Biological Databases

P.M. Selzer et al., *Applied Bioinformatics*, https://doi.org/10.1007/978-3-319-68301-0_2

2.1 Biological Knowledge is Stored in Global Databases

The most important basis for applied bioinformatics is the collection of sequence data and its associated biological information. For example, with genome sequencing projects such data are generated daily in very large quantities worldwide. In order to use these data appropriately, a structured filing system of the data is necessary, yet the data should also be accessible to those interested. Annually, the journal *Nucleic Acids Research* [nar] dedicates an entire issue (first issue in January) to all available biological databases that are recorded in tabular form with the respective URLs. Furthermore, for a number of databases, original articles describe their functions. This database issue, which is freely accessible also on the Web, is a good starting point for working with biological databases. Depending on the kind of data included, different categories of biological databases can be distinguished. Primary databases contain primary sequence information (nucleotide or protein) and accompanying annotation information regarding function, bibliographies, cross references to other databases, and so forth. Secondary biological databases, however, summarize the results from analyses of primary protein sequence databases. The aim of these analyses is to derive common features for sequence classes, which in turn can be used for the classification of unknown sequences (annotation). In addition, all other databases that save biological or medical information, for example, literature databases, are frequently classified as secondary databases.

The use of relational database systems (e.g., Oracle, MS Access, Informax, DB2) and their ability to manage large data sets would seem to make them ideal for the structured filing of data, yet these systems have not gained acceptance so far in the field of biological databases. Rather, sequence data and their accompanying information are usually filed in the form of flat file databases, that is, structured ASCII text files. This is for historical reasons and because ASCII text files offer the advantage of conferring the ability to manipulate data without requiring an expensive and complicated database system. ASCII text files also make data exchange between scientists relatively simple. One drawback, however, is that searching for certain keywords within a data set is both laborious and time-consuming. To minimize this disadvantage, various systems have been developed that can index flat file–based databases, that is, they come with an index register similar to that of a book, thus accelerating keyword-based searches.

2.2 Primary Databases

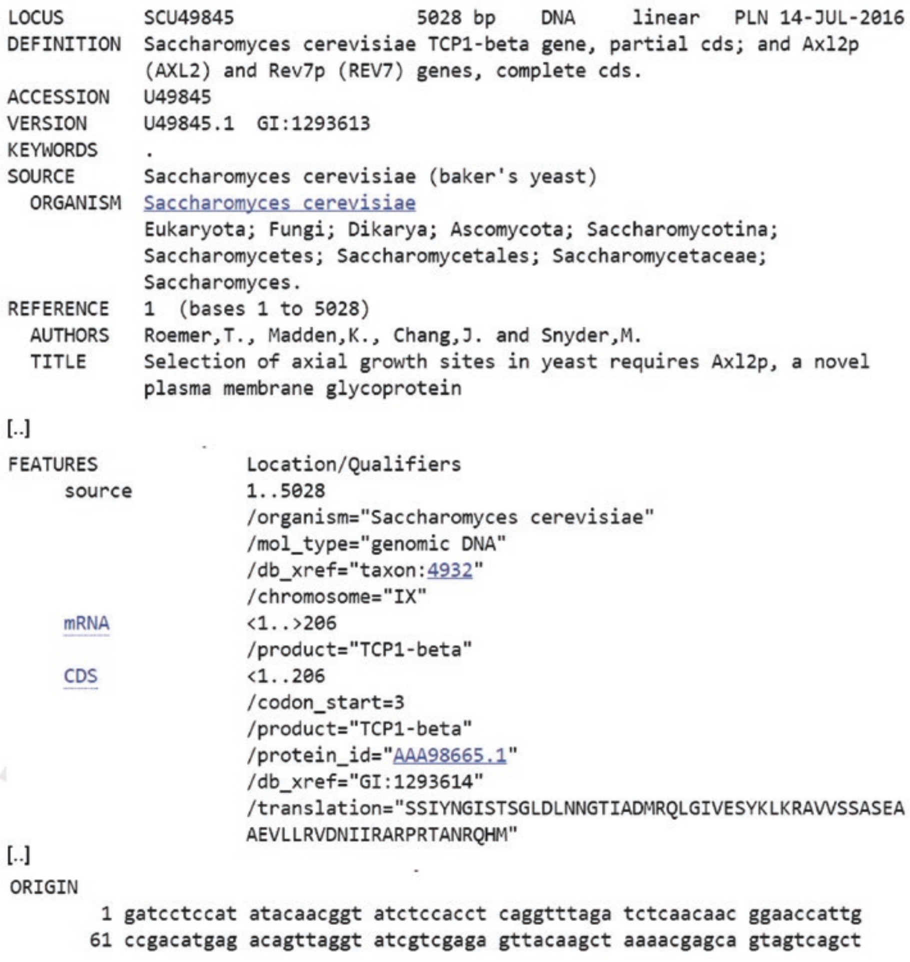
2.2.1 Nucleotide Sequence Databases

2.2.1.1 GenBank

The GenBank database [genbank] is perhaps the best-known nucleotide sequence database available at the U.S. National Center for Biotechnology Information (NCBI) [ncbi]. GenBank is a public sequence database, which in its present version (217.00, December 2016) contains roughly 199 million sequence entries. Sequences can be entered into GenBank by anyone via a Web page [bankit] or by e-mail [sequin] when working with larger sequence sets. Prior entry of sequence data into either GenBank or one of its associated databases, for example the European Nucleotide Archive (ENA) or the DNA

Database of Japan (DDBJ), is a prerequisite for the publication of new sequences in any scientific journal. Each single database entry is provided with a unique identification tag, the accession number (AN). The AN is a permanent record that remains unchanged even if changes are subsequently made to the database record. In some cases, a new AN can be assigned to an existing number if, for example, an author adds a new database record that combines existing sequences. Even then the old AN is retained as a secondary number. The AN is the only way to absolutely verify the identity of a sequence or database entry.

Figure 2.1 shows a GenBank entry. The entry has been shortened at some points and these are indicated by [...]. The required structuring of the database record is performed via defined keywords. Each entry starts with the keyword LOCUS followed by a locus name. Like the AN, the locus name is also unique; however, unlike the AN, it may change after revisions of the database. The locus name consists of eight characters,



```
LOCUS       SCU49845                 5028 bp    DNA     linear   PLN 14-JUL-2016
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
[..]
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
     mRNA            <1..>206
                     /product="TCP1-beta"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
[..]
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
```

Fig. 2.1 Database record of GenBank database. The entry was shortened at some points, as indicated by [...]

including the first letter of the genus and species names, in addition to a six-digit AN. Newer entries have an eight-digit AN. In such cases, the locus name is identical to the AN. On the same line following the locus name, the length of the sequence is given. A sequence must have at least 50 base pairs to be entered into GenBank. This requirement was introduced only relatively recently, and therefore, some older entries do not fulfill this criterion. Column 3 denotes the type of molecule of the sequence entry. Every GenBank entry must contain coherent sequence information of a single molecule type, that is, an entry cannot contain sequence information of both genomic DNA and RNA. The last column in the LOCUS line gives the date of the last entry modification. The end of the database record starts with the keyword ORIGIN. In newer entries, this field remains empty. The actual sequence information begins on the following line and may contain many lines. A detailed description of all keywords is found on the GenBank sample page [gb-sample].

■ Entrez

Query of the GenBank database is carried out via the NCBI Entrez system [entrez], which is used to query all NCBI-associated databases (NCBI Resource Coordinators 2016). Because search terms can be combined by means of logical operators (AND, OR, NOT) and single search terms restricted to certain database fields, Entrez is an important and effective tool for the execution of both simple and complicated searches. The restriction of search terms to single database fields is generally performed by a field ID placed after the term: `search term[field-id]`. For example, the search for a sequence from *Saccharomyces cerevisiae* with a length of between 3260 and 3270 base pairs would require the following search syntax: (Saccharomyces cerevisiae`[ORGN]) AND 3260:3270[SLEN]`. Representative field IDs for performing searches in GenBank are listed in ▣ Table 2.1. Complete instructions for the use of Entrez are found on the Entrez help page [entrez-help]. To simplify the construction of complex queries, the *advanced search* was introduced. To use this search, follow the link beneath the Entrez search field. Field IDs and logical operators can be selected from list boxes and the respective query is constructed automatically and entered into the search text field. For better readability in this case, the field IDs are entered with their full name. The latter does also work in the generic search; it is therefore no longer necessary to remember the abbreviated field IDs.

