

# Bioinformatics Blue Print of Genes

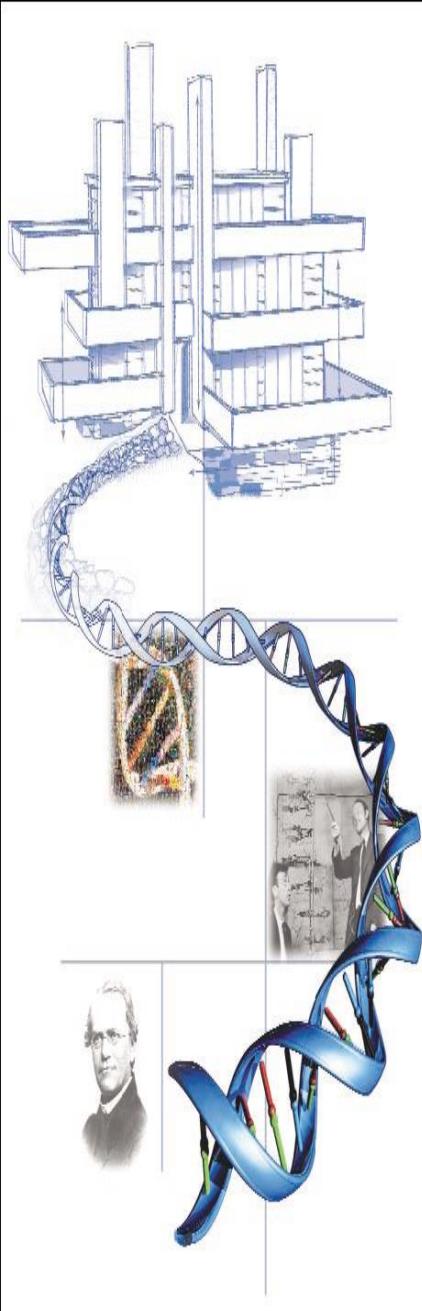
Prof. Fatchiyah, M.Kes. Ph.D

Dept. of Biology  
School of Math. and Natural Sciences  
Brawijaya University

# What is Bioinformatics

- The use of computers to collect, analyze, and interpret biological information at the molecular level.

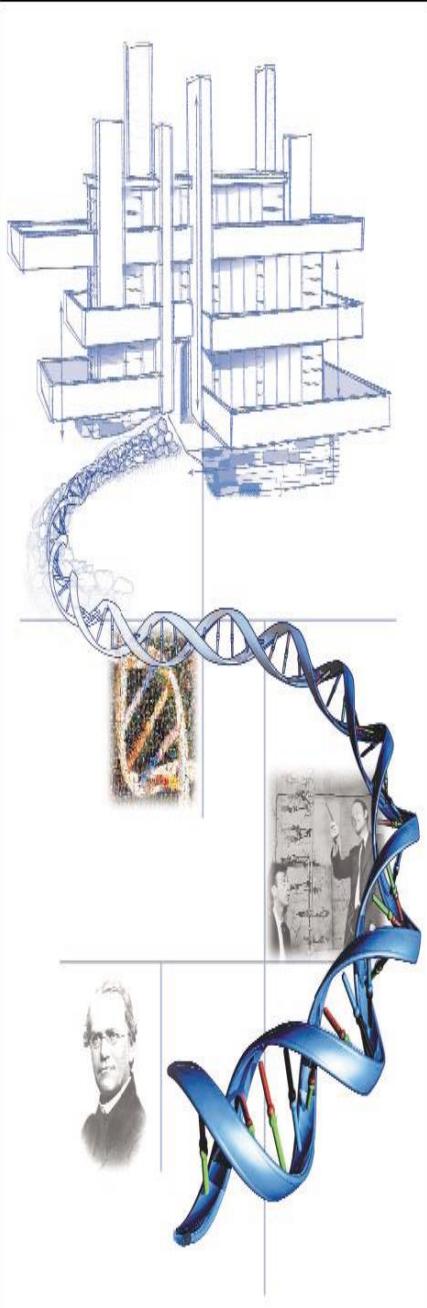
*"The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information."*
- A set of software tools for molecular sequence analysis



## The National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>)

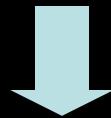
NCBI established on November 4' 1988, as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH)

The GenBank sequence database is an annotated collection of all publicly available nucleotide sequences and their protein translations.

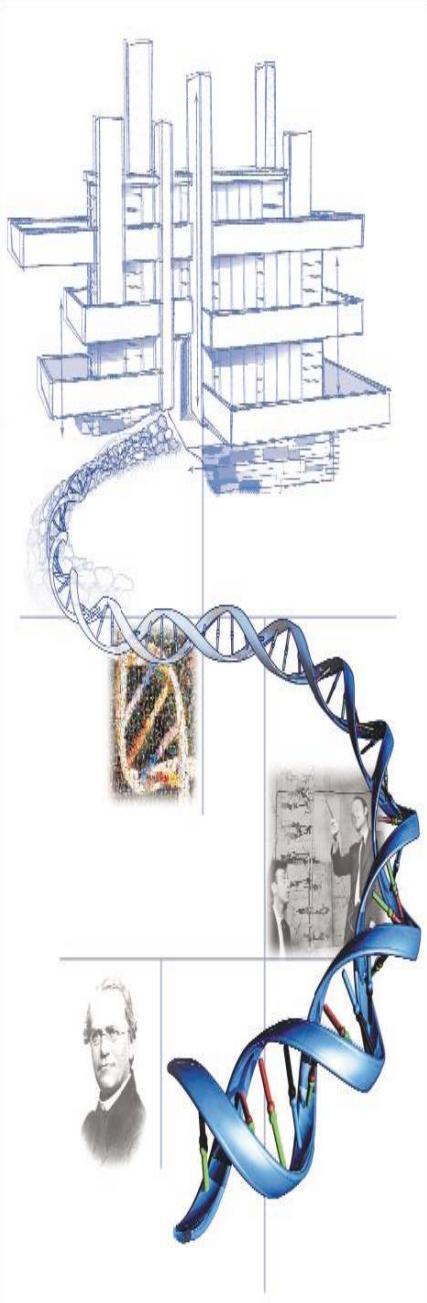


Initially, GenBank was built and maintained at Los Alamos National Laboratory (LANL).

NCBI began accepting direct submissions to GenBank in 1993 and received data from LANL until 1996



Currently, NCBI receives and processes about 20,000 direct submission sequences per month, in addition to the approximately 200,000 bulk submissions that are processed automatically.



NCBI has a multi-disciplinary research group:

Molecular Biologists

Computer Scientists

Biochemists

Mathematicians

Research Physicians

Structural Biologists

# *What is Bioinformatics ?*

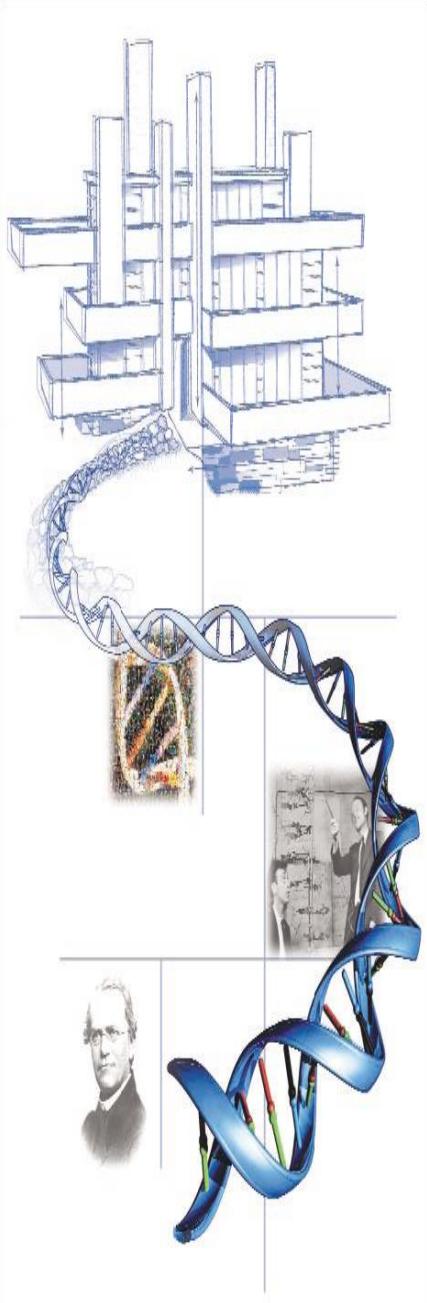


*Bioinformatics*

||

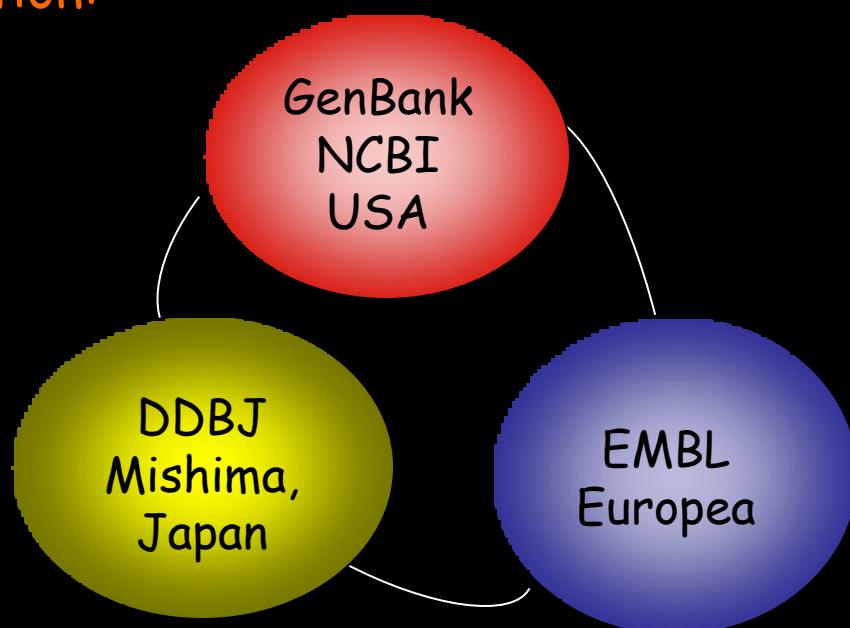
*Molecular Biology + Information Technology  
& Biotechnology (IT)*

