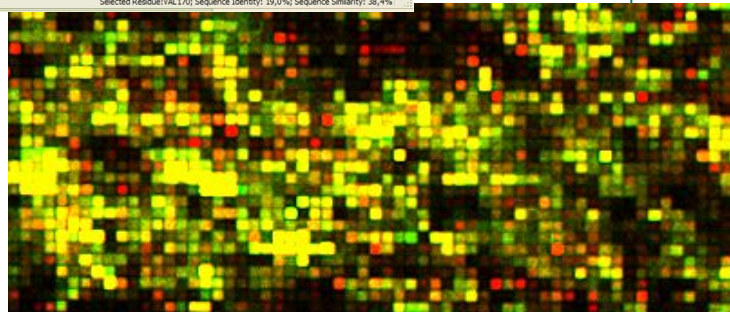
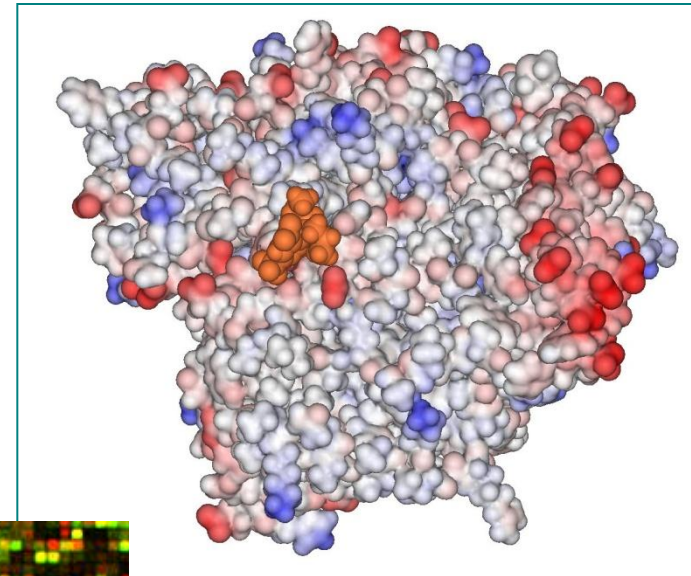
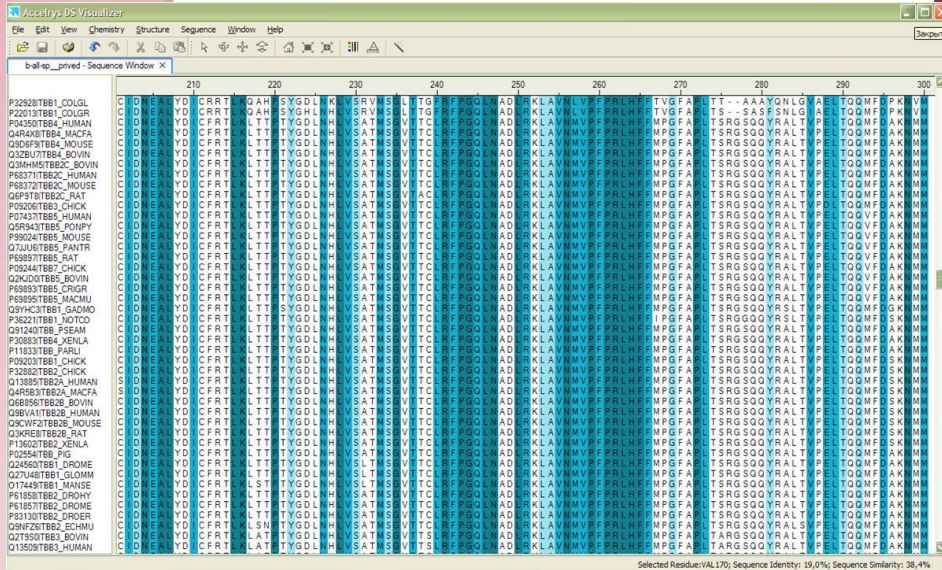


# БІОІНФОРМАТИК

к.б.н. Нидорко О.  
О.



# Банки (бази) даних – це:

## Колекції

### структурованих

### індексованих

дає можливість проведення пошуку за заданими критеріями (зокрема, містить таблицю заголовків та інші дескриптори – англ. поняття “searchable”)

### періодично


### оновлюємих

поповнюються новими даними, які включаються в нові випуски (релізи) банка

### крос-реферованих


містить перехресні гіперпосилання на інформацію в інших банках даних. Останнє дає змогу взаємної інтеграції різнорідних, але взаємопов'язаних даних та об'єднує існуючі банки даних в єдину інформаційну систему

## даних



## Банки даних обов'язково містять в себе також набір програмних інструментів, які забезпечують :

- доступ до банку даних та виконання запитів (пошукових та ін.);
- оновлення інформації в банку;
- додавання нової інформації;
- видалення помилкової/застарілої інформації.



Найчастіше, та сама інформація існує в різних форматах у різних базах даних, і різні сервери надають ті самі дані, але різними більш-менш дружелюбними стосовно користувача способами.

Вибір бази даних залежить від характеру розв'язуваної проблеми й від персональних переваг користувача.

Вибір сервера може навіть завісити від часу доби й завантаженості (кількості користувачів)



## Еволюція баз даних

- Books, articles | 1968 -> 1985
- Computer tapes | 1982 -> 1992
- Floppy disks | 1984 -> 1990
- CD-ROM | 1989 -> ?
- FTP | 1989 -> ?
- On-line services | 1982 -> 1994
- WWW | 1993 -> ?
- DVD | 2001 -> ?

# Структурна класифікація банків даних

Всі існуючі БД можуть бути класифіковані певним чином, зокрема їх підрозділяють:

на **первинні** та **вторинні** (похідні) БД,

на **архівні**, **куровані** та **автоматизовані** БД.

Також інколи в окремий клас виділяють інтегровані бази даних.

# Архівні БД

Архівні БД характеризуються тим, що вся відповідальність за інформацію, яка міститься в цих базах, лежить на дослідниках, що її тут розміщують.

Достовірність цієї інформації визначається добросовісністю самих дослідників; фахівці, що організовують і підтримують ці БД не несуть відповідальності за їхній вміст. Типовими прикладами архівних БД є GenBank, ProteinDataBank.

