NCBI Molecular Biology Resources

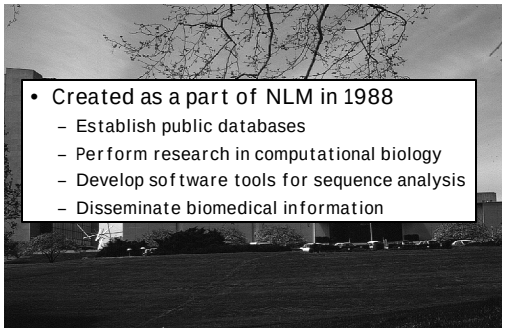
American Society for Microbiology

Peter Cooper
National Center for Biotechnology Information

May 16, 2003

NCBI

The National Center for Biotechnology Information



- Created as a part of NLM in 1988
  - Establish public databases
  - Perform research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information

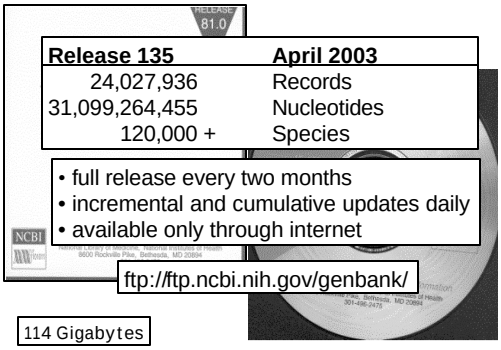
NCBI

Molecular Databases

- Primary
  - Data provided by experimentalist
  - Record maintained by submitter
  - GenBank
- Derivative
  - Value added to primary data
    - compilation
    - curation
    - assembly
  - Record maintained by database staff
  - NCBI Reference Sequences (RefSeq)
  - Protein Data

NCBI

GenBank: NCBI's Primary Sequence Database



Release 135	April 2003
24,027,936	Records
31,099,264,455	Nucleotides
120,000 +	Species

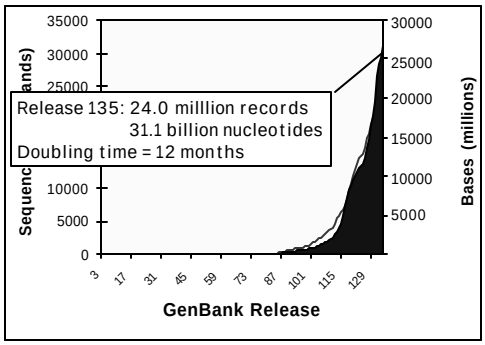
- full release every two months
- incremental and cumulative updates daily
- available only through internet

ftp://ftp.ncbi.nih.gov/genbank/

114 Gigabytes

NCBI

The Growth of GenBank



NCBI

Traditional GenBank Divisions

- Direct Submissions (Sequin and BankIt)
- Accurate
- Well characterized

BCT	Bacterial and Archeal
INV	Invertebrate
MAM	Mammalian (ex. ROD and PRI)
PHG	Phage
PLN	Plant and Fungal
PRI	Primate
ROD	Rodent
SYN	Synthetic (vectors, synth. genes)
VRL	Viral
VRT	Other Vertebrate

NCBI

Traditional GenBank Records

```
LOCUS       AF105152                 1584 bp    mRNA    linear   VRT 11-FEB-1999
DEFINITION  Danio rerio rhodopsin mRNA, complete cds.
ACCESSION   AF105152
VERSION     AF105152.1  GI:4262520
KEYWORDS    .
SOURCE      Danio rerio (zebrafish)
  ORGANISM  Danio rerio
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
REFERENCE
  AUTHORS
  TITLE
  JOURNAL
  MEDLINE
   PUBMED
REFERENCE
  AUTHORS
  TITLE
  JOURNAL
```

