking Coding regions

Database Similarity Searching

Protein Sequences

Translation to amino acids

Protein Sequence analysis

Pair wise Sequence Comparison

Multiple Sequence Alignment

Prediction of Function

Phylogeny

Motifs and Patterns

Structure prediction

Structure analysis

Sites where the integration among databases and between databases and software is developed

- USA
  NCBI (National Center for Biotechnology Information)
- Europe
  EBI (European Bioinformatics Institute,Hinxton, UK)
- Japan
  NIG (National Institute of Genetics)

Biological database history

1965

M. Dayhoff *et al*. published "Atlas of Protein Sequences and Strucures"

1982

EMBL started the DNA sequence collection

1983

Genbank started the DNA sequence collection

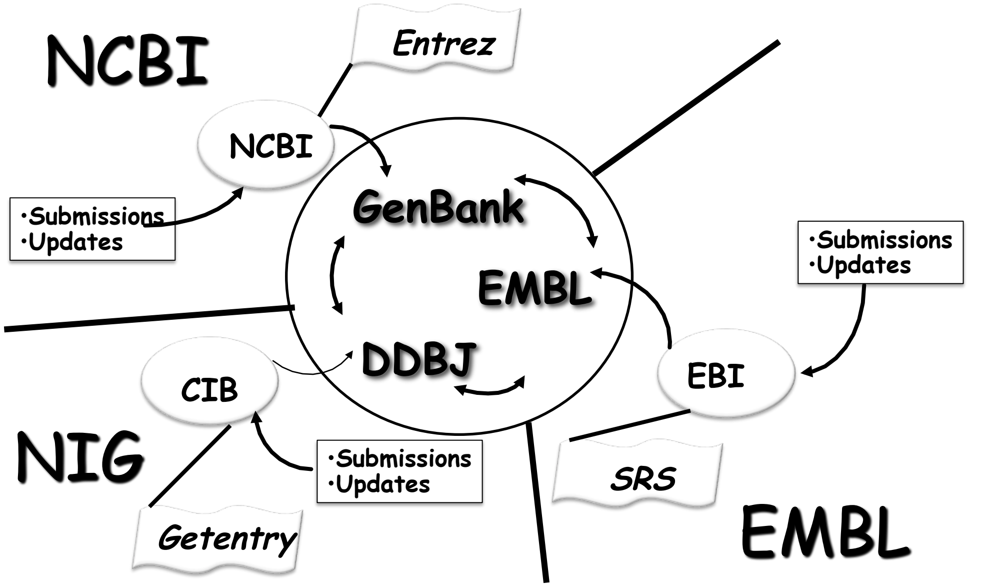
1984

DNA sequence databases of Japan

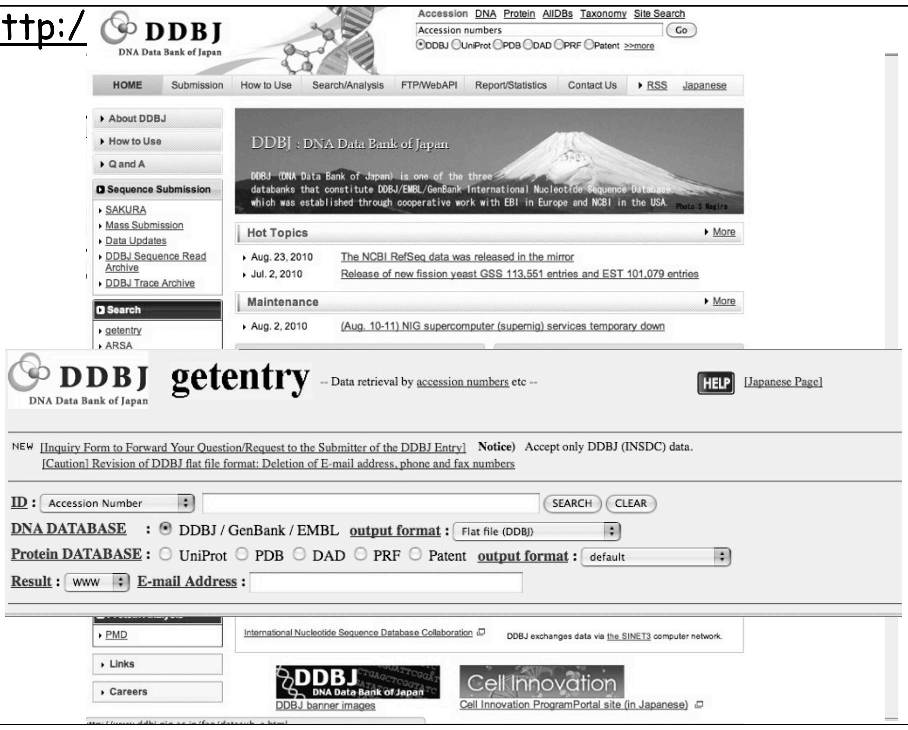
1988

Embl/GenBank/DDBJ agreed on common format for data elements

The International Sequence Database Collaboration



http:/



DDBJ : DNA Data Bank of Japan

