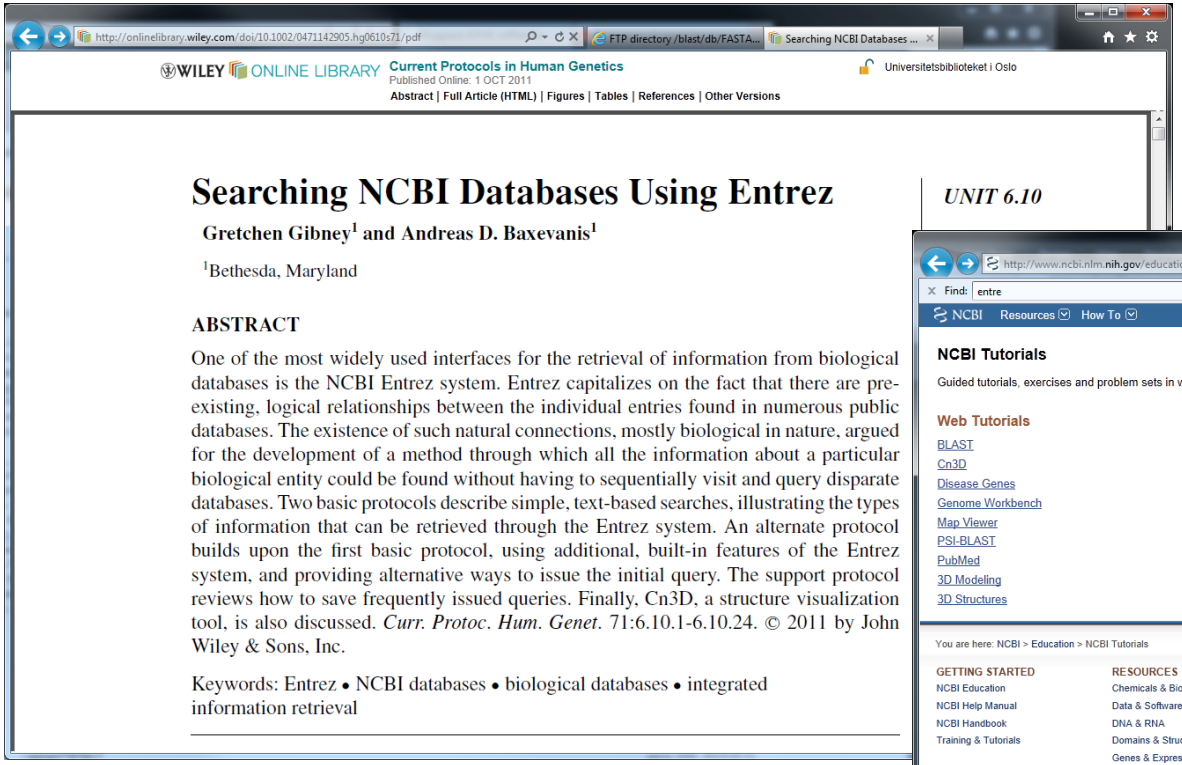
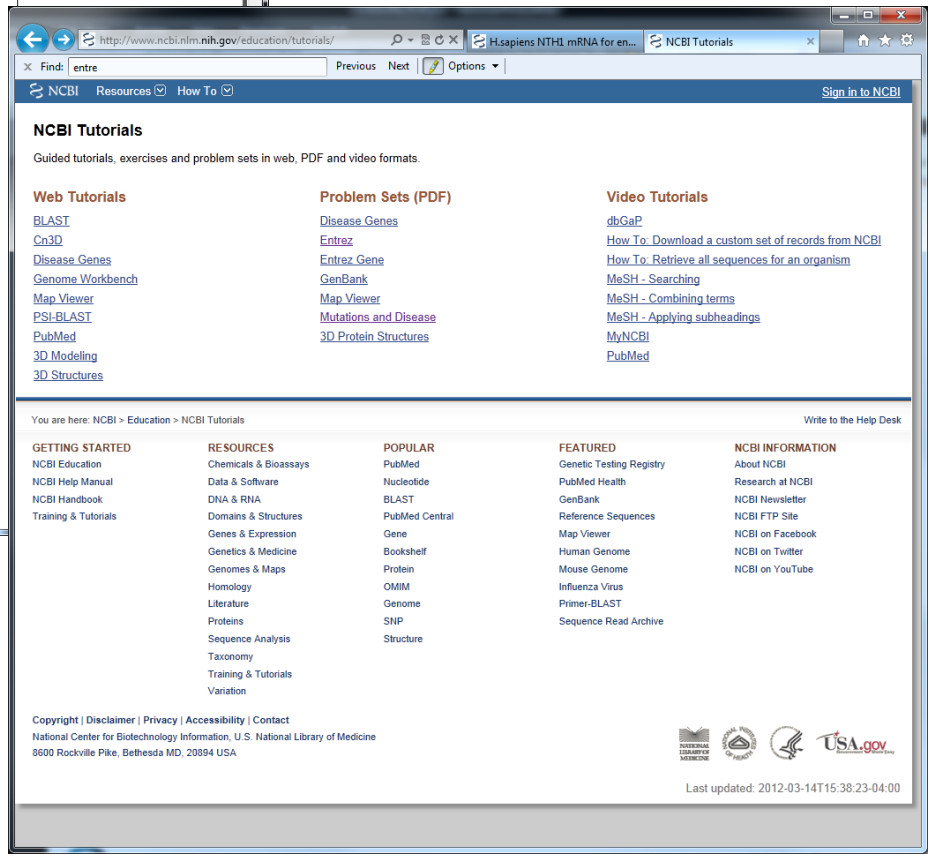
Jon K. Lærdahl,
Structural Bioinformatics

Learn to use Entrez!





Curr. Protoc. Hum. Genet. **71**, 6.10.1 (2011)

