C2 ANNOTATED SEQUENCE DATABASES

Key Notes

Primary sequence databases

The three primary sequence databases are GenBank (NCBI), the Nucleotide Sequence Database (EMBL) and the DNA Databank of Japan (DDBJ). These are repositories for raw sequence data, but each entry is extensively annotated and has a features table to highlight the important properties of each sequence. The three databases exchange data on a daily basis.

Subsidiary sequence databases

Particular types of sequence data are stored in subsidiaries of the main sequence databases. For example, ESTs are stored in dbEST, a division of GenBank. There are also subsidiary databases for GSSs and unfinished genomic sequence data.

Submission of sequences

Sequences may be submitted to any of the three primary databases using the tools provided by the database curators. Such tools include WebIn and BankIt, which can be used over the Internet, and Sequin, a stand-alone application.

SWISS-PROT and TrEMBL

SWISS-PROT is a collection of confirmed protein sequences with annotations relating to structure, function and protein family assignments. The related database TrEMBL is a translation of all coding sequences in the primary nucleic acid databases. The entries in TrEMBL are less extensively annotated than those in SWISS-PROT, but are moved to SWISS-PROT when reliable annotations become available.

Database interrogation

All the databases discussed in this topic can be searched by sequence similarity. However, detailed text-based searches of the annotations are also possible using tools such as Entrez. The simplest way to cross-reference between the primary nucleotide sequence databases and SWISS-PROT is to search by accession number, as this provides an unambiguous identifier of genes and their products.

Related topics

- Useful bioinformatics sites on the WWW (A3)
- Sequencing DNA, RNA and proteins (B1)
- File formats (C1)
- Genome and organism-specific databases (C3)
- Miscellaneous databases (C4)
- Data retrieval with Entrez and DBGET/LinkDB (D1)
- Data retrieval with SRS (Sequence Retrieval System) (D2)
- Database searches: FASTA and BLAST (E3)

Primary sequence databases

The **primary sequence databases** are repositories for raw sequence data, and can be accessed freely over the World Wide Web (WWW). There are three such

databases, comprising the **International Nucleotide Sequence Database Collaboration**. These are **GenBank**, maintained by the **National Center for Biotechnology Information (NCBI)**, the **Nucleotide Sequence Database** maintained by the **European Molecular Biology Laboratory (EMBL)**, and the **DNA Databank of Japan (DDBJ)**. New sequences can be deposited in any of the databases since they exchange data on a daily basis.

The databases contain not only sequences but also extensive annotations. As an example, *Fig. 1* shows part of a GenBank file, in this case for the human gene *BTEB*. Much of the introductory part of the file is self-explanatory, containing information such as the locus name, the accession number, the source species, literature references and the date of submission. An important section of the file is the **features table**, which describes interesting features of the sequence. Since

```
LOCUS       HUMBTEB      4859 bp    mRNA            PRI       07-FEB-1999
DEFINITION  Human mRNA for GC box binding protein, complete cds.
ACCESSION   D31716
VERSION     D31716.1  GI:505081
KEYWORDS    GC box binding protein; zinc finger.
SOURCE      Homo sapiens germline cDNA to mRNA, clone_lib:placenta.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1
            (.....)
REFERENCE   2  (bases 1 to 4859)
  AUTHORS   Ohe,N., Yamasaki,Y., Sogawa,K., Inazawa,J., Ariyama,T., Oshimura,M.
            and Fujii-Kuriyama,Y.
  TITLE     Chromosomal localization and cDNA sequence of human BTEB, a GC box
            binding protein
  JOURNAL   Somat. Cell Mol. Genet. 19 (5), 499-503 (1993)
  MEDLINE   94120483
COMMENT     Submitted (31-May-1994) to DDBJ by:
            Yoshiaki Fujii-Kuriyama
            (.....)
FEATURES             Location/Qualifiers
     source          1..4859
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /clone_lib="placenta"
                     /germline
     gene            1265..1999
                     /gene="BTEB"
     CDS             1265..1999
                     /gene="BTEB"
                     /note="three-times repeated zinc finger motif"
                     /codon_start=1
                     /product="GC box binding protein"
                     /protein_id="BAA06524.1"
                     /db_xref="GI:1060891"
                      translation="MSAAAYMDFVAAQCLVSISNRAAVPEHGVAPDAERLRLPEREVT
                     KEHGDPGDTWKDYCTLVTIAKSLLDLNKYRPIQTPSVCSDSLESPDEDMGSDSDVTTE
                     SGSSPSHSPEERQDPGSAPSPLSLLHPGVAAKGKHASEKRHKCPYSGCGKVYGKSSHL
                     KAHYRVHTGERPFPCTWPDCLKKFSRSDELTRHYRTHTGEKQFRCPLCEKRFMRSDHL
                     TKHARRHTEFHPSMIKRSKKALANAL"
BASE COUNT     1285 a   1111 c   1193 g   1270 t
ORIGIN      Chromosome 9, q13.
        1 cacgttgggt gacataatgg ggtttttta attatagatt cacactgcat ttattcatca
      (.....)
     4801 ttcaccattg tggaatgatg ccctggcttt aaggtttagc tccacatcat gcttctctt
//
```