# Куровані

## БД

- Вміст записів курованих БД визначається спеціальними експертами (кураторами), які безпосередньо формують інформаційне наповнення цих банків даних. Надійність/достовірність інформації в курованих БД значно вище, ніж в архівних. Найбільш відомим прикладом курованої БД був банк SwissProt, яка містив записи щодо амінокислотних послідовностей (зараз це частина UniProtKnowledgeBase, що анотується і перевіряється вручну

# Автоматичні

## БД

- Вміст автоматичних БД, як видно з назви, генерується за допомогою комп'ютерних програм і веб-сервісів на основі інформації, що міститься в архівних (рідше керованих) БД. Типовим прикладом автоматичної БД була база амінокислотних послідовностей TrEMBL, записи в якій формувалися автоматично на основі нуклеотидних послідовностей (мРНК або кДНК), розміщених в банку нуклеотидних послідовностей EMBL.



# Інтегровані

## БД

Інтегровані бази даних містять різноманітну інформацію (архівну, керовану, згенеровану автоматично), яка підбирається за принципом систематизованого опису певних біологічних об'єктів. Типовими прикладами інтегрованих баз даних є спеціалізовані геномні бази, кожна з яких присвячена окремому біологічному виду: TAIR (геном резушки *Arabidopsis thaliana*), FlyBase (геном дрозоді) та ін.


# Первинні БД

Під первинними базами даних, як правило, розуміють бази, які містять безпосередні результати молекулярно-біологічних експериментів, зокрема дані щодо послідовностей біополімерів (білків та нуклеїнових кислот) та їх просторових структур (в атомарному масштабі).

## Вторинні

### БД

Вторинні або, похідні БД містять т.зв. **процесовану** інформацію, тобто, інформацію, яка виникає в результаті обробки і аналізу вмісту первинних баз даних, які відбуваються за певними правилами. Таким чином, ці БД містять відфільтровану інформацію про певні властивості біологічних молекул. Прикладами вторинних БД є бази даних структурної класифікації білків SCOP та CATH, бази даних білкових доменів SMART, Pfam, ProSite, геномна база даних Ensembl та ін.



Незалежно від типу банку даних, записи/статті банку містять певні поля (метадані), що дозволяють індексувати вміст банка даних за певними критеріями, здійснювати запити до банка та забезпечувати обмін інформацією між різнорідними банками даних.

# Типові

**Accession Number** □ **Унікальний** ідентифікатор статті, дозволяє формувати швидкі запити до неї

**Source** та/або **Taxonomy** □ описує біологічний об'єкт та його систематичне положення,  
**Annotation** □ короткий опис того, що власне міститься в статті

**Reference** □ посилання на літературні та інші джерела інформації

**KeyWords** □ ключові поняття та терміни, що мають безпосереднє відношення до статті

**Cross-reference** □ посилання на інші бази даних, які містять суміжну інформацію.



# Основные биоинформатические базы данных

- **Основные БД последовательностей: EMBL, GeneBank, UniProt, SwissProt.**

Производные PFAM, PROSITE, INTERPRO, *dbEST*, *dbSNP*.....

- **БД 3D-структур: PDB.**

Производные *SCOP*, *CATH*, *RNABase*.....

- **БД и энциклопедии, в которых подробно описаны функции генов и их продуктов : KEGG, BIOSYS, ENZYME, TC-DB, REACTOME.....**

- **Онтологии : GO, OBO, HUGO.....**

# Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- 3D structure
- Protein domain/family (----> tools)
- Proteomics (2D gel, Mass Spectrometry)
- Metabolism
- Bibliography
- 'Others' (Microarrays, Protein protein interaction...)

# Sequence databases

1. DNA/RNA
2. Proteins

## Ideal minimal content of a sequence database entry

- Sequences !!
- Accession number (AC) (unique identifier)
- Taxonomic data
- References
- ANNOTATION/CURATION
- Keywords
- Cross-references
- Documentation

## Sequence Databases: some « technical » definitions

- Data storage management:
  - flat file: text file, human readable
  - relational database (e.g., Oracle, Postgres)
  - object oriented database
  
- Format:
  - Fasta, RAW
  - GCG
  - NBRF/PIR
  - MSF....



# Sequence database : example

## SWISS-PROT (protein db) (flat file)

### Accession number

ID EPO\_HUMAN STANDARD; PRT; 193 AA.  
AC P01588; Q9UHA0; Q9UEZ5; Q9UDZ0;  
DT 21-JUL-1986 (Rel. 01, Created)  
DT 21-JUL-1986 (Rel. 01, Last sequence update)  
DT 20-AUG-2001 (Rel. 40, Last annotation update)  
DE Erythropoietin precursor.

### Taxonomy

GN EPO.  
OS Homo sapiens (Human).  
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.  
OX NCBI\_TaxID=9606;

### Reference

RN [1]  
RP SEQUENCE FROM N.A.  
RX MEDLINE=85137899; PubMed=3838366;  
RA Jacobs K., Shoemaker C., Rudersdorf R., Neill S.D., Kaufman R.J.,  
RA Mufson A., Seehra J., Jones S.S., Hewick R., Fritsch E.F.,  
RA Kawakita M., Shimizu T., Miyake T.;  
RT "Isolation and characterization of genomic and cDNA clones of human  
erythropoietin."  
RL Nature 313:806-810(1985).

### Annotations (comments)

....  
CC -!- FUNCTION: ERYTHROPOIETIN IS THE PRINCIPAL HORMONE INVOLVED IN THE  
CC REGULATION OF ERYTHROCYTE DIFFERENTIATION AND THE MAINTENANCE OF  
A  
CC PHYSIOLOGICAL LEVEL OF CIRCULATING ERYTHROCYTE MASS.  
CC -!- SUBCELLULAR LOCATION: SECRETED.  
CC -!- TISSUE SPECIFICITY: PRODUCED BY KIDNEY OR LIVER OF ADULT MAMMALS  
CC AND BY LIVER OF FETAL OR NEONATAL MAMMALS.  
CC -!- PHARMACEUTICAL: Available under the names Epogen (Amgen) and  
CC Procrit (Ortho Biotech).

### Cross-references

...  
DR EMBL; X02158; CAA26095.1; -.  
DR EMBL; X02157; CAA26094.1; -.  
DR EMBL; M11319; AAA52400.1; -.  
DR EMBL; AF053356; AAC78791.1; -.  
DR EMBL; AF202308; AAF23132.1; -.  
DR EMBL; AF202306; AAF23132.1; JOINED.

### Keywords

....