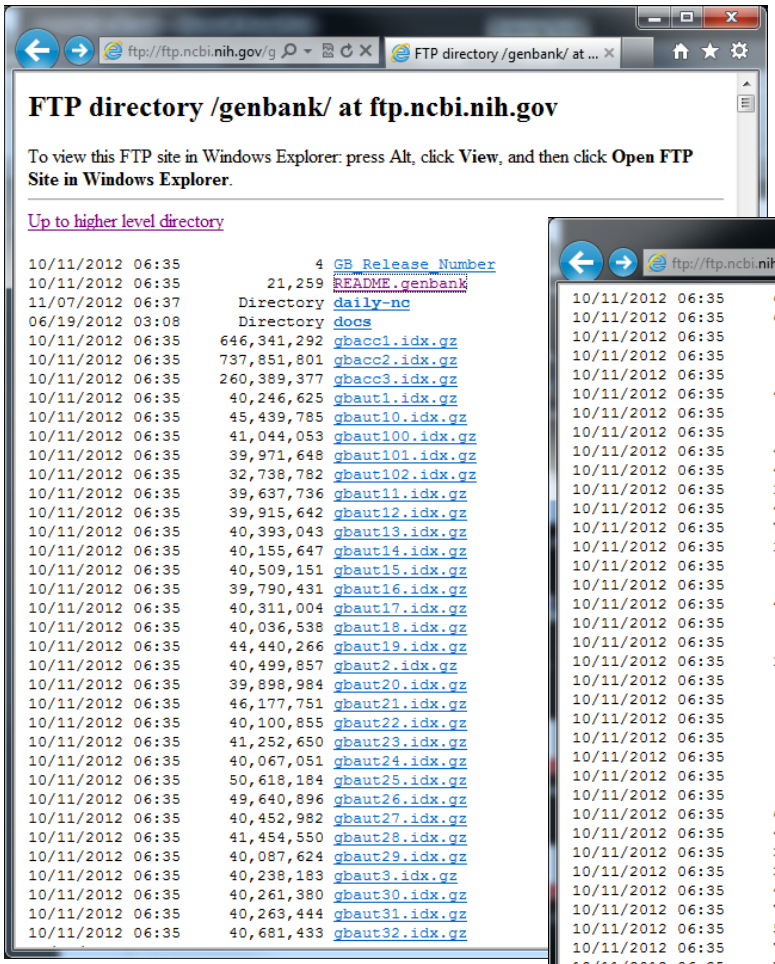
NCBI Tutorials

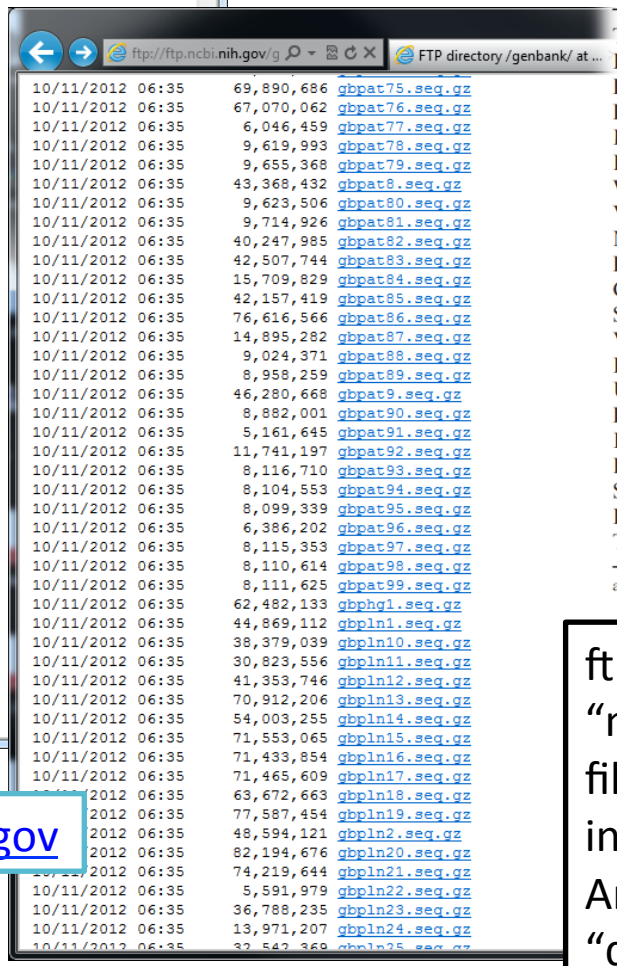
Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

NCBI ftp site





Nucleic Acids Research, 2012, Vol. 40, Database issue **D49**

Table 1. Growth of GenBank divisions (nucleotide base pairs)

Division	Description	Release 185 (8/2011)	Annual increase (%)[a]
TSA	Transcriptome shotgun data	1 874 047 448	370.1
ENV	Environmental samples	2 553 693 157	48.2
PHG	Phages	62 579 756	44.0
PAT	Patented sequences	11 154 487 762	30.9
BCT	Bacteria	6 975 597 755	30.8
INV	Invertebrates	2 535 336 197	24.5
WGS	Whole-genome shotgun data	208 315 831 132	23.1
VRL	Viruses	1 180 083 600	21.6
MAM	Other mammals	807 098 397	18.8
PLN	Plants	4 741 991 057	17.4
GSS	Genome survey sequences	20 770 772 329	12.6
SYN	Synthetic	156 218 063	9.6
VRT	Other vertebrates	2 705 250 711	6.8
EST	Expressed sequence tags	39 018 185 344	6.0
UNA	Unannotated	125 912	4.7
PRI	Primates	6 116 546 725	2.9
ROD	Rodents	4 396 957 541	2.3
HTC	High-throughput cDNA	662 320 919	0.4
STS	Sequence tagged sites	635 872 683	0.3
HTG	High-throughput genomic	24 324 068 445	0.2
TOTAL	All GenBank sequences	338 987 064 933	18.2

[a]Measured relative to Release 179 (8/2010).

ftp://ftp.ncbi.nih.gov

ftp = file transfer protocol
"network protocol for transfer of files from one host to another on the internet"
An ftp site is very basically a "directory with files on the internet"

Oslo universitetssykehus

Jon K. Lærdahl,
Structural Bioinformatics

```
LOCUS       HSNTH1H1                1030 bp    RNA     linear   PRI 10-JAN-1997
DEFINITION  H.sapiens NTH1 mRNA for endonuclease III homologue.
ACCESSION   Y09687
VERSION     Y09687  GI:1772973
KEYWORDS    endonuclease III; homologue; NTH1.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1030)
  AUTHORS   Rognes,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-NOV-1996) T. Rognes, University of Oslo, Institute of
            Medical Microbiology, The National Hospital, N-0027 Oslo, NORWAY
REFERENCE   2  (bases 1 to 1030)
  AUTHORS   Luna,L., Bjoras,M., Rognes,T., Hoff,E. and Seeberg,E.
  JOURNAL   Unpublished
FEATURES             Location/Qualifiers
     source          1..1030
                     /organism="Homo sapiens"
                     /mol_type="unassigned RNA"
                     /db_xref="taxon:9606"
                     /dev_stage="adult"
     gene            1..912
                     /gene="NTH1"
     CDS             <1..912
                     /gene="NTH1"
                     /codon_start=1
                     /product="endonuclease III homologue 1"
                     /protein_id="1772974"
                     /db_xref="GI:1772974"
                     /translation="TSALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSH
                     SPVKRPRKAQRLRVAYEGSDSEKGEAEPLKVPVWEPQDWQQQLVNIRAMRNKKDAPVD
                     HLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVDSILQTDD
                     ATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMA
                     VAWGTVSGIAVDTHVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGF
                     GQQTCLPVHPRCHACLNQALCPAAQGL"
ORIGIN
        1 acgagcgcct tgagcgcgag gatgctgacc cggagccgga gcctgggacc cggggctggg
       61 ccgcgggggt gtagggagga gcccgggcct ctccggagaa gagaggctgc agcagaagcg
      121 aggaaaagcc acagccccgt gaagcgtccg cggaaagcac agagactgcg tgtggcctat
      181 gagggctcgg acagtgagaa aggtgaggct gagcccctca aggtgccagt ctgggagccc
      241 caggactggc agcaacagct ggtcaacatc cgtgccatga ggaacaaaaa ggatgcacct
      301 gtggaccatc tggggactga gcactgctat gactccagtg ccccccaaa ggtacgcagg
      361 taccaggtgc tgctgtcact gatgctctcc agccaaacca aagaccaggt gacggcgggc
      421 gccatgcagc gactgcgggc gcggggcctg acggtggaca gcatcctgca gacagatgat
      481 gccacgctgg gcaagctcat ctaccccgtc ggtttctgga ggagcaaggt gaaatacatc
      541 aagcagacca gcgccatcct gcagcagcac tacggtgggg acatcccagc ctctgtggcc
      601 gagctggtgg cgctgccggg tgttgggccc aagatggcac acctggctat ggctgtggcc
      661 tggggcactg tgtcaggcat tgcagtggac acgcatgtgc acagaatcgc caacaggctg
      721 aggtggacca agaaggcaac caagtcccca gaggagaccc gcgccgccct ggaggagtgg
      781 ctgcctaggg agctgtggca cgagatcaat ggactcttgg tgggcttcgg ccagcagacc
      841 tgtctgcctg tgcaccctcg ctgccacgcc tgcctcaacc aagccctctg cccggccgcc
      901 cagggtctct gatggccgca tggctctggc cgaggtgccg ctgtggccac cgtctgtgaa
      961 gtggctttac gcttcaggaa gccacgcctg ttgaataaag ctttggtgtg tttgcaaaaa
     1021 aaaaaaaaaa
//
```

Accession number

▸ **DDBJ/EMBL/GenBank Accession Prefix Format**

The format for GenBank Accession numbers are:

Nucleotide:	1 letter + 5 numerals OR 2 letters + 6 numerals
Protein:	3 letters + 5 numerals
WGS:	4 letters + 2 numerals for WGS assembly version + 6-8 numerals
MGA:	5 letters + 7 numerals

The International Nucleotide Sequence Database Collaboration DDBJ/EMBL/GenBank all receive sequence submissions, assign accessions, and exchange data so that all three groups represent the total collection. The accession assignment process is managed by prior agreement within the collaboration on which group will 'own' which accession prefix. This list of accession number prefixes should be used as a guide. There are cases where these assignments are not adhered to. For instance, there are ESTs and GSSs from GenBank that have the prefix for Direct submissions.

