Vysoké učení technické v Brně

Fakulta informačních technologií

Databáze biologických dat

Databáze spravované Evropským institutem bioinformatiky

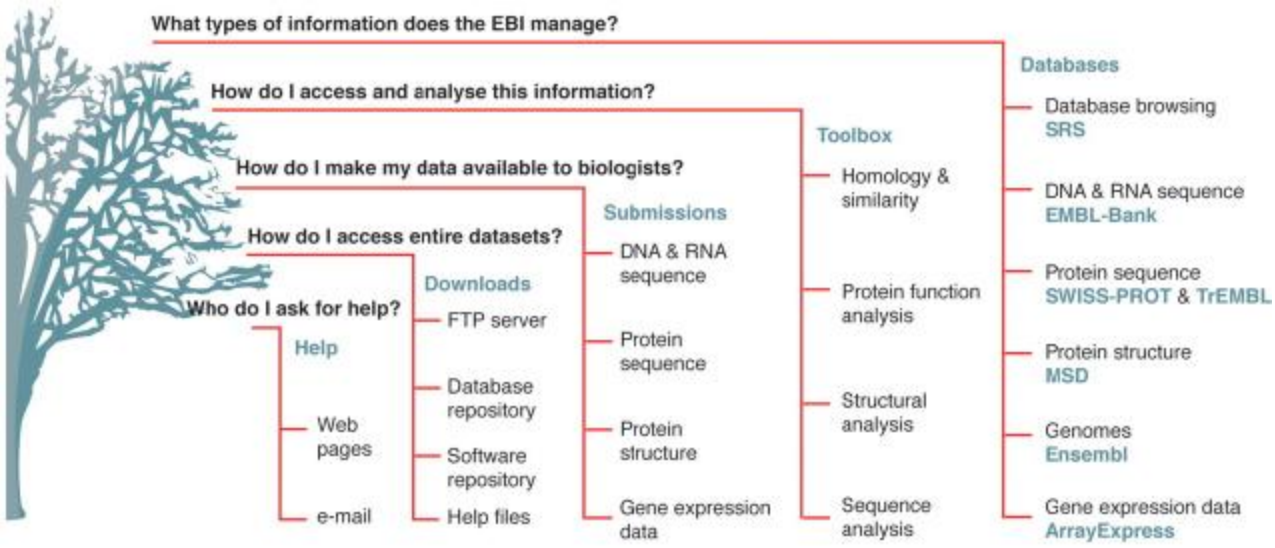
Brno 2005 **Ivana Rudolfová**

Obsah

1. Úvod 3
2. Databáze sekvencí DNA a RNA 4
2.1 Přístup k databázi EMBL-Bank 4
2.2 Vkládání dat do databáze EMBL-Bank 4
2.3 Struktura uložených dat 4
2.3.1 Třídy dat 5
2.3.2 Divize databáze 5
2.3.3 Struktura záznamu 5
2.3.4 Příklad databázového záznamu 6
2.3.5 Formát řádků databázového záznamu 8
3. Databáze sekvencí proteinů 14
3.1 Databáze Swiss-Prot 14
3.2 Databáze TrEMBL 15
3.3 Přístup k datům v databázi UniProtKB 16
3.4 Vkládání dat do databáze Swiss-Prot 16
3.5 Struktura uložených dat 16
3.5.1 Třídy dat 16
3.5.2 Struktura záznamu 16
3.5.2.1 Odlišnosti mezi formátem řádků použitých v obou databázích 17
3.5.2.2 Řádky vyskytující se pouze v databázi UniProtKB 18
3.5.2.3 Řádky vyskytující se pouze v databázi EMBL-Bank 18
4. Závěr 19
Literatura 20

1. Úvod

Tato práce je zaměřena na analýzu databází biologických dat, které spravuje Evropský institut bioinformatiky. Evropský institut bioinformatiky (European Bioinformatics Institut - EBI) je nezisková akademická organizace, která je součástí evropské laboratoře molekulární biologie (European Molecular Biology Laboratory – EMBL). EBI je centrum pro výzkum a služby v oblasti bioinformatiky. Institut spravuje databáze biologických dat zahrnujících informace o nukleových kyselinách, sekvencích proteinů a makromolekulárních strukturách. Úkolem této organizace je poskytnout veřejnosti data z oblasti molekulární biologie a genomického výzkumu, jejichž objem v současné době velmi rychle roste, volně je zpřístupnit vědecké komunitě a podpořit tak další rozvoj v této oblasti. EBI vytváří, spravuje a poskytuje databáze biologických dat a poskytuje informace, jak ukládat a získávat potřebná data. Institut spravuje 6 hlavních databází, které odráží způsoby, jimiž biologové získávají informace o tom, jak fungují buňky a celé organismy. Jsou v nich uloženy informace o sekvencích DNA a RNA (EMBL-Bank), sekvencích proteinů (SWISS-PROT a TrEMBL), struktuře proteinů (MSD), genomech (Ensembl) a expresi genů (ArrayExpress) [1]. Všechny databáze zahrnují kromě vlastních dat i další informace týkající se struktury genů, transkripčních mechanismů a funkcí proteinů. Všechny informace jsou integrovány z mnoha různých zdrojů. Dále zde existuje velké množství odkazů mezi jednotlivými databázemi, což umožňuje snadné hledání vztahů mezi molekulami různých typů. Institut také poskytuje velké množství nástrojů, které umožňují nejen ukládání a získávání biologických informací, ale také nástroje, které umožňují hledat vztahy mezi biologickými daty. Základní služby a databáze, které poskytuje organizace EBI shrnuje následující obrázek:



Kromě těchto základních databází EBI spravuje a poskytuje celou řadu dalších databází biologických dat (www.ebi.ac.uk/Databases/).
Tato práce se zabývá databázemi sekvencí nukleotidů o proteinových sekvencí. Jsou zde popsány tyto databáze, způsob přístupu k těmto databázím, možné způsoby vkládání dat do těchto databází a struktura dat uložených v databázích.

2. Databáze sekvencí DNA a RNA

Informace o sekvencích DNA a RNA poskytuje EBI v databázi EMBL-Bank. Tato databáze je základním evropským zdrojem informací o sekvencích nukleotidů. Tato databáze je vytvářena a udržována na základě spolupráce s organizacemi GenBank (USA) a DNA Database of Japan (DDBJ). Každá ze tří organizací sbírá část dat, která jsou ve světě produkována, a mezi organizacemi dochází denně k výměně nových a opravených dat (doplnění nových informací o datech…). Vzhledem k tomu, že publikace článků zabývajících se sekvencemi nukleotidů v odborných časopisech je podmíněna výskytem příslušných informací v databázi EMBL, jsou data v této databázi aktuální. Databáze tak odráží aktuální stav výzkumu v této oblasti.