*To understand living organisms  
with Bio-information*

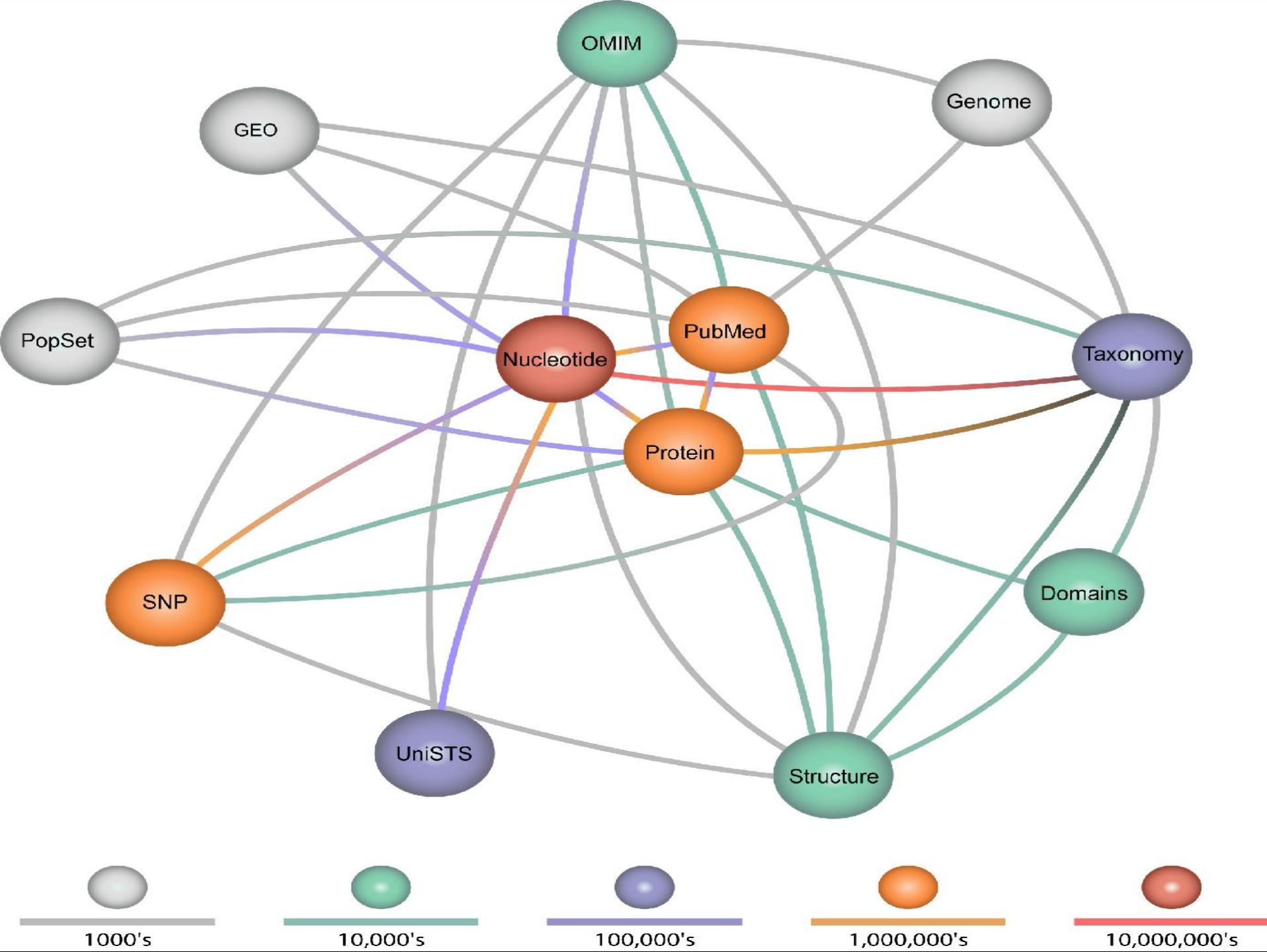


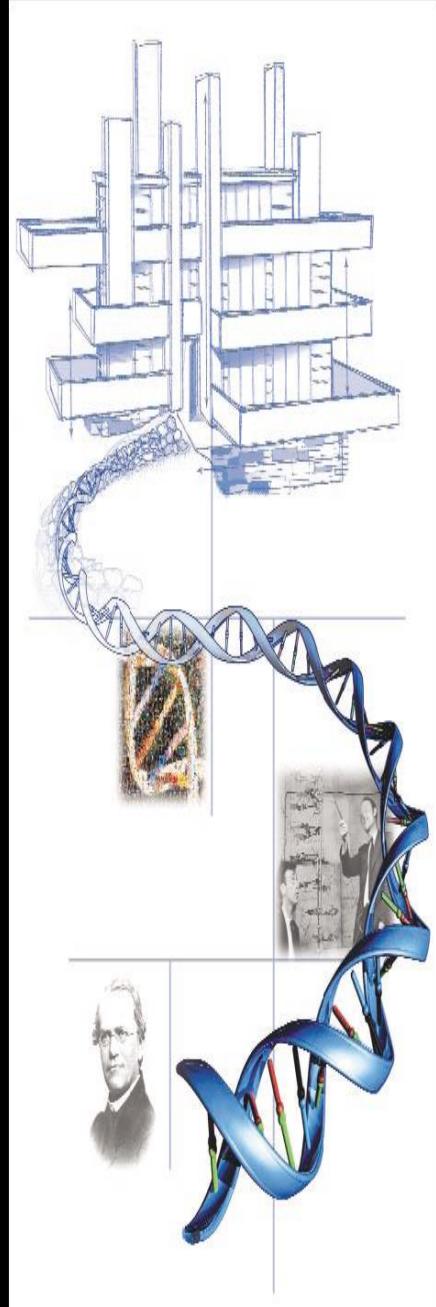
## Internationally Networking Collaboration

In the mid-1990s, the GenBank database became part of the International Nucleotide Sequence Database Collaboration:



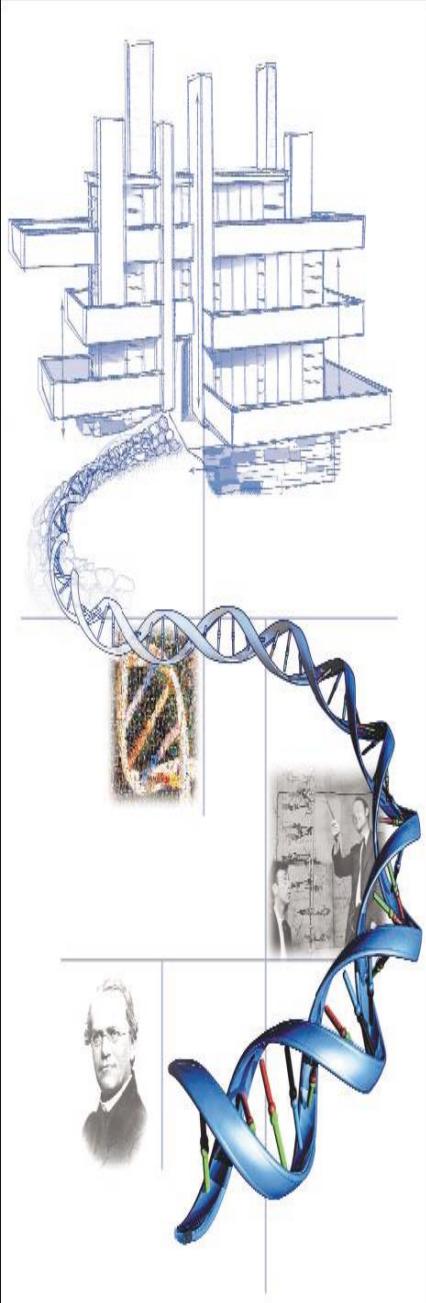
NCBI investigators maintain ongoing collaborations with several institutes within NIH and also with numerous academic and government research laboratories



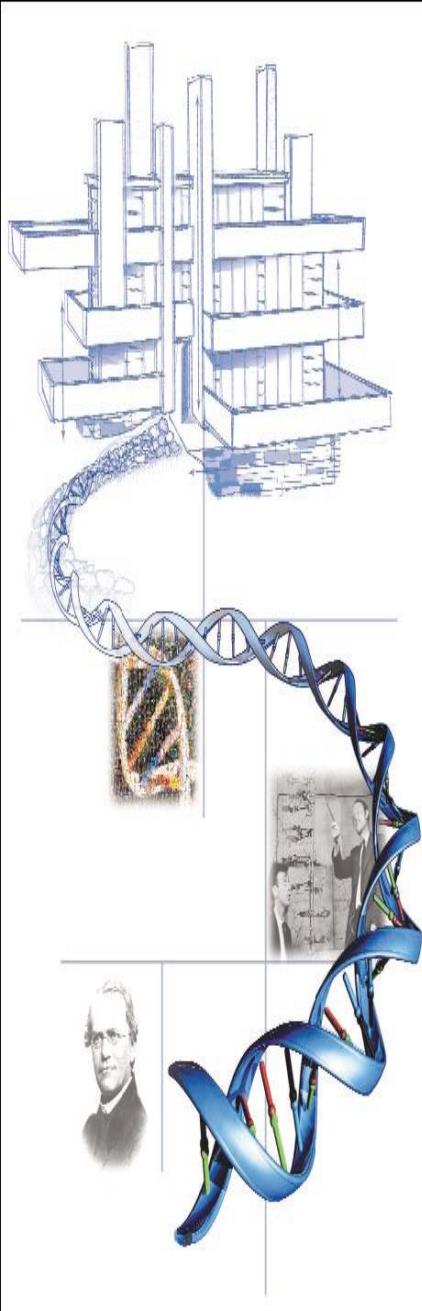


**Genomic era is now a reality**





1. Gregor Mendel's discovery of laws of heredity, in the early 20<sup>th</sup> century. Recognition of DNA as material genetics
2. The discovery of the double-helical structure of DNA, in 1953 is a landmark event. And elucidation of the genetics code
3. Development of sequencing and DNA Recombinant technologies
4. Establishment of increasingly automatable methods for DNA Sequencing
5. Human Genome Project (HGP) began in 1990



