

GENOME RESEARCH

Genome Res. 1997 7: 952-955

Access the most recent version at doi:[10.1101/gr.7.10.952](https://doi.org/10.1101/gr.7.10.952)

References

This article cites 8 articles, 4 of which can be accessed free at:
<http://www.genome.org/cgi/content/full/7/10/952#References>

Article cited in:

<http://www.genome.org/cgi/content/full/7/10/952#otherarticles>

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#)

Notes

To subscribe to *Genome Research* go to:
<http://www.genome.org/subscriptions/>



Database Divisions and Homology Search Files: A Guide for the Perplexed

B.F. Francis Ouellette and Mark S. Boguski

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894 USA

The exponential growth of DNA sequence data has become a challenge for both end users and database curators alike. When one of us (M.S.B.) was finishing graduate school, GenBank® (release 42) contained a mere 6.7 Mb in 9700 sequences. However, as we write this, GenBank (Benson et al. 1997) has topped 1000 Mb in >1.6 million sequences (release 102). (Information on GenBank releases is available at <ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt>). The National Center for Biotechnology Information (NCBI) and its partners in the international database collaboration—the DNA Database of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL)—all strive to collect, manage, and distribute this data in the most efficient and usable manner possible. These organizations also provide homology search, database query, and information retrieval services that serve the general molecular biology community as well as more specialized users. Unfortunately, it is easy to become confused about the many ways in which the data are made available for downloading, homology searching, and more general information retrieval purposes. We hope to clarify some of these issues here, with an emphasis on the manner in which high-throughput genomic sequence is processed, distributed, and made available for BLAST searching. We will emphasize services provided through NCBI but also note comparable services at European Bioinformatics Institute and the slight differences between GenBank, DDBJ, and the EMBL Data Library.

Divisions of the Nucleotide Sequence Databases

The nucleotide sequence databases were originally organized around loosely defined taxonomic groupings that reflected research trends and sequencing

activity of a former era. These divisions are not as biologically relevant today, but so many public and private software systems have been developed to process these divisions that the databases must be conservative when contemplating changes in the structure of data distributions. The current divisional structures of GenBank, EMBL, and DDBJ are shown in Table 1. The reader will note that not all of these divisions are taxonomically based and that certain “functional” divisions have been added over time. Notably, in recent years, new divisions were added for EST and STS data because

these sequences differed from traditional GenBank entries in many ways, including the way in which people computed on the data (Boguski et al. 1993). The newest functional division, the High Throughput Genomic (HTG) Sequence Division, is described below. Additional information is available at <http://www.ncbi.nlm.nih.gov/HTGS>.

HTG

Although the issue is still a matter of some controversy (Adams and Venter 1996; Bentley 1996), a consortium of

Table 1. Database Divisions

| Sequence division | Database |
|---|---------------------|
| <i>Organismal</i> | |
| BCT Bacterial | DDBJ, GenBank |
| PRO Prokaryotic | EMBL |
| FUN Fungal | EMBL |
| HUM Human | DDBJ, EMBL |
| PRI Primate | DDBJ, EMBL, GenBank |
| ROD Rodent | DDBJ, EMBL, GenBank |
| MAM Other mammalian | DDBJ, EMBL, GenBank |
| VRT Other vertebrate | DDBJ, EMBL, GenBank |
| INV Invertebrate | DDBJ, EMBL, GenBank |
| PLN Plant | DDBJ, EMBL, GenBank |
| ORG Organelle | EMBL |
| VRL Viral | DDBJ, EMBL, GenBank |
| PHG Phage | DDBJ, EMBL, GenBank |
| RNA Structural RNA | DDBJ, EMBL, GenBank |
| SYN Synthetic and chimeric | DDBJ, EMBL, GenBank |
| UNA Unannotated | DDBJ, EMBL, GenBank |
| <i>Functional</i> | |
| EST Expressed sequence tag | DDBJ, EMBL, GenBank |
| STS Sequence tagged site | DDBJ, EMBL, GenBank |
| GSS Genome survey | DDBJ, EMBL, GenBank |
| HTG High-throughput genomic | DDBJ, EMBL, GenBank |
| PAT Patent | DDBJ, EMBL, GenBank |
| CON ^a Virtual contigs of segmented sequences | DDBJ, EMBL, GenBank |