- Archival
- Redundant

```
LOCUS       AF109368                 1581 bp    mRNA    linear   VRT 09-AUG-2001
DEFINITION  Danio rerio rhodopsin mRNA, complete cds.
ACCESSION   AF109368
VERSION     AF109368.1  GI:4581736
KEYWORDS    .
SOURCE      Danio rerio (zebrafish)
  ORGANISM  Danio rerio
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Ostariophysi;
            Cypriniformes; Cyprinidae; Danio.
REFERENCE   1  (bases 1 to 1581)
  AUTHORS   Vihtelic,T.S., Doro,C.J. and Hyde,D.R.
  TITLE     Cloning and characterization of six zebrafish photoreceptor opsin
            cDNAs and immunolocalization of their corresponding proteins
  JOURNAL   Vis. Neurosci. 16 (3), 571-585 (1999)
  MEDLINE   99277479
   PUBMED   10349976
REFERENCE   2  (bases 1 to 1581)
  AUTHORS   Vihtelic,T.S., Doro,C.J. and Hyde,D.R.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-NOV-1998) Biological Sciences, University of Notre
            Dame, 264 Galvin Life Science Bldg., Notre Dame, IN 46556, USA
```

NCBI

Bulk GenBank Divisions

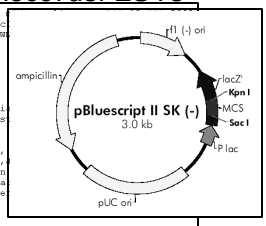
- Batch Submission and htg (email and ftp)
- Inaccurate
- Poorly Characterized

EST	Expressed Sequence Tag
STS	Sequence Tagged Site
GSS	Genome Survey Sequence
HTG	High Throughput Genomic

NCBI

Bulk GenBank Records: ESTs

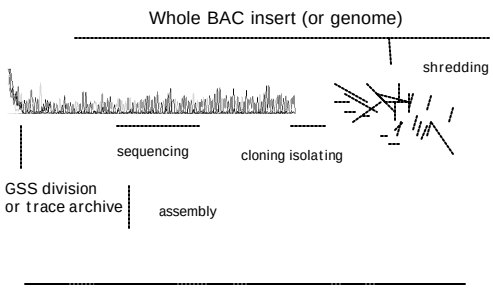
```
LOCUS       BG738754                 530 bp
DEFINITION  fp62d08.yl Zebrafish adult retina c
            IMAGE:4786599 5' similar to TR:Q9PW
            sequence.
ACCESSION   BG738754
VERSION     BG738754.1  GI:14088443
KEYWORDS    EST.
SOURCE      Danio rerio (zebrafish)
  ORGANISM  Danio rerio
            Eukaryota; Metazoa; Chordata; Crani
            Actinopterygii; Neopterygii; Teleos
            ; Cyprinidae; Danio.
REFERENCE   1  (bases 1 to 530)
  AUTHORS   Clark,M., Johnson,S.L., Lehrach,H.,
            ,S., Hillier,L., Kucaba,T., Martin,
            ,K., Steptoe,M., Theising,B., Allen
            Swaller,T., Gibbons,M., Pape,D., Ha
            Kohn,S., Shin,T., Jackson,Y., Carde
            and Wilson,R.
  TITLE     WashU Zebrafish EST Project 1998
  JOURNAL   Unpublished (1998)
COMMENT     Contact: Stephen L. Johnson
            Washington University School of
            4444 Forest Park Parkway, Box 85
            Tel: 314 286 1800
            Fax: 314 286 1810
            Email: zbrafish@watson.wustl.edu
            Library constructed by: Chandra
            Sequencing by: Washington University Genome Sequencing Center Clone
            distribution: RessourcenZentrumPrimarDatenbank, Berlin, Germany
            (web address: www.rzpd.de)
            Seq primer: T3 ET from Amersham
            High quality sequence stop: 423.
```



- first pass, single read cDNA
- largest GenBank division
- gbdiv_est[Properties]

NCBI

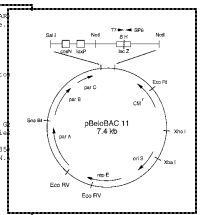
Genome Sequencing - HTG, GSS,(WGS)