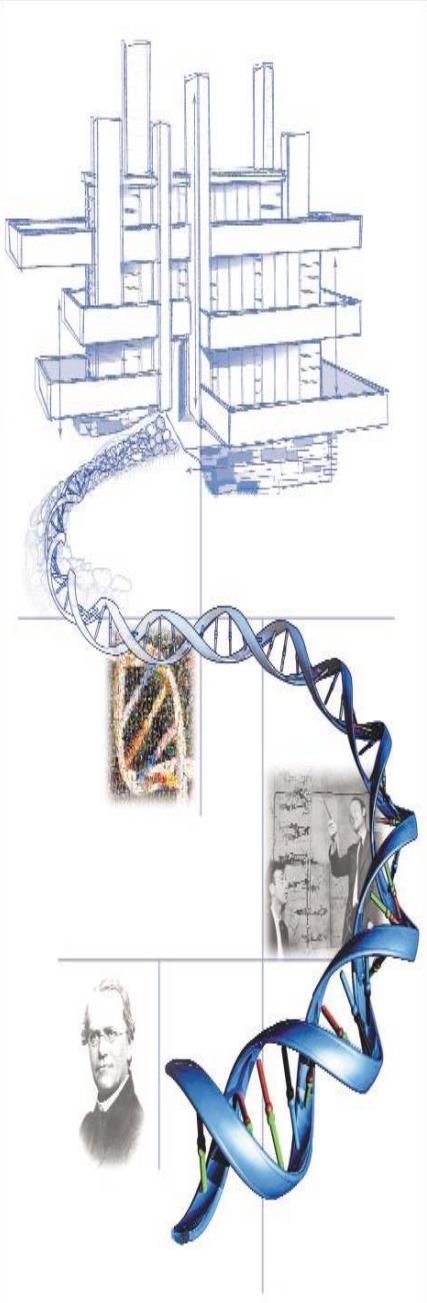
## Human Genome Project (HGP):

Provide in the understanding of gene structure, genetics variation and comparative genomics

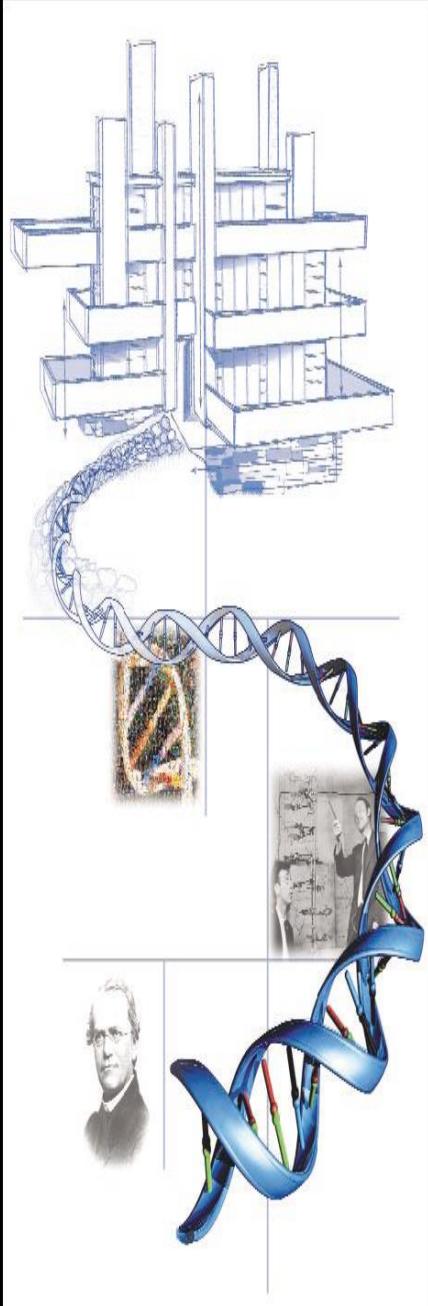
Provide the basis for 'sequence-based biology'

Appreciation of ethical, legal, and social issues surrounding the availability of human sequence data

## Resources



1. Databases that integrate sequences with curate information and other large data sets, as well as tools for effective mining of the data
2. Reference sets of coding sequences for example, full-length cDNA sequences and corresponding clones, oligonucleotide primers, and microarrays
3. Collections of knockouts and knock-downs of all genes in selected animals to accelerate the development of models of disease
4. Cohort populations for studies designed to identify genetic contributors to health and to assess the effect of individual gene variants on disease risk, including a 'healthy' cohort.



## Technology Development

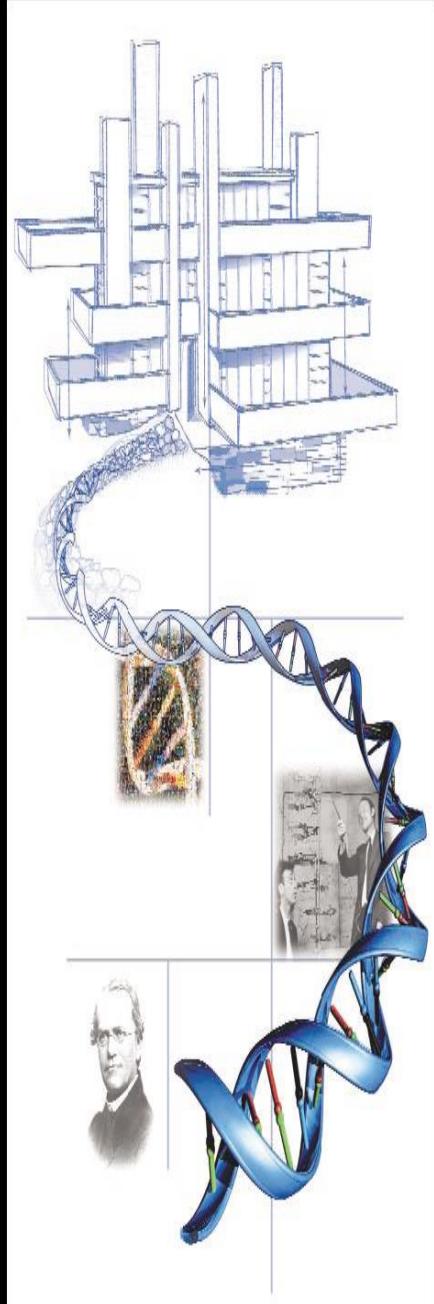
The Human Genome Project was aided by several 'breakthrough' technological developments

- Sanger DNA sequencing and its automation
- DNA-based genetic markers
- Large-insert cloning systems
- The polymerase chain reaction

'evolutionary' technology

Capillary-based sequencing and methods for genotyping single-nucleotide polymorphisms

Nanotechnology and micro-fluidics



## Further technologies .....

Sequencing and genotyping technologies to reduce costs further and increase access to a wider range of investigators

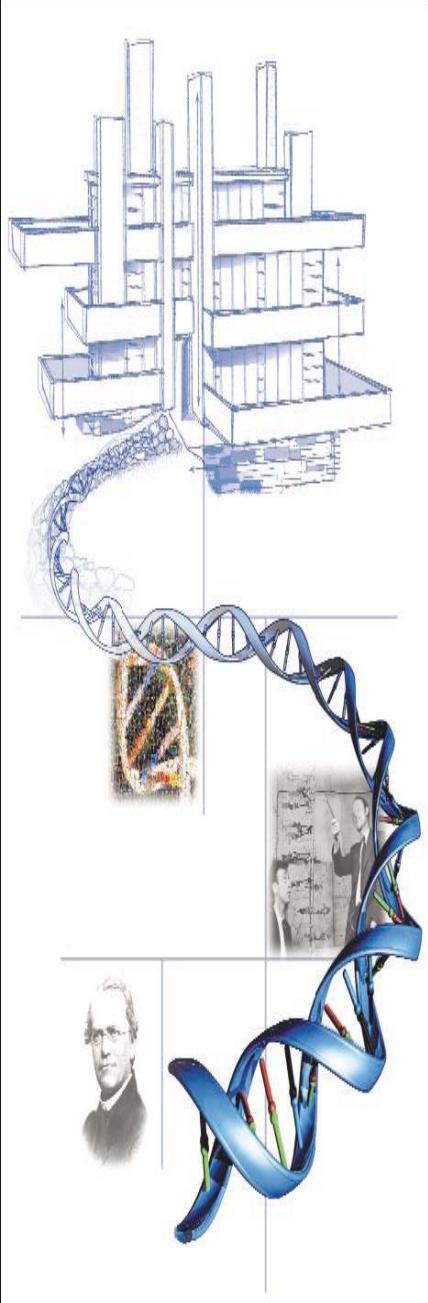
Identification and validation of functional elements that do not encode protein

*In vivo*, real-time monitoring of gene expression and the localization, specificity, modification and activity/kinetics of gene products in all relevant cell types

Modulation of expression of all gene products using, for example, large-scale mutagenesis, small-molecule inhibitors and knock-down approaches (such as RNA-mediated inhibition)

Etc.

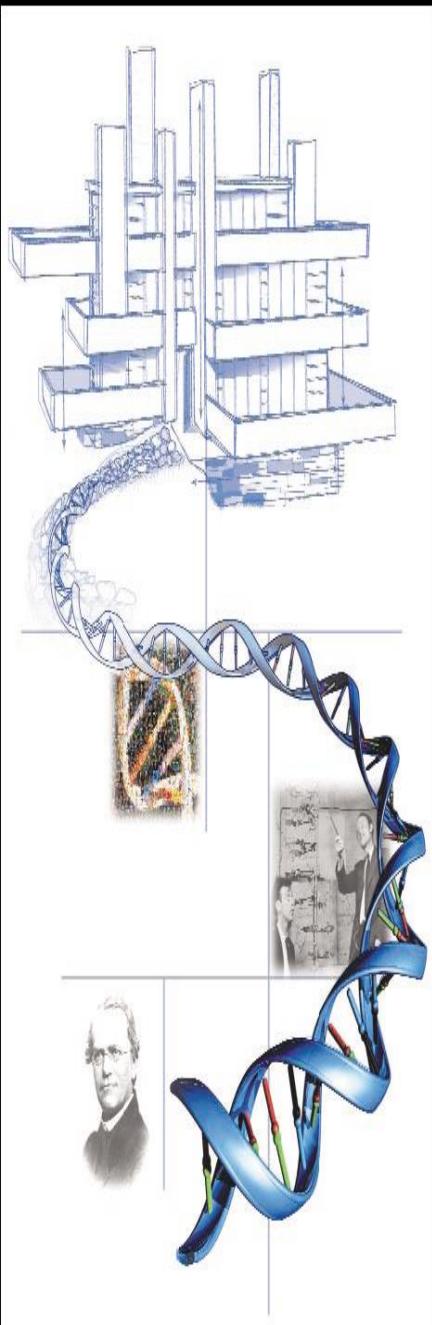
# Computational Biology Branch (CBB)

