PIUG Boston Biotechnology Meeting
12 February 2007

Which Date?: Finding the Publicly Accessible Date in a Sequence Prior Art Search

Kenneth L. Hoppe
Pfizer Inc

Publicly Accessible

- 35 U.S.C. 102 (a) and (b) MPEP 2132.
- "Printed Publication"
- “An electronic publication, including an on-line database or Internet publication, is considered to be a "printed publication" within the meaning of 35 U.S.C. 102(a) and (b) **provided the publication was accessible to persons concerned with the art to which the document relates** . . . “ (MPEP 2128)

Publicly Accessible

- Date of Availability
- “Prior art disclosures on the Internet or on an on-line database are considered to be publicly available as of **the date the item was publicly posted**.” (MPEP 2128)

Sequence Databases

- CAS REGISTRY
- GENBANK / EMBL / DDBJ
- SwissProt / TrEMBL / UniProt
- RefSeq

Sequencing

- 1951 – Insulin B-chain 30 a.a. – Sanger et al.
- 1965 – yeast RNA 77 bases – Holly et al.
- 1967 – Protein Auto – Edman and Begg
- **1977** – **DNA** – Sanger et al.
- **1977** – **DNA** – Maxam and Gilbert
- 1986 – Automation DNA – Leroy Hood

First Sequence Collections

- “Atlas of Protein Sequence and Structure” First Edition published 1965 – 65 entries
- Other scattered collections

EMBL-Bank

- First release in **1982**
- Worlds first publicly available database for nucleic acids
- Find sequence published in literature
- Re-key each base by hand
- First distributed on 9-track magnetic tape
- http://www.ebi.ac.uk/

GENBANK

- **1982:** Bolt, Beranek and Newman & Los Alamos National Laboratory
- **1987:** IntelliGenetics, Inc & Los Alamos National Laboratory
- **Oct 1992*:** National Center for Biotechnology Information (NCBI)
- http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi

DDBJ

- DNA DataBank of Japan
- Started in **1984**
- Began activities in earnest in **1986**
- Collaboration EMBL & GENBANK
- http://www.ddbj.nig.ac.jp/

International Nucleotide Sequence Database Collaboration

- GENBANK & EMBL collaborated early
- Formalized in February **1987**
- Included DDBJ
- INSD = GENBANK / EMBL / DDBJ

NCBI

- National Center for Biotechnology Information
- Created by Congress in **1988**
- Responsible for GENBANK **1992***

SwissProt & RefSeq Dates

- These are Secondary resources
- Human or Computer curated
- Good for Literature citations

SwissProt / TrEMBL / UniProt

- Released in **1986** by Amos Bairoch
- Protein Identification Resource (PIR) database in EMBL line-oriented format
- TrEMBL introduced in **1996**
- SwissProt contains high-quality annotation, is non-redundant and cross-referenced to many other databases.
- http://us.expasy.org/

NCBI RefSeq

- Released **1999**
- RefSeq provides a biologically non-redundant collection of DNA, RNA, and protein sequences. Each sequence represents a single molecule from a specific organism. Reference Sequences are manually curated and annotated to provide synthesis of information (and sequence records), similar to a review article in the literature.
- Accession number = XX_123456

Common Search Tools

- 1970 – Needleman & Wunsch - Global
- 1981 – Smith & Waterman - Local
- 1988 – FASTA (Pearson & Lipman) - Local
- 1990 – BLAST (Altshul et al.) - Local

GENBANK / EMBL / DDBJ

- “collection of all publicly available DNA sequences.”

Example 1 - Not in GENBANK

- Dog Apolipoprotein-C**I** DNA
- SwissProt cited 1989 journal article
- DNA not in GENBANK until 2005 with a different reference
- In CAS REGISTRY 1990
- CAPLUS record for 1989 journal article

GENBANK Date

- Dates on the Record are not always the public release date.
- NCBI “Sequence Revision History”
  http://www.ncbi.nlm.nih.gov/entrez/sutils/girevhist.cgi
- EMBL “Sequence Version Archive”
  http://www.ebi.ac.uk/cgi-bin/sva/sva.pl