NCBI

Bulk GenBank Records: GSS

```
LOCUS       BX247846                 551 bp    DNA     linear   GSS 14-MAR
DEFINITION  Danio rerio genomic clone DKEY-235N9, genomic survey sequence.
ACCESSION   BX247846
VERSION     BX247846.1  GI:28170522
KEYWORDS    GSS.
SOURCE      Danio rerio (zebrafish)
  ORGANISM  Danio rerio
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleosto
            Actinopterygii; Neopterygii; Teleostei; Ostariophysi;
            Cypriniformes; Cyprinidae; Danio.
REFERENCE   1  (bases 1 to 551)
  AUTHORS   Humphray,S.J., Huckle,E. and Durham,J.L.
  TITLE     Direct Submission
  JOURNAL   Submitted (13-MAR-2003) The Sanger Institute, Wellcome Trust G
            Campus, Hinxton, Cambridgeshire, CB10 1SA, UK. E-mail enquiries
            humquery@sanger.ac.uk Unpublished
COMMENT     This sequence was generated from the SP6 end of BAC 235N9. 235
            part of the Daniokey BAC Library created by R. Plasterk and N.
            Keygene. Further details:
            http://www.sanger.ac.uk/Projects/D_rerio/.
FEATURES             Location/Qualifiers
     source          1..551
                     /organism="Danio rerio"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:7955"
                     /clone="DKEY-235N9"
                     /tissue_type="Testis"
                     /note="vector pIndigoBAC-536"
BASE COUNT      205 a    71 c   110 g   165 t
ORIGIN
        1 ctttatttaa attttctact gatgaaaaca aaacacgcaa t
       61 aaactaagtc aattaatagg tgaagagggt ttataaatga c
      121 acttgcagac tggaaatgtg gagaaactgg aaataccacca g
      181 gtctggtgtt ctgttgcagg aaattaagac aaaacaggag a
      241 ggatcattaa ttaacacatc agtgagatca acaatcaagt a
      301 atagdttgga ataaggagaa agggggtggt ggtgggaata g
      361 aactgagtaa tttctgttgc atcaacdttt taatttccca g
      421 aaactgcatc atgactcaga ctttgatatg aaggataaga ggacattttt aatatttgtg
      481 tgatgacaga aaattcagaa attcattaaa acgaacataa gaaatatacc acgacatgtc
      541 agcgtataaa t
//
```

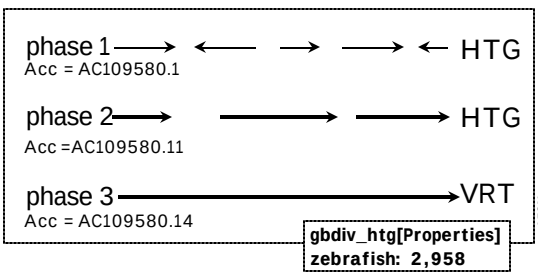


- first pass, single read gDNA
- surveys of BAC libraries
- gbdiv_gss[Properties]
- zebrafish GSS: 159, 024

NCBI

HTG Division: High Throughput Genome

Zebrafish PAC Clone



gbdiv_htg[Properties]
zebrafish: 2,958

40,000 to > 350,000 bp

NCBI

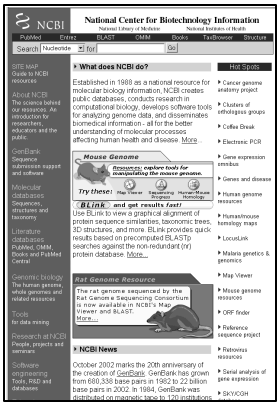
RefSeq: NCBI's Derivative Sequence Database

- **Curated transcripts and proteins [NM_, NP_]**
  - reviewed
  - human, mouse, rat, cow, fruit fly, zebrafish, arabidopsis *C.elegans*
- **Model transcripts and proteins [XM_, XP_]**
- **Assembled Genomic Regions [NT_, NW_]**
  - draft human genome
  - mouse genome
- **Chromosome records [NC_]**
  - microbial
  - organelle