2.1 Přístup k databázi EMBL-Bank

Dva hlavní způsoby přístupu k databázi EMBL-Bank jsou přístup pomocí nástroje pro dotazování, nazvaného Sequence Retrieval System (SRS), dostupného na www stránkách EBI (srs.ebi.ac.uk), nebo je možné data získat z FTP serveru EBI. Dále jsou čtvrtletně vydávány nové verze databáze a jsou distribuovány na DVD. Je možné získat data z každé ze spolupracujících databází nezávisle na tom, do které z databází byla data uložena. EBI poskytuje také další různé nástroje pro analýzu sekvencí a umožňuje tak například porovnávat sekvence vzhledem k sekvencím uloženým v EMBL-Bank. [2]

2.2 Vkládání dat do databáze EMBL-Bank

Většina vědeckých časopisů v dnešní době vyžaduje vložení informací o sekvencích nukleotidů do jedné z databází EMBL, GenBank nebo DDBJ před publikací článků, které se jimi zabývají. Díky tomu je zaručeno, že informace o sekvencích budou dostupné všem vědcům, už v době, kdy bude článek publikován. EMBL-Bank přidělí každému novému záznamu o sekvenci přístupové číslo, které jednoznačně a trvale identifikuje záznamy dané sekvence. Toto číslo také autoři článků zveřejňují ve svých příspěvcích a pomocí tohoto čísla mohou vkládat nové informace do záznamu o dané sekvenci. Po vložení dat do databáze mohou autoři data ihned zveřejnit, nebo počkat s jejich zveřejněním až do doby, kdy bude článek skutečně publikován.
Preferovaný způsob vkládání informací do EMBL databáze je pomocí www systému *Webin*, který je dostupný na adrese http://www.ebi.ac.uk/submission/webin.html. Tento systém umožňuje vkládání informací o sekvencích interaktivní formou, vyplňováním několika www formulářů. Pro vytvoření databázového záznamu o sekvenci jsou vyžadovány tyto údaje:

- informace o poskytovateli údajů
- datum, kdy mají být data zveřejněna
- vlastní informace o sekvenci (sekvence a další informace o ní)
- reference na citace v literatuře
- informace o hlavních částech sekvence (kódující regiony, regulační signály, …)

Informace o sekvencích je také možné zasílat pomocí emailu, ale data musí být poskytována ve speciálním formátu. Při vkládání většího množství podobných sekvencí, EBI doporučuje kontaktovat správce databáze, kteří pomohou s vkládáním těchto dat do databáze. [2]

2.3 Struktura uložených dat

Databáze je tvořena záznamy o sekvencích nukleotidů. Každý záznam odpovídá jedné sekvenci nukleotidů, která byla vložena do databáze nebo publikována v literatuře. V některých případech záznam shrnuje informace z více publikovaných článků zabývajících se stejnou sekvencí nukleotidů. Naopak jeden článek může být zdrojem více záznamů v databázi, zejména při porovnávání homologních sekvencí různých organismů.

2.3.1 Třídy dat

Třída každého záznamu je indikována na prvním řádku (ID) záznamu o sekvenci. Všechny distribuované a veřejně přístupné záznamy jsou třídy *standard*. Vnitřně databáze používá i další třídy dat.

2.3.2 Divize databáze

Záznamy, které tvoří databázi jsou shlukovány do divizí. Shlukování je založeno převážně na taxonomii. Výjimku tvoří divize EST (expressed sequence tags), kam patří sekvence, ke kterým obvykle neexistují téměř žádné doplňující informace a byly získány pomocí jednodušších metod. Tyto sekvence jsou také obvykle považovány za sekvence nižší kvality. Další výjimky tvoří divize STS (sequence tagged site), HTG (HighThroughput Geonome Sequence), HTC (High-Throughput cDNAs), GSS (genome survey sequences) a CON (construct). Do těchto divizí patří záznamy o sekvencích, které byly získány speciálními metodami. Kromě označení divize obsahuje záznam o sekvenci i taxonomickou klasifikaci organismu, pro který byla sekvence získána. Každý záznam v databázi přísluší právě do jedné divize, která je specifikována pomocí tříznakové zkratky na ID řádku záznamu o sekvenci. Seznam všech divizí a jejich kódy shrnuje následující tabulka.

Division	Code
Bacteriophage	PHG
Construct/Contig	CON
Expressed Sequence Tags	EST
Fungi	FUN
Genome Survey	GSS
High-Throughput cDNAs	HTC
High-Throughput Genome	HTG
Human	HUM
Invertebrates	INV
Organelles	ORG
Other Mammals	MAM
Other Vertebrates	VRT
Mus musculus	MUS
Plants	PLN
Prokaryotes	PRO
Rodents	ROD
Synthetic	SYN
Sequence Tagged Sites	STS
Unclassified	UNC
Viruses	VRL

2.3.3 Struktura záznamu

Záznamy v databázi jsou strukturovány tak, aby byly snadno zpracovatelné počítačovými programy a zároveň byly srozumitelné pro člověka. Všechny popisy dat, klasifikace i poznámky jsou v běžné angličtině. Jestliže je to možné, jsou použity symboly a označení, které se běžně používají v molekulární biologii.
Každý záznam v databázi obsahuje řádky různého typu. Každý typ řádku má definovaný formát a obsahuje určité informace o sekvenci pomocí specifikovaných typů dat. Na začátku každého řádku záznamu je dvouznakový kód, který určuje typ řádku, a tedy i informace, které tento řádek obsahuje.V současné době jsou v záznamech používány tyto typy řádků:

```
ID - identification             (begins each entry; 1 per entry)
AC - accession number           (>=1 per entry)
SV - sequence version           (1 per entry)
DT - date                       (2 per entry)
```

```
DE - description                (>=1 per entry)
KW - keyword                    (>=1 per entry)
OS - organism species           (>=1 per entry)
OC - organism classification    (>=1 per entry)
OG - organelle                  (0 or 1 per entry)
RN - reference number           (>=1 per entry)
RC - reference comment          (>=0 per entry)
RP - reference positions        (>=1 per entry)
RX - reference cross-reference  (>=0 per entry)
RG - reference group            (>=0 per entry)
RA - reference author(s)        (>=0 per entry)
RT - reference title            (>=1 per entry)
RL - reference location         (>=1 per entry)
DR - database cross-reference   (>=0 per entry)
CC - comments or notes          (>=0 per entry)
AH - assembly header            (0 or 1 per entry)
AS - assembly information       (0 or >=1 per entry)
FH - feature table header       (2 per entry)
FT - feature table data         (>=2 per entry)
XX - spacer line                (many per entry)
SQ - sequence header            (1 per entry)
CO - contig/construct line      (0 or >=1 per entry)
bb - (blanks) sequence data     (>=1 per entry)
// - termination line           (ends each entry; 1 per entry)
```

V jednom záznamu se nemusí vyskytovat všechny typy řádků a naopak v jednom záznamu se jeden typ řádku může opakovat několikrát. Každý záznam začíná řádkem ID (identification line) a je ukončen řádkem //. Řádky se vyskytují v záznamu v tom pořadí, v jakém byly uvedeny, s výjimkou řádku XX, který slouží jako oddělovač a může být použit kdekoliv mezi řádky ID a SQ.
Informace o struktuře a formátu záznamů byly převzaty z [3].