# Sequence database: example (cont.)

Annotations  
(features)

```

FT   SIGNAL                1     27
FT   CHAIN                 28    193
FT   PROPEP                190   193
FT   DISULFID              34    188
FT   DISULFID              56     60
FT   CARBOHYD              51     51
FT   CARBOHYD              65     65
FT   CARBOHYD             110    110
FT   CARBOHYD             153    153
FT   VARIANT               131    132
FT
FT
FT   VARIANT               149    149
FT
FT   CONFLICT              40     40
FT   CONFLICT              85     85
FT   CONFLICT             140    140
**
** ##### INTERNAL SECTION #####
**CL 7q22;
SQ   SEQUENCE   193 AA;  21306 MW;  C91F0E4C26A52033 CRC64;
      MGVHECPAWL WLLLSLLSLP LGLPVLGAPP RLICDSRVLE RYLLEAKEAE NITTGCAEHC
      SLNENITVPD TKVNFYAWKR MEVGQQAVEV WQGLALLSEA VLRGQALLVN SSQPWEPLQL
      HVDKAVSGLR SLTTLLRALG AQKEAISPPD AASAAPLRTI TADTFRKLFY VYSNFLRGKL
      KLYTGEACRT GDR
//

```

Sequence

# Sequence database: example

...The fasta format:

> My\_Sequence\_Name

```
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAE  
NITTGCAEHCSLNENITVPDTKVNIFYAWKRMEVGQQAVEVWQGLALLSEA  
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTLLRALGAQKEAISPPD  
AASAAPLRTITADTFRKLFVYSNFLRGKCLKLYTGEACRTGDR
```

...The RAW format:

```
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAE  
NITTGCAEHCSLNENITVPDTKVNIFYAWKRMEVGQQAVEVWQGLALLSEA  
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTLLRALGAQKEAISPPD  
AASAAPLRTITADTFRKLFVYSNFLRGKCLKLYTGEACRTGDR
```

# Database I: nucleotide sequences

- The 3 main public nucleic acid sequence databases are EMBL (Europe)/GenBank (USA) /DDBJ (Japan)  
« different views of the same data set » within 2 to 3 days
- EMBL: since 1982
- Specialized databases for the different types of RNAs (i.e. tRNA, rRNA, tmRNA, uRNA, etc...)
- 3D structure (DNA and RNA) - □ PDB
- Others: Aberrant splicing db; Eukaryotic promoter db (EPD); RNA editing sites, Multimedia Telomere Resource .....

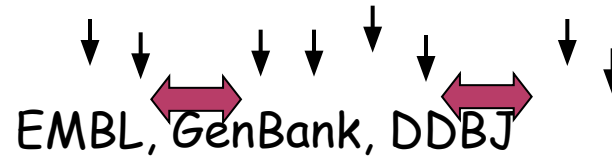
# Real life of a protein sequence ...

Data not submitted to public databases, delayed or cancelled...

cDNAs, ESTs, genomes, ...



*with or without  
annotated CDS  
provided by authors*



**CDS**  
**CoDing Sequence**  
portion of DNA/RNA translated into protein  
(from Met to STOP)



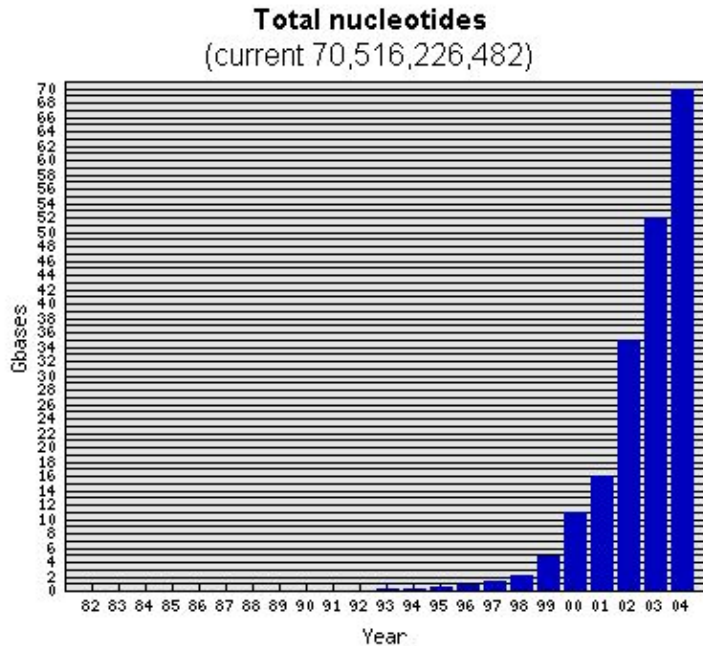
# International Nucleotide Sequence Database Collaboration

(EMBL/GenBank/DDBJ)

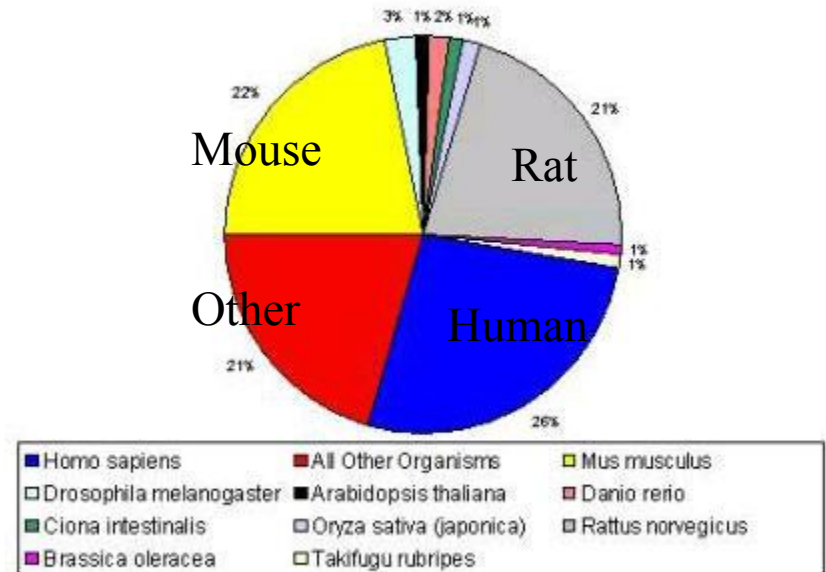
- Serve as **archives**
- Contain all **public** sequences derived from:
  - Genome projects (> 80 % of entries)
  - Sequencing centers (cDNAs, ESTs...)
  - Individual scientists ( 15 % of entries)
  - Patent offices (i.e. European Patent Office, EPO)
- Currently: 106,533,156,756 bases in 108,431,692 sequence records

# The tremendous increase in nucleotide sequences (1980-2004)

More than 50'000 species, but...



1980: 80 genes fully sequenced !



Human/Mouse/Rat:  
Organisms with the highest  
redundancy !