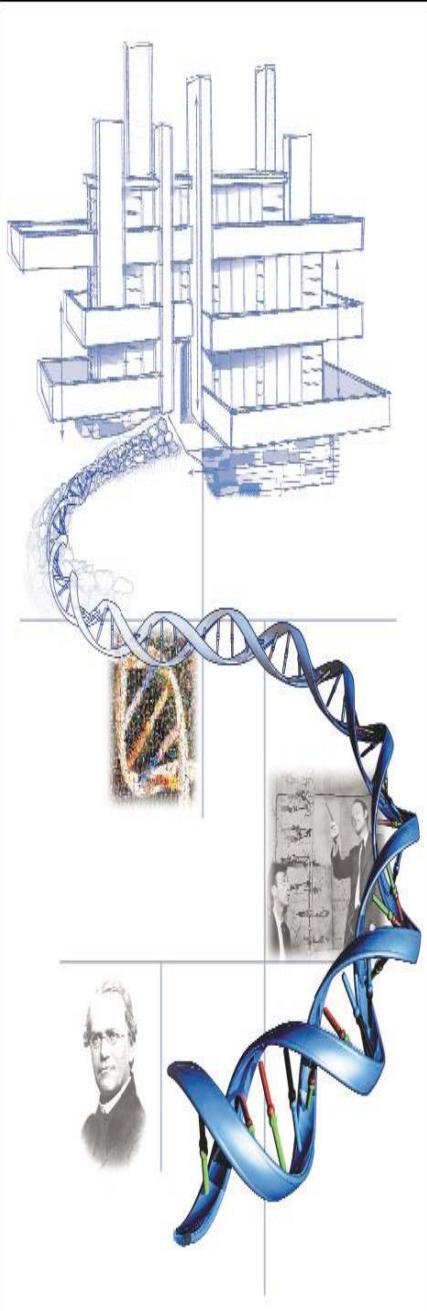


1. New approaches to solving problems, such as the identification of different features in a DNA sequence, the analysis of gene expression and regulation, the elucidation of protein structure and protein-protein interactions, the determination of the relationship between genotype and phenotype, and the identification of the patterns of genetic variation in populations and the processes that produced those patterns
2. Improved database technologies to facilitate the integration and visualization of different data types, for example, information about pathways, protein structure, gene variation, chemical inhibition and clinical information/phenotypes
3. Improved knowledge management systems and the standardization of data sets to allow the coalescence of knowledge across disciplines: biologist, chemists, mathematicians, & computer scientists

## Ethical, legal and social implications (ELSI)

- The development of models of genomics research that use attention to these ELSI issues for enhancing the research, rather than viewing such issues as impediments
- The continued development of appropriate and effective genomics research methods and policies that promote the highest levels of science and of protecting human subjects
- The establishment of crosscutting tools, analogous to the publicly accessible genomic maps and sequence databases that have accelerated other genomics research (examples of such tools might include searchable databases of genomic legislation and policies from around the world, or studies of ELSI aspects of introducing clinical genetic tests)
- The evaluation of new genetic and genomic tests and technologies, and effective oversight of their implementation, to ensure that only those with confirmed clinical validity are used for patient care.

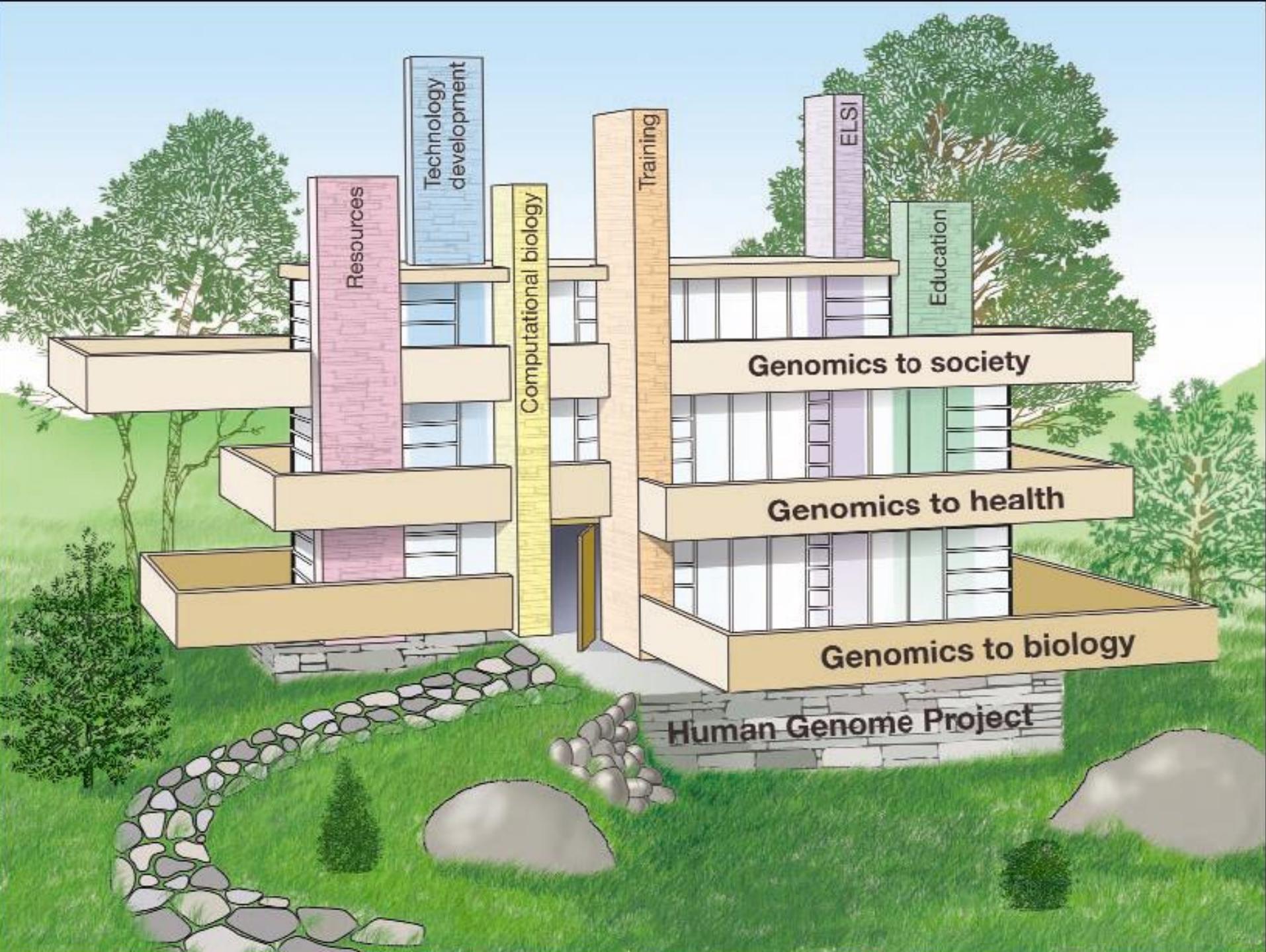




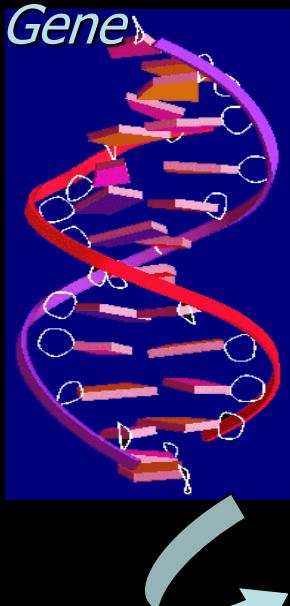
## Genomics to society:

The time is right to develop and apply large-scale genomic strategies to improve human health.

1. Develop policy options for the uses of genomics in medical and non-medical settings.
2. Understand the relationships between genomics, race and ethnicity, and the consequences of uncovering these relationships
3. Understand the consequences of uncovering the genomic contributions to human traits and behaviour
4. Assess how to define the ethical boundaries for uses of genomics

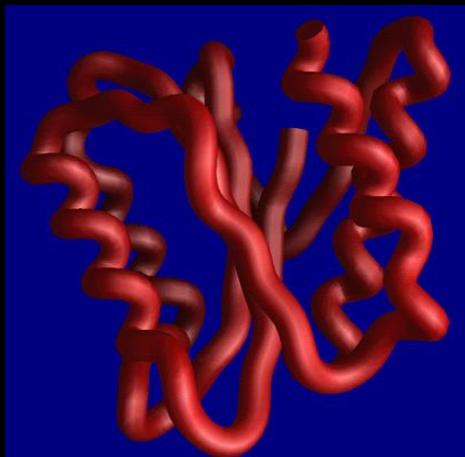


# The Flow of Biotechnology Information



## > DNA sequence

AATTCAATGAAAATCGTATACTGGTCTGGTACCGGCAACAC  
TGAGAAAAATGGCAGAGCTCATCGCTAAAGGTATCATCGAA  
TCTGGTAAAGACGTCAACACCATAACGTGTCTGACGTTA  
ACATCGATGAAC TGCTGAACGAAGATATCCTGATCCTGGG  
TTGCTCTGCCATGGCGATGAAGTTCTCGAGGAAAGCGAA  
TTTGAACCGTTCATCGAACAGAGATCTCTACCAAAATCTCTG  
GTAAGAACGGTTGCGCTGTTGGTTCTACGGTTGGGGCGA  
CGGTAAGTGGATGCGTGACTTCGAAGAACGTATGAACGGC  
TACGGTTGCGTTGTTGAGACCCCCGCTGATCGTTCAGA  
ACGAGCCGGACGAAGCTGAGCAGGACTGCATCGAACATTGG  
TAAGAACAGATCGCGAACATCTAGTAGA



## > Protein sequence

MKIVYWSGTGNTEKMAELIAKGI  
IESGKDVTNTINVSDVNI  
DELLNEDILILGCSAMGDEVLEESE  
FEPFIEEISTKISGK  
KVALFGSYGWGDGKWMRDFEERM  
NGYGVVVETPLIVQNE  
PDEAEQDCIEFGKKIANI