2.3.4 Příklad databázového záznamu

```
ID   TRBG361    standard; mRNA; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
SV   X56734.1
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   15-MAR-1999 (Rel. 59, Last updated, Version 9)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; rosids;
OC   eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolieae; Trifolium.
XX
RN   [5]
RP   1-1859
RX   MEDLINE; 91322517.
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.).";
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
```

```
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ databases.
RL   M.A. Hughes, UNIVERSITY OF NEWCASTLE UPON TYNE, MEDICAL SCHOOL, NEWCASTLE
RL   UPON TYNE, NE2 4HH, UK
XX
DR   GOA; P26204.
DR   MENDEL; 11000; Trirp;1162;11000.
DR   SWISS-PROT; P26204; BGLS_TRIRP.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1859
FT                   /db_xref="taxon:3899"
FT                   /mol_type="mRNA"
FT                   /organism="Trifolium repens"
FT                   /tissue_type="leaves"
FT                   /clone_lib="lambda gt10"
FT                   /clone="TRE361"
FT   CDS             14..1495
FT                   /db_xref="GOA:P26204"
FT                   /db_xref="SWISS-PROT:P26204"
FT                   /note="non-cyanogenic"
FT                   /EC_number="3.2.1.21"
FT                   /product="beta-glucosidase"
FT                   /protein_id="CAA40058.1"
FT                   /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                   DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                   VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                   IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                   EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                   IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT   mRNA            1..1859
FT                   /evidence=EXPERIMENTAL
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
     aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt     60
     cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag    120
     tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga    180
     aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata    240
     tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta    300
     caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc    360
     ttggccaaga atactcccaa agggaaagtt gagcggaggc ataaatcacg aaggaatcaa    420
     atattacaac aaccttatca acgaactatt ggctaacggt atacaaccat ttgtaactct    480
     ttttcattgg gatcttcccc aagtcttaga agatgagtat ggtggtttct taaactccgg    540
     tgtaataaat gattttcgag actatacgga tctttgcttc aaggaatttg gagatagagt    600
     gaggtattgg agtactctaa atgagccatg ggtgtttagc aattctggat atgcactagg    660
     aacaaatgca ccaggtcgat gttcggcctc caacgtggcc aagcctggtg attctggaac    720
     aggaccttat atagttacac acaatcaaat tcttgctcat gcagaagctg tacatgtgta    780
     taagactaaa taccaggcat atcaaaaggg aaagataggc ataacgttgg tatctaactg    840
     gttaatgcca cttgatgata atagcatacc agatataaag gctgccgaga gatcacttga    900
     cttccaattt ggattgttta tggaacaatt aacaacagga gattattcta agagcatgcg    960
     gcgtatagtt aaaaaccgat tacctaagtt ctcaaaattc gaatcaagcc tagtgaatgg   1020
     ttcatttgat tttattggta taaactatta ctcttctagt tatattagca atgccccttc   1080
     acatggcaat gccaaaccca gttactcaac aaatcctatg accaatattt catttgaaaa   1140
     acatgggata cccttaggtc caagggctgc ttcaatttgg atatatgttt atccatatat   1200
```

```
     gtttatccaa gaggacttcg agatcttttg ttacatatta aaaataaata taacaatcct   1260
     gcaattttca atcactgaaa atggtatgaa tgaattcaac gatgcaacac ttccagtaga   1320
     agaagctctt ttgaatactt acagaattga ttactattac cgtcacttat actacattcg   1380
     ttctgcaatc agggctggct caaatgtgaa gggtttttac gcatggtcat ttttggactg   1440
     taatgaatgg tttgcaggct ttactgttcg ttttggatta aactttgtag attagaaaga   1500
     tggattaaaa aggtacccta agctttctgc ccaatggtac aagaactttc tcaaaagaaa   1560
     ctagctagta ttattaaaag aactttgtag tagattacag tacatcgttt gaagttgagt   1620
     tggtgcacct aattaaataa aagaggttac tcttaacata tttttaggcc attcgttgtg   1680
     aagttgttag gctgttattt ctattatact atgttgtagt aataagtgca ttgttgtacc   1740
     agaagctatg atcataacta taggttgatc cttcatgtat cagtttgatg ttgagaatac   1800
     tttgaattaa aagtcttttt ttatttttt aaaaaaaaaa aaaaaaaaaa aaaaaaaaa    1859
//
```

2.3.5 Formát řádků databázového záznamu

Každý řádek záznamu obsahuje na začátku dvouznakový kód, který určuje typ řádků. Tento kód je následován třemi mezerami, takže aktuální informace začínají na každém řádku na pozici znaku číslo 6.

ID řádek

ID (IDentification) řádek je vždy prvním řádkem záznamu. Má následující základní formát:

```
ID   entryname  dataclass; [circular] molecule; division; sequencelength BP.
```

Entryname: identifikátor záznamu, skládající se z alfanumerických znaků. Identifikátor začíná písmenem a používají se pouze velká písmena.
Dataclass: určuje třídu záznamu, zveřejňované sekvence jsou třídy standard
Molecule type: typ molekuly uložené sekvence. Je-li známo, že jde o cyklickou molekulu, je možné to specifikovat použitím modifikátoru *circular,* není-li molekula označena tímto modifikátorem, neznamená to, že tato molekula je známa jako necyklická.
Database division: určuje divizi, do které je záznam přiřazen
Sequence length: celkový počet bází sekvence. Do tohoto počtu jsou zahrnuty i báze, které nejsou přesně specifikovány (označené jako *N – nonspecified*).
Řádek ID z uvedeného příkladu záznamu má tedy následující význam: záznam je označen identifikátor TRBG361, záznam je třídy standard, jedná se o molekulu mRNA, záznam patří do divize PLN – Plants (jedná se tedy o sekvenci některé rostliny) a sekvence má délku 1859 bází.

AC řádek

AC (ACcession number) řádek obsahuje seznam všech přístupových čísel, které jsou asociovány s danou sekvencí. Tento řádek se v záznamu vyskytuje vždy minimálně jeden. Každé přístupové číslo nebo interval přístupových čísel je ukončeno středníkem. Přístupová čísla slouží k identifikaci sekvencí a měla by být citována v literatuře, která se zabývá příslušnými sekvencemi. Kromě primárních přístupových čísel existují i sekundární přístupová čísla. Jsou použita proto, že může docházet k slučování nebo rozdělování záznamů. Pokud těmto záznamům budou přidělena nová přístupová čísla, jediná možnost jak nalézt původní záznamy je ponechat novým záznamům i původní přístupová čísla. Stanou se z nich sekundární přístupová čísla.

SV řádek

SV (Sequence Version) řádek upřesňuje verzi sekvence a vždy je v záznamu pouze jeden. Sekvenci nukleotidů je možné přesně identifikovat dvojicí: přístupové_číslo.verze_sekvence. Tuto dvojici obsahuje řádek SV. Přístupové číslo zůstává stejné, verze sekvence se zvyšuje s každou změnou sekvence.