▣ **Table 2.1** Field IDs to restrict search terms to certain database fields in the Entrez system

Field ID	Database field
ACC	Accession number
AU	Author name
DP	Publication date
GENES	Gene name
ORGN	Scientific and common name of the organism
PT	Publication type, e.g., review, letter, technical publication
TA	Journal name, official abbreviation, or ISSN number

EMBL and DDBJ

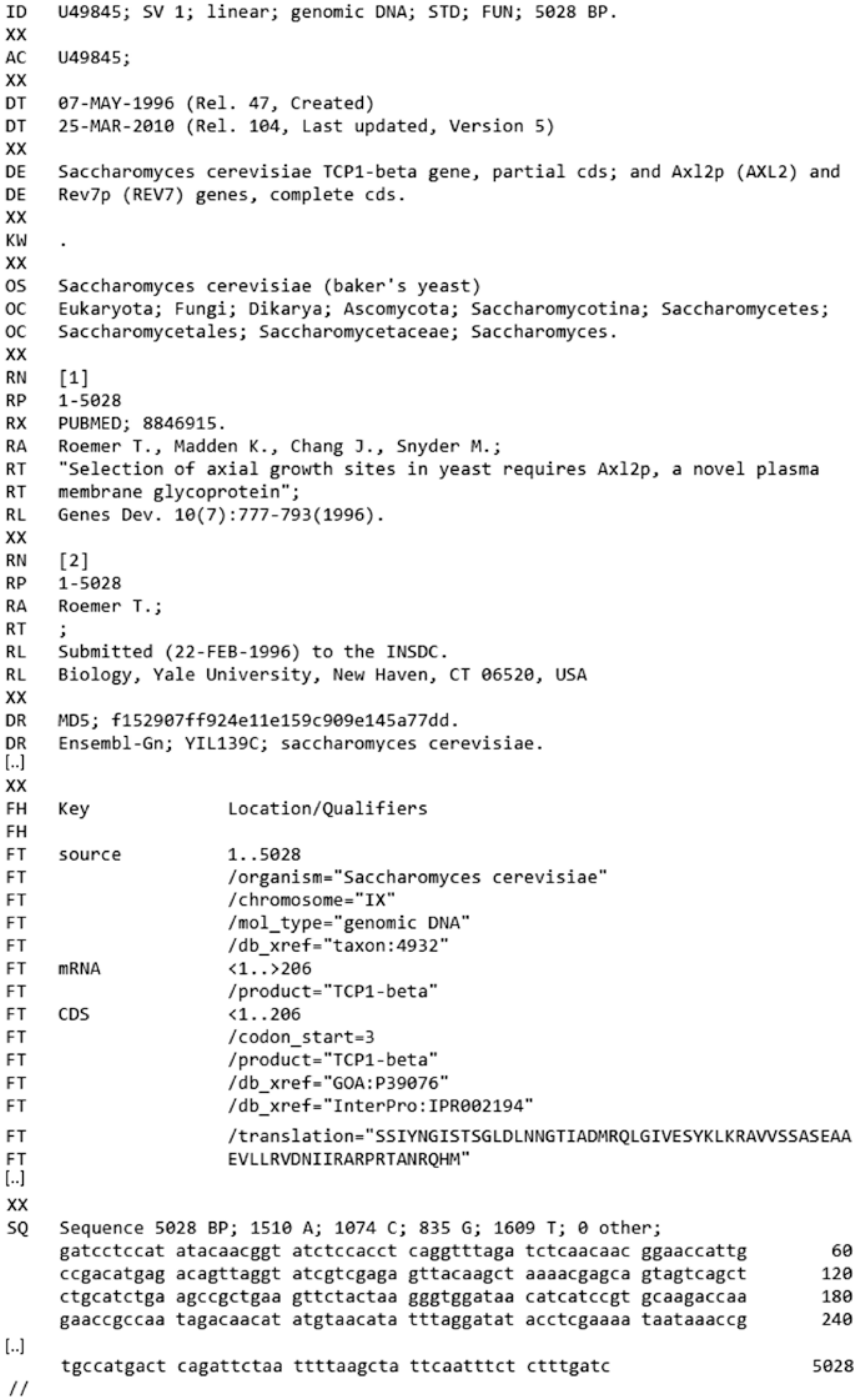
The European counterpart to GenBank is the ENA [ena], located at the European Bioinformatics Institute (EBI) [ebi]. Another primary nucleotide sequence database, the DDBJ [ddbj], is operated by the National Institute of Genetics (NIG) [nig] in Japan and is the primary nucleotide sequence database for Asia. The three database operators, NCBI, EBI, and NIG, compose the International Nucleotide Sequence Database Collaboration and synchronize their databases every 24 h. A query of all three individual databases is therefore not necessary, nor is it required to enter a new nucleotide sequence into all three databases.

While the database format of the DDBJ is identical to that of the NCBI, that of the ENA differs somewhat. ◘ Figure 2.2 shows an entry in the EMBL database. The most obvious difference is the use of two-letter codes instead of full keywords. Furthermore, there are small changes in the organization of the individual data fields. For example, the date of the last modification is not listed in the field ID (corresponding to the LOCUS field in GenBank) but appears in the field DT (database field). A complete description of the EMBL format can be found on the ENA manual page [ebi-manual].

ENA Online Retrieval

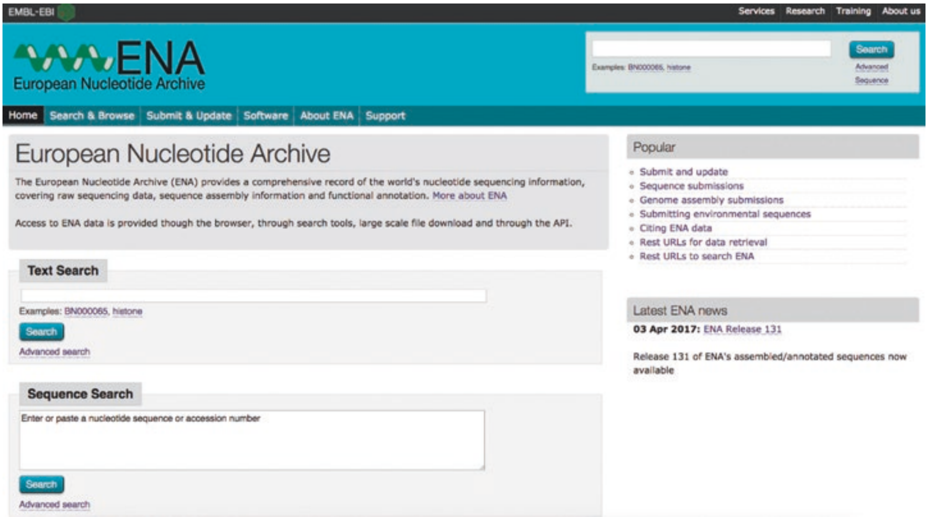
The ENA offers several search forms. First is a simple search, which allows for text searches as well as for sequence retrieval (◘ Fig. 2.3). For text search, it is possible to search for accession numbers and for simple free text. The search is not limited to certain database fields and does not allow to restrict the search to certain text fields as the Entrez system does. Instead, all database entries that randomly contain the search term are retrieved. To use this kind of parameter, to search for a sequence from *S. cerevisiae* with a sequence length of 3270 base pairs for instance, the advanced search must be used. It can be reached by following the corresponding link beneath the simple search text field.

The advanced search form (◘ Fig. 2.4) starts with several rather coarse-grained categories of the database fields. Once one of these categories is selected, additional text fields and option boxes are displayed that make it possible to restrict the search to individual database fields or groups thereof. To retrieve our aforementioned *S. cerevisiae* sequence, we must select the category *Sequence* and enter the search term `Saccharomyces cerevisiae` into the field *Taxon*. The comparison operator is set to equal. Use of the other two operators does, of course, make sense only if we compare numerical values. In the field *Base count*, `3270` is entered and the comparison operator is set to `less than or equal to (<=)`. While entered, all entries are translated into a query simultaneously, which is displayed in the gray text field at the head of the page. The retrieval is started by hitting the *Search* button. Unfortunately, this search form does not allow one to search for a range like we did in the NCBI Entrez example for the sequence length. However, it is possible to build the query in the query builder without a range first and then edit the resulting query manually. To do so, we click on the hyperlink *Edit Query* on the right of the text search field. Now we can modify the preconstructed query and add an additional restriction for the field ID *base_count* with a logical *AND*. The resulting query now is `tax_eq(4932) AND (base_count > = 3260 AND base_count <= 3270)`. Sometimes it is necessary to use brackets to influence the precedence of the logical operators. Here this would not have been necessary; however, we used the brackets for readability reasons. If we had been interested in a S. cerevisiae sequence that is either shorter than 3260 base pairs or longer than 3270 base pairs, we



```
ID   U49845; SV 1; linear; genomic DNA; STD; FUN; 5028 BP.
XX
AC   U49845;
XX
DT   07-MAY-1996 (Rel. 47, Created)
DT   25-MAR-2010 (Rel. 104, Last updated, Version 5)
XX
DE   Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p (AXL2) and
DE   Rev7p (REV7) genes, complete cds.
XX
KW   .
XX
OS   Saccharomyces cerevisiae (baker's yeast)
OC   Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
XX
RN   [1]
RP   1-5028
RX   PUBMED; 8846915.
RA   Roemer T., Madden K., Chang J., Snyder M.;
RT   "Selection of axial growth sites in yeast requires Axl2p, a novel plasma
RT   membrane glycoprotein";
RL   Genes Dev. 10(7):777-793(1996).
XX
RN   [2]
RP   1-5028
RA   Roemer T.;
RT   ;
RL   Submitted (22-FEB-1996) to the INSDC.
RL   Biology, Yale University, New Haven, CT 06520, USA
XX
DR   MD5; f152907ff924e11e159c909e145a77dd.
DR   Ensembl-Gn; YIL139C; saccharomyces cerevisiae.
[..]
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..5028
FT                   /organism="Saccharomyces cerevisiae"
FT                   /chromosome="IX"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:4932"
FT   mRNA            <1..>206
FT                   /product="TCP1-beta"
FT   CDS             <1..206
FT                   /codon_start=3
FT                   /product="TCP1-beta"
FT                   /db_xref="GOA:P39076"
FT                   /db_xref="InterPro:IPR002194"
FT                   /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEAA
FT                   EVLLRVDNIIRARPRTANRQHM"
[..]
XX
SQ   Sequence 5028 BP; 1510 A; 1074 C; 835 G; 1609 T; 0 other;
     gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg        60
     ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct       120
     ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa       180
     gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg       240
[..]
     tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc                   5028
//
```

Fig. 2.2 Database record of EMBL database. The entry has been shortened at some points as indicated by [...]