CC Data kindly reviewed (24-FEB-1986) by K. Jacobs

FH Key Location/Qualifiers

FH

FT source I..3398

FT /db\_xref=taxon:9606

FT /organism=Homo sapiens

FT mRNA join(397..627,1194..1339,1596..1682,2294..2473,2608..3327)

FT CDS join(615..627,1194..1339,1596..1682,2294..2473,2608..2763)

FT /db\_xref=SWISS-PROT:P01588

FT /product=erythropoietin

FT /protein\_id=CAA26095.1

FT /translation=MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLLE

FT AKEAENITTGCAEHCSLNENITVPTDKVNFYAWKRMEVGQQAQAVEVWQGLALLSEAVLRG

FT QALLVNSSQPWEPLQLHVDKAVSGLRSLTLLLRALGAQKEAISPPDAASAAPLRTITAD

FT TFRKLFVYSNFLRGKLLKLYTGEACRTGDR

FT mat\_peptide join(1262..1339,1596..1682,2294..2473,2608..2763)

FT /product=erythropoietin

FT sig\_peptide join(615..627,1194..1261)

FT exon 397..627

FT /number=1

FT intron 628..1193

FT /number=1

FT exon 1194..1339

FT /number=2

FT intron 1340..1595

FT /number=2

FT exon 1596..1682

FT /number=3

FT intron 1683..2293

FT /number=3

FT exon 2294..2473

FT /number=4

FT intron 2474..2607

FT /number=4

FT

CDS  
CoDing Sequence  
(proposed by submitters)

Annotation

(Prediction or  
experimentally determined)

sequence

## EMBL/GenBank/DDBJ

Sort of sequence museum, where sequences are preserved for eternity as they were determined, **interpreted and published originally by their authors**  
(primary sequence repository)

The authors have full authority over the content of the entries they submit !  
(exception: TPA, since january 2003)

# EMBL/GenBank/DDBJ

- Unexpected information you can find in these db:

```
FT      source          1..124
FT              /db_xref="taxon:4097"
FT              /organelle="plastid:chloroplast"
FT              /organism="Nicotiana tabacum"
FT              /isolate="Cuban cahibo cigar, gift from
FT              President Fidel Castro"
```

- Or:

```
FT      source          1..17084
FT              /chromosome="complete mitochondrial genome"
FT              /db_xref="taxon:9267"
FT              /organelle="mitochondrion"
FT              /organism="Didelphis virginiana"
FT              /dev_stage="adult"
FT              /isolate="fresh road killed individual"
FT              /tissue_type="liver"
```

## The second generation of nucleotide sequence databases

### Gene-centric databases

All the sequence information relevant to a given gene is made accessible at once

i.e. Locus Link/RefSeq

### Genome-centric databases

Information about gene sequence, relative position, strand orientation, biochemical functions...

Information management systems that are able to connect specialized sequence collection and **browsing tools**

i.e. Ensembl, TIGR

Working with whole genome databases:

## Genome-centric databases

« Browsing resources »

Remark: Genome-centric databases give usually access to several genomes, but some are « specialized » in particular organisms, i.e. TIGR: bacteria and plants

Search:  for

e.g. [human gene BRCA2](#) or [rat X:100000..200000](#) or [coronary heart disease](#)

### Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.








Click on a link below to go to the species' home page.

**Popular genomes** ([Log in to customize this list](#))




### New to Ensembl?

Did you know you can:

-  [Learn how to use Ensembl](#)  
with our video tutorials and walk-throughs
-  [Add custom tracks](#)  
using our new Control Panel
-  [Upload and analyse your data](#)  
and save it to your Ensembl account
-  [Search for a DNA or protein sequence](#)  
using BLAST or BLAT
-  [Fetch only the data you want](#)  
from our public database, using the Perl API
-  [Download our databases via FTP](#)  
in FASTA, MySQL and other formats
-  [Mine Ensembl with BioMart](#)  
and export sequences or tables in text, html, or Excel format

#### Did you know...?

A preliminary assembly of the Sheep (*Ovis aries*) genome is now available on our pre! site, <http://pre.ensembl.org/sheep>





# Database 2: protein sequences

- **UNIPROT:**
- **PIR-PSD:** Protein Information Resources

-> UniProt

- **Genpept:** « proteomic » version of GenBank (~TrEMBL)
- Many specialized protein databases for specific families or groups of proteins.

Examples: **AMSDb** (antibacterial peptides), **GPCRDB** (7 TM receptors), **IMGT** (immune system) **YPD** (Yeast) etc.

Swiss-Prot -> ExPASy  
([www.expasy.org](http://www.expasy.org));

Since 1986  
TrEMBL -> EBI (European Bioinformatics Institute)  
([www.ebi.ac.uk/trembl/](http://www.ebi.ac.uk/trembl/)).

Since 1996

# In a UniProt entry, you can expect to find:

- All the names of a given protein (and of its gene);
- Its biological origin with links to the taxonomic databases;
- A selection of references;
- A summary of what is known about the protein: function, alternative products, PTM, active sites, tissue expression, disease, etc....;
- Numerous cross-references;
- Selected keywords;
- A description of important sequence features: domains, variations, etc.;
- A (often corrected) protein sequence and the description of various isoforms/variants.

# View « by default » on the ExPASy server

## NiceProt View of SWISS-PROT: **Q75144**

[\[General\]](#)
[\[Name and origin\]](#)
[\[References\]](#)
[\[Comments\]](#)
[\[Cross-references\]](#)
[\[Keywords\]](#)
[\[Features\]](#)
[\[Sequence\]](#)
[\[Tools\]](#)

General information about the entry	
Entry name	ICOL_HUMAN
Primary accession number	Q75144
Secondary accession numbers	Q5I-R01; Q2HD1E
Entered in SWISS-PROT in	Release 38, July 1999
Sequence was last modified in	Release 43, October 2001
Annotations were last modified in	Release 41, March 2002

Name and origin of the protein	
Protein name	ICOL ligand [Precursor]
Synonyms	B7 homolog 2 B7-H2 CD280 B7-related protein 1 BTRP 1
Gene name	ICOL1 or B7H2 or BTRP1 or KIAA0653
From	<a href="#">Homo sapiens (Human)</a> [TaxID 9606]
Taxonomy	Eucaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primatei; Hominidae; Homo