DT řádek

DT (DaTe) řádek obsahuje informaci o tom, kdy se záznam poprvé objevil v databázi a kdy byl naposled modifikován. Každý záznam obsahuje dva DT řádky, každý z nich obsahuje jedno datum. Řádky mají následující formát:

```
DT   DD-MON-YYYY (Rel. #, Created)
DT   DD-MON-YYYY (Rel. #, Last updated, Version #)
```

Číslo Rel. označuje číslo verze databáze, ve které se poprvé objevila vložená nebo modifikovaná sekvence. Číslo verze se objevuje pouze u řádku s datem poslední změny záznamu. Jestliže záznam ještě nebyl od svého vložení do databáze modifikován, záznam bude mít také 2 řádky, které budou obsahovat stejné datum.
Záznam z uvedeného příkladu byl tedy do databáze vložen 12. 9. 1991 a poprvé se objevil ve verzi databáze číslo 29. Naposledy byl tento záznam modifikován 15. 3. 1999, současná verze je 9. verze záznamu a poprvé se objevila v 59. verzi databáze.

DE řádek

DE (DEscription) řádek obsahuje základní popisné informace o sekvenci. Může obsahovat jména genů, které jsou danou sekvencí kódovány, a další informace, které jsou užitečné pro identifikaci sekvence. Formát řádku je následující:

```
DE   description
```

Pro popis se používá běžná angličtina a popis není nijak strukturován. Často záznam obsahuje více těchto řádků. První řádek obvykle obsahuje stručný popis, který může být použitý samostatně. Standardy používané pro tento řádek spolu s příklady jsou dostupné na adrese: http://www.ebi.ac.uk/embl/Documentation/de_line_standards.html

KW řádek

KW (Key Words) řádek obsahuje informace, které mohou být použity pro generování vzájemných referencí mezi sekvencemi, založených na funkční, strukturální nebo jiné podobnosti záznamů. Jeden záznam v databázi často může vyžadovat více KW řádků, na jednom řádku se může vyskytovat více klíčových slov, které jsou oddělena středníkem. Za posledním klíčovým slovem následuje tečka. Klíčové slovo se může skládat z více slov, nikdy nemůže být rozděleno do dvou řádků. Formát tohoto řádku je následující:

```
KW   keyword[; keyword ...].
```

Klíčová slova jsou seřazena abecedně a jejich uspořádání neurčuje žádnou hierarchii. Jestliže záznam neobsahuje žádná klíčová slova bude obsahovat řádek tvaru:

```
KW   .
```

OS a OC řádky

OS (Organism Species) řádek obsahuje specifikaci organismu, pro který byla příslušná sekvence získána. Obvykle je nejprve použito latinské druhové a rodové jméno a za ním v závorkách následuje běžný anglický název, pokud je známý. Řádek má formát:

```
OS   Genus species (name)
```

Jestliže jde o organismus, který vznikl křížením organismů, budou zde popsány oba organismy.
OC (Organism Classification) řádek obsahuje taxonomickou klasifikaci organismu, který byl zdrojem dané sekvence. Formát řádku je následující:

```
OC   Node[; Node...].
```

První uzel představuje nejobecnější zařazení organismu a další uzly více specifikují daný organismus. Klasifikace může být rozdělena do více řádků, přičemž 1 uzel nemůže být

rozdělen do více řádků. Jednotlivé uzly jsou odděleny středníkem a za posledním uzlem následuje tečka.

OG řádek

OG (OrGanelle) řádek specifikuje organelu, pro kterou byla sekvence získána, jestliže nejde o sekvenci z jádra buňky. Tento řádek se vyskytuje pouze u záznamů pro sekvence, které nepocházejí z jádra buňky. Tento řádek může obsahovat jednu z hodnot: *Mitochondrion, Chloroplast, Kinetoplast, Cyanelle, Plastid* nebo jméno plasmidu. Například pokud se jedná o sekvenci chloroplastu rostliny *Euglena gracilit*, budou mít řádky OS, OC a OG tvar:

```
OS   Euglena gracilis (green algae)
OC   Eukaryota; Planta; Phycophyta; Euglenophyceae.
OG   Chloroplast
```

Řádky obsahující odkazy na literaturu

Tyto řádky poskytují informace o článcích, pro které byla data do databáze vložena. Pro každý odkaz se zde vyskytuje jeden blok řádků, které jsou vždy v následujícím pořadí: RN, RC, RP, RX, RG, RA, RT, RL. Jestliže v jednom záznamu je více odkazů, bude se zde tento blok řádků vyskytovat vícekrát. V každém bloku bude právě jeden řádek RN, řádky RA, RT a RL se zde vyskytují minimálně jedenkrát a řádky RC, RP, RX a RG se zde nemusí vyskytovat vůbec.

RN (Reference Number) řádek obsahuje jednoznačné číslo, které identifikuje citaci uvnitř jednoho záznamu. Toto číslo může být použito v poznámkách nebo v řádcích FT uvnitř záznamu. Číslo je vždy uvedeno v hranatých závorkách a čísla citací uvnitř jednoho záznamu nemusí vytvářet posloupnost. Číslo přiřazené jedné citaci není možné měnit ani není možné vkládat další citace se stejným číslem. Řádek má formát:

```
RN   [n]
```

RC (Reference Comment) řádek může obsahovat poznámky k dané citaci. Poznámky jsou psány v angličtině

RP (Reference Positron) řádek je volitelný řádek, který je použit, jestliže více intervalů bází dané sekvence se objevuje v dané citaci. Tento řádek indikuje, který interval (které intervaly) sekvence se objevuje v dané publikaci.

RX (reference cross-reference) řádek je opět volitelný. Může obsahovat odkaz na externí citace nebo zdroje abstraktů. Jestliže například citovaný článek je zařazený i v databázi MEDLINE, potom řádek RX bude odkazovat na příslušný identifikátor v databázi MEDLINE. Řádek má následující formát:

```
RX  resource_identifier; identifier.
```

Položka resource_identifier je zkratka na soubor dat, do kterého se odkazujeme. Může nabývat následujících hodnot:

```
Resource ID    Fullname
-----------    -----------------------------------
MEDLINE        MEDLINE bibliographic database (NLM)
PUBMED         PUBMED bibliographic database (NLM)
DOI            Digital Object Identifier (International DOI Foundation)
```

RG (Reference Group) řádek obsahuje seznam pracovních skupin (konsorcií), která se podílela na vzniku záznamu.

RA (Reference Author) řádek obsahuje seznam autorů citovaného článku (nebo jiné publikace). Jsou zde uvedeni všichni autoři ve stejném pořadí, jako v odkazovaném článku. Pro každého autora je uvedeno příjmení a první písmena ostatních jmen následovaná tečkou. Jména autorů jsou oddělena čárkou a seznam je ukončený středníkem. Záznam pro jednoho autora musí být uložen na jednom řádku.

RT (Reference Title) řádek obsahuje název článku (nebo jiné publikace). Název je uveden co nejpřesněji vzhledem k omezením daným znakovou sadou počítačů. Název je uveden v uvozovkách a může být rozdělen do více řádků. Poslední řádek je ukončen středníkem. Jestliže název obsahuje řecká písmena, bude místo nich použit hláskovaný tvar. Podobně budou nahrazeny například i horní nebo dolní indexy.