Allocation of Accession Prefixes

Nucleotide Accession Prefixes

Prefix	Database	Type
BA,DF,DG	DDBJ	CON division
AN	EMBL	CON division
CH,CM,DS,EM, EN,EP,EQ,FA, GG,GL,JH,KB	NCBI	CON division
C,AT,AU,AV,BB, BJ,BP,BW,BY,CI, CJ,DA,DB,DC, DK,FS,FY,HX, HY	DDBJ	EST

Jon K. Lærdahl,
Structural Bioinformatics

```
LOCUS       HSNTH1H1              1030 bp    RNA     linear   PRI 10-JAN-1997
DEFINITION  H.sapiens NTH1 mRNA for endonuclease III homologue.
ACCESSION   Y09687
VERSION     Y09687  GI:1772973
KEYWORDS    endonuclease III; homologue; NTH1.
```

Accession number

The corresponding protein identifier is CAA70865

DDBJ/EMBL/GenBank Accession Prefix Format

The format for GenBank Accession numbers are:

Nucleotide:	1 letter + 5 numerals OR 2 letters + 6 numerals
Protein:	3 letters + 5 numerals
WGS:	4 letters + 2 numerals for WGS assembly version + 6-8 numerals
MGA:	5 letters + 7 numerals

The International Nucleotide Sequence Database Collaboration DDBJ/EMBL/GenBank all receive sequence submissions, assign accessions, and exchange data so that all three groups represent the total collection. The accession assignment process is managed by prior agreement within the collaboration on which group will 'own' which accession prefix. This list of accession number prefixes should be used as a guide. There are cases where these assignments are not adhered to. For instance, there are ESTs and GSSs from GenBank that have the prefix for Direct submissions.

Allocation of Accession Prefixes

Nucleotide Accession Prefixes

Prefix	Database	Type
BA,DF,DG	DDBJ	CON division
AN	EMBL	CON division
CH,CM,DS,EM, EN,EP,EQ,FA, GG,GL,JH,KB	NCBI	CON division
C,AT,AU,AV,BB, BJ,BP,BW,BY,CI, CJ,DA,DB,DC, DK,FS,FY,HX, HY	DDBJ	EST
F	EMBL	EST
H,N,T,R,W,AA,AI, AW,BE,BF,BG, BI,BM,BQ,BU, CA,CB,CD,CF, CK,CN,CO,CV, CX,DN,DR,DT, DV,DW,DY,EB, EC,EE,EG,EH, EL,ES,EV,EW, EX,EY,FC,FD, FE,FF,FG,FK, FL,GD,GE,GH, GO,GR,GT,GW, HO,HS,JG,JK, JZ	GenBank	EST
D,AB	DDBJ	Direct submissions
V,X,Y,Z,AJ,AM, [illegible],FN,HE,HF, [illegible]G,FO	EMBL	Direct submissions
[illegible]J,AF,AY,DQ,EF, EU,FJ,GQ,GU, HM,HQ,JF,JN, JQ,JX	GenBank	Direct submissions
AP	DDBJ	Genome project data
BS	DDBJ	Chimpanzee genome data
AL,BX,CR,CT, CU,FP,FQ,FR	EMBL	Genome project data
AE,CP,CY	GenBank	Genome project data
AG,DE,DH,FT, GA	DDBJ	GSS
B,AQ,AZ,BH,BZ, CC,CE,CG,CL, CW,CZ,DU,DX, ED,EI,EJ,EK, ER,ET,FH,FI, GS,HN,HR,JJ, JM,JS,JY	GenBank	GSS
AK	DDBJ	cDNA projects
AC,DP	GenBank	HTGS
E,BD,DD,DI,DJ, DL,DM,FU,FV, FW,FZ,GB,HV, HW	DDBJ	Patents
A,AX,CQ,CS,FB, GM,GN,HA,HB, HC,HD,HH,HI, JA,JB,JC,JD,JE	EMBL	Patents (nucleotide only)
I,AR,DZ,EA,GC, GP,GV,GX,GY, GZ,HJ,HK,HL	GenBank	Patents (nucleotide)
G,BV,GF	GenBank	STS
BR	DDBJ	TPA
BN	EMBL	TPA
BK	GenBank	TPA
HT,HU	DDBJ	TPA CON division
BL,GJ,GK	GenBank	TPA CON division
EZ,HP,JI,JL, JO,JP,JR,JT, JU,JV,JW,KA	GenBank	TSA
FX	DDBJ	TSA
S	GenBank	From journal scanning
AD	GenBank	From GSDB
AH	GenBank	Segmented set header
AS	GenBank	Other - not currently being used
BC	GenBank	MGC project
BT	GenBank	FLI-cDNA projects
J,K,L,M	GenBank	from GSDB direct submissions
N	GenBank and DDBJ	N0-N2 were used intially by both groups but have been removed from circulation, N2-N9 are ESTs
AAAA-AZZZ	GenBank	WGS
BAAA-BZZZ	DDBJ	WGS
CAAA-CZZZ	EMBL	WGS
EAAA-EZZZ	DDBJ	WGS TPA
DAAA-DZZZ	GenBank	WGS TPA
GAAA-GZZZ	GenBank	TSA
AAAAA-AZZZZ	DDBJ	MGA

Protein Accession Prefixes

Prefix	Database	Type
BAA-BZZ	DDBJ	Protein ID
CAA-CZZ	EMBL	Protein ID
AAA-AZZ	GenBank	Protein ID
AAE	GenBank	Protein ID for Patents (note that there are also some patent proteins with AAA and AAC

http://www.ncbi.nlm.nih.gov/Sequin/acc.html

GenBank accession numbers

- Each GenBank record has a ***sequence*** with its ***annotations***
- Each record has a unique identifier (accession number) that is shared between GenBank, DDBJ, and ENA
- The accession number does not change even if the sequence changes
  - Changes are tracked by an integer extension of the accession number
  - Initial version is ".1"
  - Each version of a sequence get a new unique NCBI identifier called a ***GI*** number.

```
LOCUS       AAH00391                 305 aa            linear   PRI 06-NOV-2003
DEFINITION  NTHL1 protein, partial [Homo sapiens].
ACCESSION   AAH00391
VERSION     AAH00391.2  GI:38197140
```

Link to AHH00391 record

How to access GenBank data?

- Use Entrez?
- Download from the ftp site?
- Use sequence searching (BLAST)?

- As for most databases, there are several ways to access the data – use the method that suits your purpose

Jon K. Lærdahl,
Structural Bioinformatics

FASTA format

```
>gi|5102658|emb|CAB45242.1| AP endonuclease XTH2, putative [Homo sapiens]
MLRVVSWNINGIRRPLQGVANQEPSNCAAVAVGRILDELDADIVCLQETKVTRDALTEPLAIVEGYNSYF
SFSRNRSGYSGVATFCKDNATPVAAEEGLSGLFATQNGDVGCYGNMDEFTQEELRALDSEGRALLTQHKI
RTWEGKEKTLTLINVYCPHADPGRPERLVFKMRFYRLLQIRAEALLAAGSHVIILGDLNTAHRPIDHWDA
VNLECFEEDPGRKWMDSLLSNLGCQSASHVGPFIDSYRCFQPKQEGAFTCWSAVTGARHLNYGSRLDYVL
GDRTLVIDTFQASFLLPEVMGSDHCPVGAVLSVSSVPAKQCPPLCTRFLPEFAGTQLKILRFLVPLEQSP
VLEQSTLQHNNQTRVQTCQNKAQVRSTRPQPSQVGSSRGQKNLKSYFQPSPSCPQASPDIELPSLPLMSA
LMTPKTPEEKAVAKVVKGQAKTSEAKDEKELRTSFWKSVLAGPLRTPLCGGHREPCVMRTVKKPGPNLGR
RFYMCARPRGPPTDPSSRCNFFLWSRPS
```

- Single letter codes
- nt or aa
- Single line with description starts with «>»