Fig. 1. GenBank entry for the human gene BTEB. Some information has been deleted from the file for the sake of brevity and is indicated thus (.....)

GenBank is a nucleic acid repository, the fact that there is a protein-coding region is a feature in the entry. Note that *BTEB* is a very simple gene that has no introns. If there were introns, the CDS (coding sequence) feature would be more complicated: the entry would be extended to indicate the base positions of the exons, delimited by commas. For example, if there was a second exon encoding a further 20 amino acids residues, the CDS feature would read as follows: 1265..1999,2100..2159.

Subsidiary sequence databases

The main sequence databases have a number of subsidiaries for the storage of particular types of sequence data. For example, **dbEST** is a division of GenBank which is used to store **expressed sequence tags (ESTs)**, and an example entry in dbEST is shown in *Fig. 2*. Other divisions of GenBank include **dbGSS**, which is used to store single-pass **genomic survey sequences (GSSs)**, **dbSTS**, which is used to store **sequence tagged sites (STSs**; unique genomic sequences that can be used as physical markers) and the **HTG (high-throughput genomic) division**, which is used to store unfinished genomic sequence data. These types of sequences are discussed in more detail in Topic B1.

Submission of sequences

The robustness of data submitted to the primary sequence databases is important in the context of bioinformatics software. Clearly, the integrity of the scientists who submit the data is not readily checked by computers but errors must be avoided in database consistency. It is essential that the data are submitted in a supported format and that the submission is carried out by means of software provided by the database curators. Examples are **WebIn** provided by EMBL (www.ebi.ac.uk/embl/Submission) and **BankIt** provided by the NCBI (http://www.ncbi.nlm.nih.gov/BankIt/), each of which can be used to submit sequences to the databases over the WWW. A powerful stand-alone software tool, **Sequin**, is provided by the NCBI and can be used on UNIX, PC/Windows and Macintosh systems for sequence submission for those with no WWW access (http://www.ncbi.nlm.nih.gov/Sequin/index.html).

SWISS-PROT and TrEMBL

SWISS-PROT and the related database **TrEMBL** (Translated EMBL) are repositories for annotated protein sequences. *Figure 3* shows the SWISS-PROT entry for the BTEB protein, corresponding to the GenBank entry in *Fig. 1*. The entry contains large numbers of annotations, including a **features table** before the sequence. Each line begins with two letters, many of which are self-explanatory, for example ID (identity), AC (accession number), DT (date), DE (description), GN (gene name), CC (comment). Continuation lines are indicated by the symbols -!- at the start of a section and indents thereafter (this is shown in the CC field in *Fig. 3*). Characteristic features of SWISS-PROT entries include the DR (reference), KW (key words) and FT (features) fields. It is the presence of these careful and extensive annotations that makes SWISS-PROT so popular with biochemists. For example, in *Fig. 3*, there is a fairly comprehensive description of the protein and its function but also (in the DT field) cross-references to the relevant entries in the secondary databases PROSITE, PRINTS and Pfam (Topic F2).