References	
[1]	SEQUENCE FROM NUCLEIC ACID (ISOFORM 1). <b>TISSUE</b> =Derivative cell. MEDLINE=2077816, PubMed=11003515, [NCBI ExPASy, EBI, Israel, Japan] Wang S, Zhu G, Chavesol A L, Dong H, Tomoda K, Hiji, Cher L, "Characterization of T cells by B7-H2, a B7-like molecule that binds ICOS". <a href="#">Blood 95:2308-2313(2000)</a>
[2]	SEQUENCE FROM NUCLEIC ACID (ISOFORM 1), AND CHARACTERIZATION <b>TISSUE</b> =Peripheral blood mononuclear. MEDLINE=20465019, PubMed=11307763, [NCBI ExPASy, EBI, Israel, Japan] Yoshimura S K, Zhang M, Bando T, Horai T, Zhang S L, Maser S, Schenerson M, Boone T, Brinkley D, Du T, Dehney J, Han H, Hui A, Kobayashi T, Matsuoka K, Whittow J S, Coburn M A, "Characterization of a new human B7-related protein, B7H2 (ICOL1)." <a href="#">Immunol 124:1439-1447(2000)</a>
[3]	SEQUENCE FROM NUCLEIC ACID (ISOFORM 2) <b>TISSUE</b> =Leukocyte. MEDLINE=20126021, PubMed=10557606, [NCBI ExPASy, EBI, Israel, Japan] Luo V, Wang W, Finerty S K, Bean K M, Spadina V, Foster L A, Leonard J Z, Hunter S B, Zoller K, Thomas J L, Miyashita J S, Jacobs R A, Collier M, "Identification of GLD-0, a novel B7-like protein that functionally binds to ICOS receptor". <a href="#">J Immunol 164:1624-1627(2000)</a>
[4]	SEQUENCE FROM NUCLEIC ACID. <b>TISSUE</b> =Brain. MEDLINE=98403880, PubMed=49734811, [MCE, ExPASy, EBI, Israel, Japan] Ichihara K, L, Nagata T, Szymanski M, Okura H, Tachibana A, Kohno H, Nishimura N, Ohara O, "Prediction of the coding sequences of unidentified human genes. X: The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro". <a href="#">EMBL Rep 5:169-176(1993)</a>
[5]	SEQUENCE FROM NUCLEIC ACID (ISOFORM 2) <b>TISSUE</b> =T-helper-2-lymphoblastoid cell. "GEO molecules and user thereof". Feature number W00_21296, 29 MAR 2001.

Comments	
•	<b>FUNCTION</b> LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND CYTOKINE SECRETION, INDUCES ALSO B-CELL PROLIFERATION AND DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN MEDIATING LOCAL TISSUE RESPONSES TO THE IMMUNATORY CONDITIONS, AS WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY COSTIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
•	<b>SUBCELLULAR LOCATION</b> Type I membrane protein (By similarity).
•	<b>ALTERNATIVE PRODUCTS</b> AT LEAST 2 ISOFORMS, 1 (SHOWN HERE) AND 2 ARE PRODUCED BY ALTERNATIVE SPLICING.
•	<b>TISSUE SPECIFICITY</b> ISOFORM 1 IS WIDELY EXPRESSED IN BONE, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE MARROW, FOALON, COLON, LUNG, PANCREAS, TESTIS, THYROID NODULES, LEUKOCYTES, SPLEEN, THYMUS AND TONSIL. WHILE ISOFORM 2 IS DETECTED ONLY IN LYMPH NODES, LEUKOCYTES AND SPLEEN.
•	<b>INDUCTION</b> CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY TREATMENT WITH THE ALPHA IN PERIPHERAL BLOOD B-CELLS AND MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
•	<b>SIMILARITY</b> BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY - B7/MOG SUBFAMILY.
•	<b>SIMILARITY</b> CONTAINS 1 IMMUNOGLOBULIN-LIKE V-TYPE DOMAIN.
•	<b>SIMILARITY</b> CONTAINS 1 IMMUNOGLOBULIN-LIKE C2-TYPE DOMAIN.
•	<b>CAUTION</b> Ref.4 sequence differs from that shown in position 300 onward for an unknown reason.

**Copyright**  
This SWISS-PROT entry is copyright © by its producer. Through a collaboration between the Swiss Institute of Bioinformatics and the EMBL database, the European Bioinformatics Institute. There are no restrictions on the use of any parts of this entry for any purpose, provided that the original source is properly acknowledged.

Cross-references	
EMBL	AF199728, AF647391, AF289028, AAC01176, AF216749, AAK15241, ABC4453, BAA31528, AX100595, CAC36765, ALT_SEQ
MDM	605717 [NCBI/EBI]
GeneCards	ICOL1
GeneLans	ICOL1
Protein	Q75144
HUGL	KIAA0653
InterPro	IP003322, IP003095, IP003860, IP003860, IP003860
Plan	PF00047, pf. 3
SMART	SM00409, IG_1, SM00410, IG_1b, IG_1c
ProDom	[Domain structure / List of seq. sharing a leaf: 1 domain]
BLOCKS	Q75144
ProteinMap	Q75144
PIESAGE	Q75144
LIF	Q75144
ModBase	Q75144
SWISS 2DPAGE	GET SEQUENCE ON 2DPAGE

Keywords	
B-cell activation, Immune response, Glycoprotein, Immunoglobulin domain, Signal, Transmembrane, Multigene family, Alternative splicing	
Features	
Key	From To Length Description
REGION	1 18 18 POTENTIAL
CHAIN	19 302 284 ICOL1 (CHAIN)
DOMAIN	19 356 238 INTRACELLULAR (POTENTIAL)
TRANSMEM	257 277 21 POTENTIAL
DOMAIN	278 302 25 CYTOPLASMIC (POTENTIAL)
DOMAIN	30 150 91 IG-LIKE V-TYPE DOMAIN
DOMAIN	151 223 73 IG-LIKE C2-TYPE DOMAIN
DISTILED	37 113 POTENTIAL
DISTILED	230 216 POTENTIAL
CARBOHYD	20 20 N-LINKED (GLYCAN...) (POTENTIAL)
CARBOHYD	137 137 N-LINKED (GLYCAN...) (POTENTIAL)
CARBOHYD	173 173 N-LINKED (GLYCAN...) (POTENTIAL)
CARBOHYD	184 184 N-LINKED (GLYCAN...) (POTENTIAL)
CARBOHYD	225 225 N-LINKED (GLYCAN...) (POTENTIAL)
VERSITL7	100 101 GIN -> ENHANCEMENT (BY ISOFORM 2)