Example 2 - GENBANK

NCBI Record

```
LOCUS       DQ294229                2268 bp    mRNA    linear   INV 03-AUG-2006
DEFINITION  Glossina morsitans morsitans TC1715 yolk protein 3 mRNA, complete
            cds.
ACCESSION   DQ294229
VERSION     DQ294229.1  GI:83944681
KEYWORDS    .
SOURCE      Glossina morsitans morsitans
  ORGANISM  Glossina morsitans morsitans
            Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
            Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
            Hippoboscoidea; Glossinidae; Glossina.
REFERENCE   1  (bases 1 to 2268)
  AUTHORS   Attardo,G.M., Strickler-Dinglasan,P., Perkin,S.A.H., Caler,E.,
            Bonaldo,M.F., Soares,M.B., El-Sayeed,N. and Aksoy,S.
  TITLE     Analysis of fat body transcriptome from the adult tsetse fly,
            Glossina morsitans morsitans
  JOURNAL   Insect Mol. Biol. 15 (4), 411-424 (2006)
REFERENCE   2  (bases 1 to 2268)
  AUTHORS   Attardo,G.M., Strickler-Dinglasan,P., Perkin,S.A.H., Soares,M.B.,
            El-Sayeed,N. and Aksoy,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (16-NOV-2005) Public Health and Epidemiology, Yale
            University Medical School, 60 College Street, New Haven, CT 06520,
```

August Issue

NCBI History

83944681	1	Dec 28 2005 12:07 AM

Accession DQ294229 was first seen at NCBI on Dec 28 2005 12:07 AM

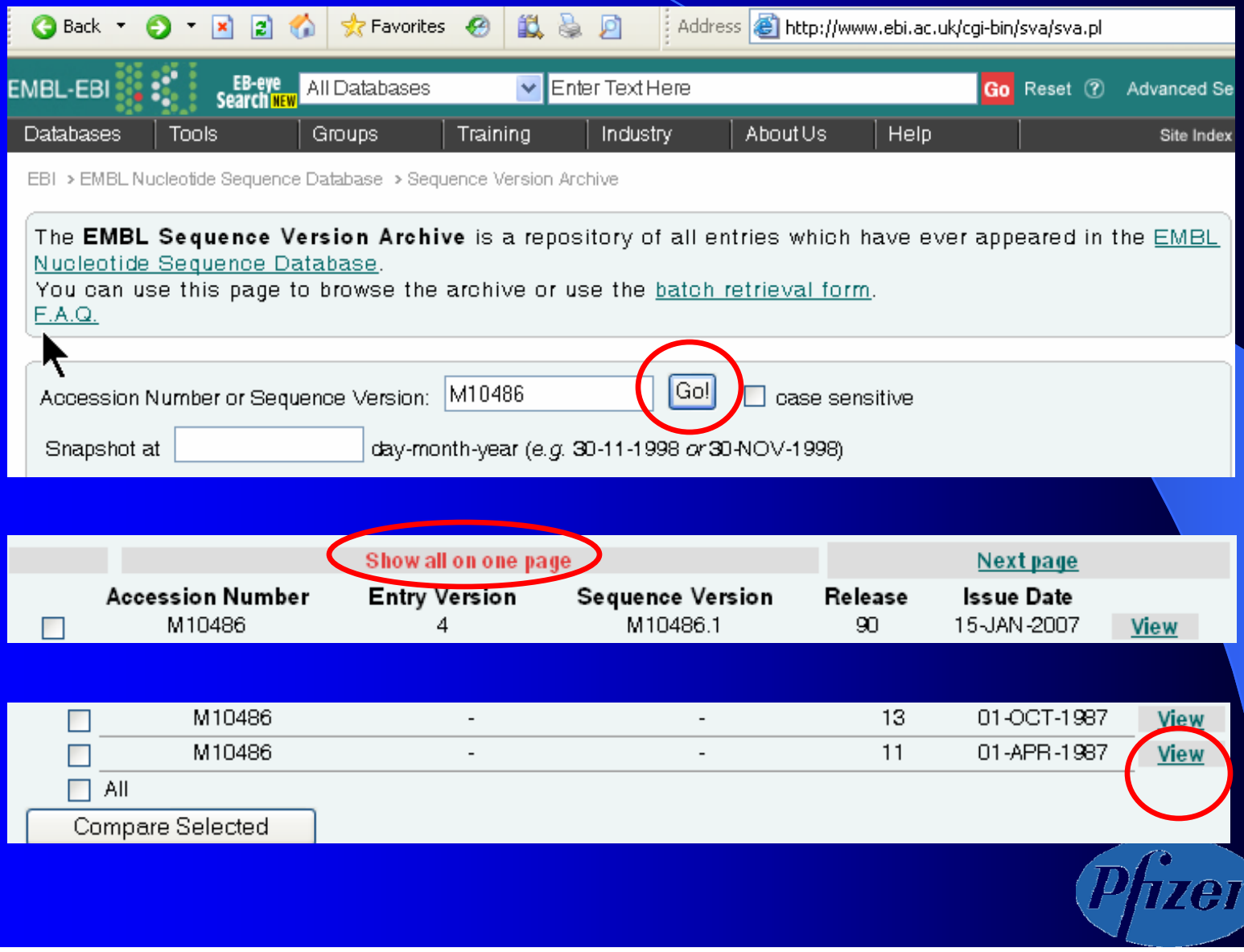
EMBL History

DQ294229	1	DQ294229.1	86	28-DEC-2005	View
All					

Example 3 - GENBANK Date?