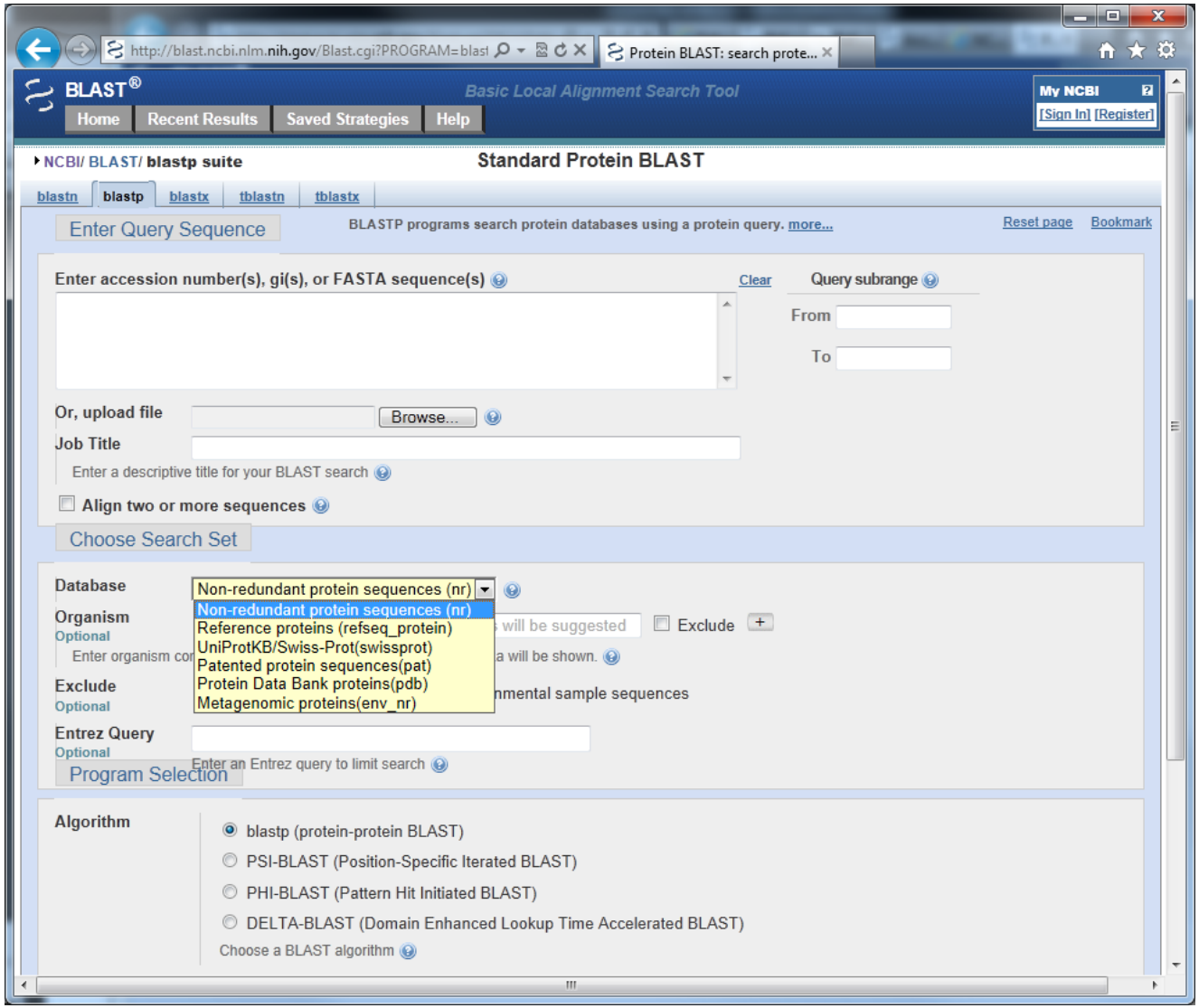
```
>gi|1772974|emb|CAA70865.1| endonuclease III homologue 1 [Homo sapiens]
TSALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEA
EPLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAG
AMQRLRARGLTVDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGP
KMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGFGQQT
CLPVHPRCHACLNQALCPAAQGL
>gi|84618885|emb|CAJ31885.1| methylpurine-DNA glycosylase [Bacillus cereus]
MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFLFIGIQTPERRQLLKDVIQIHTLPDPKDFRIIVRELWD
LPEREFQAAALDMMQKYKKYINETHIPFLEELIVTKSWWDTVDSIVPTFLGNIFLQHPELISAYIPKWIA
SDNIWLQRAAILFQLKYKQKMDEELLFWVIGQLHSSKEFFIQKAIGWVLREYAKTKPDVVWEYVQNNELA
PLSRREAIKHIKENYGINNEKIGETLS
>gi|559882|emb|CAA86405.1| threonine synthase [Schizosaccharomyces pombe]
MSSQVSYLSTRGGSSNFSFEEAVLKGLANDGGLFIPSEIPQLPSGWIEAWKDKSFPEIAFEVMSLYIPRS
EISADELKKLVDRSYSTFRHPETTPLKSLKNGLNVLELFHGPTFAFKDVALQFLGNLFEFFLTRKNGNKP
EDERDHLTVVGATSGDTGSAAIYGLRGKKDVSVFILFPNGRVSPIQEAQMTTVTDPNVHCITVNGVFDDC
QDLVKQIFGDVEFNKKHHIGAVNSINWARILSQITYYLYSYLSVYKQGKADDVRFIVPTGNFGDILAGYY
AKRMGLPTKQLVIATNENDILNRFFKTGRYEKADSTQVSPSGPISAKETYSPAMDILVSSNFERYLWYLA
LATEAPNHTPAEASEILSRWMNEFKRDGTVTVRPEVLEAARRDFVSERVSNDETIDAIKKIYESDHYIID
PHTAVGVETGLRCLEKTKDQDITYICLSTAHPAKFDKAVNLALSSYSDYNFNTQVLPIEFDGLLDEERTC
IFSGKPNIDILKQIIEVTLSREKA
```

Jon K. Lærdahl,
Structural Bioinformatics

RefSeq

- The INSDC databases (GenBank, ENA, and DDBJ) is an archival repository of all sequences
  - one enormous mess?
- NCBI builds the RefSeq (Reference Sequence) database from the INSDC data
  - an attempt to "tidy up the data"
  - collection of genomic, transcript and protein sequence records
  - significant reduction in redundancy (less records describing exactly the same "thing")
  - only one record (ideally) for each natural biological molecule (*i.e.* Protein, RNA, or DNA sequence)
  - only "model organisms" (more than 16,000, while GenBank has >250,000)
  - maintained by combined approach of automated analyses (mainly) and manual curation in pipelines

http://nar.oxfordjournals.org/content/40/D1/D130.abstract



The “non-redundant” database contains a lot of redundancy

RefSeq is the attempt to make the nr database less redundant

Extremely confusing naming....

RefSeq can also be searched with Entrez or BLAST, and data can be downloaded from ftp site

Jon K. Lærdahl,
Structural Bioinformatics

RefSeq record features

```
LOCUS       NP_899834                229 aa            linear   BCT 20-JAN-2012
DEFINITION  DNA alkylation repair enzyme [Chromobacterium violaceum ATCC
            12472].
ACCESSION   NP_899834
VERSION     NP_899834.1  GI:34495619
DBLINK      Project: 58001
            BioProject: PRJNA58001
DBSOURCE    REFSEQ: accession NC_005085.1
KEYWORDS    .
SOURCE      Chromobacterium violaceum ATCC 12472
  ORGANISM  Chromobacterium violaceum ATCC 12472
            Bacteria; Proteobacteria; Betaproteobacteria; Neisseriales;
            Neisseriaceae; Chromobacterium.
REFERENCE   1  (residues 1 to 229)
  AUTHORS   Vasconcelos,A.T.R., de Almeida,D.F., Almeida,F.C., de
            ..................

COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AAQ57843.
            Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..229
                     /organism="Chromobacterium violaceum ATCC 12472"
                     /strain="ATCC 12472"
                     /db_xref="ATCC:12472"
                     /db_xref="taxon:243365"
     Protein         1..229
                     /product="DNA alkylation repair enzyme"
                     /calculated_mol_wt=25598
     Region          29..219
                     /region_name="AlkD_like_1"
                     /note="A new structural DNA glycosylase containing
                     HEAT-like repeats; cd07064"
                     /db_xref="CDD:132881"
     Region          30..219
                     /region_name="DNA_alkylation"
                     /note="DNA alkylation repair enzyme; pfam08713"
                     /db_xref="CDD:204039"
     Site            order(108,112,146,177..178,185)
                     /site_type="active"
                     /db_xref="CDD:132881"
     CDS             1..229
                     /gene="alkD"
                     /locus_tag="CV_0164"
                     /coded_by="NC_005085.1:173742..174431"
                     /transl_table=11
                     /db_xref="GeneID:2550257"
ORIGIN
        1 mdridalraq ltaaadpara pamraymrgq fdflgvaapa rrkaavawik shdtagpdvw
       61 ltlaerlwqe perefqyval dllarhaael paailprlla lvtakswwdt vdglaawvig
      121 glvrgrrelq temdtlagds dfwlrrvail hqlywkrdtd agrlfrycsa naadpeffir
      181 kaigwalrey aytdaeavrg fvasaalspl srrealkriq qnplpakea
//
```