DDBJ (DNA Data Bank of Japan) is one of the three databanks that constitute DDBJ/EMBL/GenBank International Nucleotide Sequence Database, which was established through cooperative work with EBI in Europe and NCBI in the USA. Photo:S Nagira

Hot Topics More

- Aug. 23, 2010 The NCBI RefSeq data was released in the mirror
- Jul. 2, 2010 Release of new fission yeast GSS 113,551 entries and EST 101,079 entries

Maintenance More

- Aug. 2, 2010 (Aug. 10-11) NIG supercomputer (supernig) services temporary down

DDBJ **getentry** -- Data retrieval by accession numbers etc --

NEW [Inquiry Form to Forward Your Question/Request to the Submitter of the DDBJ Entry] **Notice)** Accept only DDBJ (INSDC) data.
[Caution] Revision of DDBJ flat file format: Deletion of E-mail address, phone and fax numbers

ID : Accession Number

DNA DATABASE : ● DDBJ / GenBank / EMBL **output format :** Flat file (DDBJ)

Protein DATABASE : ○ UniProt ○ PDB ○ DAD ○ PRF ○ Patent **output format :** default

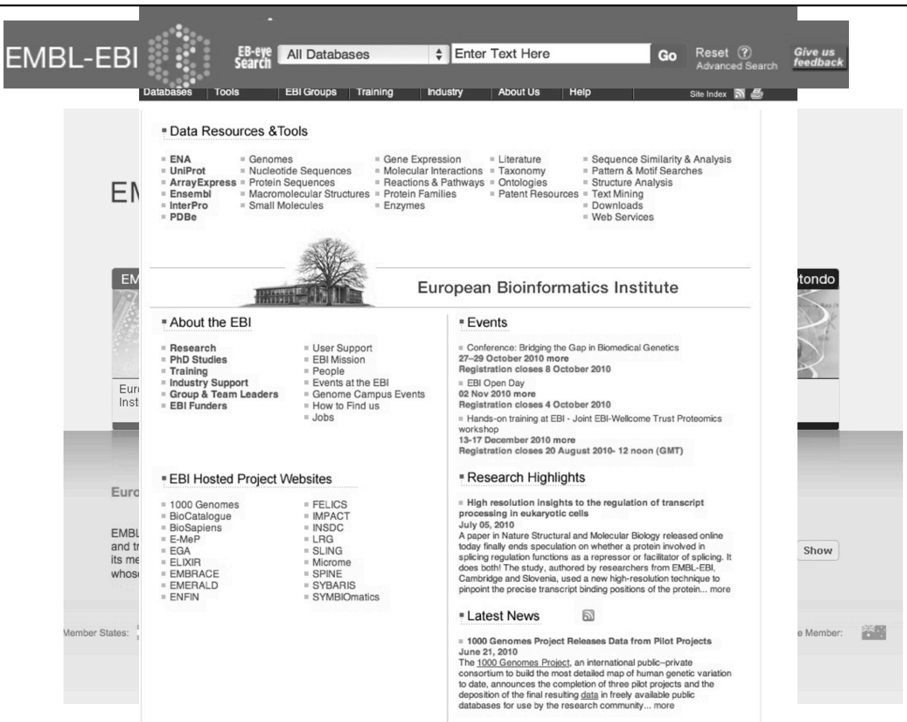
Result : www **E-mail Address :**

International Nucleotide Sequence Database Collaboration

DDBJ exchanges data via the SINET3 computer network.