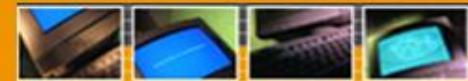
# Biology Information on the Internet



- **Introduction to Databases**
- **Searching the Internet for Biology Information.**
  - General Search methods
  - Biology Web sites
- **Introduction to Genbank file format.**
- **Introduction to Entrez and Pubmed**
- **Ref: Chapters 1,2,5,6 of “Bioinformatics”**

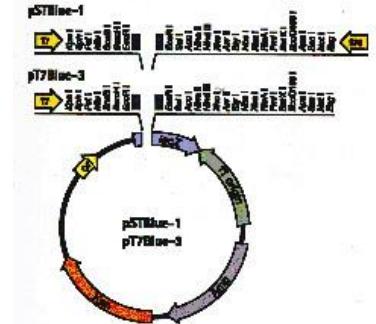


# **Free unrestricted access for all**



## The Wellcome Trust

*The door to discovery is wide open*



### Genome browsers

Ensembl

[www.ensembl.org](http://www.ensembl.org)

University of California Santa Cruz

<http://genome.cse.ucsc.edu>

MGD the Jackson Laboratory

[www.informatics.jax.org](http://www.informatics.jax.org)

### Genome Databases

European Bioinformatics Institutes

[www.ebi.ac.uk](http://www.ebi.ac.uk)

GenBank

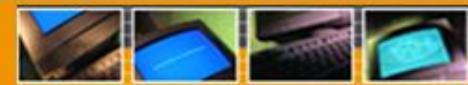
[www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)

DNA Data Bank of Japan

[www.ddbj.nig.ac.jp](http://www.ddbj.nig.ac.jp)



# Refseq and LocusLink



- Attempt to produce **1 mRNA, 1 protein, and 1 genomic gene for each frequently occurring allele of a protein expressing gene.**
- [www.ncbi.nlm.nih.gov/LocusLink](http://www.ncbi.nlm.nih.gov/LocusLink)
- **Special non-genbank Accession numbers**
  - **NM\_nnnnnn mRNA refseq**
  - **NP\_nnnnnn protein refseq**
  - **NC\_nnnnnn refseq genomic contig**
  - **NT\_nnnnnn temporary genomic contig**
  - **NX\_nnnnnn predicted gene**



# Browsing on GenBank (NCBI)

Firefox Ad4BP/SF-1 - Nucleotide - NCBI BLAST: Basic Local Alignment Search ... +

www.ncbi.nlm.nih.gov/nuccore/?term=Ad4BP%2FSF1

UtilityCHEST To Do List Planning Tools Calculators Converters Language Tools 69°F Malang, Indonesia Listen to the Radio More...

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Ad4BP/SF-1 Save search Limits Advanced Search Help

Display Settings:  Summary, 20 per page, Sorted by Default order

Found 30757 nucleotide sequences. Nucleotide (30) EST (30727)

There were some problems retrieving the sequence. GI: 11527548

There were some problems retrieving the sequence. GI: 168479586

Results: 1 to 20 of 30

<< First < Prev Page 1 of 2 Next > Last >>

Rattus norvegicus gene for Ad4BP/SF-1, partial cds  
1. Accession: GI: 11527548  
GenBank FASTA Graphics Related Sequences

Eubalepharis macularius Ad4BP/Sf-1 mRNA for Ad4BP, complete cds  
2. Accession: GI: 168479586  
GenBank FASTA Graphics

Rattus norvegicus nuclear receptor subfamily 5, group A, member 1 (Nr5a1), mRNA  
3. 2,182 bp linear mRNA  
Accession: NM\_001191099.1 GI: 300797823  
GenBank FASTA Graphics Related Sequences

Rattus norvegicus strain BN/ScNUiedMCW chromosome 2, Draft 5.0

Send to:  Filter your results:  
All (30)  
Bacteria (0)  
INSDC (GenBank) (13)  
mRNA (21)  
RefSeq (17)

Manage Filters

Top Organisms [Tree]  
Rattus norvegicus (8)  
Oryzias latipes (5)  
Bos taurus (4)  
Mus musculus (3)  
Homo sapiens (3)  
All other taxa (7)  
More...

Find related data

9:48 PM 2/16/2014

# Header

# GeneBank Record



□ 1: NM\_000457. Homo sapiens hep...[gi:21361184]

LOCUS HNF4A  
DEFINITION Homo sapiens hepatocyte nuclear factor 4,  
ACCESSION NM\_000457  
VERSION NM\_000457.2  
KEYWORDS .  
SOURCE human.  
ORGANISM [Homo sapiens](#)  
REMARKS Eukaryota; Metazoa; Chordata; Craniota; Vertebrata; Euteleostomi;  
Mammalia; Eutheria; Primates; Catarrhini; Hominidae;  
REFERENCE 1 (bases 1 to 2289)  
AUTHORS Bell, G.I., Xiang, K.S., Newman, M.V., Wu, S.H., Wright, J.,  
Fajans, S.S., Spielman, R.S. and Cox, N.J.  
TITLE Gene for non-insulin-dependent diabetes mellitus (maturity-onset diabetes of the young subtype) is linked to DNA polymorphisms on human chromosome 20q  
JOURNAL Proc. Natl. Acad. Sci. USA 88:1138-1142 (1991)  
MEDLINE 91142197  
PUBMED 1899928

## Accession Number

Sequence Length

Version Number

TITLE Cloning and sequencing of cDNAs encoding the human hepatocyte nuclear factor 4 indicate the presence of two isoforms  
JOURNAL [Proc. Natl. Acad. Sci. USA](#) 89:1138-1142 (1991)  
MEDLINE 91142197  
PUBMED 1899928

REFERENCE 3 (bases 1 to 2289)  
AUTHORS Drewes, T., Senkel, S., Holewa, B. and Ryffel, G.U.

modification date

Molecule Type

Locus Name

Modification Date

GenBank Division

# GeneBank Record

FEATURES	Location/Qualifiers
source	1..2289 /organism="Homo sapiens" /db_xref="taxon:9606" /chromosome="20" /map="20q12-q13.1" /tissue_type="kidney" /clone_lib="cDNA library"
<u>gene</u>	1..2289 /gene="HNF4A" /note="TCF; HNF4; MODY; MODY1; NR2A1; TCF14; NR2A21" /db_xref="LocusID: <a href="#">3172</a> " /db_xref="MIM: <a href="#">600281</a> "
.	complement(86) /allele="C" /allele="G" /db_xref="dbSNP: <a href="#">1063240</a> "
<u>variation</u>	104..1618 /gene="HNF4A" /function="transcription factor" /note="transcription factor-14; hepatic nuclear factor 4, alpha" /codon_start=1 /evidence=experimental /product="hepatocyte nuclear factor 4, alpha" /protein_id=" <a href="#">NP_000448.2</a> " /db_xref="GI:21361185" /db_xref="LocusID: <a href="#">3172</a> " /db_xref="MIM: <a href="#">600281</a> "
<u>CDS</u>	/translation="MRLSKTLVMDMADYSAAALDPAYTTLFENVQVLTMGN DLLPLR LARLRHPLRHWSISGGVDSSPQGDTSPSEGTLNAPNSLGV SALCAICGDRATGKHY GASSCDGCKGFRRSRVKNHMYSCRFSRQC VVDKD KRQN QCRY CRLKKC FRAGMKKEAV QNERDRISTRRSSYEDSSLPSIN ALLQAEVLSRQITSPVSGINGDIRAKKIASIADVC ESMK EQLLV LVEWAKYI PAFCE PLDDQV ALLRAHAGEHLLLGATKRS MVFKD VL LLG NDYIVPRHCPELAEMS RVSIRILDELVLPFQELQIDDNEYAYLKAIIFDFPD AKGLSD PGKIKRRLRSQVQSLEDYINDRQYDSRGFGE LLLLPTLQSITUQMIEQIQF IKLFG MAKIDNLLQEMLLGGSPSDAPHAHPLHPHLMQE HMG TNVIVANTMP THLSNGQMCEW

**CDS (Coding DNA Sequence)**

**Accession Number for Protein**

# PROTEIN SEQUENCE



- mRNA <1..>1380
- sig\_peptide 33..104
- CDS 33..1283
  - /codon\_start=1
  - /product="alpha-1-antitrypsin"
  - /protein\_id="CAA44840.1"
  - /db\_xref="PID:g42"
  - /db\_xref="GI:42"
  - /db\_xref="SWISS-PROT:P34955"
- /translation="MALSITRGLLLLAALCCLAPISLAGVLQGHAVQETDDTSHQEAACHKIAPNLANFAFSIYHHHLAHQSNTSNIFFSPVSIASAFAMLSLGAKGNTHTEILKGLGFNLTELAEAEIHKGFQHLLHTLNQPNHQLQLTTGNGLFINESAKLVDTFLEDVKNLYHSEAFSINFRDAEEAKKKINDYVEKGSHGKIVELVKVLDPTVFAVLVNYISFKGKWEKPFEVKHTTERDFHVDEQTTVKVPMMNRGLGMFDLHYCDKLASWVLLDYVGNTACFILPDLGKLQQLEDKLNNELLAKFLEKKYASSANLHLPKLSISETYDLKSVLGDVGITEVFSDRADLSGITKEQPLKVSKALHKAALTIDEKGTEAVGSTFLEAIPMSLPPDVEFNRPFLCILYDRNTKSPLFVGKVVNPTQA"
- mat\_peptide 105..1280
  - /product="alpha-1-antitrypsin"
- polyA\_signal 1343..1348
- polyA\_site 1368