European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA

Access to ENA data is provided though the browser, through search tools, large scale file download and through the API.

Text Search

Examples: BN000065, histone

Sequence Search

Enter or paste a nucleotide sequence or accession number

Popular

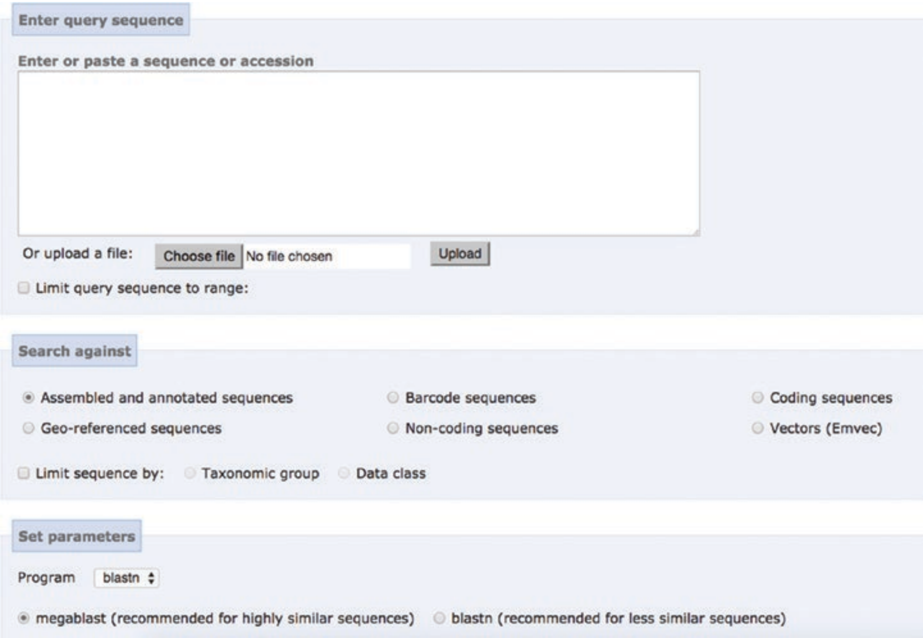
- Submit and update
- Sequence submissions
- Genome assembly submissions
- Submitting environmental sequences
- Citing ENA data
- Rest URLs for data retrieval
- Rest URLs to search ENA

Latest ENA news

03 Apr 2017: ENA Release 131

Release 131 of ENA's assembled/annotated sequences now available

Fig. 2.3 Home page of ENA with simple search fields for text and sequence retrieval (Courtesy EMBL-EBI)


Enter query sequence

Enter or paste a sequence or accession

Or upload a file: Choose file No file chosen Upload

□ Limit query sequence to range:

Search against

● Assembled and annotated sequences ○ Barcode sequences ○ Coding sequences
○ Geo-referenced sequences ○ Non-coding sequences ○ Vectors (Emvec)

□ Limit sequence by: ○ Taxonomic group ○ Data class

Set parameters

Program: blastn

● megablast (recommended for highly similar sequences) ○ blastn (recommended for less similar sequences)

Fig. 2.4 Advanced ENA search form (Courtesy EMBL-EBI)

would have had to use brackets to override the logical operator precedence. The query would have resulted in `tax_eq(4932) AND (base_count <= 3260 OR base_count >= 3270)`.

In addition to a text search, the ENA also allows for sequence searches using sequence comparisons. Basically, this is a BLAST search, which can either be carried out using standard BLAST parameters or which makes it possible to tweak BLAST parameters on the advanced search page. BLAST searches will be discussed in detail in the following chapter, so we will not cover this in more detail here.

2.2.2 Protein Sequence Databases

2.2.2.1 UniProt

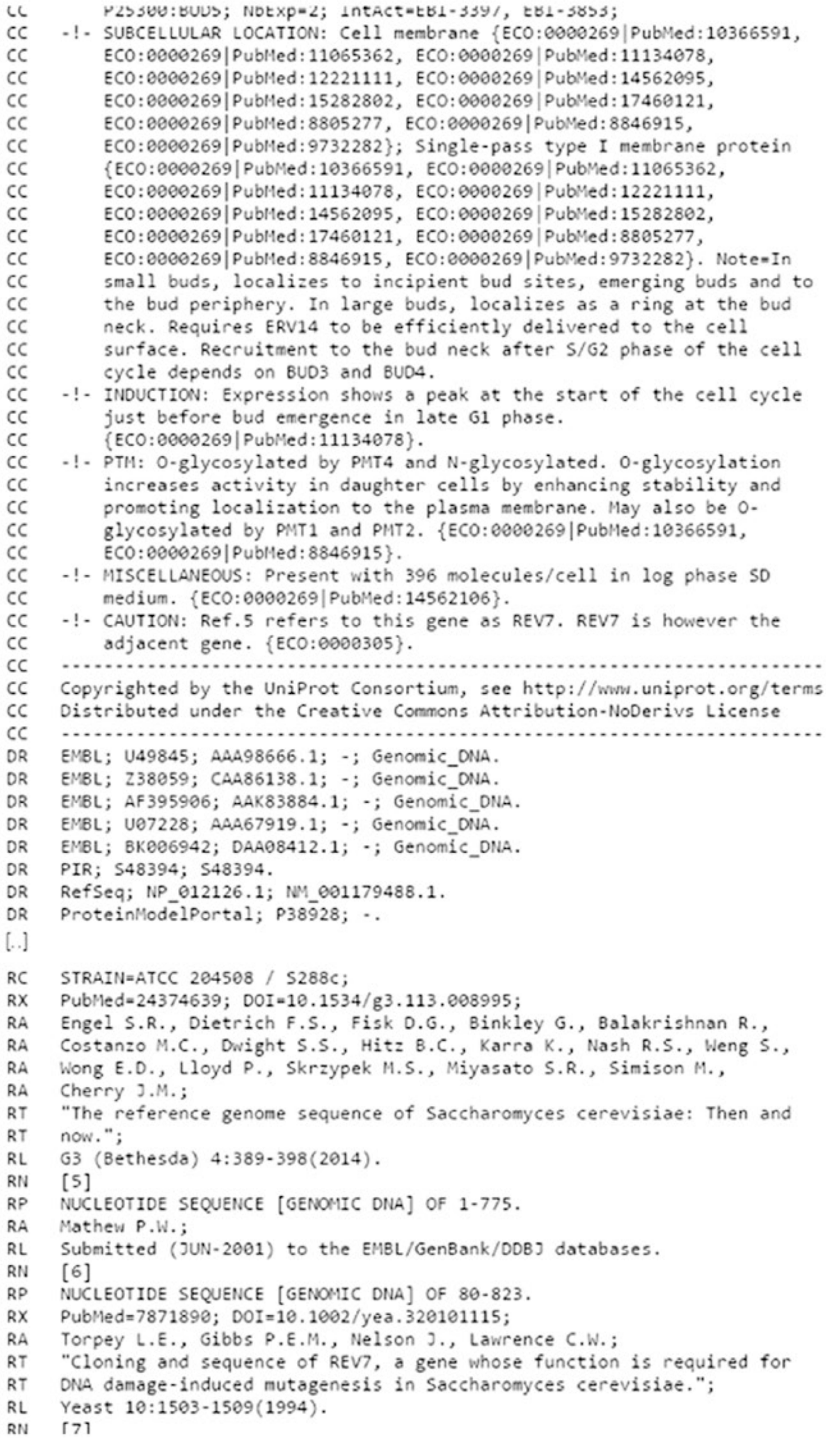
The information available for proteins continues to grow rapidly. Besides sequence information, expression profiles can be examined, secondary structures predicted, and biological/biochemical function(s) analyzed. All these data are stored in databases, some of which are quite specialized. Therefore, it can be time consuming to collect all the relevant information regarding any given protein. For this reason, EBI, the Swiss Institute of Bioinformatics (SIB), and Georgetown University have built a consortium with the aim of developing a central catalog for protein information. The result is the Universal Protein Resource (UniProt) [uniprot] (UniProt Consortium 2016), which unites the information in the three protein databases Swissprot, TrEMBL, and Protein Information Resource (PIR). UniProt consists of three parts, the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters Database (UniRef), and the UniProt Archive (UniPArc), a collection of protein sequences and their history.

Protein sequences and their annotations are stored in the UniProt Knowledgebase (UniProtKB), which is divided into two realms. First is the UniProtKB/TrEMBL realm, which contains automatically annotated sequences, and there is the UniProtKN/SwissProt realm, where manually curated and annotated sequences are stored. UniProtKB/TrEMBL currently (June 2016) contains approx. 65 million entries and is thus around 120 times larger than the realm UniProtKB/SwissProt, which contains approx. 550,000 entries. Because of the manual curation, the UniprotKB/SwissProt realm is regarded as one of the most important protein databases. Quite often, it is also referred to as the gold standard of protein annotation.

The SwissProt database existed long before the UniProt database was founded and was located at the SIB. Because the team of specialists at the SIB was overwhelmed with the flood of new sequences being entered into the databases, a supplement to the SwissProt database, the TrEMBL database, was introduced. TrEMBL stands for translated EMBL and contained all protein translations of the EMBL database, which had not yet been manually curated. The EMBL database is the predecessor of the ENA. All entries in TrEMBL (today UniProtKB/TrEMBL) are annotated automatically, that is, the quality of the annotations is not comparable to that of UniProtKB/SwissProt annotations.