RL (Reference Location) řádek obsahuje informace o tom, kde je možné uvedený článek nalézt. Informace je uvedena tak, jak je běžné pro citace literatury. Obvykle je zde uveden název časopisu, číslo vydání, rozsah stránek a rok vydání příspěvku. Pro název časopisu se používají zkratky definované standardem ISO. Řádek má následující formát:

```
RL   journal vol:pp-pp(year).
```

Jestliže článek ještě nebyl publikován, ale byl přijat, bude doplněn pouze název časopisu, kde bude článek publikován. Případně další informace, pokud jsou již známé. Jinak budou čísla nahrazena nulami. Jestliže byl článek publikován v knize bude použita jiná varianta řádku RL. První řádek RL bude potom obsahovat označení *(in)* a dále bude následovat seznam autorů článku, název knihy, rozsah stránek místo a rok vydání knihy. Pokud se jedná o elektronickou publikaci bude mít řádek tvar:

```
RL   (er) Free text
```

DR řádek

DR (Databáze Cross-reference) řádek obsahuje odkazy do dalších databází, které obsahují informace vztahující se k uvažovanému záznamu. Například jestliže záznam o proteinu, který může vzniknout překladem dané sekvence, je uložen v databázi SWISS-PROT, potom řádek DR bude obsahovat odkaz na příslušný záznam v databázi SWISS-PROT. Řádek má následující formát:

```
DR  database_identifier; primary_identifier; secondary_identifier.
```

První položkou je identifikátor databáze, který tvoří zkratka názvu databáze. Za ní následují primární identifikátor záznamu v cílové databázi a případně i sekundární identifikátor, pokud existuje.

AH a AS řádek

Tyto řádky se vyskytují pouze u TPA záznamů (Third Party Annotation). U těchto záznamů se vyskytují povinně a obsahují informace o tom, z jakých částí byla TPA sekvence složena, z kterých primárních sekvencí. AH (Assembly Header) řádek poskytuje názvy sloupců pro následující AS řádek. AH řádek neobsahuje žádná data a má následující tvar:

```
AH   TPA-SPAN       PRIMARY_IDENTIFIER     PRIMARY_SPAN     COMP
```

AS (ASsembly Information) řádek poskytuje informace o složení TPA sekvence. Obsahuje informace o tom, který interval bází primární sekvence tvoří jednotlivé intervaly bází TPA sekvence.
Záznam může mít například následující tvar:

```
AH   TPA-SPAN       PRIMARY_IDENTIFIER     PRIMARY_SPAN     COMP
AS   1-426          AC004528.1             18665-19090
AS   427-526        AC001234.2             1-100            c
AS   527-1000       TI55475028             not_available
```

Sloupec TPA-SPAN identifikuje interval v TPA sekvenci, sloupec PRIMARY_IDENTIFIER obsahuje identifikátor primární sekvence, sloupec PRIMARY_SPAN identifikuje interval primární sekvence a sloupec COMP obsahuje hodnotu c, jestliže v sekvenci TPA byl použit řetězec komplementární k primární sekvenci. V uvedeném příkladu tedy budou báze 1 – 426 v sekvenci TPA tvořit báze 18665 – 19090 ze sekvence s identifikátorem AC004528.1.

OC řádek

Tento řádek se vyskytuje pouze u záznamů patřících do divize CON (Constructed, Contig). Tyto záznamy reprezentují celé chromozómy, genomy nebo jiné dlouhé sekvence složené ze segmentů, kterým odpovídají záznamy v databázi. Záznamy neobsahují přímo sekvence jako takové, ale obsahují informace jak z ostatních sekvencí, jejichž záznamy jsou v databázi, složit dlouhé sekvence. Tyto informace zahrnují identifikátory jednotlivých sekvencí ve tvaru přístupové_číslo.verze a popis části sekvence (např. interval bází), která je použita pro složení dlouhé sekvence. Tato data, která jsou nezbytná pro vytvoření dlouhé sekvence, jsou obsažena v řádcích CO. Tyto řádky mohou mít následující tvar:

```
CO   join(Z99104.1:1..213080,Z99105.1:18431..221160,Z99106.1:13061..209100,
CO   Z99107.1:11151..213190,Z99108.1:11071..208430,Z99109.1:11751..210440,
CO   Z99110.1:15551..216750,Z99111.1:16351..208230,Z99112.1:4601..208780,
CO   Z99113.1:26001..233780,Z99114.1:14811..207730,Z99115.1:12361..213680,
CO   Z99116.1:13961..218470,Z99117.1:14281..213420,Z99118.1:17741..218410,
CO   Z99119.1:15771..215640,Z99120.1:16411..217420,Z99121.1:14871..209510,
CO   Z99122.1:11971..212610,Z99123.1:11301..212150,Z99124.1:11271..215534)
```

První řádek CO obsahuje klíčové slovo *join* a za ním v závorce jsou uvedeny identifikátory sekvencí a popis použitých bází. Informace o jednotlivých sekvencích jsou odděleny čárkou. Jestliže složená sekvence obsahuje mezeru nedefinované délky, bude mezi identifikátory okolních sekvencí vložen výraz *gap()*. Vzhledem k tomu, že délka této mezery není definována, nezapočítává se do celkové délky sekvence. Mezery definované délky budou reprezentovány výrazem *gap(#)*, kde # je délka mezery. Zde se délka mezery zahrnuje do celkové délky sekvence.

FH řádek

FH (Feature Header) řádek se používá pouze pro zvýšení srozumitelnosti záznamu pro člověka. Řádky tohoto typu se vyskytují v záznamu vždy dva (pokud záznam obsahuje řádky FT) a neobsahují žádná data. Obsahují pouze hlavičku pro následující řádky FT. Řádky mají vždy následující tvar:

```
FH   Key             Location/Qualifiers
FH
```

Druhý řádek slouží pouze jako oddělovač.

FT řádky

FT (Feature Table) řádky poskytují mechanismus pro popis jednotlivých částí sekvence. Oblasti nebo místa (features) dané sekvence, které jsou určitým způsobem zajímavá, jsou popsána v této tabulce. Obecně jsou zde popsána místa, kterými se zabývají příslušné citované články. Často jsou to místa, která reprezentují určité signály. FT řádky patří mezi nejčastěji upravované řádky tabulky. Záznamy v nich se mění s novými poznatky o sekvencích. Pro každou popsanou oblast sekvence se zde nachází označení této oblasti (key, obvykle odvozeno od funkce oblasti), umístění oblasti v rámci sekvence a kvalifikátory, které obsahují další užitečná data o sekvenci. Pro jednotlivé typy oblastí existují povinné a volitelné kvalifikátory. Podrobnější informace o struktuře této tabulky lze nalézt v dokumentu *The DDBJ / EMBL / Gen Bank Feature Table: Definition* [4].