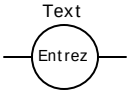
`srcdb_refseq[Properties]`

NCBI

Web Access



Text

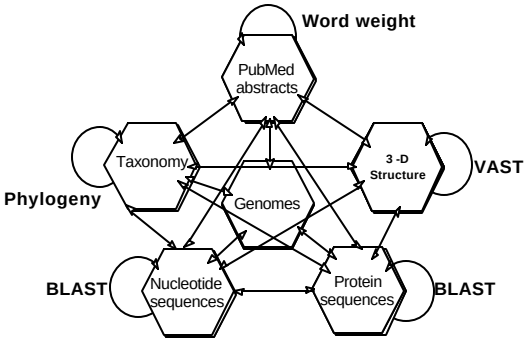


Sequence

Structure

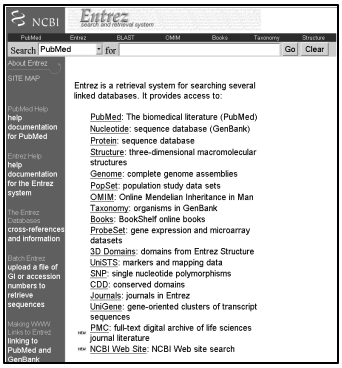
NCBI

Entrez: Database Integration



NCBI

WWW Entrez



NCBI

Sequence Similarity Searching

Basic Local Alignment Search Tool

NCBI

Why do we need similarity searching?

- ✦ Identification and annotation
  - Incomplete or no annotations (GenBank)
  - Incorrectly annotated sequences
- ✦ Evolutionary relationships
  homologous molecules **may** have similar functions

NCBI

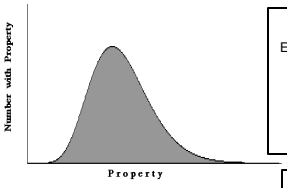
Basic **L**ocal **A**lignment **S**earch **T**ool

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- All combinations (DNA/Protein) query and database.
  - DNA vs DNA
  - DNA translation vs Protein
  - Protein vs Protein
  - Protein vs DNA translation
  - DNA translation vs DNA translation
- www, standalone, and network clients

NCBI

Local Alignment Statistics

High scores of local alignments between two random sequences follow Extreme Value Distribution



For ungapped alignments:

Expected number with score S or greater

$E = Kmne^{-\lambda S}$

or

$E = mn2^{-S'}$

K = scale for search space
λ = scale for scoring system
S'= bitscore = (λS - lnK)/ln2

http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

NCBI

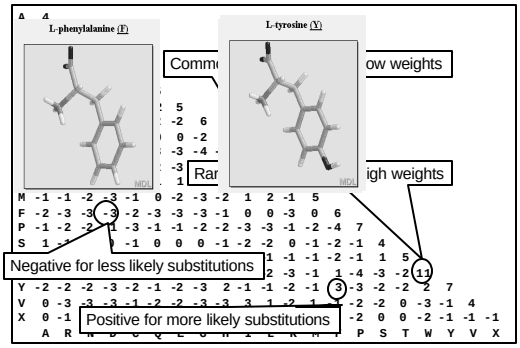
Scoring Systems

	A	G	C	T
A	+1	–3	–3	-3
G	–3	+1	–3	-3
C	–3	–3	+1	-3
T	–3	–3	–3	+1

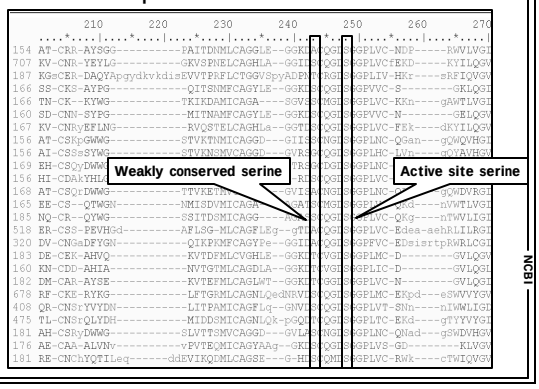
- **Position Independent Matrices**
  - Nucleic Acids – identity matrix
  - Proteins
    - PAM Matrices (Percent Accepted Mutation)
      - Implicit model of evolution
      - Higher PAM number all calculated from PAM1
      - PAM250 widely used
    - BLOSUM Matrices (BLOck SUbstitution Matrices)
      - Empirically determined from alignment of conserved blocks
      - Each includes information up to a certain level of identity
      - BLOSUM62 widely used
- **Position Specific Score Matrices (PSSMs)**
  - PSI and RPS BLAST