All RefSeq accession numbers have an underbar at position 3 – GenBank records never have this

Jon K. Lærdahl,
Structural Bioinformatics

RefSeq accession numbers and types of molecules

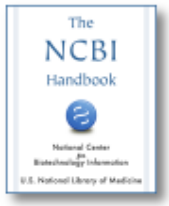
Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGS[a]
NW_	Genomic	Contig or scaffold, primarily WGS[a]
NS_	Genomic	Environmental sequence
NZ_[b]	Genomic	Unfinished WGS
NM_	mRNA	
NR_	RNA	
XM_[c]	mRNA	Predicted model
XR_[c]	RNA	Predicted model
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_[c]	Protein	
XP_[c]	Protein	Predicted model, associated with an XM_ accession
ZP_[c]	Protein	Predicted model, annotated on NZ_ genomic records

[a] Whole Genome Shotgun sequence data.

[b] An ordered collection of WGS sequence for a genome.

[c] Computed.

From: Chapter 18, The Reference Sequence (RefSeq) Database



The NCBI Handbook [Internet].
McEntyre J, Ostell J, editors.
Bethesda (MD): National Center for Biotechnology Information (US); 2002-.

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

```
LOCUS       NP_899834                229 aa            linear   BCT 20-JAN-2012
DEFINITION  DNA alkylation repair enzyme [Chromobacterium violaceum ATCC
            12472].
ACCESSION   NP_899834
VERSION     NP_899834.1  GI:34495619
DBLINK      Project: 58001
            BioProject: PRJNA58001
DBSOURCE    REFSEQ: accession NC_005085.1
KEYWORDS    .
SOURCE      Chromobacterium violaceum ATCC 12472
  ORGANISM  Chromobacterium violaceum ATCC 12472
            Bacteria; Proteobacteria; Betaproteobacteria; Neisseriales;
            Neisseriaceae; Chromobacterium.
REFERENCE   1  (residues 1 to 229)
  AUTHORS   Vasconcelos,A.T.R., de Almeida,D.F., Almeida,F.C., de
            ..................

COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AAQ57843.
            Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..229
                     /organism="Chromobacterium violaceum ATCC 12472"
                     /strain="ATCC 12472"
                     /db_xref="ATCC:12472"
                     /db_xref="taxon:243365"
     Protein         1..229
                     /product="DNA alkylation repair enzyme"
                     /calculated_mol_wt=25598
     Region          29..219
                     /region_name="AlkD_like_1"
                     /note="A new structural DNA glycosylase containing
                     HEAT-like repeats; cd07064"
                     /db_xref="CDD:132881"
     Region          30..219
                     /region_name="DNA_alkylation"
                     /note="DNA alkylation repair enzyme; pfam08713"
                     /db_xref="CDD:204039"
     Site            order(108,112,146,177..178,185)
                     /site_type="active"
                     /db_xref="CDD:132881"
     CDS             1..229
                     /gene="alkD"
                     /locus_tag="CV_0164"
                     /coded_by="NC_005085.1:173742..174431"
                     /transl_table=11
                     /db_xref="GeneID:2550257"
ORIGIN
        1 mdridalraq ltaaadpara pamraymrgq fdflgvaapa rrkaavawik shdtagpdvw
       61 ltlaerlwqe perefqyval dllarhaael paailprlla lvtakswwdt vdglaawvig
      121 glvrgrrelq temdtlagds dfwlrrvail hqlywkrdtd agrlfrycsa naadpeffir
      181 kaigwalrey aytdaeavrg fvasaalspl srrealkriq qnplpakea
//
```

Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGS[a]
NW_	Genomic	Contig or scaffold, primarily WGS[a]
NS_	Genomic	Environmental sequence
NZ_[b]	Genomic	Unfinished WGS
NM_	mRNA	
NR_	RNA	
XM_[c]	mRNA	Predicted model
XR_[c]	RNA	Predicted model
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_[c]	Protein	
XP_[c]	Protein	Predicted model, associated with an XM_ accession
ZP_[c]	Protein	Predicted model, annotated on NZ_ genomic records

NP_899834 is a protein, a translation of the NC_005085 genomic sequence from nucleotide (nt) 173742 – 174431, *i.e.* 690 nt (229 codons & stop codon)

Jon K. Lærdahl,
Structural Bioinformatics

```
LOCUS       NP_899834                229 aa            linear   BCT 20-JAN-2012
DEFINITION  DNA alkylation repair enzyme [Chromobacterium violaceum ATCC
            12472].
ACCESSION   NP_899834
VERSION     NP_899834.1  GI:34495619
DBLINK      Project: 58001
            BioProject: PRJNA58001
DBSOURCE    REFSEQ: accession NC_005085.1
KEYWORDS    .
SOURCE      Chromobacterium violaceum ATCC 12472
  ORGANISM  Chromobacterium violaceum ATCC 12472
            Bacteria; Proteobacteria; Betaproteobacteria; Neisseriales;
            Neisseriaceae; Chromobacterium.
REFERENCE   1  (residues 1 to 229)
  AUTHORS   Vasconcelos,A.T.R., de Almeida,D.F., Almeida,F.C., de
            ..................

COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AAQ57843.
            Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..229
                     /organism="Chromobacterium violaceum ATCC 12472"
                     /strain="ATCC 12472"
                     /db_xref="ATCC:12472"
                     /db_xref="taxon:243365"
     Protein         1..229
                     /product="DNA alkylation repair enzyme"
                     /calculated_mol_wt=25598
     Region          29..219
                     /region_name="AlkD_like_1"
                     /note="A new structural DNA glycosylase containing
                     HEAT-like repeats; cd07064"
                     /db_xref="CDD:132881"
     Region          30..219
                     /region_name="DNA_alkylation"
                     /note="DNA alkylation repair enzyme; pfam08713"
                     /db_xref="CDD:204039"
     Site            order(108,112,146,177..178,185)
                     /site_type="active"
                     /db_xref="CDD:132881"
     CDS             1..229
                     /gene="alkD"
                     /locus_tag="CV_0164"
                     /coded_by="NC_005085.1:173742..174431"
                     /transl_table=11
                     /db_xref="GeneID:2550257"
ORIGIN
        1 mdridalraq ltaaadpara pamraymrgq fdflgvaapa rrkaavawik shdtagpdvw
       61 ltlaerlwqe perefqyval dllarhaael paailprlla lvtakswwdt vdglaawvig
      121 glvrgrrelq temdtlagds dfwlrrvail hqlywkrdtd agrlfrycsa naadpeffir
      181 kaigwalrey aytdaeavrg fvasaalspl srrealkriq qnplpakea
//
```

Each RefSeq records has a comment block indicating the ***status*** of the record

STATUS	Definition
GENOME ANNOTATION	This identifies RefSeq records provided by the NCBI Genome Annotation process. These records are provided via automated processing and are not subject to individual review or revision between builds (see description of the assembly and annotation process). The mRNA records are identified based on alignments of other mRNAs to the genomic sequence and the proteins are conceptual translations of these mRNAs. These model transcripts and proteins may differ from pre-existing curated RefSeq (accession prefix NM, NR, NP) or GenBank records because they correspond to the genomic sequence.
INFERRED	*Not curated.* Inferred by genome sequence analysis with no direct same-species support for the product. Support for the record may include a combination of orthologous or paralogous protein homology and alignments of transcripts from related genes. A portion of the sequence may be defined by *ab initio* prediction.
MODEL	*Not curated.* The RefSeq record is predicted by a whole-genome computational genome annotation pipeline. The record may represent an *ab initio* prediction, or may have some level of transcript or protein homology support.
PREDICTED	*Not curated.* Automatically provided based on GenBank sequence data; limited or partial support for the transcript or protein. A portion of the transcript or protein may reflect an *ab initio* annotation prediction that was submitted to GenBank.
PROVISIONAL	*Not curated.* Automatically provided based on GenBank sequence data; there is support for the transcript and protein. This is the default status code applied to some genomes for which there is no clear information about the method used to define the sequence.
REVIEWED	*Curated.* The RefSeq record has been reviewed to provide the preferred sequence standard and to add additional functional descriptive information and feature annotation, as relevant.
VALIDATED	*Curated.* The RefSeq record has undergone an initial review to provide the preferred sequence standard.
WGS	*Not curated.* The RefSeq record represents a collection of whole genome shotgun (WGS) sequences. This status code is applied to genomic records.