◘ Figure 2.5 shows an entry in the UniProtKB/SwissProt database. At first glance the entry is similar to an ENA entry. Indeed, the two database formats are related. Both database schemes use two-letter identifiers, and most identifiers are identical for the two databases. Some identifiers, however, are modified for the UniProtKB and some are added. The raw database entry as shown in ◘ Fig. 2.5 is rarely found. Most times, a graphical version is presented by UniProtKB, as shown in ◘ Fig. 2.6.

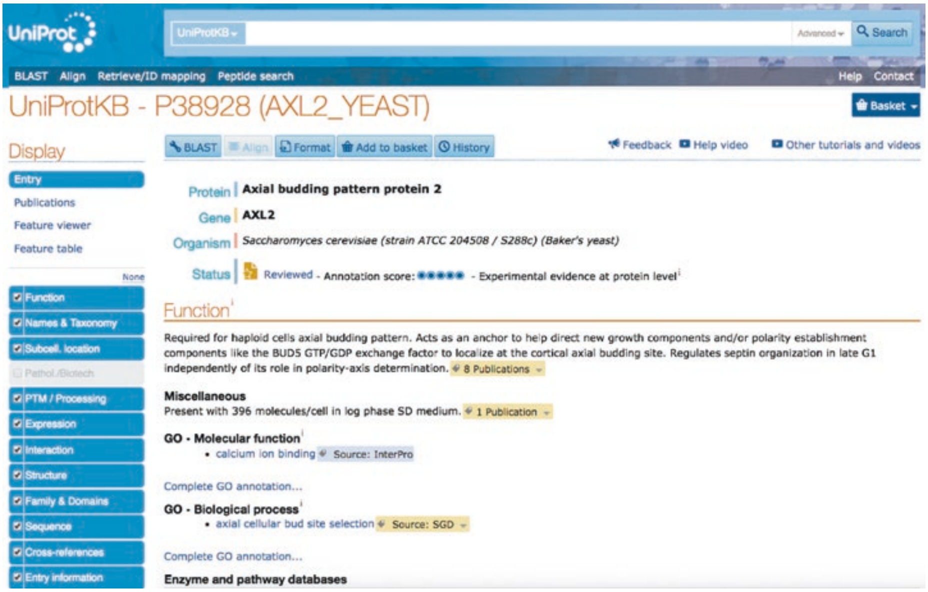
The UniProtKB can be queried using simple full text search or using complex queries with logical operators (◘ Fig. 2.7). For a simple full text search, the search term can simply be entered in the text field at the top of the page. For complex searches, an advanced search form is used. The search is initiated by clicking on the hyperlink



```
CC         P25300:BUD5; NbExp=2; IntAct=EBI-3397, EBI-3853;
CC     -!- SUBCELLULAR LOCATION: Cell membrane {ECO:0000269|PubMed:10366591,
CC         ECO:0000269|PubMed:11065362, ECO:0000269|PubMed:11134078,
CC         ECO:0000269|PubMed:12221111, ECO:0000269|PubMed:14562095,
CC         ECO:0000269|PubMed:15282802, ECO:0000269|PubMed:17460121,
CC         ECO:0000269|PubMed:8805277, ECO:0000269|PubMed:8846915,
CC         ECO:0000269|PubMed:9732282}; Single-pass type I membrane protein
CC         {ECO:0000269|PubMed:10366591, ECO:0000269|PubMed:11065362,
CC         ECO:0000269|PubMed:11134078, ECO:0000269|PubMed:12221111,
CC         ECO:0000269|PubMed:14562095, ECO:0000269|PubMed:15282802,
CC         ECO:0000269|PubMed:17460121, ECO:0000269|PubMed:8805277,
CC         ECO:0000269|PubMed:8846915, ECO:0000269|PubMed:9732282}. Note=In
CC         small buds, localizes to incipient bud sites, emerging buds and to
CC         the bud periphery. In large buds, localizes as a ring at the bud
CC         neck. Requires ERV14 to be efficiently delivered to the cell
CC         surface. Recruitment to the bud neck after S/G2 phase of the cell
CC         cycle depends on BUD3 and BUD4.
CC     -!- INDUCTION: Expression shows a peak at the start of the cell cycle
CC         just before bud emergence in late G1 phase.
CC         {ECO:0000269|PubMed:11134078}.
CC     -!- PTM: O-glycosylated by PMT4 and N-glycosylated. O-glycosylation
CC         increases activity in daughter cells by enhancing stability and
CC         promoting localization to the plasma membrane. May also be O-
CC         glycosylated by PMT1 and PMT2. {ECO:0000269|PubMed:10366591,
CC         ECO:0000269|PubMed:8846915}.
CC     -!- MISCELLANEOUS: Present with 396 molecules/cell in log phase SD
CC         medium. {ECO:0000269|PubMed:14562106}.
CC     -!- CAUTION: Ref.5 refers to this gene as REV7. REV7 is however the
CC         adjacent gene. {ECO:0000305}.
CC     -----------------------------------------------------------------------
CC     Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC     Distributed under the Creative Commons Attribution-NoDerivs License
CC     -----------------------------------------------------------------------
DR     EMBL; U49845; AAA98666.1; -; Genomic_DNA.
DR     EMBL; Z38059; CAA86138.1; -; Genomic_DNA.
DR     EMBL; AF395906; AAK83884.1; -; Genomic_DNA.
DR     EMBL; U07228; AAA67919.1; -; Genomic_DNA.
DR     EMBL; BK006942; DAA08412.1; -; Genomic_DNA.
DR     PIR; S48394; S48394.
DR     RefSeq; NP_012126.1; NM_001179488.1.
DR     ProteinModelPortal; P38928; -.
[..]

RC     STRAIN=ATCC 204508 / S288c;
RX     PubMed=24374639; DOI=10.1534/g3.113.008995;
RA     Engel S.R., Dietrich F.S., Fisk D.G., Binkley G., Balakrishnan R.,
RA     Costanzo M.C., Dwight S.S., Hitz B.C., Karra K., Nash R.S., Weng S.,
RA     Wong E.D., Lloyd P., Skrzypek M.S., Miyasato S.R., Simison M.,
RA     Cherry J.M.;
RT     "The reference genome sequence of Saccharomyces cerevisiae: Then and
RT     now.";
RL     G3 (Bethesda) 4:389-398(2014).
RN     [5]
RP     NUCLEOTIDE SEQUENCE [GENOMIC DNA] OF 1-775.
RA     Mathew P.W.;
RL     Submitted (JUN-2001) to the EMBL/GenBank/DDBJ databases.
RN     [6]
RP     NUCLEOTIDE SEQUENCE [GENOMIC DNA] OF 80-823.
RX     PubMed=7871890; DOI=10.1002/yea.320101115;
RA     Torpey L.E., Gibbs P.E.M., Nelson J., Lawrence C.W.;
RT     "Cloning and sequence of REV7, a gene whose function is required for
RT     DNA damage-induced mutagenesis in Saccharomyces cerevisiae.";
RL     Yeast 10:1503-1509(1994).
RN     [7]
```

Fig. 2.5 Database entry in UniProtKB/SwissProt in raw format. The entry is shortened at various places, marked by [..] (Courtesy UniProt Consortium)



UniProtKB - P38928 (AXL2_YEAST)

Protein | Axial budding pattern protein 2

Gene | AXL2

Organism | *Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)*

Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

Function[i]

Required for haploid cells axial budding pattern. Acts as an anchor to help direct new growth components and/or polarity establishment components like the BUD5 GTP/GDP exchange factor to localize at the cortical axial budding site. Regulates septin organization in late G1 independently of its role in polarity-axis determination. 8 Publications

Miscellaneous

Present with 396 molecules/cell in log phase SD medium. 1 Publication

GO - Molecular function[i]

- calcium ion binding Source: InterPro

Complete GO annotation...

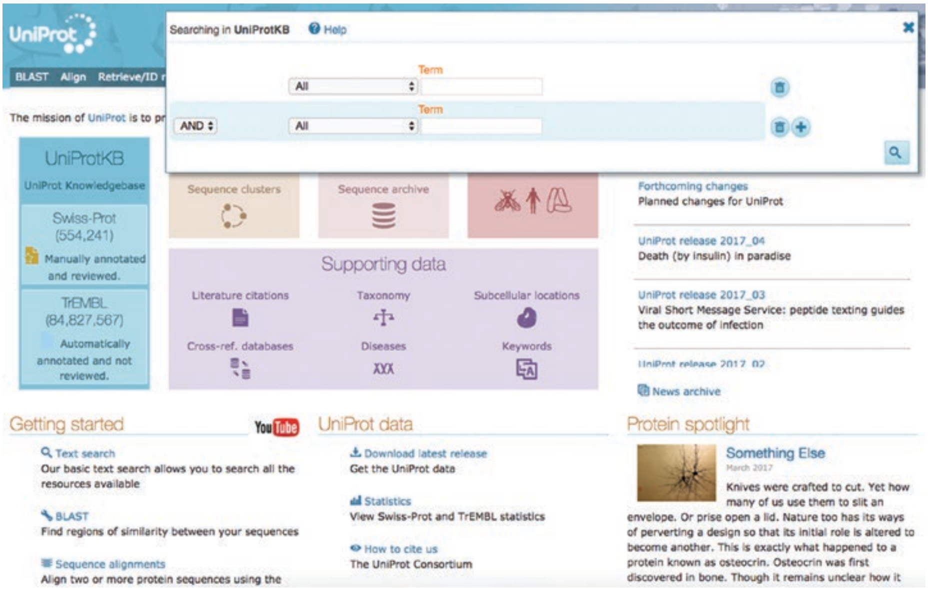
GO - Biological process[i]

- axial cellular bud site selection Source: SGD

Complete GO annotation...