DDBJ banner images

Cell Innovation ProgramPortal site (in Japanese)



Data Resources &Tools

- **ENA**
- **UniProt**
- **ArrayExpress**
- **Ensembl**
- **InterPro**
- **PDBe**
- Genomes
- Nucleotide Sequences
- Protein Sequences
- Macromolecular Structures
- Small Molecules
- Gene Expression
- Molecular Interactions
- Reactions & Pathways
- Protein Families
- Enzymes
- Literature
- Taxonomy
- Ontologies
- Patent Resources
- Sequence Similarity & Analysis
- Pattern & Motif Searches
- Structure Analysis
- Text Mining
- Downloads
- Web Services

European Bioinformatics Institute

About the EBI

- **Research**
- **PhD Studies**
- **Training**
- **Industry Support**
- **Group & Team Leaders**
- **EBI Funders**
- User Support
- EBI Mission
- People
- Events at the EBI
- Genome Campus Events
- How to Find us
- Jobs

EBI Hosted Project Websites

- 1000 Genomes
- BioCatalogue
- BioSapiens
- E-MeP
- EGA
- ELIXIR
- EMBRACE
- EMERALD
- ENFIN
- FELICS
- IMPACT
- INSDC
- LRG
- SLING
- Microme
- SPINE
- SYBARIS
- SYMBIOmatics

Events

- Conference: Bridging the Gap in Biomedical Genetics
27–29 October 2010 more
Registration closes 8 October 2010
- EBI Open Day
02 Nov 2010 more
Registration closes 4 October 2010
- Hands-on training at EBI - Joint EBI-Wellcome Trust Proteomics workshop
13-17 December 2010 more
Registration closes 20 August 2010- 12 noon (GMT)

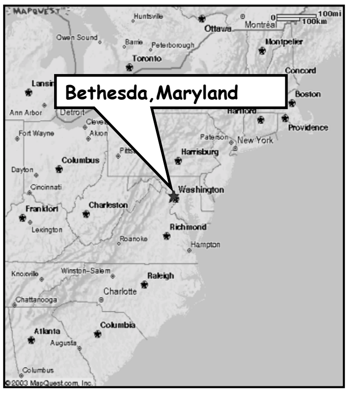
Research Highlights

- **High resolution insights to the regulation of transcript processing in eukaryotic cells**
July 05, 2010
A paper in Nature Structural and Molecular Biology released online today finally ends speculation on whether a protein involved in splicing regulation functions as a repressor or facilitator of splicing. It does both! The study, authored by researchers from EMBL-EBI, Cambridge and Slovenia, used a new high-resolution technique to pinpoint the precise transcript binding positions of the protein... more