SQ řádek

SQ (SeQuence header) řádek označuje začátek dat sekvence (výpis pořadí nukleotidů v sekvenci). Obsahuje také souhrnné informace o sekvenci – celkový počet bázových párů v sekvenci (tento jednotný formát se používá i pro jednoduché řetězce, kde nejde o bázové páry, ale o jednotlivé báze) a počet jednotlivých bází (A, C, G, T) v sekvenci. Počet jiných bází než A, C, G nebo T je shrnut do skupiny *other*. Na začátku řádku je obsaženo klíčové slovo *Sequence*, které opět slouží pouze pro zvýšení čitelnosti záznamu. Řádek může mít například následující tvar:

```
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
```

Řádky obsahující vlastní sekvenci nukleotidů

Kódem těchto řádků jsou 2 mezery, proto se zdá, že tyto řádky jsou bez kódu. Sekvence je zapsána po 60 bází na jeden řádek. Vždy po deseti bázích je vložena jedna mezera. První báze je na šesté pozici v řádku stejně jako u ostatních řádků. Báze jsou vždy zapsány od konce 5' ke konci 3' a pokud je to možné, používá se nekódující řetězec. Na pozicích 73 – 80 každého řádku se nachází číslo báze pro snadnější orientaci a snadnější vyhledávání oblastí v sekvenci. Čísla jsou zarovnána vpravo a obsahují číslo poslední báze na daném řádku. Písmena použitá pro jednotlivé báze odpovídají doporučením komise IUPAC-IUB a shrnuje je následující tabulka:

Code	Base	Description
G	Guanine	
A	Adenine	
T	Thymine	
C	Cytosine	
R	Purine	(A or G)
Y	Pyrimidine	(C or T or U)
M	Amino	(A or C)
K	Ketone	(G or T)
S	Strong interaction	(C or G)
W	Weak interaction	(A or T)
H	Not-G	(A or C or T) H follows G in the alphabet
B	Not-A	(C or G or T) B follows A
V	Not-T (not-U)	(A or C or G) V follows U
D	Not-C	(A or G or T) D follows C
N	Any	(A or C or G or T)

CC řádek

Tento řádek obsahuje nestrukturované poznámky týkající se záznamu, které nelze zařadit do jiných typů řádků. Mohou v nich být zaznamenány libovolné užitečné informace.

XX řádek

XX (spacer) řádek neobsahuje žádná data ani poznámky a slouží jako oddělovač jednotlivých částí záznamu. Důvodem k jeho použití je opět větší čitelnost a srozumitelnost záznamu. Prázdné řádky jsou označeny tímto kódem, aby nedocházelo k jejich záměně s řádky obsahujícími vlastní pořadí nukleotidů sekvence.

// řádek

Tento řádek také neobsahuje žádná data ani poznámky, ale označuje konec záznamu.

3. Databáze sekvencí proteinů

Přepis genů (jejich sekvence jsou v databázi EMBL) do jejich produktů je značně komplikovaný proces, který může být ovlivněn vnějšími faktory. Dochází k různým transformacím, vynechání určitých částí genů, a tak z jednoho genu může vzniknout větší množství produktů. Vzhledem k tomu, že velká část sekvencí genomů je již známá, zaměřují se dnes vědci spíše na oblast produktů (proteinů = bílkovin), které jsou jimi kódovány. Je nezbytné, aby vědcům byla snadno dostupná data, která jsou o proteinech již známá. Z těchto důvodů institut EBI spravuje databáze sekvencí proteinů, v nichž jsou známá data popsána jednotným způsobem a mohou tak být snadno využívána.
Centrální databází proteinových sekvencí, kterou spravuje EBI, je databáze UniProt Knowledgebase (UniProtKB). Je zde uloženo velké množství informací o sekvencích a jejich funkcích v jednotné a přesné formě. Hlavním úkolem je poskytnout vědcům vysoce kvalitní, přesná, klasifikovaná a srozumitelná data. Tuto databázi tvoří dvě hlavní sekce: Swiss-Prot a TrEMBL. Sekce SwissProt obsahuje ručně vložené záznamy o sekvencích, které byly popsané v literatuře nebo se jedné o zkontrolované výsledky získané počítačovou analýzou. Sekce TrEMBL obsahuje záznamy o sekvencích, které byly získány počítačovou analýzou, ale ještě nebyly manuálně popsány. Následující podkapitoly se zabývají hlavními rysy těchto dvou sekcí, další dvě podkapitoly shrnují možnosti přístupu k datům v databázi a způsob vkládání dat do databáze. Poslední podkapitola se zabývá strukturou databáze. Struktura databáze UniProtKB je velmi podobná struktuře databáze EMBL, proto zde již nebude tak detailně popisována.

3.1 Databáze Swiss-Prot

Swiss-Prot je databáze proteinových sekvencí, která obsahuje i popisy uložených sekvencí. Byla založena roku 1986 a od roku 1987 je společně spravována skupinou vědců kolem Amose Bairocha ze Švýcarského institutu bioinformatiky (Swiss Institute of Bioinformatics - SIB) a jednou z částí instituce EBI – EMBL Data Library. Databáze Swiss-Prot se stejně jako databáze EMBL skládá ze záznamů o sekvencích. Jednotlivé záznamy se skládají z řádků různého typu, které mají svůj specifikovaný formát. Formát databáze je navržen tak, aby byl co nejvíce podobný formátu databáze EMBL-Bank.Čtyři hlavní rysy této databáze tvoří:

- popis dat
- minimalizace redundance
- integrace s ostatními databázemi
- dokumentace

Popis dat

Každý záznam v databázi Swiss-Prot může kromě základních vlastních dat obsahovat další data popisující danou sekvenci. Mezi základní data patří:

- vlastní sekvence aminokyselin
- informace o citacích v literatuře
- taxonomická data (o organismu, z kterého byla sekvence získána)

Další popisná data potom mohou blíže specifikovat:

- funkci proteinu
- posttranslační modifikace
- zajímavé oblasti nebo místa sekvence (např. vazební místa)
- sekundární strukturu sekvence (alfa-helix, skládaný list…)
- kvartérní strukturu
- podobnosti s dalšími proteiny
- choroby spojené s odchylkami proteinu
- varianty sekvence

Cílem je zahrnout do databáze co největší množství dat, která jsou o jednotlivých sekvencích známa. Zaměstnanci organizace spravující databázi procházejí články publikované

v odborných časopisech a nové poznatky vkládají do databáze. Popisná data se vyskytují především v řádcích CC (comment lines), FT (feature table) a KW (keyword lines).

Minimalizace redundance

Některé databáze obsahují pro jednu proteinovou sekvenci více záznamů, odpovídajících jednotlivým citacím v literatuře. V databázi Swiss-Prot jsou pokud možno data o jedné sekvenci shrnuta do jednoho záznamu, aby se předešlo redundanci dat. Jestliže dochází ke konfliktům mezi daty pro jednu sekvenci, je tato informace obsažena v řádcích FT (feature table) příslušného záznamu.

Integrace s ostatními databázemi

Pro biology je důležitá integrace mezi třemi hlavními databázemi týkajícími se sekvencí: sekvence nukleotidů, proteinové sekvence a databáze struktury proteinů. Je důležitá také integrace s ostatními specializovanými databázemi. V databázi Swiss-Prot se v současné době používají odkazy na dalších asi padesát databází. Reference jsou poskytovány ve formě ukazatelů na záznamy v jednotlivých databázích. Vzhledem k tomu, že Swiss-Prot používá odkazy do velkého počtu dalších databází, vytváří určitý centrální bod mezi databázemi biologických dat.