Sequence information					
Length: 302 AA [This is the length of the unprocessed precursor]	Molecular weight: 33349 Da [This is the MW of the unprocessed precursor]				
Checksum: 64793421855E334A [This is a checksum on the sequence]					
10	20	30	40	50	60
MLDGLDGLT	LLFSLDADT	QKQVAVNV	QKVELDCACT	EGSNFMDLV	VYVWQKSK
70	80	90	100	110	120
IVNTRIPQI	SSMIDVQYK	ENRSLDAS	ELDFNSLFL	IVTPDQKQ	FNQGLDLSL
130	140	150	160	170	180
GFQVLSVEV	TLVAVNSV	VYVAVRPS	QDELTPQS	...	...
190	200	210	220	230	240
DQALDQIVT	LVNGLVYV	QVLRDPTD	VNIQDQIV	LLQQLTSS	LTQNLQED
250	260	270	280	290	300
CTTNSVSTG	RSVAVTIL	LVSLDQVA	VATQDQKQ	GLQVYAVK	RSVPTFLTG

Feature aligner

Feature table viewer

Sequence

Q75144.FA:FASTA format

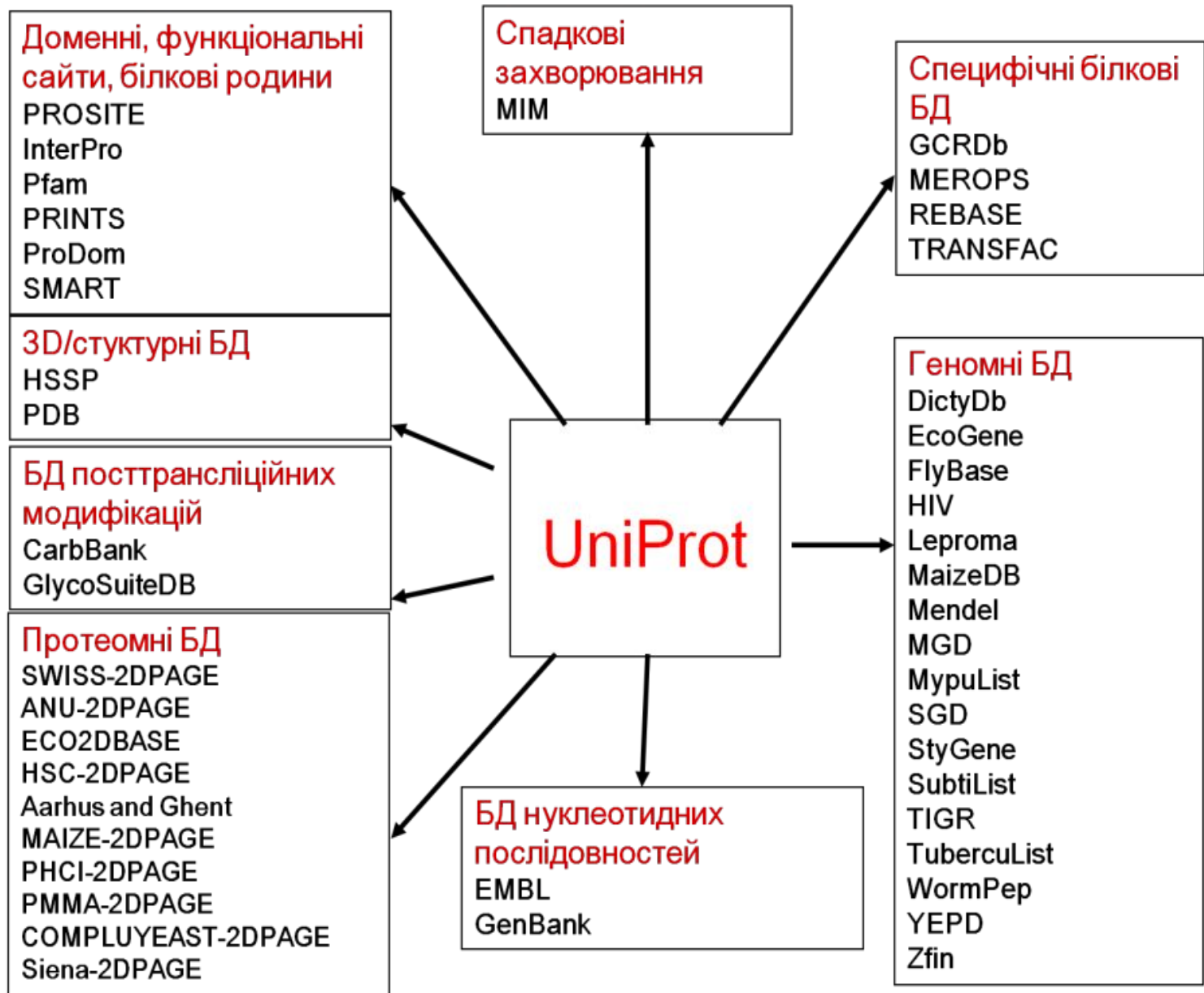
# Annotation/Curation (Comment lines)

- **Function(s) and role(s); enzymes:**
  - a. **Catalytic activity (if EC number)**
  - b. **Cofactor**
  - c. **Enzyme regulation**
  - d. **Pathway**
- **Subunit (Protein/protein interactions)**
- **Subcellular location**
- **Alternative products (alt. splicing, alt. initiation, RNA editing)**
- **Tissue specificity (Northern and Western results)**
- **Developmental stage**
- **Induction**
- **Domain**
- **Post-translational modifications (PTM)**
- **Mass spectrometry**
- **Polymorphisms**
- **Disease**
- **Pharmaceutical**
- **Miscellaneous**
- **Similarities**
- **Caution**
- **Database (specialized cross-references)**

# Annotation/Curation (Comment lines)

Information is derived from:

- Publications;
- Databases;
- Personal communication;
- Prediction;
- Brain storming...





# Cross-references

Cross-references		ICE8_HUMAN Q14790
	X98172; CAA66853.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	X98173; CAA66854.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	X98174; CAA66855.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	X98175; CAA66856.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	X98176; CAA66857.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	X98177; CAA66858.1; ALT_SEQ	[EMBL / GenBank / DDBJ] [CoDingSequence]
	X98178; CAA66859.1; ALT_SEQ	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U58143; AAC50602.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	U60520; AAC50645.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF102146; AAD24962.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF102139; AAD24962.1; JOINED	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF102140; AAD24962.1; JOINED	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF102141; AAD24962.1; JOINED	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF102142; AAD24962.1; JOINED	[EMBL / GenBank / DDBJ] [CoDingSequence]
EMBL	AF102143; AAD24962.1; JOINED	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF102144; AAD24962.1; JOINED	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF102145; AAD24962.1; JOINED	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF009620; AAB70913.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AB038985; BAB32555.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AB038982; BAB32555.1; JOINED	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AB038983; BAB32555.1; JOINED	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AB038984; BAB32555.1; JOINED	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF380342; AAK57437.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF422925; AAL87628.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF422926; AAL87629.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF422927; AAL87630.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF422928; AAL87631.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF422929; AAL87632.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	BC028223; AAH28223.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
	PDB	1QDU; 10-JUL-00. [ExpASY / RCSB]
	MEROPS	C14.009; -
	Genew	HGNC:1509; CASP8.
	CleanEx	HGNC:1509; CASP8.
	MM	601763 [NCBI / EBI]. 607271 [NCBI / EBI].
	GeneCards	CASP8
	GeneLynx	CASP8; Homo sapiens.
	SOURCE	CASP8; Homo sapiens.
	Ensembl	Q14790; Homo sapiens. [Entry / Contig view]

Examples of implicit links to GenBank and DDBJ added 'on the fly' by the ExpASY server

ADN  
(Index of low redundancy)

3D



genomic





# Sequence description:

Derived from:

- Publications;
- Databases;
- Personal communication;
- Prediction.