Enzyme and pathway databases

■ **Fig. 2.6** Database entry in UniProtKB/SwissProt in graphical format (Courtesy UniProt Consortium)



The mission of UniProt is to pr

Swiss-Prot (554,241) Manually annotated and reviewed.

TrEMBL (84,827,567) Automatically annotated and not reviewed.

Forthcoming changes
Planned changes for UniProt

UniProt release 2017_04
Death (by insulin) in paradise

UniProt release 2017_03
Viral Short Message Service: peptide texting guides the outcome of infection

UniProt release 2017_02

News archive

Text search
Our basic text search allows you to search all the resources available

BLAST
Find regions of similarity between your sequences

Sequence alignments
Align two or more protein sequences using the

Download latest release
Get the UniProt data

Statistics
View Swiss-Prot and TrEMBL statistics

How to cite us
The UniProt Consortium

Something Else
March 2017
Knives were crafted to cut. Yet how many of us use them to slit an envelope. Or prise open a lid. Nature too has its ways of perverting a design so that its initial role is altered to become another. This is exactly what happened to a protein known as osteocrin. Osteocrin was first discovered in bone. Though it remains unclear how it

■ **Fig. 2.7** Home page of UniProt with text field for a simple full text search, overlaid with advanced search form (Courtesy UniProt Consortium)

Advanced on the right of the text field. In the advanced search form the field IDs and the corresponding logical operators can be selected from drop-down menus. When started, the search query is displayed in the text field and can be tweaked manually if necessary.

UniRef is a nonredundant sequence database that allows for fast similarity searches. The database exists in three versions: UniRef100, UniRef90, and UniRef50. Each database allows for the searching of sequences that are 100%, ≥ 90%, or ≥50% identical. The size of the database changes accordingly, making similarity searches, for example with BLAST, much faster.

2.2.2.2 NCBI Protein Database

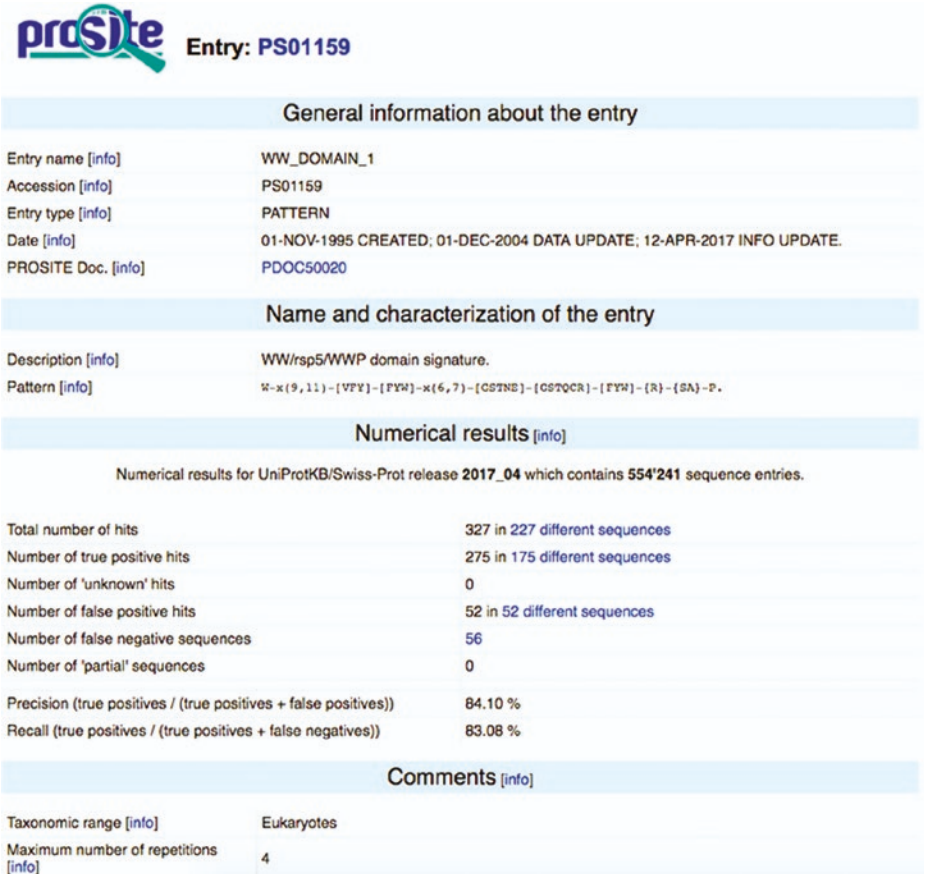
Another well-known protein sequence database is maintained at the NCBI. This database, however, is not a single database but a compilation of entries found in other protein sequence databases. For example, the NCBI database contains entries from Swissprot, the PIR database [pir], the Protein Data Bank (PDB) database [pdb], protein translations of the GenBank database, and several other sequence databases. Its format corresponds to that of GenBank, and queries are carried out analogously to those in GenBank via the Entrez system of NCBI.

2.3 Secondary Databases

2.3.1 Prosite

An important secondary biological database is Prosite [prosite] (Sigrist et al. 2012), which resides at the SIB [expasy]. Classification of proteins in Prosite is determined using single conserved motifs, i.e., short sequence regions (10–20 amino acids) that are conserved in related proteins and usually have a key role in the protein's function. The search for such sequence motifs in unknown proteins can provide a first hint of an affiliation to a protein family or function.

A motif is derived from multiple alignments (▶ Chap. 3) and saved in the database as a regular expression (◘ Fig. 2.8). This is a formalized pattern for the description of a sequence of characters. In a regular expression in Prosite, individual amino acids are represented by a one-letter code and separated by hyphens. If a position can contain more than one residue, then these are written in square brackets. Positions that can be filled by any amino acid are marked by a lowercase letter *x*. Repetitions of the same amino acid are indicated in full brackets, followed by the number of repetitions. A typical regular expression in Prosite would have the following form: `[GSTNE]-[GSTQCR]-[FYW]-{ANW}-x(2)-P`. This regular expression has seven amino acid positions. The first amino acid can be glycine, serine, threonine, asparagine, or glutamate; the second position glycine, serine, threonine, glutamine, cysteine or arginine; and the third position phenylalanine, tyrosine, or tryptophan. Position four can be any amino acid except alanine, asparagine, and tryptophan. In positions five and six, any amino acid



prosite Entry: PS01159

General information about the entry

Entry name [info]	WW_DOMAIN_1
Accession [info]	PS01159
Entry type [info]	PATTERN
Date [info]	01-NOV-1995 CREATED; 01-DEC-2004 DATA UPDATE; 12-APR-2017 INFO UPDATE.
PROSITE Doc. [info]	PDOC50020

Name and characterization of the entry

Description [info]	WW/rsp5/WWP domain signature.
Pattern [info]	W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-{R}-{SA}-P.

Numerical results [info]

Numerical results for UniProtKB/Swiss-Prot release **2017_04** which contains **554'241** sequence entries.

Total number of hits	327 in 227 different sequences
Number of true positive hits	275 in 175 different sequences
Number of 'unknown' hits	0
Number of false positive hits	52 in 52 different sequences
Number of false negative sequences	56
Number of 'partial' sequences	0
Precision (true positives / (true positives + false positives))	84.10 %
Recall (true positives / (true positives + false negatives))	83.08 %

Comments [info]

Taxonomic range [info]	Eukaryotes
Maximum number of repetitions [info]	4

Fig. 2.8 *NiceSite* view of Prosite database record PS01159 (Printed with permission of Swiss Institute for Bioinformatics)

can follow, and position seven is occupied by proline. The Prosite user manual [prosite-manual] contains a complete description of the Prosite database as well as the syntax of the regular Prosite expressions. The Expasy Prosite Web server [prosite] offers different possibilities to query the Prosite database. Besides searching for keywords, one can examine a sequence for the presence of Prosite motifs. Furthermore, using the algorithm ScanProsite, Prosite offers the possibility to search Swissprot, TrEMBL, and PDB for protein sequences that contain a user-defined pattern.

2.3.2 PRINTS

The PRINTS database [prints] (Attwood et al. 2003) uses fingerprints to classify sequences. Fingerprints consist of several sequence motifs, represented in the PRINTS database by short, local, ungapped alignments (▶ Chap. 3). The PRINTS database takes advantage of the fact that proteins usually contain functional regions that result in several sequence motifs per protein. By using fingerprints the sensitivity of the analysis increases, i.e., it is possible to evaluate the affiliation of a protein to a protein family even in the absence of one of the surveyed motifs. Besides information on how to derive a fingerprint and judge its quality, PRINTS also offers cross references to entries in related databases, permitting access to more information regarding a given protein family. Like Prosite, PRINTS contains information about each protein family and, if available, the biological function of each motif in the fingerprint. Querying the database on the PRINTS Web server [prints] can be carried out via a keyword search. However, it can be more interesting to search for fingerprints in protein sequences. Like the Prosite server, the PRINTS server offers tools for sequence analysis.