http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status

Jon K. Lærdahl,
Structural Bioinformatics

GenBank versus RefSeq

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

http://www.ncbi.nlm.nih.gov/books/NBK21105/#ch1.Appendix_GenBank_RefSeq_TPA_and_UniP

Jon K. Lærdahl,
Structural Bioinformatics

NCBI Gene database

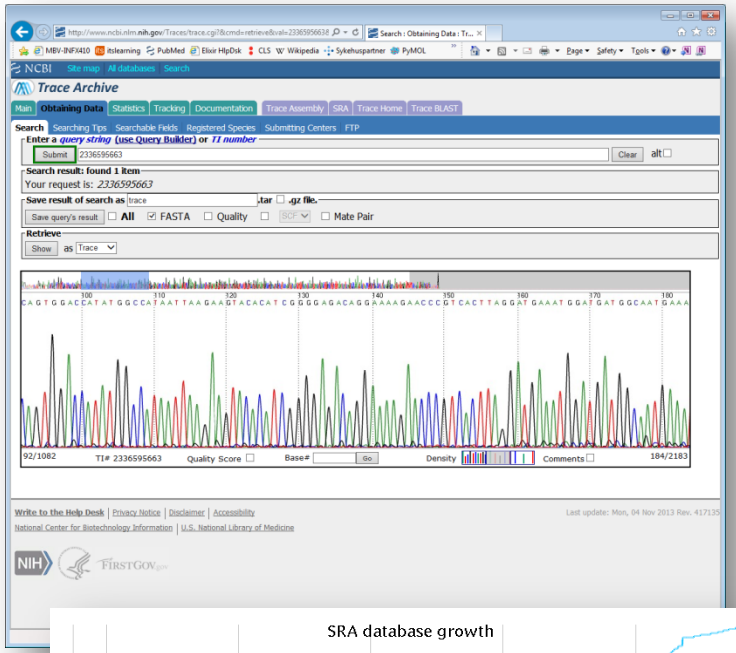
- Searched with Entrez
- A good place to start looking at new genes
- Lots of links to other databases!

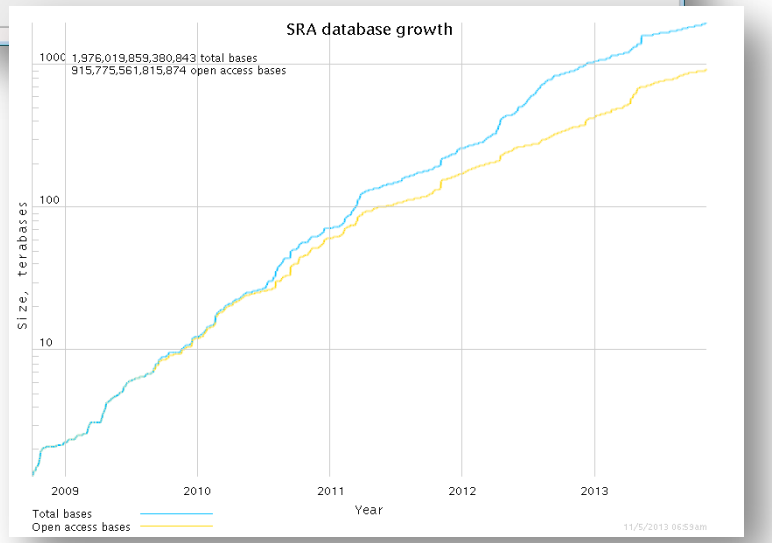
http://www.ncbi.nlm.nih.gov/gene/4968

Jon K. Lærdahl,
Structural Bioinformatics

NCBI «Trace Archives»

- Trace Archive
  - Repository of raw data sequencing traces from gel and capillary electrophoresis sequencers
  - >2 billion traces
- Sequence Read Archive (SRA)
  - Data from high-throughput sequencing (454, Illumina, IonTorrent, SOLiD, etc.)
  - Now 1.6 Pbases (1.6×10^{15}) open access sequences
  - At present >1 Tbase added daily





http://nar.oxfordjournals.org/content/40/D1/D54.abstract

1 Pbp ≈ 100,000 human genomes

UniProt

- Database of protein sequences and functional annotations – “a single worldwide database of protein sequence and function” (2002)
- UniProt consortium
  - EMBL-EBI
  - Swiss Institute of Bioinformatics (SIB)
    - Swiss-Prot (Amos Bairoch, 1986)
    - TrEMBL (Translated EMBL Nucleotide Sequence Data Library, 1996)
  - Protein Information Resource (PIR)
    - roots in Margaret Dayhoff's *Atlas of Protein Sequence and Structure* (1965)
- http://www.uniprot.org

Jon K. Lærdahl,
Structural Bioinformatics

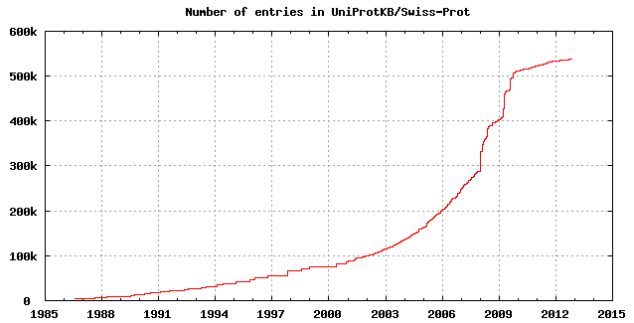
UniProt

- Four major components
  - the UniProt Knowledgebase (UniProtKB)
    - UniProtKB/Swiss-Prot
    - UniProtKB/TrEMBL
  - the UniProt Archive (UniParc)
  - the UniProt Reference Clusters (UniRef)
  - the UniProt Metagenomic and Environmental Sequence database (UniMES)
- Database release frequency: 4 weeks

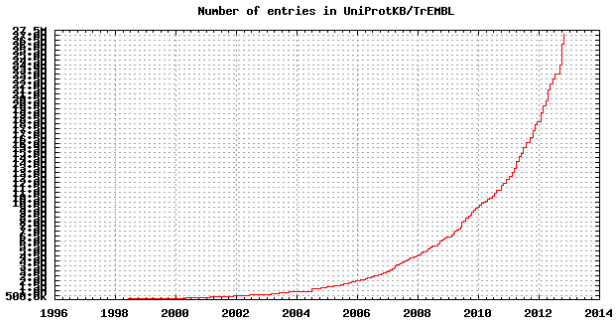
Jon K. Lærdahl,
Structural Bioinformatics

UniProtKB

- Two sections
  - UniProtKB/Swiss-Prot (Release 2013_10 of Oct 16, 2013)
    - 541,561 sequences
    - 192,480,382 amino acids
    - from 223,284 references



  - UniProtKB/TrEMBL (Release 2013_10 of Oct 16, 2013)
    - 44,746,523 sequences
    - 14,225,235,989 amino acids



UniProtKB

- UniProtKB/Swiss-Prot
  - high-quality data
    - information from literature
    - computational analysis by expert biocurators
  - ***manually annotated/curated***
  - annotation regularly reviewed
  - non-redundant
  - sequences from same gene in same organism merged and differences between sequences are identified
    - alternative splicing, natural variation, incorrect initiation sites, incorrect exon boundaries, frame shifts, unidentified conflicts?
  - ***a very good place to start working on a new protein!***
- UniProtKB/TrEMBL
  - translations of annotated coding sequences submitted to the ENA/GenBank/ DDBJ nucleotide sequence resources (INSDC)
  - sequences are experimental *(e.g.* mRNA/cDNA) or generated by gene prediction programs
  - computational (only) analysis and automatic annotation
  - also data from Ensembl, RefSeq, and CCDS not submitted to INSDC, as well as PDB database sequences (3D structures)