^aThis division, which will appear in future database releases, is designed to contain instructions for assembly of segmented sequence records.

large-scale sequencing centers and their funding agencies have reached a consensus agreement (the "Bermuda Principles") regarding data produced in publicly funded projects. This agreement states that "unfinished" sequence data be released as soon as it is "usable" for homology searching and other types of sequence analysis. Usable data are currently defined as all sequences existing in contigs of >2 kb. Preliminary data such as these can be generated quite rapidly as they usually represent automated assemblies of single-pass, shotgun sequences. However, conversion to the "finished" state (complete contiguity with an error rate of 10^{-4} or less) may take considerably longer; hence, the motivation to release unfinished but usable sequence earlier. This process of data generation and public release is entirely different from traditional GenBank data submission, and the international collaborators have devised and implemented a system to accommodate this new paradigm. Unfinished sequences are submitted to and stored in the HTG Division, and each record is plainly labeled to indicate the preliminary nature of the data. An example is given in Figure 1.

HTG records contain sequences derived from a single genomic clone, and the entire set receives a single GenBank accession number that remains with the sequence as it progresses to the finished state. When declared finished by the submitting laboratory, these records move into the traditional repositories of finished data—the organismal divisions of GenBank—and are placed according to the biological source of the sequence. Thus, finished human sequences are distributed in the Primate (PRI) Division of GenBank (or the HUM Division for EMBL and DDBJ), whereas finished nematode and *Arabidopsis* sequences are found in the Invertebrate (INV) and Plant (PLN) Divisions, respectively (Table 2). It may seem rather coarse to lump *Homo sapiens* with other primates and *Caenorhabditis elegans* with other invertebrates; but this legacy of the earlier history of GenBank is irrelevant in the face of meta-information retrieval systems such as NCBI's Entrez that, in conjunction with NCBI's Taxonomy Database, permits one to explore and retrieve sequence records for any of the ~25,000 biological species in GenBank. Furthermore, new versions of the BLAST soft-

ware permit homology searches based on inclusive taxonomy parameters (Zhang and Madden 1997).

Homology Search Files at NCBI and EBI

The divisional structures of GenBank, DDBJ, and EMBL Data Library were primarily designed for the purposes of efficient data distribution and file storage. For homology search purposes, there are other, more practical and desirable ways to organize the sequence data. For example, unfinished data such as EST and HTG sequences always need to be analyzed with error-tolerant software (such as BLASTX or TBLASTN) (Altschul et al. 1994). On the other hand, finished (accurate and annotated) data may have coding features that can be automatically converted to conceptual translations in a protein database where BLASTP provides a more sensitive and specific search tool. Thus, it is inefficient to combine finished and unfinished data in a single file for homology search purposes. It is also undesirable to combine qualitatively different types of data in a single search file. STSs, for example, have their own division of GenBank, and homology searching is not the most appropriate method for querying these data (Schuler 1997).

Another important consideration in the construction of homology search files is the issue of sequence redundancy (Altschul et al. 1994). GenBank, DDBJ, and EMBL Data Library are historical archives and may contain many, nearly identical versions of the same sequence. The "nr" (for nonredundant) data set (Altschul et al. 1994) is NCBI's attempt to provide a more streamlined, yet comprehensive, collection of sequences for homology search purposes. nr includes finished (but not unfinished) HTG records (Table 2). Another important example is the "month" data set that provides a rolling month view of new GenBank entries. month is provided so that one does not have to repeatedly search previously examined portions of nr to identify matches to new sequences that have appeared since the last search was performed. month includes newly finished HTG records. Unfinished (phase 1 and phase 2) HTG data are accessible for BLAST searching at NCBI by specifying the htgs database (Table 2).