*Protein Sequence*



# Gene Sequence

BASE COUNT    489 a    723 c    653 g    424 t  
ORIGIN

1 gcagagaggg cactgggagg aggcaagtggg agggcggagg gcggggccct tcgggggtggg  
61 cgcggcagggt agggcaggtg gcccgcggcg ggaggcaggg agaatgcacat tctccaaaac  
121 cctcggtcgac atggacatgg ccgactacag tgctgcactg gacccagcct acaccaccc  
181 ggaattttag aatgtgcagg ttgtgacgt gggcaatgtat ttgttgcgc tgcgtctcg  
241 cagattgagg catccccctcc gacatcactg gagcatatct ggaggggtgg acagttctcc  
301 acagggagac acgtccccat cagaaggcac caacctaaca gcgcggcaaca gcctgggtgt  
361 cagcgccttg tgtgccatct gcgggggaccg ggccacgggc aaacactacg gtgcctcgag  
421 ctgtgacggc tgcaagggtct tttccggag gagcgtgcgg aagaaccaca tgtactctcg  
481 cagatttagc cggcagtgcc tggtggacaa agacaagagg aaccagtgcgc gctactgcag  
541 gctcaagaaa tgcttccggg ctggcatgaa gaaggaagcc gtccagaatg agcggggaccg  
601 gatcagcact cgaaggtaa gctatgagga cagcagcctg ccctccatca atgcgctct  
661 gcaggcggag gtcctgtccc gacagatcac ctcccccgtc tccgggatca acggcgacat  
721 tcgggcgaag aagattgcca gcatcgcaga tgtgtgttag tccatgaagg agcagctgt  
781 ggttctcggt gagtgggcca agtacatccc agctttctgc gagctcccccc tggacgacca  
841 ggtggccctg ctcagagccc atgctggcga gcacctgctg ctcggagccca ccaagagatc  
901 catggtgttc aaggacgtgc tgctcttagg caatgactac attgtccctc ggcactgccc  
961 ggagctggcg gagatgagcc gggtgtccat acgcatactt gacgagctgg tgctggccctt  
1021 ccaggagctg cagatcgatg acaatgagta tgcctacctc aaagccatca tcttctttga  
1081 cccagatgcc aaggggctga gcgatccagg gaagatcaag cggctgcgtt cccaggtgca  
1141 ggtgagctt gaggactaca tcaacgaccg ccagtatgac tgcgtggcc gctttggaga  
1201 gctgctgctg ctgctgcucca cttgcagag catcacgtgg cagatgatcg agcagatcc  
1261 gttcatcaag ctttcggca tggccaaatgat tgacaaacctg ttgcaggaga tgctgctgg  
1321 agggcccccc agcgatgcac cccatgccc ccacccctg caccctcacc ttagtcggaga  
1381 acatatggga accaacgtca tcgttgccaa cacaatgccc actcacctca gcaacggaca  
1441 gatgtgttag tggcccccgcac ccaggggaca ggcagccacc cctgagaccc cacagccctc  
1501 accgccagggt ggctcagggt ctgagcccta taagctcctg ccggggcccg tcgcccacaa  
1561 cgtcaagccc ctctctgcctt tccccccagcc gaccatcacc aagcagggaaat ttatcttaga  
1621 agccgcgtggg gcttgggggc tccactggct ccccccagcc ccctaagaga gcacctggtg  
1681 atcacgtggt cacggcaaag gaagacgtga tgccaggacc agtcccagag caggaatggg

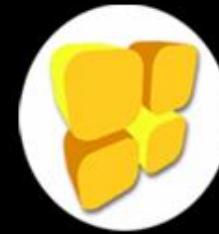
Gene Sequence

Initiation Codon

CDS (Coding DNA Sequence)

Stop Codon (UAG)

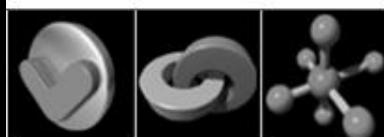
# BLAST Sequence Analysis Tool



***Basic Local Alignment Search Tool (BLAST) (1, 2) is the tool most frequently used for calculating sequence similarity. BLAST comes in variations for use with different query sequences against different databases.***

***All BLAST applications, as well as information on which BLAST program to use and other help documentation, are listed on the BLAST homepage [<http://www.ncbi.nlm.nih.gov/BLAST/>].***

***To understand how BLAST works, its output, and how both the output and program itself can be further manipulated or customized, rather than on how to use BLAST [<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>] or interpret BLAST results.***



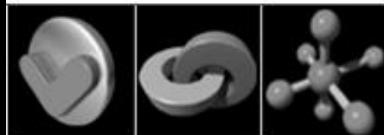
# Function of BLAST Analysis



*The comparison of nucleotide or protein sequences from the same or different organisms is a very powerful tool in molecular biology.*

*By finding similarities between sequences, scientists can infer the function of newly sequenced genes, predict new members of gene families, and explore evolutionary relationships.*

*Now that whole genomes are being sequenced, sequence similarity searching can be used to predict the location and function of protein-coding and transcription regulation regions in genomic DNA.*



# FASTA Format

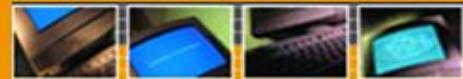
>identifier descriptive text  
nucleotide or amino-acid  
sequence on multiple lines if needed:

*MOST important  
data format!!!*

Example:

>gi|41|emb|X63129.1|BTA1AT B. taurus mRNA for alpha-1-anti-trypsin  
GACCAGCCCTGACCTAGGACAGTGAATCGATAATGGCACTCTC  
CATCACGCGGGCCTTGCTGCTGGC ....

# Modified FASTA Format



- 1) A few tools follow the convention that lower case sequences are masked. (repeat masker, some versions of **blast**, megablast, blastz)
- 2) A few analysis tools (like **CLUSTAL**) want a simplified identifier on the defline. So they can have a short string for the alignment.

```
>X63129.1
GACCAGCCCTGACCTAGGACAGTGAATCGATAATGGCACTCTC
CATCACGCGGGGCCTTCTGCTGCTGGC ....
```



# Blast Assembly

Firefox ▾ Ad4BP/SF-1 - Nucleotide - NCBI × BLAST: Basic Local Alignment Search ... × Inbox (257) - fatchiya@gmail.com - G... × +

blast.ncbi.nlm.nih.gov/Blast.cgi

Search UTILITYCHEST To Do List Planning Tools Calculators Converters Language Tools 69°F Malang, Indonesia

Ask More My NCBI [Sign In] [Register]

NCBI/ BLAST Home

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

BLAST finds regions of similarity between biological sequences. [more...](#)

New DELTA-BLAST, a more sensitive protein-protein search [Go](#)

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

<input type="checkbox"/> <a href="#">Human</a>	<input type="checkbox"/> <a href="#">Oryza sativa</a>	<input type="checkbox"/> <a href="#">Gallus gallus</a>
<input type="checkbox"/> <a href="#">Mouse</a>	<input type="checkbox"/> <a href="#">Bos taurus</a>	<input type="checkbox"/> <a href="#">Pan troglodytes</a>
<input type="checkbox"/> <a href="#">Rat</a>	<input type="checkbox"/> <a href="#">Danio rerio</a>	<input type="checkbox"/> <a href="#">Microbes</a>
<input type="checkbox"/> <a href="#">Arabidopsis thaliana</a>	<input type="checkbox"/> <a href="#">Drosophila melanogaster</a>	<input type="checkbox"/> <a href="#">Apis mellifera</a>

Basic BLAST

Choose a BLAST program to run.

<a href="#">nucleotide blast</a>	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontiguous megablast
<a href="#">protein blast</a>	Search protein database using a protein query <i>Algorithms:</i> blastp, psi-blast, phi-blast, delta-blast
<a href="#">blastx</a>	Search protein database using a translated nucleotide query
<a href="#">tblastn</a>	Search translated nucleotide database using a protein query
<a href="#">tblastx</a>	Search translated nucleotide database using a translated nucleotide query

Your Recent Results [New!](#)

[All Recent results...](#)

News

**BLAST 2.2.29+ released**

A new version of the stand-alone BLAST+ applications is available.  
Mon, 06 Jan 2014 12:00:00 EST

[More BLAST news...](#)

Tip of the Day

[More tips...](#)

Bioinformatic... e Firefox Skype PROLiNK HSP... bioinformatic... BLAST: Basic ... 7:55 PM 2/16/2014

# BLAST Analysis

**Copy the Sequence**

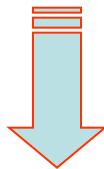
"MASQGTRSYE QMET GGERQNATE IRASUGRMISGIGRFYIQMC  
TELKLSDYEGRIL IQNS IT IERMULSAFDERPMPYLEEHPSAGKDPKETGGP IVYRRDG  
KWWREL ILHDKEE IRR IWRQANNGEDATAGLTHLMIWHEMLNDATYQRTRALURTGMD  
PRMC3LMQGSTLPPR3GA&AAUKGUGTMUMELIRMIKRGINDRNURGENGERTRIA  
YERMCN ILKGKFQT&AQRMIDQURESRMPGNMEIEDLIFLARSLILRGSVAMHSCL  
PACUVYGL&UASGYD FERE GYSLUG IDP FRLLQNSQFSL IRPNENPAHKSQLUWMACH  
3AAFEDLRUSS FIRGTRUUPRGQLSTRGUQ IASNEMMEAMDNTLELRSKYWAIRTRS  
GGNTNQQRAS AGQIS IQPT FSVQRNLP FERAT IMAAFTGNT EGRT 3DMRTE I IRMME 3  
ARPEDU3FQGRGUFE LDEKATNP IUP3FDMMNNE GSYFFGDN&EYDN"

ORIGIN

1 gtagataatc actcacccgag tgacatcaac atccatgggt ctcaaggcac caaacgatct  
61 tatgaacaga tggaaactgg tggagaaege cagaatgtca ctgagatctag ggcatctgtt  
121 ggaagaatga ttatgtggcat cgggagggttc tacatcacaga tgtgcacaga actcaaaactc  
181 agtgactatg aagggaggct tattccagaaac agcataacaa tagagagaat ggtactctct  
241 gcattttatgt aacgaaggaa cagataactg gaagaacacc ccagtgcggg gaaggaccgg  
301 aaaaaaactg gaggtccaat ttataggagg agagacggga aatgggtgag agagctgatt  
361 ctgcacgaca aagaggagat caggaggatt tggcgbcagg cgaacaatgg agaggacgca  
421 actgtggtc tcacccacct gatgatatgg catccaaatc taatgtatgc cacatatacg  
481 agaaccgagag ctcttgtacg tactggatg gacccccagga tgtgctctct gatgcacaggg  
541 tcaactctcc cgaggaggatc tggagctgcc ggtgcaggag tggagggggtt agggacaatg  
601 gtaatggagc tgattcggat gataaaaacga gggatcaacg accggaaattt ctggagaggg  
661 gaaaatggaa gaagaacaag gattgtcatat gagagaatgt gcaacatct caaaggaa  
721 ttccaaacag cagcacaaaaag agcaatgtatg gatcaggtgc gagagagcag aaatccctggg  
781 aatgtgtaaaa ttgaagatct catttttctg gcacggctcg cactcatct gagaggatca  
841 gttagccata agtcctgttt gcctgttgtt gtgtacggac ttgtctgtggc cagtggatat  
901 gactttgaga gagaagggtt ctctctgttt ggaatagatc ttccctgttt gettcaaaaac  
961 agccaggatct ttatgtctat tagacccaaat gagaatccgg cacataagag tcaattatgt  
1021 tggatggcat gccactctgc agcattttag gaccctttagag ttcacagttt catcagaggg  
1081 acaagagtgg tccccaaagagg acagctatcc accagagggg ttcaatgtc ttcaatgtag  
1141 aacatggaaag caatggactc caacacttt gaactgagaa gtaatatgtt ggctataaga  
1201 accagaagcg gaggaaacac caaccagcg aggcatctg caggccagat aagcatccag  
1261 cccactttct cggtagacag aacaccttca ttccgaaagag cgaccattat ggcagccatc  
1321 acaggaaata ctgagggcag aacgtctgac atgagaactg aatcataag aatgtatggaa  
1381 agtggccagac cagaagatgt gtcatteccag gggcggggag tttcgagct ctggacggaa  
1441 aaggcaacga acccgatctg gccttccctt gacatgtata atgaaggatc ttatcttc  
1501 ggagacaaatg cagaggagta tgacaaatca agaaaaatac

[PubMed](#)[All Databases](#)[BLAST](#)[OMIM](#)

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.