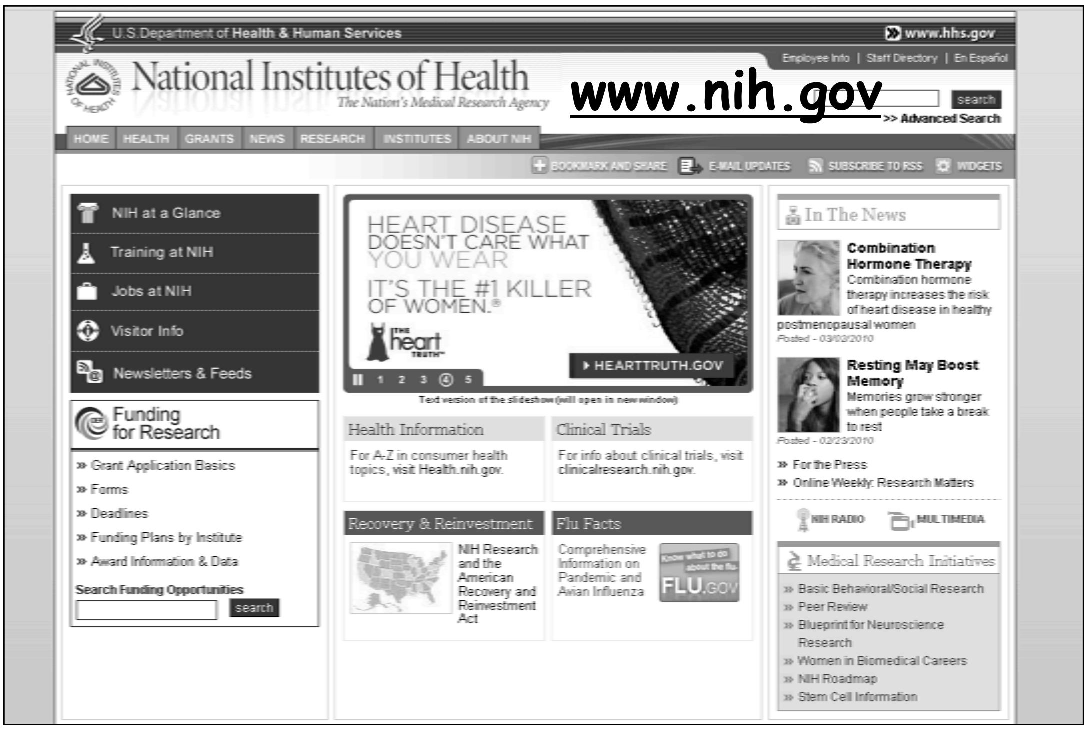
Latest News

- **1000 Genomes Project Releases Data from Pilot Projects**
June 21, 2010
The 1000 Genomes Project, an international public–private consortium to build the most detailed map of human genetic variation to date, announces the completion of three pilot projects and the deposition of the final resulting data in freely available public databases for use by the research community... more

NCBI: The National Center for Biotechnology Information



Created by US Congress in 1988, NCBI is part of the National Library of Medicine at National Institutes of Health.


U.S. Department of Health & Human Services
www.hhs.gov
Employee Info | Staff Directory | En Español
National Institutes of Health
The Nation's Medical Research Agency
search
>> Advanced Search
HOME
HEALTH
GRANTS
NEWS
RESEARCH
INSTITUTES
ABOUT NIH
BOOKMARK AND SHARE
E-MAIL UPDATES
SUBSCRIBE TO RSS
WIDGETS
NIH at a Glance
Training at NIH
Jobs at NIH
Visitor Info
Newsletters & Feeds
Funding for Research
» Grant Application Basics
» Forms
» Deadlines
» Funding Plans by Institute
» Award Information & Data
Search Funding Opportunities
search
HEART DISEASE DOESN'T CARE WHAT YOU WEAR
IT'S THE #1 KILLER OF WOMEN.®
the heart truth
HEARTTRUTH.GOV
1 2 3 4 5
Text version of the slideshow (will open in new window)
Health Information
Clinical Trials
Recovery & Reinvestment
Flu Facts
Know what to do about the flu
FLU.GOV
In The News
» For the Press
» Online Weekly: Research Matters
NIH RADIO
MULTIMEDIA
Medical Research Initiatives
» Basic Behavioral/Social Research
» Peer Review
» Blueprint for Neuroscience Research
» Women in Biomedical Careers
» NIH Roadmap
» Stem Cell Information

www.nih.gov

For A-Z in consumer health topics, visit Health.nih.gov.

For info about clinical trials, visit clinicalresearch.nih.gov.

NIH Research and the American Recovery and Reinvestment Act

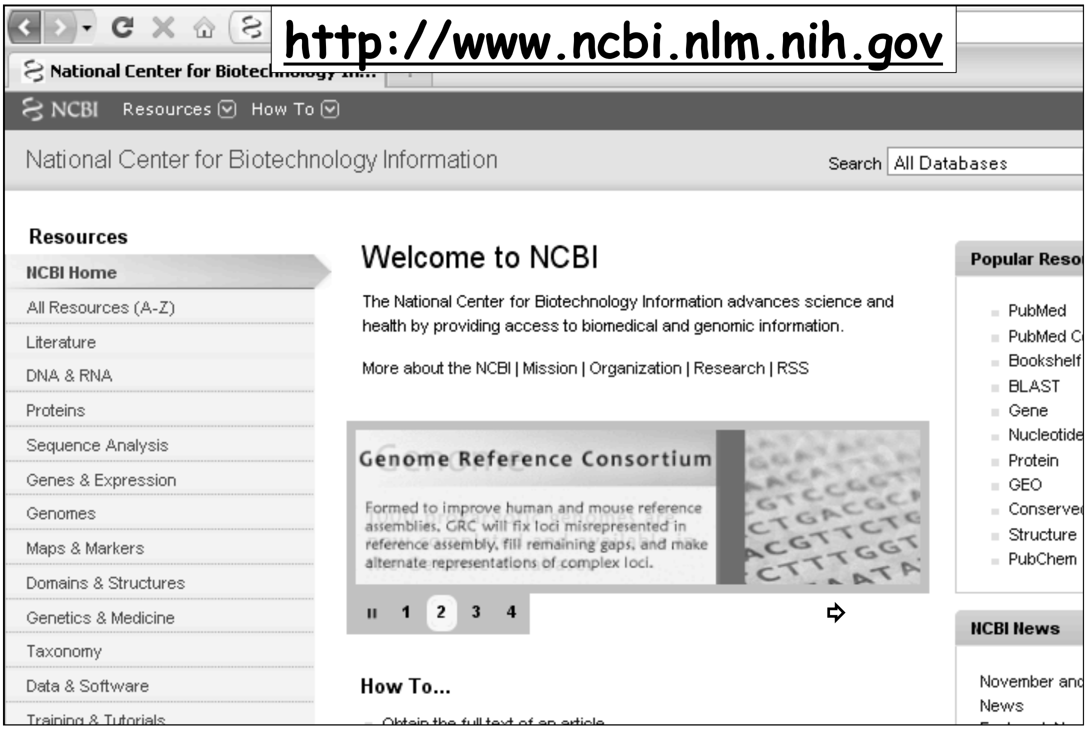
Comprehensive Information on Pandemic and Avian Influenza