As described previously, there are

slight variations in the divisional structures of the three collaborating databases (Table 1). There are also differences in the ways in which the sequence data are made available for homology searching. One important example of this is the EMBL "ALL" database (emall) that combines both finished and unfinished HTG sequences for FASTA searching (Table 2).

DDBJ, EMBL, and GenBank must be conservative in contemplating changes to the divisional structures of the databases. However, these organizations can be and have been more flexible in producing specialized collections for homology searching. Thus, the user community should view the databases listed in Table 2 as subject to changes and improvements, driven by the ever-increasing quantity and variety of new sequence data.

Other Ways to Access Data

Entrez is a meta-information system that has been described in detail elsewhere (Schuler et al. 1996; Benson et al. 1997) and allows the user to query an extensive information space characterized by six divisions: (1) DNA sequences; (2) protein sequences; (3) maps and genomes; (4) macromolecular structures; (5) biomedical literature; and (6) taxonomy. Regarding DNA sequences in Entrez, all data in GenBank, regardless of Division, are available, including unfinished HTG records. These data may be queried using accession numbers, nucleotide sequence identifiers (NIDs), authors' names, and a variety of other key words, as well as by accessing pre-computed homology search results—a concept referred to as neighboring. In the near future, NCBI hopes to make available BLASTX neighbors through its Entrez service. This would allow users to access sequence similarities between even unfinished HTG records and the proteins they may encode.

Summary

All of the data in GenBank (and EMBL and DDBJ) are made available in a variety of ways, tailored to particular uses such as efficient data submission, distribution, and sequence homology searching. Unfortunately this can be somewhat confusing for contributors, data managers, and end users, all of whom

Insight/Outlook

A

LOCUS HSAC000003 120000 bp DNA HTG 20-SEP-1996
 DEFINITION *** SEQUENCING IN PROGRESS *** Chromosome 17 genomic sequence;
 HTGS

phase 1, 6 unordered pieces.

ACCESSION AC000003
 NID g1556454
 KEYWORDS HTG; HTGS_PHASE1.
 SOURCE human.

COMMENT ***
 *** WARNING: Phase 1 High Throughput Genome Sequence ***

* This sequence is unfinished. It consists of 6 contigs for
 * which the order is not known; their order in this record is
 * arbitrary. In some cases, the exact lengths of the gaps
 * between the contigs are also unknown; these gaps are presented
 * as runs of N as a convenience only. When sequencing is complete,
 * the sequence data presented in this record will be replaced
 * by a single finished sequence with the same accession number.

* 1 22526: contig of 22526 bp in length
 * 22527 23035: gap of unknown length
 * 23036 33919: contig of 10884 bp in length
 * 33920 34427: gap of unknown length
 * 34428 61877: contig of 27450 bp in length
 * 61878 62385: gap of unknown length
 * 62386 65891: contig of 3506 bp in length
 * 65892 66399: gap of unknown length
 * 66400 102207: contig of 35808 bp in length
 * 102208 102715: gap of unknown length
 * 102716 120000: contig of 17285 bp in length.

FEATURES Location/Qualifiers
 source 1..120000
 /organism="Homo sapiens"
 /clone="104_H_12"
 /clone_lib="CITB978SK-B"
 /chromosome="17"

BASE COUNT 31822 a 28286 c 27634 g 29488 t 2770 others
 ORIGIN

l accntccac attacactcg

B

LOCUS AC000003 121910 bp DNA HTG 10-JUN-1997
 DEFINITION *** SEQUENCING IN PROGRESS *** Genomic sequence from Human 17;
 HTGS

phase 2, 2 ordered pieces.

ACCESSION AC000003
 NID g2182283
 KEYWORDS HTG; HTGS_PHASE2.
 SOURCE human.

COMMENT The Staden databases, finishing information, and all
 chromatographic files used in the assembly of this clone are
 available from our anonymous ftp site.