SWISS-PROT provides the most up-to-date and extensively annotated information on protein sequences and its quality reflects its active management by human curators. TrEMBL (translated EMBL) is another database in the same format. The entries in TrEMBL are derived from translation of all coding sequences in the EMBL Nucleotide Sequence Database that are not already in

```
LOCUS       T48601          355 bp    mRNA            EST       06-FEB-1995
DEFINITION  yb01a01.s1 Stratagene placenta (#937225) Homo sapiens cDNA clone
            IMAGE:69864 3' similar to similar to gb:S71043_rna1 IG ALPHA-2
            CHAIN C REGION (HUMAN), mRNA sequence.
ACCESSION   T48601
VERSION     T48601.1  GI:650461
KEYWORDS    EST.
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 355)
  AUTHORS   {....}
TITLE     Generation and analysis of 280,000 human expressed sequence tags
JOURNAL   Genome Res. 6 (9), 807-828 (1996)
MEDLINE   97044478
COMMENT     Other_ESTs: yb01a01.r1
            Contact: Wilson RK
            Washington University School of Medicine
            4444 Forest Park Parkway, Box 8501, St. Louis, MO 63108
            Tel: 314 286 1800
            Fax: 314 286 1810
            Email: est@watson.wustl.edu
            High qality sequence stops: 277
            Source: IMAGE Consortium, LLNL
            This clone is available royalty-free through LLNL ; contact the
           IMAGE Consortium (info@image.llnl.gov) for further information.
           Seq primer: -21m13
            High quality sequence stop: 277.
FEATURES             Location/Qualifiers
     source          1..355
                     /organism="Homo sapiens"
                     /db_xref="GDB:490761"
                     /db_xref="taxon:9606"
                     /clone="IMAGE:69864"
                     /sex="male"
                     /clone_lib="Stratagene placenta (#937225)"
                     /lab_host="SOLR cells (kanamycin resistant)"
                     /note="Organ: placenta; Vector: pBluescript SK-; Site_1:
                     EcoRI; Site_2: XhoI; Cloned unidirectionally.  Primer:
                     Oligo dT. Caucasian. Average insert size: 1.2 kb; Uni-ZAP
                     XR Vector; ~5' adaptor sequence: 5' GAATTCGGCACGAG 3' ~3'
                     adaptor sequence: 5' CTCGAGTTTTTTTTTTTTTTTT 3'"
BASE COUNT       62 a    117 c     98 g     69 t      9 others
ORIGIN
        1 ggcggctcag tagcaggtgc cgtccacctc cgccatgaca acagacacat tgacatgggt
       61 gggtttacca ccaagcgtcc gatggtcttc tgtgtgaagg ccagccaggc gcctccatgg
      121 caccatgcag gagaaggnct ccccttctct ccagtcctcg gctgccacgc gcagtatgct
      181 ggtcacacga aggtcgtggt gccctggctg gntcctncan ggatgcccaa gtcaggtact
      241 tntcgcgggg cagctcctgt gacccctgca gccagcgaac cagcacgtcc ttggggcttn
      301 aagcngcgct accaggcact tcaaccgttc nccagcttcg ttcagggcca ncttc
//
```

Fig. 2. GenBank (dbEST) entry for a human EST clone. Some information has been deleted from the file for the sake of brevity and is indicated thus {.....}

SWISS-PROT. As further data ensure the reliability of annotations, TrEMBL entries are moved to SWISS-PROT.