## Features

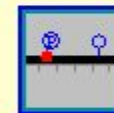
ICOL\_HUMAN, O75144

Key	From	To	Length	Description
SIGNAL	<u>1</u>	<u>18</u>	18	POTENTIAL.
CHAIN	<u>19</u>	<u>302</u>	284	ICOS LIGAND.
DOMAIN	<u>19</u>	<u>256</u>	238	EXTRACELLULAR (POTENTIAL).
TRANSMEM	<u>257</u>	<u>277</u>	21	POTENTIAL.
DOMAIN	<u>278</u>	<u>302</u>	25	CYTOPLASMIC (POTENTIAL).
DOMAIN	<u>30</u>	<u>120</u>	91	IG-LIKE V-TYPE DOMAIN.
DOMAIN	<u>151</u>	<u>223</u>	73	IG-LIKE C2-TYPE DOMAIN.
DISULFID	<u>37</u>	<u>113</u>		POTENTIAL.
DISULFID	<u>158</u>	<u>216</u>		POTENTIAL.
CARBOHYD	<u>70</u>	<u>70</u>		N-LINKED (GLCNAC...) (POTENTIAL).
CARBOHYD	<u>137</u>	<u>137</u>		N-LINKED (GLCNAC...) (POTENTIAL).
CARBOHYD	<u>173</u>	<u>173</u>		N-LINKED (GLCNAC...) (POTENTIAL).
CARBOHYD	<u>186</u>	<u>186</u>		N-LINKED (GLCNAC...) (POTENTIAL).
CARBOHYD	<u>225</u>	<u>225</u>		N-LINKED (GLCNAC...) (POTENTIAL).
VARSP LIC	<u>300</u>	<u>302</u>		GHV -> ESWNLLLLLS (IN <u>ISOFORM 2</u> ).

PTM



[Feature aligner](#)



[Feature table viewer](#)

# Databases 3: 'genomics'

- Contain informations on gene chromosomal location (mapping) and nomenclature, and provide links to sequence databases; *has usually no sequence*;
- Exist for most organisms important in life science research; usually species specific.
- Examples: TAIR (Arabidopsis), FlyBase (Drosophila), MaizeDB (maize), SubtiList (B.subtilis), etc.;

# Databases 4: mutation/polymorphism

- Contain informations on sequence variations linked or not to genetic diseases;
- Mainly human but: OMIA - Online Mendelian Inheritance in Animals
- **General db:**
  - OMIM
  - HMGD - Human Gene Mutation db
  - SVD - Sequence variation db
  - HGBASE - Human Genic Bi-Allelic Sequences db
  - dbSNP - Human single nucleotide polymorphism (SNP) db
- **Disease-specific db:** most of these databases are either linked to a single gene or to a single disease;
  - p53 mutation db
  - ADB - Albinism db (Mutations in human genes causing albinism)
  - Asthma and Allergy gene db
  - ....

# Mutation/polymorphism: definitions

- **SNPs:** single nucleotide polymorphisms; occur approximately once every 100 to 300 bases  
(distinction between sequencing error and polymorphism !)
- **c-SNPs:** coding single nucleotide polymorphisms (Single Nucleotide Polymorphisms within cDNA sequences)
- **SAPs:** single amino-acid polymorphisms
- Missense mutation: -> SAP
- Nonsense mutation: -> STOP
- Insertion/deletion of nucleotides -> frameshift...

# Database 5: protein domain/family

## Protein domain/family: some definitions

- Most proteins have « modular » structures
- Estimation: ~ 3 domains / protein



# Protein domain/family databases

- Contains biologically significant « pattern / profiles/ HMM » formulated in such a way that, with appropriate computational tools, it can rapidly and reliably determine to which known family of proteins (if any) a new sequence belongs to
- Used as a tool to identify the function of uncharacterized proteins translated from genomic or cDNA sequences (« functional diagnostic »)
- Either manually curated (i.e. PROSITE, PfamA, PRINTS, SMART, TIGRFAM etc.) or automatically generated (i.e. PfamB, ProDom, DOMO)



# Protein domain/family db

PROSITE	Patterns / Profiles
ProDom	Aligned motifs (PSI-BLAST) (Pfam B)
PRINTS	Aligned motifs
Pfam	HMM (Hidden Markov Models)
SMART	HMM
TIGRfam	HMM

I  
n  
t  
e  
r  
p  
r  
o

DOMO	Aligned motifs
BLOCKS	Aligned motifs (PSI-BLAST)
CDD	Pfam and SMART

-> A Conserved Domain Database and Search Service

# Prosite <http://www.expasy.org/prosite/>

- Created in 1988 (SIB)
- Contains functional domains fully annotated, based on two methods: patterns and profiles
- Entries are deposited in PROSITE in two distinct files:
  - Pattern/profiles with the list of all matches in SWISS-PROT
  - Documentation

# PFAM (HMMs): an entry

<http://www.sanger.ac.uk/Software/Pfam/>



Protein families database of alignments and HMMs



EPO\_TPO

Home Search by Browse by ftp iPfam Help



Figure 1: 1eer  
Complex (cytokine/receptor)

Crystal structure of human erythropoietin complexed to its receptor at 1.9 angstroms

1buu

Accession number: PF00758

## Erythropoietin/thrombopoietin

[Add Annotation](#)

This family forms **structural complexes** with other Pfam families, to view them click [here](#)

### INTERPRO description (entry [IPRO01323](#))

Erythropoietin, a plasma glycoprotein, is the primary physiological mediator of erythropoiesis [PUBMED:3773894](#). It is involved in the regulation of the level of peripheral erythrocytes by stimulating the differentiation of erythroid progenitor cells, found in the spleen and bone marrow, into mature erythrocytes [PUBMED:3346214](#). It is primarily produced in adult kidneys and foetal liver, acting by attachment to specific binding sites on erythroid progenitor cells, stimulating their differentiation [PUBMED:2877922](#). Severe kidney dysfunction causes reduction in the plasma levels of erythropoietin, resulting in chronic anaemia - injection of purified erythropoietin into the blood stream can help to relieve this type of anaemia. Levels of erythropoietin in plasma fluctuate with varying oxygen tension of the blood, but androgens and prostaglandins also modulate the levels to some extent [PUBMED:2877922](#). Erythropoietin glycoprotein sequences are well conserved, a consequence of which is that the hormones are cross-reactive among mammals, i.e. that from one species, say human, can stimulate erythropoiesis in other species, say mouse or rat [PUBMED:1420369](#).