Dokumentace

Databáze Swiss-Prot je distribuována s velkým množstvím indexových souborů a dalších souborů s dokumentací. Podrobné informace o distribuovaných souborech lze nalézt v příloze D dokumentu: UniProt knowledgebase user manual [5].

3.2 Databáze TrEMBL

TrEMBL je sekce databáze UniProtKB, která obsahuje záznamy o sekvencích vytvořené počítačem. Obsahuje proteinové sekvence pro všechny kódující oblasti sekvencí uložených v databázích DDBJ/EMBL/GenBank a proteinové sekvence získané z literatury nebo vložené do UniProtKB, které ještě nabyly integrované do sekce Swiss-Prot. Vzhledem k tomu, že sekce Swiss-Prot vyžaduje kvalitní popis sekvence, který nemusí být ihned dostupný, sekce TrEMBL umožňuje zveřejnit sekvence, které ještě nejsou detailně popsané.
Pro záznamy, které jsou uloženy v databázích DDBJ/EMBL/GenBank, jsou nalezeny proteiny, které jsou jednotlivými sekvencemi kódovány. Tyto proteinové sekvence jsou vloženy do databáze TrEMBL společně s informacemi, které je možné získat ze záznamu o sekvenci nukleotidů. Kvalita záznamu potom závisí na množství dat obsažených v databázích DDBJ/EMBL/GenBank. Takto vytvořená data mohou být upravena následujícími procesy

- automatický popis dat
- odstranění redundance
- přisuzování důkazů

Takto upravené záznamy zůstávají v sekci TrEMBL, dokud nejsou manuálně popsány a integrovány do sekce Swiss-Prot.

Automatický popis dat

Tento proces umožňuje zlepšit popis dat v databázi TrEMBL, která čekají na manuální popis. Informace jsou převzaty z dobře charakterizovaných záznamů v databázi Swiss-Prot a doplněny do záznamů o proteinech v TrEMBL, které patří do stejné skupiny proteinů definované v databázi InterPro. InterPro je databáze rodin proteinů, domén a funkčních míst. Tímto procesem je možné zvýšit kvalitu dat dostupných v databázi TrEMBL.

Odstranění redundance

Sekvence, které pocházejí ze stejného organismu, a které jsou zcela shodné, jsou sloučeny do jednoho záznamu pro snížení redundance dat.

Přisuzování důkazů
Záznamy v databázi TrEMBL pocházejí z různých zdrojů, které zahrnují databáze sekvencí nukleotidů, data z různých specifických programů, data získaná automatickým popisem i data vložená do databáze manuálně. Pro uživatele má informace o zdroji dat velký význam, proto byl v databázi TrEMBL zaveden systém přisuzování důkazů. Tento systém také umožňuje automatickou změnu dat v databázi UniProtKB, jestliže došlo ke změně zdrojových dat.

3.3 Přístup k datům v databázi UniProtKB

Pro prohlížení dat v databázi UniProtKB je možné použít nástroj SRS (Sequence Retrieval System), který je možné použít i pro přístup k datům v databázi EMBL. Tento systém je poskytován institutem EBI a je dostupný na jeho www stránkách [6]. Data je také možné získat z FTP serveru ftp://ftp.ebi.ac.uk/databases. Stejně jako u databáze EMBL-Bank i u databáze UniProtKB jsou nové verze této databáze distribuovány na DVD a lze si je vyžádat na adrese support@ebi.ac.uk. Další dokumentaci k této databázi je možné získat na www adresách: www.ebi.ac.uk/swissprot a www.ebi.ac.uk/trembl. [1]

3.4 Vkládání dat do databáze Swiss-Prot

Data do databáze Swiss-Prot lze vkládat opět pomocí formulářů na stránkách organizace EBI: www.ebi.ac.uk/swissprot/Submission. Do databáze TrEMBL se data nevkládají, ale jsou automaticky generována z databází DDBJ/EMBL/GenBank. [1]

3.5 Struktura uložených dat

Databáze je tvořena záznamy o proteinových sekvencích, jejichž struktura je velmi podobná záznamům v EMBL-Bank. Každý záznam odpovídá jedné proteinové sekvenci, která byla do databáze vložena ručně nebo automaticky vygenerována z databází DDBJ/EMBL/GenBank nebo publikována v literatuře. Některé záznamy byly vytvořeny na základě několika článků, které se zabývaly stejnými sekvencemi. Záznamy, které jsou součástí sekce Swiss-Prot obsahují kromě vlastních dat o sekvenci také množství dalších popisných dat, které lépe charakterizují danou sekvenci.

3.5.1 Třídy dat

Třída každého záznamu je indikována na prvním řádku (ID) záznamu o sekvenci. V databázi UniProtKB jsou použity dvě třídy záznamů – *standard* a *preliminary*. Tyto třídy umožňují rozlišení záznamů databáze Swiss-Prot a TrEMBL. Záznamy třídy *standart* musí obsahovat takový popis dat, který je vyžadován standardy databáze Swiss-Prot (kompletní, kvalitní záznamy). Záznamy třídy *preliminary* spadají výhradně do sekce TrEMBL a nebyly pracovníky organizace ručně charakterizovány.

3.5.2 Struktura záznamu

Záznamy v databázi jsou strukturovány tak, aby byly snadno zpracovatelné počítačovými programy a zároveň byly srozumitelné pro člověka. Všechny popisy dat, klasifikace i poznámky jsou v běžné angličtině. Jestliže je to možné, jsou použity symboly, které se běžně používají v molekulární biologii.
Struktura záznamů je vytvořena tak, aby byla co nejvíce shodná se strukturou záznamů v databázi EMBL-Bank. Každý záznam v databázi opět obsahuje řádky různého typu. Každý typ řádku má definovaný formát a obsahuje určité informace o sekvenci pomocí specifikovaných typů dat. Na začátku každého řádku záznamu je dvouznakový kód, který

určuje typ řádku, a tedy i informace, které tento řádek obsahuje. V databázi Swiss-Prot se používají následující typy řádků:

Line code	Content	Occurrence in an entry
ID	Identification	Once; starts the entry
AC	Accession number(s)	Once or more
DT	Date	Three times
DE	Description	Once or more
GN	Gene name(s)	Optional
OS	Organism species	Once or more
OG	Organelle	Optional
OC	Organism classification	Once or more
OX	Taxonomy cross-reference(s)	Once or more
RN	Reference number	Once or more
RP	Reference position	Once or more
RC	Reference comment(s)	Optional
RX	Reference cross-reference(s)	Optional
RG	Reference group	Once or more (Optional if RA line)
RA	Reference authors	Once or more (Optional if RG line)
RT	Reference title	Optional
RL	Reference location	Once or more
CC	Comments or notes	Optional
DR	Database cross-references	Optional
KW	Keywords	Optional
FT	Feature table data	Optional
SQ	Sequence header	Once
(blanks)	Sequence data	Once or more
//	Termination line	Once; ends the entry

Vzhledem k tomu, že většina typů řádků byla popsána v předchozí kapitole, bude následovat pouze popis rozdílů mezi záznamy v databázích EMBL-Bank a UniProtKB.