- J. Bacteriol. 144 (1), 131-140 (1980)
- Version “Accession **M10486** was first seen at NCBI on Apr 26 1993 4:27 PM“
- NOTE: “at NCBI” not “at GENBANK”
- Record date: 26-APR-1993

Example 3 - EMBL Date?



EMBL Dates?

- What is an Entry Version? Each time an entry is modified, excluding taxonomic, database cross-reference, and journal name changes, the entry is assigned a new entry version number. The EMBL Sequence Version Archive captures all the entry changes, including changes in the flat file format. As a result multiple entries with the same entry version may coexist in the archive.

- What is a Release? The release refers to the **quarterly** EMBL release in which a flat file appeared, or was expected to appear.

- What is an **Issue Date**? The issue data is the date when the entry was **made available to the public** as part of daily update, or release.

From EMBL Sequence Version Archive FAQ.

Example 3 - EMBL

```
This flat file was issued: 01-APR-1987 Rel: 11

ID   M10486   unannotated; DNA; 266 BP.
**   Converted from GenBank entry "M10486" on 26-FEB-1987 at 17:42:07.48
XX
AC   M10486;
XX
DT   26-FEB-1987  (incorporated)
**   03-FEB-1986  (in Genbank database)
XX
DE   ColE1 IS102 insertion site.
XX
KW   .
XX
OS
OC
XX
RN   [1]  (bases 1-266)
RA   Ohtsubo H., Zenilman M., Ohtsubo E.;
RT   "Insertion element IS102 resides in plasmid pSC101";
RL   J. Bact. 144:131-140(1980).
```

Example 4 - NCBI

NCBI History

14789	1	Apr 20 1993 9:25 PM

Accession X14869 was first seen at NCBI on Apr 20 1993 9:25 PM

```
LOCUS       BAT4INH                 3121 bp    DNA     linear   PHG 16-OCT-1991
DEFINITION  Bacteriophage T4 hoc gene.
ACCESSION   X14869
VERSION     X14869  GI:14789
KEYWORDS    hoc gene; unidentified reading frame.
SOURCE      Unknown.
  ORGANISM  Unknown.
            Unclassified.
REFERENCE   1  (bases 1 to 3121)
  AUTHORS   Kaliman,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-MAR-1981) Kaliman A., Institute of Biochemistry and
            Physiology of Microorganismus, USSR Acadeny of Sciences, Pushchino,
            Moscow Region 142292, USSR.
REFERENCE   2  (bases 1 to 3121)
  AUTHORS   Kaliman,A.V., Khasanova,M.A., Kryukov,V.M., Tanyashin,V.I. and
            Bayev,A.A.
  TITLE     The nucleotide sequence of the region of bacteriophage T4
            inh(lin)-hoc genes
  JOURNAL   Nucleic Acids Res. 18, 4277-4277 (1990)
```

Example 4 - EMBL

```
This flat file was issued 01-AUG-1990 Rel: 24

ID   BAT4INH    standard; DNA; PHG; 3121 BP.
XX
AC   X14869;
XX
DT   15-SEP-1989 (annotation)
XX
DE   Bacteriophage T4 hoc gene.
XX
KW   hoc gene; unidentified reading frame.
XX
OS   Bacteriophage T4
OC   Viridae; ds-DNA nonenveloped viruses; Myoviridae.
XX
RN   [1] (bases 1-3121)
RA   Kaliman A.;
RT   ;
RL   Submitted (28-MAR-1981) to the EMBL Data Library by:
RL   Kaliman A., Institute of Biochemistry and Physiology of
RL   Microorganismus , USSR Acadeny of Sciences , Pushchino, Moscow
RL   Region 142292, USSR.
XX
RN   [2] (bases 1-3121)
RA   Kaliman A.V., Khasanova M.A., Kryuk
RA   Bayer A.A.;
RT   "The nucleotide sequence of the reg
RT   inh(lip)-hoc genes";
RL   Nucleic Acids Res. 18:4277-4277(1990).
```

25 July Issue

Summary

1. Check “history” at both NCBI & EMBL for public release date
2. Public release before publication date
3. 1980’s complex issues