 *** WARNING: Phase 2 High Throughput Genome Sequence ***

* This sequence is unfinished. It consists of 2 contigs for
 * which the order is known. The lengths of the gaps have been
 * estimated by the submitter but are not known exactly. When
 * sequencing is complete, the sequence data presented in this
 * record will be replaced by a single finished sequence
 * with the same accession number.

* 1 56538: contig of 56538 bp in length
 * 56539 56538: gap of unknown length
 * 56539 121910: contig of 65372 bp in length.

FEATURES Location/Qualifiers
 source 1..121910
 /organism="Homo sapiens"
 /clone="104H12"
 /clone_lib="Research Genetics/Cal Tech CITB978SK-B (plates
 1-194)"
 /chromosome="17"

BASE COUNT 34447 a 30318 c 27782 g 29362 t 1 others
 ORIGIN

l aagctctgg atccgtaggt

C

LOCUS AC000003 122228 bp DNA PRI 02-SEP-1997
 DEFINITION Genomic sequence from Human 17, complete sequence.

ACCESSION AC000003
 NID g2204282
 KEYWORDS HTG.
 SOURCE human.

COMMENT The Staden databases, finishing information, and all
 chromatographic files used in the assembly of this clone are
 available from our anonymous ftp site.

All repeats were identified using RepeatMasker: Smit, A.F.A. &
 Green, P. (1996-1997)
<http://ftp.genome.washington.edu/RM/RepeatMasker.html>.

FEATURES Location/Qualifiers
 source 1..122228
 /organism="Homo sapiens"
 /clone="104H12"
 /clone_lib="Research Genetics/Cal Tech CITB978SK-B (plates
 1-194)"
 /chromosome="17"
 repeat_region 261..370
 /rpt_family="MLT1B"
 repeat_region 374..510
 /rpt_family="AluJb"
 repeat_region 570..842
 /rpt_family="MLT1B"
 repeat_region 1028..1320
 /rpt_family="AluJb"
 repeat_region complement(1462..1762)
 /rpt_family="AluY"

[full set of annotations deleted for brevity]

Figure 1 An example of a genomic sequence record (DDBJ/EMBL/GenBank accession number AC000003) as it progresses from an unfinished to a finished state. (These records have been truncated for the printed journal. Full views of these sequence can be retrieved from <http://www.ncbi.nlm.nih.gov/Entrez/nucleotide.html> by entering the corresponding NID numbers (excluding the initial "g") into the query box and specifying "Sequence ID" as the search field. Using the accession number, i.e., AC000003, as the query term will always and only retrieve the latest (finished) version of the record.) (A) Phase 1 records consist of multiple sequences derived from a single genomic clone such as the insert of a cosmid vector or bacterial artificial chromosome (BAC). The entire insert is represented by a single accession number, even though at this stage it consists of multiple sequence fragments, the order and orientation of which are unknown. Such records can be identified in GenBank by the keywords HTG; HTGS_PHASE1 and are found in the HTG Division of GenBank. (B) Phase 2 records consist of ordered sequence fragments with one or more gaps and are identified by the keywords HTG; HTGS_PHASE2. (C) Phase 3 records represent finished data with no gaps and an assumed accuracy of 10–4 errors or less. When records reach this finished state, they are moved to the appropriate organismal division of GenBank, in this case the Primate (PRI) Division. The only distinctions between these records and traditional GenBank records are their size and the keyword, HTG, which indicates their origin as part of a high-throughput sequencing project. Note well that although the accession number remains constant as the genomic sequence progresses through the various stages of completion, a different nucleotide sequence identifier (NID) number is assigned to each phase (e.g. g1556454 → g2182283 → g2204282). In practice, not all laboratories employ these phase definitions and not all records go through all phases. Some records are submitted initially as finished (phase 3); others may come in initially as phase 1 and updated directly to phase 3. Also note that records tend to include more and more annotation as they progress through the process; however, this is not a requirement for finished sequence and the degree of annotation varies considerably depending upon the submitting laboratory.