2.3.3 Pfam

The Pfam database [pfam] (Finn et al. 2016) classifies protein families according to profiles. A profile is a pattern that evaluates the probability of the appearance of a given amino acid, an insertion, or a deletion at every position in a protein sequence. Conserved positions are weighted more than less conserved positions, i.e., a weighted scoring scheme. Pfam is based on sequence alignments. High-quality, manually checked alignments serve as starting points for the automatic construction of hidden Markov models (HMMs). More sequences are then automatically added to the individual alignments of the SwissProt database. The resulting alignments should represent functionally interesting structures and contain evolutionarily related sequences. Owing to the partly automatic construction of the alignments, however, it is also possible that sequence alignments will arise that have no evolutionary relationship to one other. Therefore, the results of a search against the Pfam database should be carefully reviewed.

2.3.4 Interpro

The Integrated Resource of Protein Families, Domains and Sites (Interpro) [interpro] (Mulder et al. 2007) integrates important secondary databases into a comprehensive signature database. Interpro merges the databases Swissprot, TrEMBL, Prosite, Pfam, PRINTS, ProDom, Smart, and TIGRFAMs [tigr] and thereby allows a simple and simultaneous query of these databases. The result page combines the output of the individual queries. This makes for a fast comparison of the results while considering the strengths and weaknesses of the individual databases. The Interpro Web server offers a few intuitive query facilities for text and sequence searches.

2.4 Genotype-Phenotype Databases

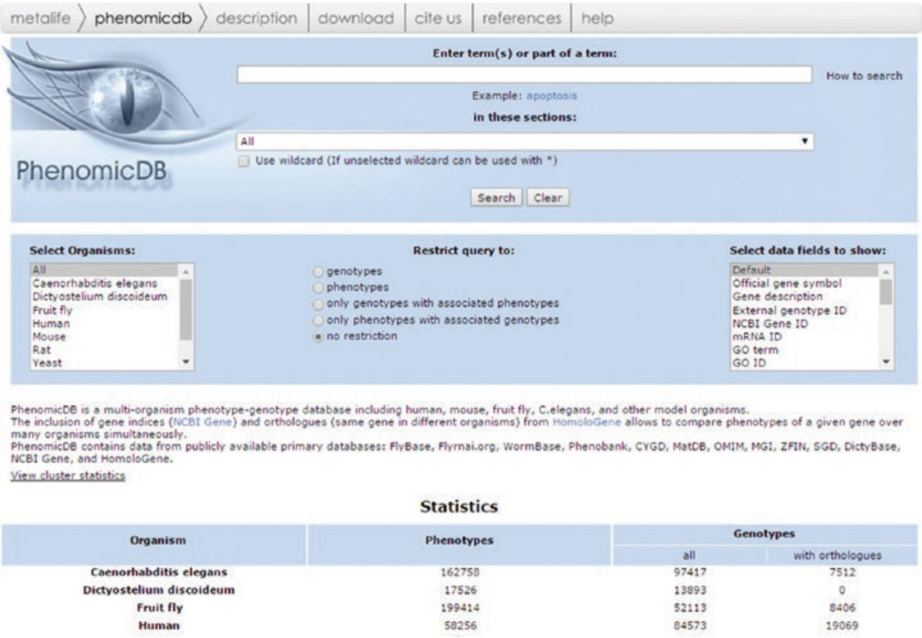
For diseases to emerge and progress, several genes or their products are frequently required. The identification of genes relevant to disease is, therefore, of vital importance in a target-based approach to rational drug development. A number of genotype-phenotype databases have been established that record relationships between genes and the biological properties of organisms. The Online Mendelian Inheritance in Man (OMIM) database of the NCBI [omim] is perhaps the best-known genotype-phenotype database. A new database of this type, dbGaP [dbgap], was also recently established at the NCBI. The data in this database come with analyses of the statistical significance of the respective genotype-phenotype relationship. The Online Mendelian Inheritance in Animals (OMIA) database [omia] at the NCBI also contains genotype-phenotype relationships of various animals, except mice and humans. For mice, the relevant database is in the Mouse Genome Database (MGD) [mgd]. Genotype-phenotype relationships of the two important model organisms, *D. melanogaster* and *C. elegans,* are recorded in FlyBase [flybase] and WormBase [wormbase], respectively. Both databases also contain much more information than just genotype-phenotype data. A detailed description of all the aforementioned databases [nar] would be beyond the scope of this book. In what follows, therefore, only a genotype-phenotype database is discussed that semantically integrates the contents of the aforementioned databases.

2.4.1 PhenomicDB

The PhenomicDB database is a multiorganism genotype-phenotype database containing data from humans and other important organisms such as the mouse, zebra fish (*Danio rerio*), fruit fly (*D. melanogaster*), nematode (*C. elegans*), baker's yeast (*S. cerevisiae*), and cress plant (*Arabidopsis thaliana*). PhenomicDB integrates data from the aforementioned and other primary genotype-phenotype databases. A complete listing of all underlying data sources can be found on the home page [phenomicdb] and in Kahraman et al. (2005).

A characteristic of PhenomicDB is that cross-organism comparisons of genotype-phenotype relationships are possible. This is accomplished by incorporating orthology data and gene indices from the database HomoloGene [homologene] at the NCBI. For example, the cause of porphyria, an inherited or acquired enzyme defect of humans, is a nonfunctional δ-aminolevulinate dehydratase. The respective gene has the symbol ALAD. As PhenomicDB indicates, a defect in the orthologous gene of baker's yeast (gene symbol: HEM2) leads to a very similar phenotype, characterized by the keywords auxotrophies, carbon and nitrogen utilization defects, carbon utilization, and respiratory deficiency. Of course, one cannot expect that distantly related organisms such as baker's yeast and humans show identical genotype-phenotype relationships in every case. Nevertheless, similar relationships can occur that might generate new hypotheses regarding disease pathogenesis or that allow the advancement of a disease model, thereby supporting the development of new drugs.

PhenomicDB is queried via a simple search interface. Search terms can be complemented automatically or manually by wildcards and restricted to certain database fields. Furthermore, it is possible to restrict the search to selected organisms. If orthologs of a given gene are found, the result page offers a hyperlink to the corresponding database record, allowing for a fast comparison of the genotype-phenotype relationships across organisms (■ Fig. 2.9). Owing to the semantic integration of the primary databases, some detail information can be lost, however, but this is compensated for by the



Enter term(s) or part of a term:

Example: apoptosis

in these sections: All

□ Use wildcard (If unselected wildcard can be used with *)

Select Organisms: All, Caenorhabditis elegans, Dictyostelium discoideum, Fruit fly, Human, Mouse, Rat, Yeast

Restrict query to:

- ○ genotypes
- ○ phenotypes
- ○ only genotypes with associated phenotypes
- ○ only phenotypes with associated genotypes
- ● no restriction

Select data fields to show: Default, Official gene symbol, Gene description, External genotype ID, NCBI Gene ID, mRNA ID, GO term, GO ID

PhenomicDB is a multi-organism phenotype-genotype database including human, mouse, fruit fly, C.elegans, and other model organisms.
The inclusion of gene indices (NCBI Gene) and orthologues (same gene in different organisms) from HomoloGene allows to compare phenotypes of a given gene over many organisms simultaneously.
PhenomicDB contains data from publicly available primary databases: FlyBase, Flyrnai.org, WormBase, Phenobank, CYGD, MatDB, OMIM, MGI, ZFIN, SGD, DictyBase, NCBI Gene, and HomoloGene.

View cluster statistics

Statistics

Organism	Phenotypes	Genotypes	
		all	with orthologues
Caenorhabditis elegans	162758	97417	7512
Dictyostelium discoideum	17526	13893	0
Fruit fly	199414	52113	8406
Human	58256	84573	19069

Fig. 2.9 Start page of PhenomicDB (Printed with permission of Metalife AG)

interconnections of the primary data and the breadth of information included. PhenomicDB can therefore be regarded as a metasearch engine for phenotypic information.