NCBI

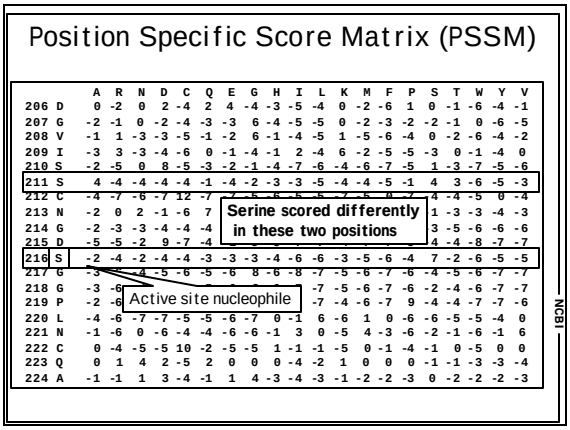
BLOSUM62



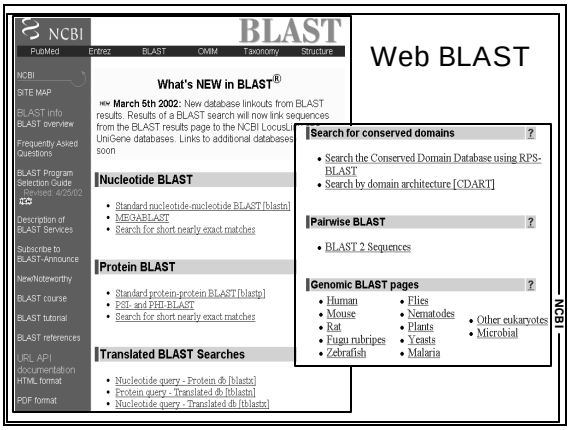
Position Specific Substitution Rates



Position Specific Score Matrix (PSSM)



Web BLAST



NCBI

BLAST Databases: Non-redundant protein

nr (non-redundant protein sequences)
- GenBank CDS translations
- NP_ RefSeqs
- Outside Protein
  - PIR, **Swiss-Prot**, PRF
- **PDB** (sequences from structures)

NCBI

BLAST Databases: Nucleic Acid

- **nr (nt)**
  - Traditional GenBank Divisions
  - NM_ and XM_RefSeqs
- **dbest**
  - EST Division
- **htgs**
  - HTG division
- **gss**
  - GSS division
- **chromosome**
  - NC_ RefSeqs

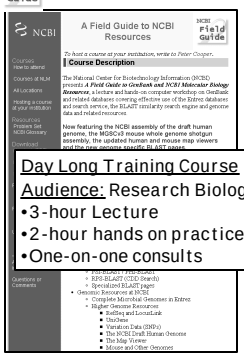
NCBI

Service Addresses

- ***General Help*** `info@ncbi.nlm.nih.gov`
- ***BLAST*** `blast-help@ncbi.nlm.nih.gov`

NCBI

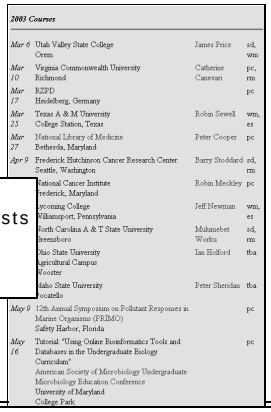
NCBI Training



Day Long Training Course

Audience: Research Biologists

- 3-hour Lecture
- 2-hour hands on practice
- One-on-one consults



2003 Courses

Mar 6	Utah Valley State College Orem	James Price	sd, wm
Mar 10	Virginia Commonwealth University Richmond	Catherine Canevari	pc, rm
Mar 17	RZPD Heidelberg, Germany		pc
Mar 25	Texas A & M University College Station, Texas	Robin Sewell	wm, es
Mar 27	National Library of Medicine Bethesda, Maryland	Peter Cooper	pc
Apr 9	Frederick Hutchinson Cancer Research Center Seattle, Washington	Barry Stoddard	sd, rm
	National Cancer Institute Frederick, Maryland	Robin Meckley	pc
	Lycoming College Williamsport, Pennsylvania	Jeff Newman	wm, es
	North Carolina A & T State University Greensboro	Mahmebet Worku	sd, rm
	Ohio State University Agricultural Campus Wooster	Ian Holford	tba
	Idaho State University Pocatello	Peter Sheridan	tba
May 9	12th Annual Symposium on Pollutant Responses in Marine Organisms (PRIMO) Safety Harbor, Florida		pc
May 16	Tutorial "Using Online Bioinformatics Tools and Databases in the Undergraduate Biology Curriculum" American Society of Microbiology Undergraduate Microbiology Education Conference University of Maryland College Park		pc

NCBI