Jon K. Lærdahl,
Structural Bioinformatics

```
ID   ALG8_CAEEL              Reviewed;         766 AA.
AC   P52887; Q5WRQ2;
DT   01-OCT-1996, integrated into UniProtKB/Swiss-Prot.
DT   03-OCT-2012, sequence version 3.
DT   31-OCT-2012, entry version 89.
DE   RecName: Full=Probable dolichyl pyrophosphate Glc1Man9GlcNAc2 alpha-1,3-glucosyltransferase;
DE            EC=2.4.1.265;
DE   AltName: Full=Asparagine-linked glycosylation protein 8 homolog;
DE   AltName: Full=Dol-P-Glc:Glc(1)Man(9)GlcNAc(2)-PP-dolichyl alpha-1,3-glucosyltransferase;
DE   AltName: Full=Dolichyl-P-Glc:Glc1Man9GlcNAc2-PP-dolichyl glucosyltransferase;
GN   ORFNames=C08H9.3;
OS   Caenorhabditis elegans.
OC   Eukaryota; Metazoa; Ecdysozoa; Nematoda; Chromadorea; Rhabditida;
OC   Rhabditoidea; Rhabditidae; Peloderinae; Caenorhabditis.
OX   NCBI_TaxID=6239;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA], AND ALTERNATIVE
RP   SPLICING.
RC   STRAIN=Bristol N2;
RX   MEDLINE=99069613; PubMed=9851916; DOI=10.1126/science.282.5396.2012;
RG   The C. elegans sequencing consortium;
RT   "Genome sequence of the nematode C. elegans: a platform for
RT   investigating biology.";
RL   Science 282:2012-2018(1998).
CC   -!- FUNCTION: Adds the second glucose residue to the lipid-linked
CC       oligosaccharide precursor for N-linked glycosylation. Transfers
CC       glucose from dolichyl phosphate glucose (Dol-P-Glc) onto the
CC       lipid-linked oligosaccharide Glc(1)Man(9)GlcNAc(2)-PP-Dol (By
CC       similarity).
CC   -!- CATALYTIC ACTIVITY: Dolichyl beta-D-glucosyl phosphate + D-Glc-
CC       alpha-(1->3)-D-Man-alpha-(1->2)-D-Man-alpha-(1->2)-D-Man-alpha-
CC       (1->3)-[D-Man-alpha-(1->2)-D-Man-alpha-(1->3)-(D-Man-alpha-(1->2)-
CC       D-Man-alpha-(1->6))-D-Man-alpha-(1->6)]-D-Man-beta-(1->4)-D-
CC       GlcNAc-beta-(1->4)-D-GlcNAc-diphosphodolichol = D-Glc-alpha-
CC       (1->3)-D-Glc-alpha-(1->3)-D-Man-alpha-(1->2)-D-Man-alpha-(1->2)-D-
CC       Man-alpha-(1->3)-[D-Man-alpha-(1->2)-D-Man-alpha-(1->3)-(D-Man-
CC       alpha-(1->2)-D-Man-alpha-(1->6))-D-Man-alpha-(1->6)]-D-Man-beta-
CC       (1->4)-D-GlcNAc-beta-(1->4)-D-GlcNAc-diphosphodolichol + dolichyl
CC       phosphate.
CC   -!- PATHWAY: Protein modification; protein glycosylation.
CC   -!- SUBCELLULAR LOCATION: Endoplasmic reticulum membrane; Multi-pass
CC       membrane protein (By similarity).
CC   -!- ALTERNATIVE PRODUCTS:
CC       Event=Alternative splicing; Named isoforms=2;
CC       Name=a;
CC         IsoId=P52887-1; Sequence=Displayed;
CC         Note=No experimental confirmation available;
CC       Name=b;
CC         IsoId=P52887-2; Sequence=VSP_012307, VSP_012308;
CC         Note=No experimental confirmation available;
          ..................
```

Flat text format of a UniProtKB record

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

```
.................
DR   InterPro; IPR021173; Dolichyl-P-Glc/endonucV.
DR   InterPro; IPR007581; Endonuclease-V.
DR   InterPro; IPR004856; Glyco_trans_ALG6/ALG8.
DR   PANTHER; PTHR12413; PTHR12413; 1.
DR   Pfam; PF03155; Alg6_Alg8; 1.
DR   Pfam; PF04493; Endonuclease_5; 1.
DR   PIRSF; PIRSF037345; Dolichyl-P-Glc/endonucV; 1.
PE   2: Evidence at transcript level;
KW   Alternative splicing; Complete proteome; Endoplasmic reticulum;
KW   Glycosyltransferase; Membrane; Reference proteome; Transferase;
KW   Transmembrane; Transmembrane helix.
FT   CHAIN         1    766       Probable dolichyl pyrophosphate
FT                                Glc1Man9GlcNAc2 alpha-1,3-
FT                                glucosyltransferase.
FT                                /FTId=PRO_0000174164.
FT   TRANSMEM      6     26       Helical; (Potential).
FT   TRANSMEM     60     80       Helical; (Potential).
FT   TRANSMEM     96    116
FT   TRANSMEM    156    176
FT   TRANSMEM    190    210
FT   TRANSMEM    228    248
FT   TRANSMEM    324    344
FT   TRANSMEM    350    370
FT   TRANSMEM    395    415
FT   TRANSMEM    423    443
FT   TRANSMEM    452    472
FT   TRANSMEM    482    502
FT   VAR_SEQ     111    123
FT
FT
FT   VAR_SEQ     124    766
FT
SQ   SEQUENCE   766 AA;  8687
     MGEVQLVLAV TAILISFKCL LI
     PPFFAYFELG LASVAHFFGF DE
     RSPRLVSRIP KKLQQNGREA CF
     YLMAALSYSI LLNFKHIYVY YA
     ASIFPFIHAS GVQGLQNIAT RL
     FDAPTYTSGL VQEYSHSVLP NV
     FCFFYFGYHV HEKAIILVTV PM
     YAICVSYFFI QLVFLKRVTL MP
     PLMAISILTA IELTGLIGAL IW
     DVKLVAGIDT SAAKLNSDMV YI
     MADFLKSVIT ERPELRPDVI LC
     ETIGMENKSK VDSFVEHCRE VY
     GIDLELSTEI CSQLLLNNTT IE
//
```

Accession

Last modified March 31, 2011

This subsection of the 'Entry information' section provides one or more accession number(s). These are stable identifiers and should be used to cite UniProtKB entries. Upon integration into UniProtKB, each entry is assigned a unique accession number, which is called 'Primary (citable) accession number'.

UniProtKB accession numbers consist of 6 alphanumerical characters in the format:

1	2	3	4	5	6
[A-N,R-Z]	[0-9]	[A-Z]	[A-Z, 0-9]	[A-Z, 0-9]	[0-9]
[O,P,Q]	[0-9]	[A-Z, 0-9]	[A-Z, 0-9]	[A-Z, 0-9]	[0-9]

Examples: A2BC19, P12345.

Entries can have more than one accession number. This can be due to two distinct mechanisms:

- a) When two or more entries are merged, the accession numbers from all entries are kept. The first accession number is referred to as the 'Primary (citable) accession number', while the others are referred to as 'Secondary accession numbers'. These are listed in alphanumerical order.
- b) If an existing entry is split into two or more entries ('demerged'), new 'primary' accession numbers are attributed to all the split entries while all original accession numbers are retained as 'secondary' accession numbers.

Example: P29358 which has been 'demerged' into P68250 and P68251.