Combination Hormone Therapy
Combination hormone therapy increases the risk of heart disease in healthy postmenopausal women
Posted - 03/02/2010

Resting May Boost Memory
Memories grow stronger when people take a break to rest
Posted - 02/23/2010

The National Center for Biotechnology Information (NCBI)

- Created as a part of the ***National Library of Medicine*** in 1988
  - Establish public databases
  - Research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information
- Tools: BLAST(1990), Entrez (1992)
- GenBank (1992)
- Free MEDLINE (PubMed, 1997)
- Other databases: dbEST, dbGSS, dbSTS, MMDB, OMIM, UniGene, GeneMap, Taxonomy, CGAP, SAGE, Gene, RefSeq



http://www.ncbi.nlm.nih.gov

National Center for Biotechnology Information

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

More about the NCBI | Mission | Organization | Research | RSS

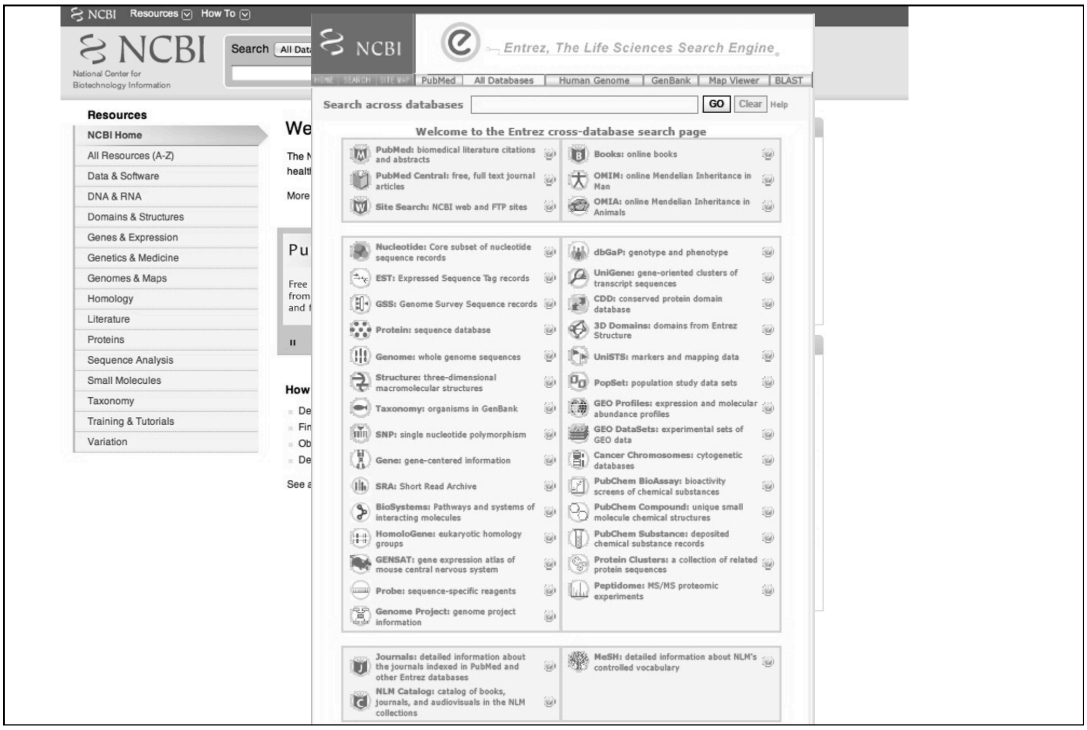
Genome Reference Consortium

Formed to improve human and mouse reference assemblies, GRC will fix loci misrepresented in reference assembly, fill remaining gaps, and make alternate representations of complex loci.