### Nucleotide

[Quickly search for highly similar sequences \(megablast\)](#)

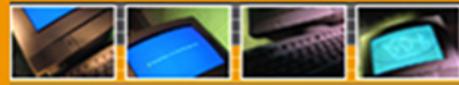
[Quickly search for divergent sequences \(discontiguous megablast\)](#)

[Nucleotide-nucleotide BLAST \(blastn\)](#)

[Search for short, nearly exact matches](#)

[Search trace archives with megablast or discontiguous megablast](#)

# FASTA Format



```
1 gaccagccct gacctaggac agtgaatcga taatggcact ctccatcacg cggggccttc  
61 tgctgctggc agccctgtgc tgcctggccc ccatctccct ggctggagtt ctccaaggac  
121 acgctgtcca agagacagat gatacatccc accaggaagc agcgtgccac aagattgcc  
181 ccaaacctggc caactttgcc ttcagcatat accaccattt ggctcatcag tccaaacacca  
241 gcaacatctt cttctcccccc gtgagcatcg cttcagcctt tgcgatgctc tccctggag  
301 ccaagggcaa cactcacact gagatcctga agggcctggg ttcaacctc actgagctg  
361 cagaggctga gatccacaaa ggcttcagc atcttctcca caccctgaac cagccaaacc
```

*nr*

*Format*

*Reset*



# Blast Result



*results of* **BLAST**

**BLASTP 2.2.1 [Apr-13-2001]**

**Reference:**

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

RID: 1009580302-26840-4362

**Query=** RAB protein  
(656 letters)

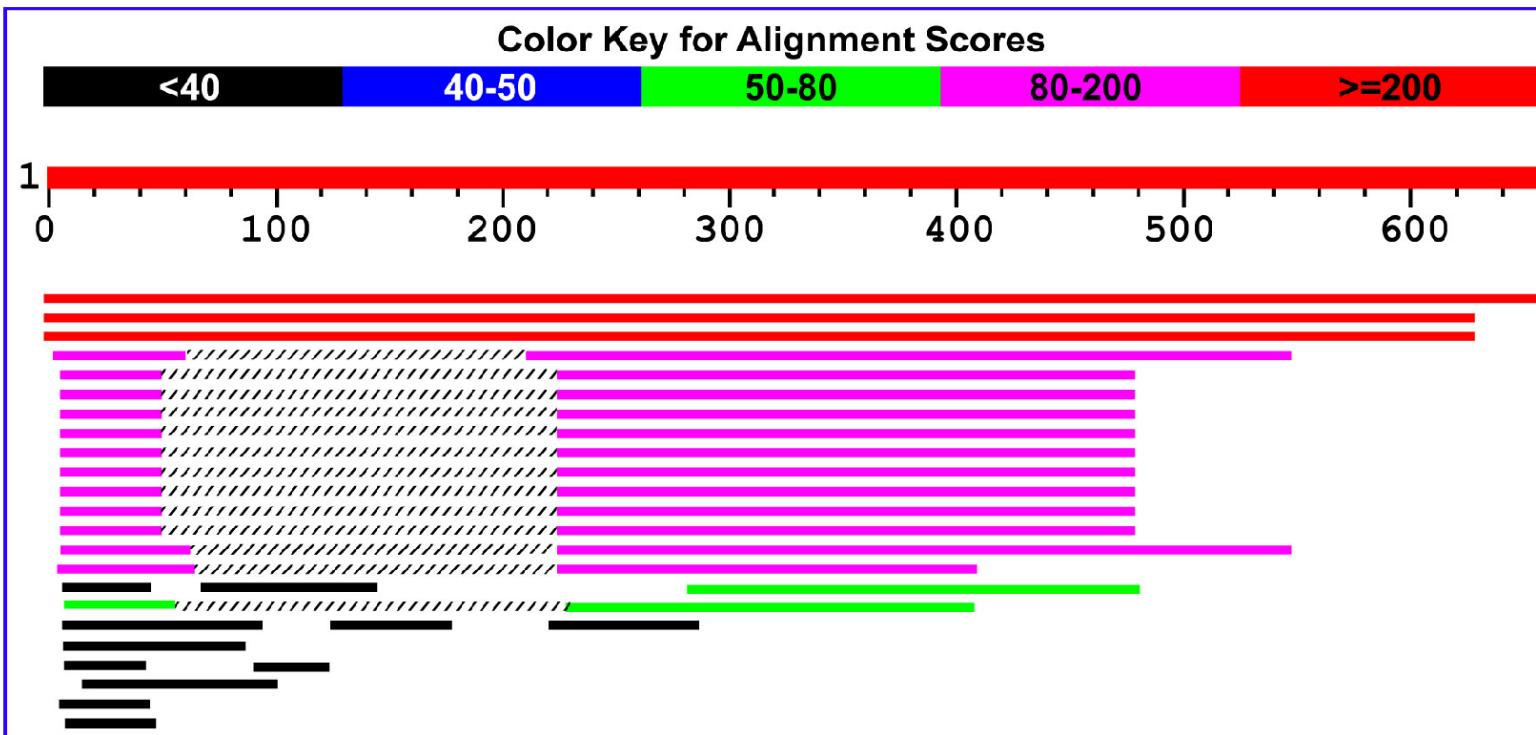
**Database:** Non-redundant SwissProt sequences  
102,387 sequences; 37,391,913 total letters

If you have any problems or questions with the results of this search please refer to the [\*\*BLAST FAQs\*\*](#)

[Taxonomy reports](#)

## Distribution of 41 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments



Sequences producing significant alignments:

(a)	(b)	Score (bits)	E Value
		(c)	(d)
<a href="#">gi 116365 sp P26374 RAE2 HUMAN</a>	Rab proteins geranylgeranyl...	<a href="#">1216</a>	0.0
<a href="#">gi 21431807 sp P24386 RAE1 HUMAN</a>	Rab proteins geranylgerany...	<a href="#">879</a>	0.0
<a href="#">gi 585775 sp P37727 RAE1 RAT</a>	Rab proteins geranylgeranyltra...	<a href="#">846</a>	0.0
<a href="#">gi 13626886 sp Q61598 GDIC MOUSE</a>	RAB GDP dissociation inhib...	<a href="#">127</a>	5e-29
<a href="#">gi 729566 sp P39958 GDI1 YEAST</a>	SECRETORY PATHWAY GDP DISSOC...	<a href="#">127</a>	5e-29
<a href="#">gi 13626813 sp O97556 GDIB CANFA</a>	Rab GDP dissociation inhib...	<a href="#">126</a>	1e-28
<a href="#">gi 13638229 sp P50397 GDIB MOUSE</a>	RAB GDP dissociation inhib...	<a href="#">125</a>	3e-28
<a href="#">gi 1707888 sp P50398 GDIA RAT</a>	RAB GDP dissociation inhibito...	<a href="#">124</a>	7e-28
<a href="#">gi 121108 sp P21856 GDIA BOVIN</a>	Rab GDP dissociation inhibit...	<a href="#">124</a>	7e-28
<a href="#">gi 21903424 sp P50396 GDIA MOUSE</a>	Rab GDP dissociation inhib...	<a href="#">124</a>	7e-28
<a href="#">gi 13626812 sp O97555 GDIA CANFA</a>	RAB GDP dissociation inhib...	<a href="#">124</a>	8e-28
<a href="#">gi 1707886 sp P31150 GDIA HUMAN</a>	Rab GDP dissociation inhibi...	<a href="#">123</a>	9e-28
<a href="#">gi 13638228 sp P50395 GDIB HUMAN</a>	Rab GDP dissociation inhib...	<a href="#">122</a>	2e-27
<a href="#">gi 1707891 sp P50399 GDIB RAT</a>	RAB GDP DISSOCIATION INHIBITO...	<a href="#">121</a>	5e-27
<a href="#">gi 1723467 sp Q10305 YD4C SCHPO</a>	Putative secretory pathway ...	<a href="#">120</a>	8e-27
<a href="#">gi 585776 sp P32864 RAEP YEAST</a>	RAB proteins geranylgeranyl...	<a href="#">97</a>	7e-20
<a href="#">gi 10720243 sp O93831 RAEP CANAL</a>	RAB proteins geranylgerany...	<a href="#">74</a>	9e-13
<a href="#">gi 2498411 sp Q49398 GLF MYCGE</a>	UDP-galactopyranose mutase	<a href="#">35</a>	0.63
<a href="#">gi 11135401 sp Q9XBQ9 STHA AZOVI</a>	Soluble pyridine nucleotid...	<a href="#">34</a>	1.0
<a href="#">gi 11135075 sp O05139 STHA PSEFL</a>	Soluble pyridine nucleotid...	<a href="#">33</a>	1.3
<a href="#">gi 11135195 sp P57112 STHA PSEAE</a>	Soluble pyridine nucleotid...	<a href="#">33</a>	1.8
<a href="#">gi 22257022 sp Q8TZJ8 RLAO PYRFU</a>	Acidic ribosomal protein P...	<a href="#">33</a>	2.1
<a href="#">gi 3915516 sp P94488 YNAJ BACSU</a>	Hypothetical symporter ynaJ	<a href="#">32</a>	3.4
<a href="#">gi 231788 sp P30599 CHS2 USTMA</a>	CHITIN SYNTHASE 2 (CHITIN-UD...	<a href="#">32</a>	3.7
<a href="#">gi 2498412 sp P75499 GLF MYCPN</a>	UDP-galactopyranose mutase	<a href="#">32</a>	4.2
<a href="#">gi 547891 sp P36225 MAP4 BOVIN</a>	Microtubule-associated prote...	<a href="#">32</a>	4.2
<a href="#">gi 586602 sp P37747 GLF ECOLI</a>	UDP-galactopyranose mutase	<a href="#">32</a>	4.6