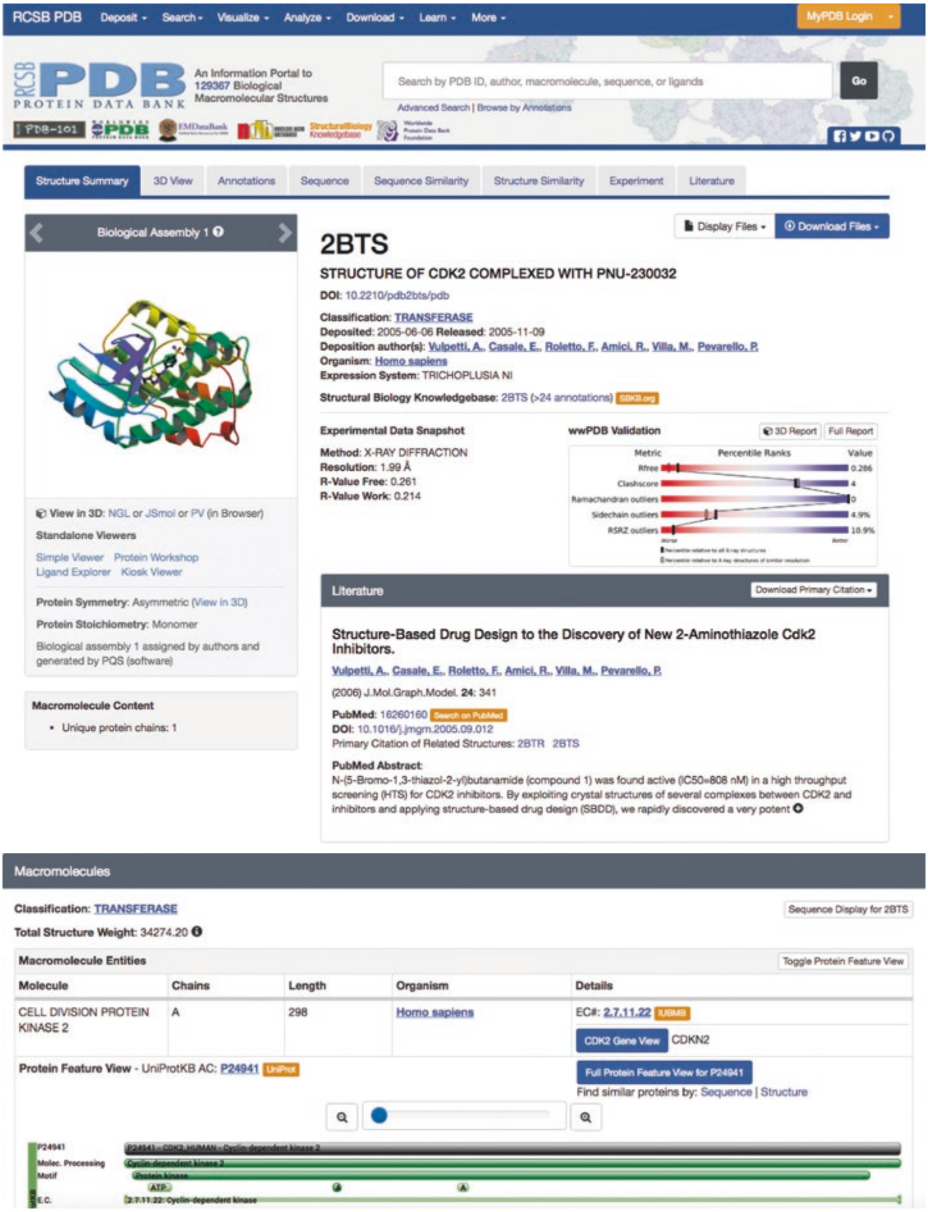
2.5 Molecular Structure Databases

2.5.1 Protein Data Bank

The PDB is a database of experimentally determined crystal structures of biological macromolecules and is coordinated by a consortium located in the USA, Europe, and Japan [wwpdb] (Berman et al. 2000). Probably the best-known Web page of the PDB is that of the Research Collaboratory for Structural Bioinformatics [pdb]. The PDB was founded at the Brookhaven National Laboratory in 1971, reflected in the frequent use of the name Brookhaven Protein Data Bank.

About 121,000 macromolecule structures are stored in the PDB database (as of July 2016). These are predominantly proteins, but also include DNA and RNA structures and protein–nucleic acid complexes. Structures of other macromolecules, for example glycopeptides and polysaccharides, constitute only a very small proportion of the total structures. As of 2002, only those crystal structures that have been solved experimentally are stored in the PDB database, whereas data of theoretical protein models are kept in their own section [pdb-models].

The PDB database offers several query options. A text-based search for a PDB ID or a keyword can be initiated on the main page. Furthermore, a number of search options



An Information Portal to 129367 Biological Macromolecular Structures

View in 3D: NGL or JSmol or PV (in Browser)

Standalone Viewers

Simple Viewer Protein Workshop
Ligand Explorer Kiosk Viewer

Protein Symmetry: Asymmetric (View in 3D)

Protein Stoichiometry: Monomer

Biological assembly 1 assigned by authors and generated by PQS (software)

Macromolecule Content

- Unique protein chains: 1

2BTS

STRUCTURE OF CDK2 COMPLEXED WITH PNU-230032

DOI: 10.2210/pdb2bts/pdb

Classification: TRANSFERASE
Deposited: 2005-06-06 Released: 2005-11-09
Deposition author(s): Vulpetti, A., Casale, E., Roletto, F., Amici, R., Villa, M., Pevarello, P.
Organism: Homo sapiens
Expression System: TRICHOPLUSIA NI

Structural Biology Knowledgebase: 2BTS (>24 annotations) SBKB.org

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 1.99 Å
R-Value Free: 0.261
R-Value Work: 0.214

wwPDB Validation

Literature

Structure-Based Drug Design to the Discovery of New 2-Aminothiazole Cdk2 Inhibitors.

Vulpetti, A., Casale, E., Roletto, F., Amici, R., Villa, M., Pevarello, P.

(2006) J.Mol.Graph.Model. **24**: 341

PubMed: 16260160
DOI: 10.1016/j.jmgm.2005.09.012
Primary Citation of Related Structures: 2BTR 2BTS

PubMed Abstract:
N-(5-Bromo-1,3-thiazol-2-yl)butanamide (compound 1) was found active (IC50=808 nM) in a high throughput screening (HTS) for CDK2 inhibitors. By exploiting crystal structures of several complexes between CDK2 and inhibitors and applying structure-based drug design (SBDD), we rapidly discovered a very potent

Macromolecules

Classification: TRANSFERASE

Total Structure Weight: 34274.20

Macromolecule Entities

Molecule	Chains	Length	Organism	Details
CELL DIVISION PROTEIN KINASE 2	A	298	Homo sapiens	EC#: 2.7.11.22 CDKN2

Protein Feature View - UniProtKB AC: P24941

Find similar proteins by: Sequence | Structure

■ **Fig. 2.10** Overview representation of PDB entry 2BTS (Printed courtesy of RCSB)

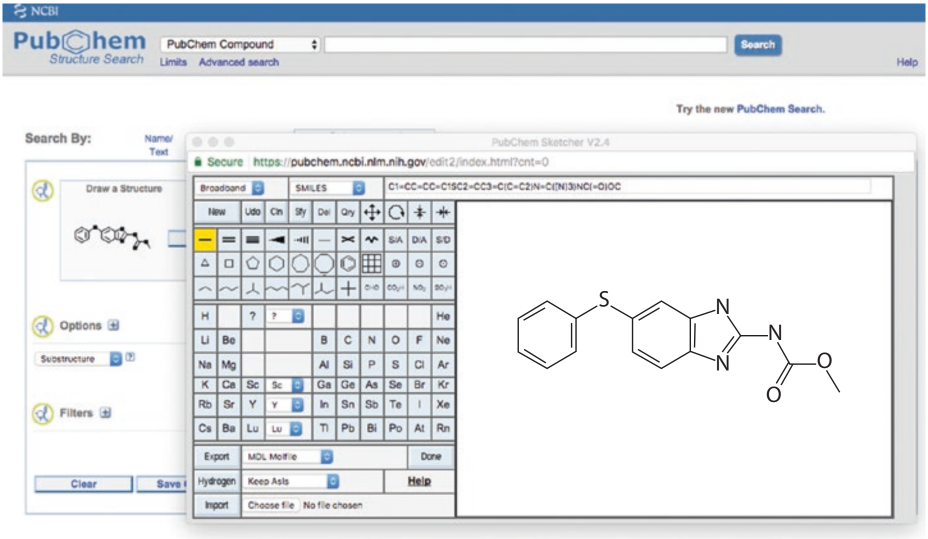
exist on the search database page, including detailed keyword and BLAST queries. A database record summarizes all of the information in the file and which is then detailed on subsequent pages. In addition, the molecular structure can be visualized by means of different applets (■ Fig. 2.10).

2.5.2 SCOP

Proteins that perform a similar biological function and are evolutionary related must have a similar structural organization, at least in the region of their active centers. It should, therefore, be possible to predict the function of an unknown protein by comparison of its structural organization with that of known proteins. Two databases, SCOP and CATH, provide such predictions. SCOP (Structural Classification Of Proteins) [scop] (Murzin et al. 1995) classifies proteins of a known structure in a hierarchical manner. The three main classifications are families, superfamilies, and folds. Families describe proteins with a clear evolutionary relationship to each other and are limited by a sequence identity that must be at least 30% greater than the total length of the proteins. Nevertheless, proteins that fall below this limit can be included in a family if relatedness can be shown owing to proven similar structures and functions. Proteins with very low sequence identities with respect to one other, even with suggested relations due to structural and functional properties, are put into superfamilies, however. Proteins that have the same arrangement of secondary structural elements in the same topology are classified into folds. It is unimportant whether the proteins have a functional relationship or whether the similarity of folds is based on physicochemical principles. Recently, a new version, the SCOP2 database [scop2] (Andreeva et al. 2014), has been developed. Instead of displaying relations in simple tree structures, networks are used to do so.

2.5.3 CATH

The CATH database [cath] (Greene et al. 2007) classifies protein structures hierarchically into four categories: Class (C), Architecture (A), Topology (T), and Homologous Superfamily (H). The classification of proteins into the Class category is mainly automatic, but it can be complemented by manual intervention when required. In the Class category, the proportion of secondary structural elements is taken into account without consideration of their arrangement or connections. Four classes of proteins are distinguished: proteins composed mainly of helices (*mainly alpha*), sheets (*mainly beta*), both helices and sheets (*alpha-beta*), and, finally, proteins with very few secondary structural elements. The Architecture category describes the arrangement of secondary structural elements to one another and is curated manually. Its categorization is performed via simple descriptors such as, for example, *barrel, sandwich,* and *beta-propeller*. In the Topology category, protein form and the interconnections of secondary structural elements are described. Its categorization is based on an algorithm that uses empirically derived parameters for domain classification. The Homologous Superfamily category encompasses homologous protein domains, i.e., domains with a common origin. The similarity of the sequences is determined by a sequence comparison followed by a structural comparison according to the classification in the Topology category. In addition to these four categories (whose first letters form the database name), a fifth category has been defined, the Sequence Families. Here, domains are classified based on high sequence identity (at least 35% identity over 60% of the length of the larger domain) and, thus, will likely possess similar functions.