How To...

Obtain the full text of an article

November and
News



Entrez, The Life Sciences Search Engine

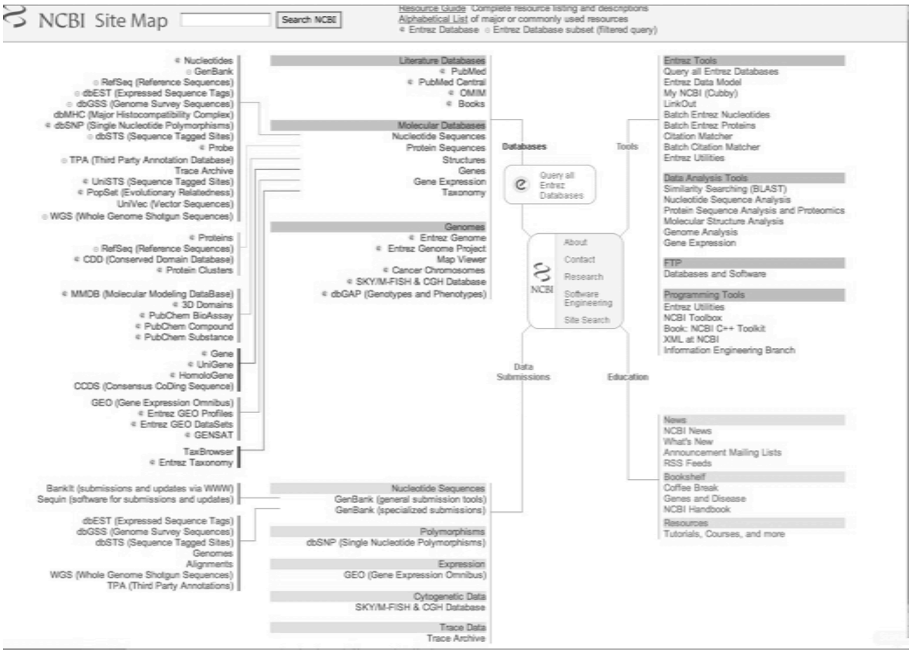
Search across databases

Welcome to the Entrez cross-database search page

PubMed: biomedical literature citations and abstracts
PubMed Central: free, full text journal articles
Site Search: NCBI web and FTP sites
Books: online books
OMIM: online Mendelian Inheritance in Man
OMIA: online Mendelian Inheritance in Animals

Nucleotide: Core subset of nucleotide sequence records
EST: Expressed Sequence Tag records
GSS: Genome Survey Sequence records
Protein: sequence database
Genome: whole genome sequences
Structure: three-dimensional macromolecular structures
Taxonomy: organisms in GenBank
SNP: single nucleotide polymorphism
Gene: gene-centered information
SRA: Short Read Archive
BioSystems: Pathways and systems of interacting molecules
HomoloGene: eukaryotic homology groups
GENSAT: gene expression atlas of mouse central nervous system
Probe: sequence-specific reagents
Genome Project: genome project information
dbGaP: genotype and phenotype
UniGene: gene-oriented clusters of transcript sequences
CDD: conserved protein domain database
3D Domains: domains from Entrez Structure
UniSTS: markers and mapping data
PopSet: population study data sets
GEO Profiles: expression and molecular abundance profiles
GEO DataSets: experimental sets of GEO data
Cancer Chromosomes: cytogenetic databases
PubChem BioAssay: bioactivity screens of chemical substances
PubChem Compound: unique small molecule chemical structures
PubChem Substance: deposited chemical substance records
Protein Clusters: a collection of related protein sequences
Peptidome: MS/MS proteomic experiments

Journals: detailed information about the journals indexed in PubMed and other Entrez databases
NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections
MeSH: detailed information about NLM's controlled vocabulary



NCBI Site Map

Search NCBI

Resource Guide Complete resource listing and descriptions
Alphabetical List of major or commonly used resources
Entrez Database Entrez Database subset (filtered query)