Table 2. Relationships Between Divisions and Homology Search Files

| Database division | BLAST databases at NCBI | FASTA databases at EBI | Location of "finished" HTG records |
|-------------------|---------------------------|------------------------|---|
| BCT | nr, ^a month | emall, emnew, ebact | |
| PRO | | emall, emnew, epro | |
| FUN | | emall, emnew, efun | |
| HUM | | emall, emnew, ehum | <i>H. sapiens</i> (EMBL) |
| PRI | nr, month | | <i>H. sapiens</i> (GenBank) |
| ROD | nr, month | emall, emnew, erod | |
| MAM | nr, month | emall, emnew, emam | |
| VRT | nr, month | emall, emnew, evrt | |
| INV | nr, month | emall, emnew, einv | <i>C. elegans</i> and <i>D. melanogaster</i> |
| PLN | nr, ^b month | emall, emnew, epln | <i>A. thaliana</i> |
| ORG | | emall, emnew, eorg | |
| VRL | nr, month | emall, emnew, evrl | |
| PHG | nr, month | emall, emnew, ephg | |
| RNA | nr, month | emall, emnew, erna | |
| SYN | nr, month | emall, emnew, esyn | |
| UNA | nr, month | emall, emnew, euna | |
| EST | dbest, ^c month | eest | |
| STS | dbsts, month | ests | |
| GSS | dbgss, month | emall, emnew | |
| HTG | htgs, month | emall, emnew | Includes all "unfinished" HTG |
| PAT | nr, month | emall, emnew, epat | |

(month) A rolling month database consisting of nucleotide or protein sequences added to nr in the last 28 days; (nr) a nonredundant nucleotide (or protein) database of all sequences, excluding ESTs, STSs, GSSs, and HTGs; (emnew) new EMBL entries since latest release; (emall) all EMBL entries, latest release + new (other FASTA database acronyms are derived from the EMBL division to which they correspond).

^aNCBI offers *ecoli* as a separate BLAST database for queries against *Escherichia coli* genome and protein sequences.

^bNCBI offers yeast as a separate BLAST database for queries against the *Saccharomyces cerevisiae* genome and protein sequences.

^cNCBI plans to split dbest into three files of human only, mouse only, and all nonhuman, non-mouse ESTs.

Tolstoshev. 1993. dbEST—Database for expressed sequence tags. *Nature Genet.* 4: 332–333.

Schuler, G.D. 1997. Sequence mapping by electronic PCR. *Genome Res.* 7: 541–550.

Schuler, G.D., J.A. Epstein, H. Ohkawa, and J.A. Kans. 1996. Entrez: Molecular biology database and retrieval system. *Methods Enzymol.* 266: 141–162.

Zhang, J. and T.L. Madden. 1997. PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res.* 7: 649–656.

have somewhat different perspectives and needs. The international database collaborators have striven to meet the various requirements of a diverse community, but new suggestions are always welcomed and may be directed to NCBI's service desk at info@ncbi.nlm.nih.gov. Information resource providers will continue to experiment with new ways in which to make sequence data more accessible and useful to the community, particularly for homology search purposes.

REFERENCES

Adams, M.D. and J.C. Venter. 1996. Should

non-peer-reviewed raw DNA sequence data release be forced on the scientific community? *Science* 274: 534–536.

Altschul, S.F., M.S. Boguski, W. Gish, and J.C. Wootton. 1994. Issues in searching molecular sequence databases. *Nature Genet.* 6: 119–129.

Benson, D.A., M.S. Boguski, D.J. Lipman, and J. Ostell. 1997. GenBank. *Nucleic Acids Res.* 25: 1–6.

Bentley, D.R. 1996. Genomic sequence information should be released immediately and freely in the public domain. *Science* 274: 533–534.

Boguski, M.S., T.M. Lowe, and C.M.