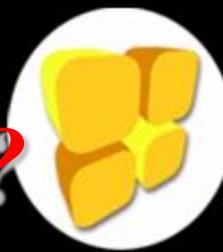
> gi|71013501|dbj|AB212055.1| Influenza A virus(A/Hong Kong/213/03(H5N1)) NP gene for nucleoprotein,  
complete cds, MDCK isolate, embryonated chicken egg  
isolate  
Length=1565

Score = 2855 bits (1440), Expect = 0.0  
Identities = 1515/1540 (98%), Gaps = 0/1540 (0%)  
Strand=Plus/Plus

Query 1	GTAGATAATTCACTCACCGAGTGACATCAACATCATGGCGTCTCAAGGCACCAAAACGATCT	60	→
Subject 13	GTAGATAATTCACTCACCGAGTGACATCAAGCATCATGGCGTCTCAAGGCACCAAAACGATCT	72	→
Query 61	TATGAACAGATGGAAACTGGTGGAGAACGCCAGAATGCTACTGAGATCAGGGCATCTGTT	120	
Subject 73	TATGAACAGATGGAAACTGGTGGAGAACGCCAGAATGCTACTGAGATCAGGGCATCTGTT	132	
★			
Query 121	GGAAGAATGATTAGTGGCATTGGGAGGTTCTACATACAGATGTGCACAGAACTCAAACTC	180	
Subject 133	GGAAGAATGGTTAGTGGCATTGGGAGGTTCTACATACAGATGTGCACAGAACTCAAACTC	192	
★			
Query 181	AGTGACTATGAAGGGAGGGCTTATCCAGAACAGCATAACAAATAGAGAGAAATGGTACTCTCT	240	
Subject 193	AGTGACTATGAAGGGAGGGCTGATCCAGAACAGCATAACAAATAGAGAGAAATGGTACTCTCT	252	
★			
Query 241	GCATTTGATGAAACGAAGGGAACAGATACCTGGAAAGAACACCCCCAGTGCGGGGAAAGGACCCG	300	
Subject 253	GCATTTGATGAAAGAAGGAACAGATACCTGGAAAGAACACCCCCAGTGCGGGGAAAGGACCCG	312	
★			
Query 301	AAGAAAATGGAGGTCCAATTATAGGAGGGAGAGACGGGAATGGGTGAGAGAGCTGATT	360	
Subject 313	AAGAAGACTGGAGGTCCAATTATCGGAGGGAGAGACGGGAATGGGTGAGAGAGCTGATT	372	
★			
Query 361	CTGCACGACAAAGAGGGAGATCAGGAGGGATTGGCGTCAAGCGAACAAATGGAGAGGGACGCA	420	
Subject 373	CTGTACGACAAAGAGGGAGATCAGGAGGGATTGGCGTCAAGCGAACAAATGGAGAGGGACGCA	432	



# *What Next? Do Genomics Analyze?*

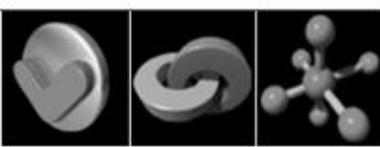


- *Identification of Open Reading Frame, encoding region of protein*
- *Gene Annotation (Prediction of bio-function)*
- *Homology of Pair Wise (or Multiple) DNA Sequences.*
- *Alignment of Sequences.*
- *Discovery of Evolutionary Relationships using Sequence Data.*
- *Predicting Protein Structure and Function.*

# Alignment



- ◆ **Alignment is the basis for finding similarity**
- ◆ **Pairwise alignment = dynamic programming**
- ◆ **Multiple alignment: protein families and functional domains**
- ◆ **Multiple alignment is "impossible" for lots of sequences**
- ◆ **Another heuristic - progressive pairwise alignment**



# Alignment of DNA Sequences



*Arabidopsis* TACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCA

*S\_cerevisiae* TATCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCAA

*Human* TACCTGGTTGATCCTGCCAGTAGCATATGCTTGTCTCAAAGATTAAGCCATGCA

*Mouse* TACCTGGTTGATCCTGCCAGTAGCATATGCTTGTCTCAAAGATTAAGCCATGCAT

**Substitution**



CLUSTAL W (1.83) multiple sequence alignment

<i>Arabidopsis</i>	TACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCA
<i>S_cerevisiae</i>	TATCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCAA
<i>Human</i>	TACCTGGTTGATCCTGCCAGTAG- CATATGCTTGTCTCAAAGATTAAGCCATGCA
<i>Mouse</i>	TACCTGGTTGATCCTGCCAGTAG- CATATGCTTGTCTCAAAGATTAAGCCATGCAT

\*\*\* \* \*\*\*\* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \*

**Deletion or insertion**

# Protein Seq. Multiple Alignment



Alignment Editor: cytochrome-c.bal

File Edit Transfer Display Help

Pos: 31

10 20 30 40 50 60

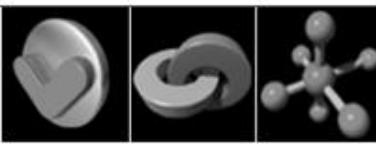
123456789012345678901234567890123456789012345678901234567890

Consensus:

	10	20	30	40	50	60	Lengt																																																	
1: horse	GDVE	GKK	I	FVQ	KCA	Q	CHT	V	E	KGG	HKT	GPNL	HGL	FGR	K	TGQ	A	P	G	F	T	Y	T	D	A	N	K	N	G	I	T	W	K																							
2: honeybee	GIP	A	G	D	P	E	K	G	K	I	F	V	Q	K	C	A	Q	C	H	T	I	E	S	G	G	K	H	K	V	G	P	N	L	Y	G	V	Y	G	R	K	T	G	Q	A	P	G	S	Y	T	D	A	N	K	G	K	
3: hippo	GDVE	GKK	I	FVQ	KCA	Q	CHT	V	E	KGG	HKT	GPNL	HGL	FGR	K	TGQ	S	P	G	F	S	Y	T	D	A	N	K	N	G	I	T	W	G																							
4: guin. pig	GDVE	GKK	I	FVQ	KCA	Q	CHT	V	E	KGG	HKT	GPNL	HGL	FGR	K	TGQ	A	G	F	S	Y	T	D	A	N	K	N	G	I	T	W	G																								
5: guanaco	GDVE	GKK	I	FVQ	KCA	Q	CHT	V	E	KGG	HKT	GPNL	HGL	FGR	K	TGQ	A	V	G	F	S	Y	T	D	A	N	K	N	G	I	T	W	G																							
6: alga	STF	A	P	P	G	B	P	A	K	G	A	K	I	F	K	T	C	A	Z	C	H	T	V	B	--	G	A	H	K	Q	G	P	M	L	N	A	G	F	E	T	S	G	T	A	A	G	F	S	Y	S	A					
7: ginkgo	ATF	S	E	A	P	P	G	D	P	K	A	G	E	K	I	F	K	T	C	A	Z	C	H	T	Z	--	G	A	H	K	Q	G	P	M	L	H	G	F	R	Q	S	G	T	A	G	S	Y	S	T							
8: yeast	PY	A	P	G	D	E	K	K	G	A	S	L	F	K	T	R	C	A	Q	C	H	T	V	E	K	G	A	N	K	V	G	P	N	L	H	G	V	F	R	K	T	G	Q	A	E	F	S	Y	T	E	A	N	K	D	R	G
9: elder	ASF	A	E	A	P	P	G	N	P	K	A	G	E	K	I	F	K	T	C	N	Q	C	H	T	V	D	--	G	A	H	K	Q	G	P	M	L	N	G	F	R	Q	S	G	T	A	G	S	Y	S	T						
10: E. viridis	QDA	E	R	G	K	K	L	F	E	S	R	A	Q	C	H	S	S	Q	K	G	V	N	S	T	G	P	A	L	Y	G	V	Y	G	R	T	S	G	V	P	G	Y	A	S	M	A	N	K	N	A	I	Y	W	E			
11: E. gracili	GDA	E	R	G	K	K	L	F	E	S	R	A	Q	C	H	S	A	Q	K	G	V	N	S	T	G	P	S	L	W	G	V	Y	G	R	T	S	G	S	V	P	G	Y	A	S	M	A	N	K	N	A	I	Y	W	E		
12: emu	GDI	E	K	K	I	F	Y	Q	K	C	S	Q	C	H	T	V	E	K	G	G	K	H	K	T	G	P	N	L	N	G	F	R	K	T	G	Q	A	E	F	S	Y	T	E	A	N	K	N	A	I	Y	W	E				

Consensus Threshold: 65 %

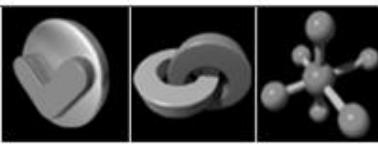
← → Compute Alignment



# Entrez



- <http://www.ncbi.nlm.nih.gov/Entrez/>
- **Tutorials:**
- [\*\*http://www.ncbi.nlm.nih.gov/Class/MLACourse/Genetics/index.html\*\*](http://www.ncbi.nlm.nih.gov/Class/MLACourse/Genetics/index.html)
- [\*\*http://www.ncbi.nlm.nih.gov/Literature/pubmed\\_search.html\*\*](http://www.ncbi.nlm.nih.gov/Literature/pubmed_search.html)
- [\*\*http://www.ncbi.nlm.nih.gov/Database.tut1.html\*\*](http://www.ncbi.nlm.nih.gov/Database.tut1.html)



# EMBL DATA FORMAT



- Embl:  
<http://www.ebi.ac.uk/Databases/>
- <http://www.ebi.ac.uk/cgi-bin/embffetch>
- Use Accession X63129