Nucleotides
GenBank
RefSeq (Reference Sequences)
dbEST (Expressed Sequence Tags)
dbGSS (Genome Survey Sequences)
dbMHC (Major Histocompatibility Complex)
dbSNP (Single Nucleotide Polymorphisms)
dbSTS (Sequence Tagged Sites)
Probe
TPA (Third Party Annotation Database)
Trace Archive
UniSTS (Sequence Tagged Sites)
PopSet (Evolutionary Relatedness)
UniVec (Vector Sequences)
WGS (Whole Genome Shotgun Sequences)

Proteins
RefSeq (Reference Sequences)
CDD (Conserved Domain Database)
Protein Clusters

MMDB (Molecular Modeling DataBase)
3D Domains
PubChem BioAssay
PubChem Compound
PubChem Substance

Gene
UniGene
HomoloGene
CCDS (Consensus CoDing Sequence)

GEO (Gene Expression Omnibus)
Entrez GEO Profiles
Entrez GEO DataSets
GENSAT

TaxBrowser
Entrez Taxonomy

BankIt (submissions and updates via WWW)
Sequin (software for submissions and updates)

dbEST (Expressed Sequence Tags)
dbGSS (Genome Survey Sequences)
dbSTS (Sequence Tagged Sites)
Genomes
Alignments
WGS (Whole Genome Shotgun Sequences)
TPA (Third Party Annotations)

Literature Databases
PubMed
PubMed Central
OMIM
Books

Molecular Databases
Nucleotide Sequences
Protein Sequences
Structures
Genes
Gene Expression
Taxonomy

Genomes
Entrez Genome
Entrez Genome Project
Map Viewer
Cancer Chromosomes
SKY/M-FISH & CGH Database
dbGAP (Genotypes and Phenotypes)

Nucleotide Sequences
GenBank (general submission tools)
GenBank (specialized submissions)

Polymorphisms
dbSNP (Single Nucleotide Polymorphisms)

Expression
GEO (Gene Expression Omnibus)

Cytogenetic Data
SKY/M-FISH & CGH Database

Trace Data
Trace Archive

Entrez Tools
Query all Entrez Databases
Entrez Data Model
My NCBI (Cubby)
LinkOut
Batch Entrez Nucleotides
Batch Entrez Proteins
Citation Matcher
Batch Citation Matcher
Entrez Utilities

Data Analysis Tools
Similarity Searching (BLAST)
Nucleotide Sequence Analysis
Protein Sequence Analysis and Proteomics
Molecular Structure Analysis
Genome Analysis
Gene Expression

FTP
Databases and Software

Programming Tools
Entrez Utilities
NCBI Toolbox
Book: NCBI C++ Toolkit
XML at NCBI
Information Engineering Branch

News
NCBI News
What's New
Announcement Mailing Lists
RSS Feeds

Bookshelf
Coffee Break
Genes and Disease
NCBI Handbook

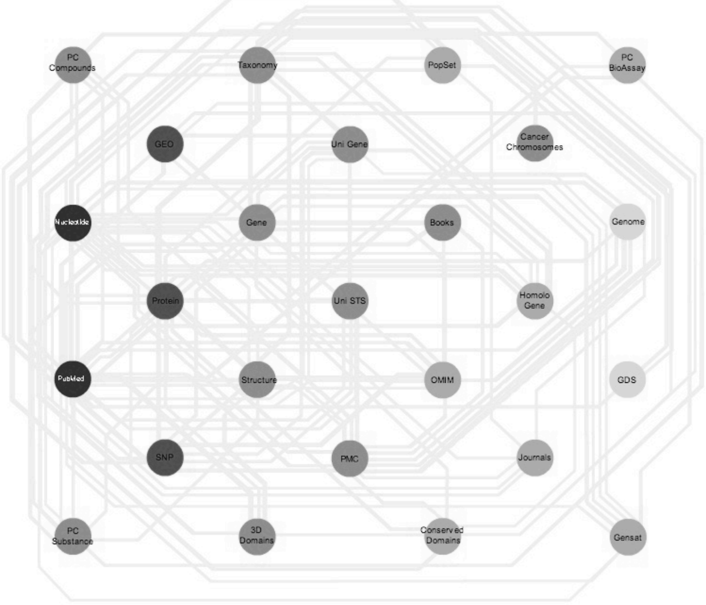
Resources
Tutorials, Courses, and more

http://www.ncbi.nlm.nih.gov/Sitemap/index.html

Databases connections

http://www.ncbi.nlm.nih.gov/Database/datamodel/index.html

Entrez databases and the connections between them



http://www.ncbi.nlm.nih.gov/Database/datamodel/index.html