Jon K. Lærdahl,
Structural Bioinformatics

UniProt

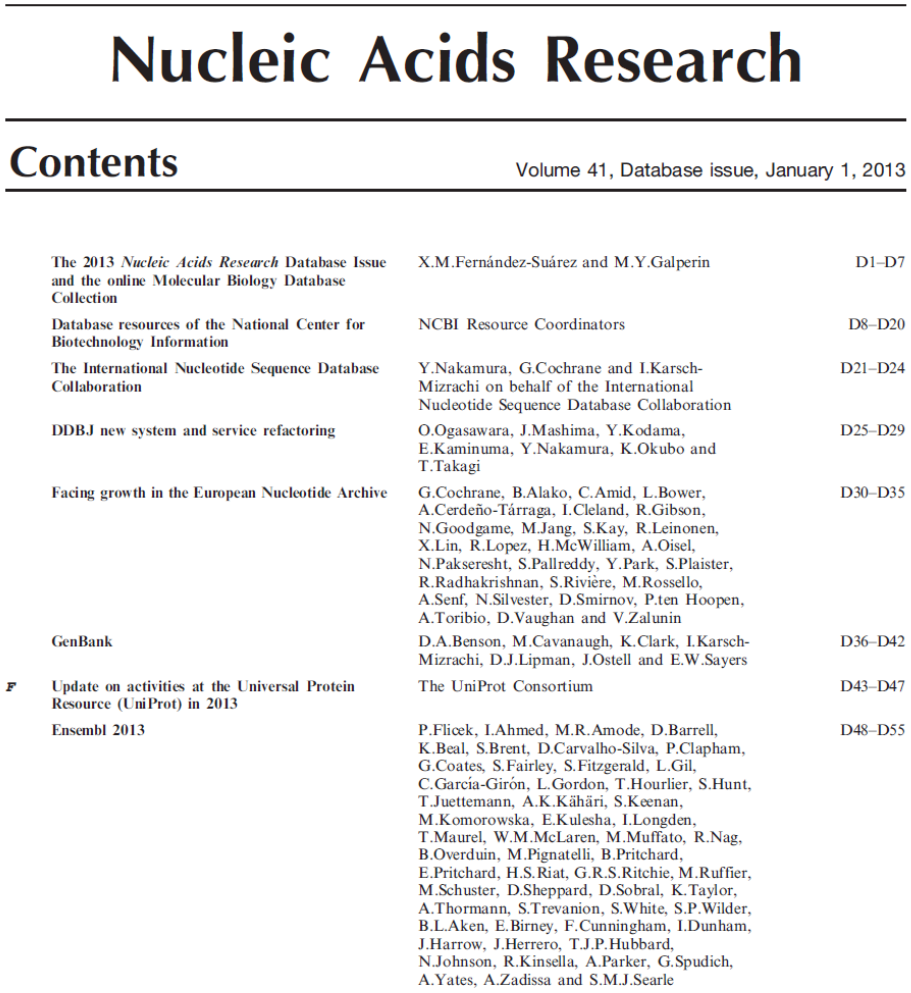
- UniProt Archive (UniParc)
  - the most comprehensive, publicly accessible, non-redundant protein sequence database available
  - new and updated protein sequences are loaded daily from “all” public databases
  - each unique sequence is stored only once and assigned a UniParc identifier
  - identifiers are stable and are never deleted or reassigned
  - UniParc identifiers can be used to uniquely identify protein sequences in any protein database
  - format of UniParc identifiers: UPI000000D8D2 (“UPI” and 10 hexadecimal numbers)
  - UniParc contains only protein sequences and database cross-references
- UniProt Reference Clusters (UniRef)
  - 3 databases of clustered sets of protein sequences from UniProtKB
  - UniRef100 database: identical sequences (from any organism) into a single entry
  - UniRef90: each entry has sequences that have at least 90% sequence identity (wrt the longest sequence)
  - UniRef50: each entry has sequences that have at least 50% sequence identity
- UniProt Metagenomic and Environmental Sequence database (UniMES)
  - developed for MES data
  - clusters of sequences (100% and >90%)

Two main types of databases?

- *Primary* databases
  - gather original data
  - curation and archiving new data
  - *e.g.* GenBank, SRA, and the 3D structure database PDB
- *Secondary* (derived databases)
  - collect data from primary databases and reformat, reannotate, recombine etc.
  - *e.g.* UniProtKB, UniParc, UniRef, or Gene Ontology Annotation (GOA)

Jon K. Lærdahl,
Structural Bioinformatics

An even better place to look for good biological databases -



Nucleic Acids Research

Contents

Volume 41, Database issue, January 1, 2013

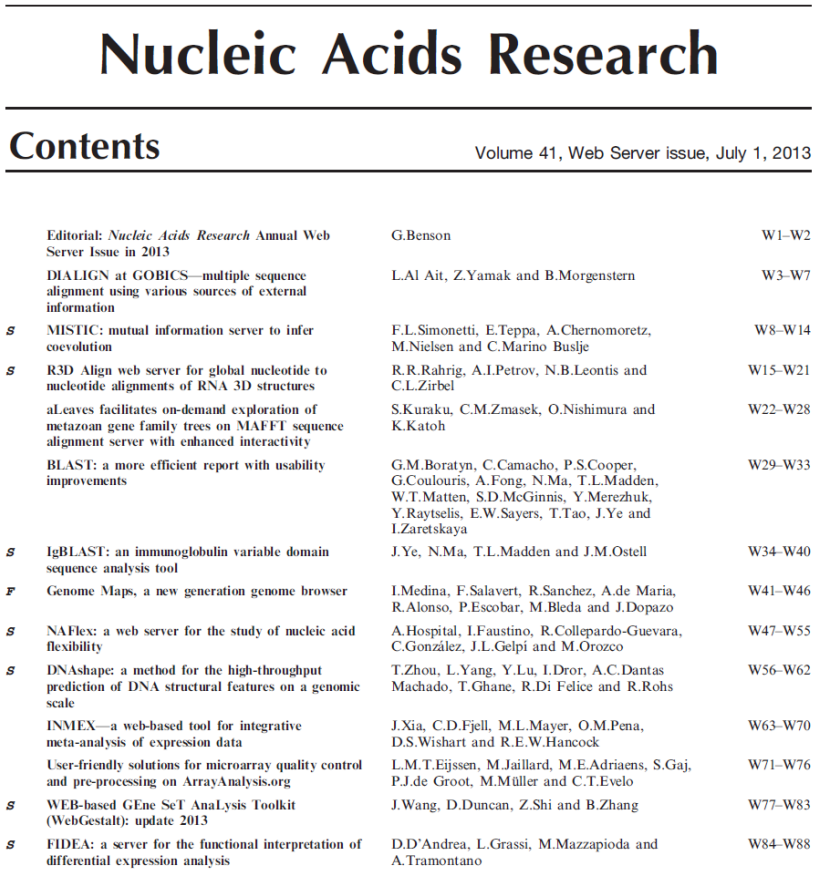
Nucleic Acids Res. Database issues

- released once every year, in January
- 21st issue (2013)
  - 58 new databases
  - 100 updates on databases previously described in *NAR*
  - 23 updates on databases previously described elsewhere

http://nar.oxfordjournals.org/content/42/D1.toc

While we are visiting *NAR:* a good place to look for bioinformatics tools

Jon K. Lærdahl,
Structural Bioinformatics



Nucleic Acids Research

Contents

Volume 41, Web Server issue, July 1, 2013

Nucleic Acids Res. Web server issues

- released once every year, in July
- 12th issue (2014)
  - 80 web servers

http://nar.oxfordjournals.org/content/42/W1.toc

If you need an article or a citation for a bioinformatics tool or database, the *NAR* web server or databases issues are often good places to look

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

Huge number of databases!

- In bioinformatics, the number of databases, tools, algorithms, and papers is enormous
  - impossible to have an overview, especially if bioinformatics is not your main research area
  - instead of trying to do everything yourself:

Get yourself a bioinformatics expert colleague or collaborator!

http://www.oxfordjournals.org/nar/database/a

NAR online Molecular Biology Database Collection, currently contains >1500 databases

Jon K. Lærdahl,
Structural Bioinformatics

Good and bad databases

- Some are exceptionally good, well maintained and often updated
  - EMBL-EBI, NCBI, Ensembl,...
  - http://string.embl.de
  - http://www.pdb.org
  - Maintained by 10s and 100s of experts...
- Species specific
  - http://www.pombase.org (*Schizosaccharomyces pombe*)
  - http://flybase.org (*Drosophila*)
  - http://ecocyc.org (*Escherichia coli* K-12 MG1655)
- Unique content
  - http://www.proteinatlas.org

- Also many have poor quality, are never updated, are unreliable

Trick is to know what is good and what is bad...

Oslo universitetssykehus

Let your favourite bioinformatician follow the field!

It is possible to spend weeks learning to become an expert on only a single resource...

We have only looked at a tiny corner here...