SMILES: C1=CC=CC=C1SC2=CC3=C(C=C2)N=C(N3)NC(=O)OC

Fig. 2.11 Two-dimensional molecular structure editor of PubChem database (Printed courtesy of NCBI)

2.5.4 PubChem

The PubChem database at the NCBI [pubchem] stores small chemical molecules and information about their biological activities. It consists of three components, PubChem Compound, PubChem Substance, and PubChem BioAssay. PubChem Compound contains approx. 91 million molecules (July 2016) together with their two-dimensional (2D) molecular structures. A query is performed graphically via a molecular structure editor that allows the drawing of the desired (partial) structure (Fig. 2.11). Furthermore, PubChem Compound makes possible a search for molecules that fulfill certain physicochemical parameters, for example, a particular molecular weight range, a given number of acceptors or donors for hydrogen bonds, and a certain logP range.

PubChem Substance permits the search for substances produced by various manufacturers, samples of unknown composition, and natural substances of unknown 2D molecular structure. The records of both databases are linked and include a link to the third database, PubChem BioAssay, if the corresponding data are present. Information on biological assays and molecules that have been tested in these systems is recorded in PubChem BioAssay, and this database can be queried via a text search in the Entrez system.

The PubChem databases have multiple applications thanks to internal and external database linking, including to PubMed. For example, with a known enzyme inhibitor it is possible to find other similar potential inhibitors. Furthermore, small chemical molecules can be identified that have different structures yet have been shown to have similar effects in a biological test system.

2.6 Exercises

? Exercise 2.1

Search for a protein (enzyme) from the organism *Bacillus subtilis* that hydrolyzes terminal nonreducing arabinofuranoside residues. To do this, use the keyword search under Entrez (► http://www.ncbi.nlm.nih.gov/entrez/). Note: hydrolysis, arabinofuranoside, hydrolases, glycosyl, terminal, nonreducing. The Advanced search link leads you to an editor and your query history, so you can modify previous searches of the same session. Possible combinations are AND, OR, NOT.

? Exercise 2.2

Locate the gene for the enzyme IABF-BACSU from ► Exercise 3.1 in the nucleotide database. If you are unable to find it, try to develop new search strategies from the results and hints provided.

? Exercise 2.3

Search for the protein with the following accession number in Entrez: P94552.

? Exercise 2.4

Search for the same accession number on the EBI home page (► http://www.ebi.ac.uk/).

? Exercise 2.5

Take a closer look at the entry from ► Exercise 2.4 and change it to the TextEntry view. Which information can you obtain from such an entry? Describe briefly the information found. It is not necessary to characterize IABF2-BACSU any further.

? Exercise 2.6

In the graphical representation, under Publications, in the panel on the left hand side, you will find a hyperlink to a publication in the journal *Microbiology*. Click on this hyperlink. What happens? Note: The hyperlink to the publication is also available in the TextEntry view.

? Exercise 2.7

In the literature, two genes for *arfI* and *arfII* are described that are homologous to α-L-arabinofuranosidase 1 and α-L-arabinofuranosidase 2. From which species are these two genes? Which other species are reported in the literature to have homologous genes that are very similar? To answer these questions, go again to the NCBI page (► http://www.ncbi.nlm.nih.gov/) and search in the PubMed database. The History function that was mentioned earlier (► Exercise 2.1) can also be used in PubMed and in all other database searches at NCBI.

? Exercise 2.8

In the PubMed database, look for a publication by an author with your own last name. How many do you find? Are there several authors with your name? If you find nothing with your name, try it with the name *Blobel*. How can you restrict the results further using the name Günther Blobel, for example? How do you explain the differences with different search strategies?

Exercise 2.9

Carry out a Prosite scan (► http://www.expasy.org/prosite) with the sequence of the database entry IABF2-BACSU. You can enter the sequence by cutting and pasting or by entering the Swissprot accession number or ID. How many patterns are found? Which ones? What information about the motifs do you obtain on the results page? How can you obtain information about the biological role of the individual motifs in a simple way?

Exercise 2.10

Go to the start page of the PRINTS database (► http://bioinf.man.ac.uk/dbbrowser/PRINTS/) and perform a Fingerprint search against the PRINTS database with the sequence IABF2_BACSU. Note that the sequence must be entered in raw format. When done, perform the same search with the sequence of the database record A1AB_HUMAN.

Exercise 2.11

Go to the Blocks server (► http://blocks.fhcrc.org/) and initiate a database search with the Blocks Searcher using the sequence P35368. Because the search can take several minutes, possibly leading to a browser timeout, you should enter your e-mail address on the form. The results of the analysis will then be sent to you by e-mail. How many hits are found?

Exercise 2.12

With the protein from ► Exercise 2.11, query the Pfam database. The proteins of the Swissprot and TrEMBL databases are already present on the Pfam server. You can therefore either retrieve the previously obtained result with the accession number or protein ID or run a new analysis by providing the sequence in the FASTA format.

Exercise 2.13

Repeat the preceding search (► Exercise 2.12) using the Interpro database.

Exercise 2.14

Retrieve the 3D structure of bovine rhodopsin (a GPCR) from the PDB database. How many entries do you find? Take a closer look at the entry with the best crystallographic resolution for the complete protein (detailed in the overview). At what temperature was the crystallization carried out, and how many cysteine bonds does the protein have?

Exercise 2.15

Is there an assay to check for the HERG channel activity of a molecule? How many compounds were tested in this assay, and how many of them were active? Use the PubChem database to answer this question.

Exercise 2.16

In how many assays was the molecule fenbendazole tested, and in how many of those assays was it active? What is fenbendazole used for, and how does the molecule differ from albendazole?

Exercise 2.17

Is there a genotype-phenotype relationship in *D. melanogaster* that resembles the human genotype-phenotype relationship responsible for coproporphyria? Use PhenomicDB to answer this question.

References

Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG (2014) SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res 42(Databaseissue):D310–D314

Attwood TK, Bradley P, Flower DR, Gaulton A et al (2003) PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 31:400–402

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242

Finn RD, Coggill P, Eberhardt RY, Eddy SR et al (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279–D285

Greene LH, Lewis TE, Addou S, Cuff A et al (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35:D291–D297

Kahraman A, Avramov A, Nashev L, Popov D et al (2005) PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. Bioinformatics 21:418–420

Kim KS, Lilburn TG, Renner MJ, Breznak JA (1998) arfI and arfII, two genes of encoding alpha-L-arabinofuranosidases in *Cytophaga xylanolytica*. Appl Environ Microbiol 64:1919–1923

Mulder NJ, Apweiler R, Attwood TK, Bairoch A et al (2007) New developments in the InterPro database. Nucleic Acids Res 35:D224–D228

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540

NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 45:D12–D17

Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Bougueleret L, Xenarios I (2012) New and continuing developments at PROSITE. Nucleic Acids Res 41:D344–D347

The UniProt Consortium (2016) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158–D169

Further Reading

bankit. http://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank
cath. http://www.cathdb.info/
dbgap. http://www.ncbi.nlm.nih.gov/gap
ddbj. http://www.ddbj.nig.ac.jp/
ebi. http://www.ebi.ac.uk/
ebi-manual. http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html
ena. http://www.ebi.ac.uk/ena/
entrez. http://www.ncbi.nlm.nih.gov/nucleotide
entrez-help. http://www.ncbi.nlm.nih.gov:80/entrez/query/static/help/helpdoc.html
expasy. http://www.expasy.org/
flybase. http://www.flybase.org/
gb-sample. http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

genbank. http://www.ncbi.nlm.nih.gov/Genbank/
homologene. http://www.ncbi.nlm.nih.gov/homologene
interpro. http://www.ebi.ac.uk/interpro/
mgd. http://www.informatics.jax.org/
nar. http://nar.oxfordjournals.org/
ncbi. http://www.ncbi.nlm.nih.gov/
nig. https://www.nig.ac.jp/nig/
omia. http://omia.angis.org.au/home/
omim. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
pdb. http://www.rcsb.org/pdb/home/home.do
pdb-models. http://www.rcsb.org/pdb/search/searchModels.do
pfam. http://pfam.xfam.org/
phenomicdb. http://www.phenomicdb.de/
pir. http://pir.georgetown.edu/pirwww/dbinfo/pir_psd.shtml
prints. http://bioinf.man.ac.uk/dbbrowser/PRINTS/
prosite. http://prosite.expasy.org/
prosite-manual. http://prosite.expasy.org/prosuser.html
pubchem. http://pubchem.ncbi.nlm.nih.gov/
scop. http://scop.mrc-lmb.cam.ac.uk/scop/
scop2. http://scop2.mrc-lmb.cam.ac.uk/
sequin. http://www.ncbi.nlm.nih.gov/Sequin/
swissprot. http://www.expasy.org/sprot/
tigr. http://maize.jcvi.org/
uniprot. http://www.uniprot.org/
wormbase. http://www.wormbase.org/
wwpdb. http://www.wwpdb.org/