3.5.2.1 Odlišnosti mezi formátem řádků použitých v obou databázích

ID řádek
Název záznamu může být tvořen 10 znaky (9 znaků v EMBL) a může začínat i číslicí. Záznamy ve Swiss-Prot neobsahují identifikátor divize. Typ molekuly u záznamů ve Swiss-Prot je vždy PRT (PRoTein). Délka molekuly je následována zkratkou AA (Amino Acids) místo zkratky BP (Base Pairs).

AC řádek
Formát tohoto řádku je stejný jako u EMBL-Bank, liší se pouze vlastní přístupová čísla. Přístupová čísla použitá v EMBL-Bank a v UniProtKB se nepřekrývají. Přístupové číslo v databázi UniProtKB je složeno ze šesti znaků, které mohou nabývat následujících hodnot:

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

Přístupová čísla v EMBL-Bank mohou být složena z 8 znaků, nebo ze šesti znaků, kde prvním znakem je písmeno (kromě O, P, Q) následované 5 číslicemi.

DT řádek
V databázi UniProtKB obsahuje každý záznam 3 DT řádky. První řádek je shodný jako v EMBL-Bank, druhý a třetí řádek obsahují informace o poslední změně sekvence a změně v popisných datech. Následují příklad řádků DT jednoho záznamu:

```
DT   21-JUL-1986 (Rel. 01, Created)
DT   23-OCT-1986 (Rel. 02, Last sequence update)
DT   01-APR-1990 (Rel. 14, Last annotation update)
```

DE řádek
Poslední DE řádek je ukončen tečkou a neobsahuje druhové určení organismu.

OS řádek
Tento řádek může obsahovat více různých organismů, jestliže se daná proteinová sekvence vyskytuje u více organismů. Poslední OS řádek je opět ukončen tečkou.

OG řádek
V databázi EMBL-Bank se rozlišují organely *Mitochondrion* a *Kinetoplast*, zatímco UniProtKB požívá pouze první označení. Podobně i u organel *Chloroplast* a *Plastid* se používá pouze první označení. OG řádek je zde ukončen tečkou.

RP a RC řádky
RP řádek se zde vyskytuje před řádkem RC.

RT řádek
Název citace je ukončen tečkou, otazníkem nebo vykřičníkem.

FT řádek
Formát těchto řádků je zcela odlišný. Tyto řádky opět obsahují informace o zajímavých oblastech nebo místech sekvence (vazební místa, posttranslační modifikace,…). Každý FT řádek obsahuje klíč, určující dané místo, počáteční a koncový bod popisované oblasti a vlastní popis. Podrobnější informace o struktuře těchto řádků lze nalézt v příslušné dokumentaci [5].

CC řádek
Řádky obsahující poznámky jsou v záznamech této databáze shrnuty do jednoho bloku, který je umístěn za blok informací o citacích v literatuře. Řádky s poznámkami mají definovaný tvar.

SQ řádek
Tento řádek obsahuje celkový počet aminokyselin v sekvenci (AA) na rozdíl od celkového počtu bázových párů. Není zde určen celkový počet jednotlivých aminokyselin v sekvenci, ale hmotnost molekuly a 64-bitové CRC pro uvedenou sekvenci.

3.5.2.2 Řádky vyskytující se pouze v databázi UniProtKB

GN řádek
GN (Gene Name) řádek obsahuje informace o genu, který sloužil jako předpis pro tvorbu dané proteinové sekvence. Je zde uveden jeden název genu a další známé názvy tohoto genu za klíčovým slovem *Synonyms.* Dále potom obsahuje označení místa daného genu v celém genomu a jméno příslušného ORF (open reading frame).

OX řádek
OX (Organism taxonomy cross-reference) řádek obsahuje odkazy na výše specifikovaný organismus do taxonomických databází. V současné době se používají odkazy do databáze NCBI.

3.5.2.3 Řádky vyskytující se pouze v databázi EMBL-Bank
V databázi UniProt se nepoužívají řádky FH, XX a SV. Řádky FH a XX neobsahují žádná data a v záznamech EMBL-Bank jsou použity pouze pro zvýšení čitelnosti záznamu. V databázi UniProtKB naopak nejsou použity pro zvýšení kompaktnosti záznamů. Řádek SV obsahuje označení verze sekvence nukleotidů a není důvod ho použít v databázi UniProtKB.

4. Závěr

Tato práce se zabývá dvěmi databázemi biologických dat: EMBL-Bank a UniProtKB. EMBL-Bank je databáze sekvencí nukleotidů a je udržována na základě spolupráce s databázemi DDBJ a GenBank. UniProt je databáze proteinových sekvencí, která se skládá ze dvou sekcí. Sekce Swiss-Prot obsahuje bohatě popsané záznamy o proteinových sekvencí a sekce TrEMBL obsahuje sekvence, které vznikl přeložením kódujících sekvencí z databází DDBJ/EMBL/GenBank. Obě tyto databáze spravuje Evropský institut bioinformatiky. Hlavním cílem tohoto institutu je poskytovat data všem vědcům, kteří o ně mají zájem. Institut se ve svých databázích snaží shrnout veškeré známé informace týkající se genomů, sekvencí DNA a RNA a proteinů. Data jsou ukládána do databází, které jsou volně přístupné.
V práci jsou uvedeny možnosti přístupů k datům uložených v těchto databázích i možnosti vkládání dat do těchto databází. Dále je zde popsána struktura obou databází.
Databáze EMBL-Bank i UniProtKB mají velmi podobnou strukturu. Jsou tvořeny záznamy o sekvencích (sekvence nukleotidů, proteinové sekvence). Každý záznam je složen z řádků různého typu. Všechny použité řádky mají specifikovaný formát a obsahují daný typ informací. Třída každého záznamu je identifikována na prvním řádku záznamu. Typy řádků používaných v obou databázích jsou popsány v kapitolách 2.3.5 a 3.5.2.
Obě databáze představují hlavní evropské zdroje biologických dat a jejich záznamy obsahují velké množství odkazů na další biologické databáze. Pro biology je zde dostupné obrovské množství dat, která jsou přístupná ve srozumitelné a přehledné formě.

Literatura

1. Brooksbank C., Camon E., Midori A. Harris, Magrane M., Maria Jesus Martin, Mulder N.,Claire O'Donovan, Parkinson H., Tuli M. A., Apweiler R., Birney E., Brazma A., Henrick K., Lopez R., Stoesser G., Stoehr P., Cameron G: **The European Bioinformatics Institute's data resources**, Nucleic Acids Research, 2003, vol. 31, p. 43-50
2. Guenter Stoesser, Mary Ann Tuli, Rodrigo Lopez Peter Sterk: **The EMBL Sequence Database**, Nucleic Acids Research, 1997, vol. 27, p. 18-24
3. **User Manual Release 82 Mar 2005**, URL: http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html
4. **The DDBJ/EMBL/GenBank Feature Table: definition**, URL: http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html
5. **UniProt knowledgebase user manual**, URL: http://www.expasy.org/sprot/userman.html
6. European Bioinformatics Institute, URL: http://www.ebi.ac.uk/