Database interrogation

Detailed queries of the text annotation in the databases discussed above can be carried out using tools like SRS and Entrez (Section D). However, a comparison of *Figs 1* and 3 shows how a user can cross-reference between these databases. Let us assume, for example, that searching GenBank with a new sequence

```
ID   BTE1_HUMAN      STANDARD;      PRT;   244 AA.
AC   Q13886; Q16196;
DT   15-DEC-1998 (Rel. 37, Created)
DT   15-DEC-1998 (Rel. 37, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   TRANSCRIPTION FACTOR BTEB1 (BASIC TRANSCRIPTION ELEMENT BINDING
DE   PROTEIN 1) (GC BOX BINDING PROTEIN 1) (KRUEPPEL-LIKE FACTOR 9).
GN   BTEB1 OR BTEB OR KLF9.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=94120483; PubMed=8291025;
RA   Ohe N., Yamasaki Y., Sogawa K., Inazawa J., Ariyama T., Oshimura M.,
RA   Fujii-Kuriyama Y.;
RT   "Chromosomal localization and cDNA sequence of human BTEB, a GC box
RT   binding protein.";
RL   Somat. Cell Mol. Genet. 19:499-503(1993).
RN   [2]
RP   SEQUENCE OF 1-31 FROM N.A.
RX   MEDLINE=94327649; PubMed=8051167;
RA   Imataka H., Nakayama K., Yasumoto K., Mizuno A., Fujii-Kuriyama Y.,
RA   Hayami M.;
RT   "Cell-specific translational control of transcription factor BTEB
RT   expression. The role of an upstream AUG in the 5'-untranslated
RT   region.";
RL   J. Biol. Chem. 269:20668-20673(1994).
CC   -!- FUNCTION: TRANSCRIPTION FACTOR THAT BINDS TO GC BOX PROMOTER
CC       ELEMENTS. SELECTIVELY ACTIVATES MRNA SYNTHESIS FROM GENES
CC       CONTAINING TANDEM REPEATS OF GC BOXES BUT REPRESSES GENES WITH
CC       A SINGLE GC BOX.
CC   -!- SUBCELLULAR LOCATION: NUCLEAR.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; D31716; BAA06524.1; -.
DR   EMBL; S72504; AAD14110.1; -.
DR   MIM; 602902; -.
DR   InterPro; IPR000822; Znf-C2H2.
DR   Pfam; PF00096; zf-C2H2; 3.
DR   PRINTS; PR00048; ZINCFINGER.
DR   SMART; SM00355; ZnF_C2H2; 3.
DR   PROSITE; PS00028; ZINC_FINGER_C2H2_1; 3.
DR   PROSITE; PS50157; ZINC_FINGER_C2H2_2; 3.
KW   Transcription regulation; DNA-binding; Nuclear protein; Repeat;
KW   Zinc-finger; Metal-binding.
FT   DOMAIN       84     116       ASP/GLU-RICH (ACIDIC).
FT   DOMAIN      143     225       ZINC FINGERS.
FT   ZN_FING     143     167       C2H2-TYPE.
FT   ZN_FING     173     197       C2H2-TYPE.
FT   ZN_FING     203     225       C2H2-TYPE.
SQ   SEQUENCE   244 AA;  27234 MW;  2D1B5A5BB9D42221 CRC64;
     MSAAAYMDFV AAQCLVSISN RAAVPEHGVA PDAERLRLPE REVTKEHGDP GDTWKDYCTL
     VTIAKSLLDL NKYRPIQTPS VCSDSLESPD EDMGSDSDVT TESGSSPSHS PEERQDPGSA
     PSPLSLLHPG VAAKGKHASE KRHKCPYSGC GKVYGKSSHL KAHYRVHTGE RPFPCTWPDC
     LKKFSRSDEL TRHYRTHTGE KQFRCPLCEK RFMRSDHLTK HARRHTEFHP SMIKRSKKAL
     ANAL
//
```

Fig. 3. SWISS-PROT entry for the human protein BTEB, equivalent to the GenBank entry shown in Fig. 1. Note the DR field provides the EMBL accession number, allowing database entries to be cross-referenced.

obtained in the laboratory identifies the gene in *Fig. 1* as particularly interesting for a research project. How do we find the corresponding entry in SWISS-PROT? Note that the IDs are different, and, although SWISS-PROT has alternative GN entries (gene names), they do not correspond to the name on the GenBank file. The way to find the correct SWISS-PROT file is to search the SWISS-PROT database for the accession number D31716. Although SWISS-PROT has its own accession number, D31716 can be found as a DR field entry. The SWISS-PROT site or one of its mirrors will successfully locate the entry with D31716 as the search string.