Thrombopoietin (TPO), a glycoprotein, is the mammalian hormone which functions as a megakaryocytic lineage specific growth and differentiation factor affecting the proliferation and maturation from their committed progenitor cells acting at a late stage of megakaryocyte development. It acts as a circulating regulator of platelet numbers.

### QuickGO

FUNCTION :	hormone activity ( <a href="#">GO:0005179</a> )
COMPONENT :	extracellular ( <a href="#">GO:0005576</a> )

For additional annotation, see the [PROSITE](#) document [PDOC00644](#) [[Expasy](#) | [SRS-UK](#) | [SRS-USA](#)]

### Alignment

Seed (7)  Full (32)

Format

Further alignment options [here](#)  
Help relating to Pfam alignments [here](#)

### Domain organisation

View 1 representative architecture  
 View architectures for 32 proteins

Zoom  pixels/aa.

## InterPro

[www.ebi.ac.uk/interpro](http://www.ebi.ac.uk/interpro)

- Search simultaneously many domain databases.
- Single set of documents linked to the various methods;
- InterPro release 8.0 contains 11007 entries, representing 2573 domains, 8166 families, 201 repeats, 26 active sites, 21 binding sites and 20 post-translational modification sites.



## IPR002967 Delta tubulin

## Protein matches

UniProtKB Matches: 89 proteins	Overview:	<a href="#">sorted by AC</a> ,	<a href="#">sorted by name</a> ,	<a href="#">of known structure</a> ,	<a href="#">proteins with splice variants</a>
	Detailed:	<a href="#">sorted by AC</a> ,	<a href="#">sorted by name</a> ,	<a href="#">of known structure</a>	<a href="#">proteins with splice variants</a>
	Table:	<a href="#">For all matching proteins</a> , <a href="#">of known structure</a>			
	<a href="#">Architectures</a> <a href="#">Accession List</a> <a href="#">Matches in BioMart</a>				

**Accession** IPR002967 Delta\_tubulin

**Type** Family

Database ID	Name	Proteins
<a href="#">PRINTS</a> <a href="#">PR01224</a>	DELATUBULIN	80
<a href="#">PANTHER</a> <a href="#">PTHR11588:SF4</a>	Delta_tubulin	73

[Signatures in BioMart](#)

## InterPro Relationships

**Parent** [IPR000217](#) Tubulin

[IPR003008](#) Tubulin/FtsZ, GTPase domain



# Databases 6: proteomics

- Contain informations obtained by 2D-PAGE: images of master gels and description of identified proteins
- Examples: SWISS-2DPAGE, ECO2DBASE, Maize-2DPAGE, Sub2D, Cyano2DBase, etc.
- Composed of image and text files
- There is currently no protein Mass Spectrometry (MS) database (not for long...)

# Databases 7: 3D structure




# Формати структурних даних

- правила та засоби зберігання даних щодо просторової структури макромолекул

- базова інформація – просторове розташування атомів в молекулі

описується за допомогою просторових координат – декартових або внутрішніх





# Формат PDB (Protein Data Bank) – один з основних форматів зберігання молекулярних даних

забезпечує стандартне представлення молекулярних структур, отриманих за допомогою рентгенівської/електронної кристалографії та ЯМР-спектроскопії

розроблений в 1971 році, підтримується будь-яким програмним забезпеченням в галузі структурної біології



**оперує декартовими  
координатами**

**всі записи прив'язані до певних  
полів**

**Остання версія керівництва по формату PDB**

—

**Atomic Coordinate Entry Format Description  
Version 3.1, July 19, 2007**

<http://www.wwpdb.org/documentation/format3.1-20070719.pdf>

Каждый файл имеет свой идентификатор или код : 1ea1

*Title:* Cytochrome P450 14 $\alpha$ -Sterol Demethylase (Cyp51) From Mycobacterium Tuberculosis In Complex With Fluconazole

*Compound:* **Mol\_Id:** 1; **Molecule:** Cytochrome P450 51-Like Rv0764C; **Chain:** A;  
**Synonym:** Cyp51, 14 $\alpha$ -Sterol Demethylase; **Engineered:** Yes; **Mutation:** Yes;  
**Other\_Details:** Cys 394 Binds Heme Iron. Fluconazole Is Bound In The Active Site Coordinating Heme Iron As A Sixth Ligand

*№ат ИмяА*

*ИмяОст*

*ИмяЦепи*

*НомерОстатка*

					<i>X</i>	<i>Y</i>	<i>Z</i>	<i>occ</i>	<i>B-factor</i>	
ATOM	1	N	ALA	A	3	-14.763	-13.683	100.347	1.00	49.82
ATOM	2	CA	ALA	A	3	-14.759	-13.806	98.856	1.00	50.45
ATOM	3	C	ALA	A	3	-13.343	-14.025	98.327	1.00	49.54
ATOM	4	O	ALA	A	3	-12.509	-13.118	98.350	1.00	48.65
ATOM	5	CB	ALA	A	3	-15.371	-12.553	98.227	1.00	51.16
ATOM	6	N	VAL	A	4	-13.088	-15.237	97.845	1.00	49.20
ATOM	7	CA	VAL	A	4	-11.781	-15.620	97.314	1.00	48.77
ATOM	8	C	VAL	A	4	-11.248	-14.700	96.215	1.00	46.92

.....

.....

# Типи записів в заголовному розділі

● **HEADER** – описує надходження банку через унікальний номер, класифікацію та дату депонування

	1	2	3	4	5	6	7
1234567890123456789012345678901234567890123456789012345678901234567890							
HEADER	MUSCLE	PROTEIN			02-JUN-93	1MYS	
HEADER	HYDROLASE	(CARBOXYLIC ESTER)			08-APR-93	2PHI	
HEADER	COMPLEX	(LECTIN/TRANSFERRIN)			07-JAN-94	1LGB	

**OBSLTE** – яке надходження замінене ПОТОЧНИМ

	1	2	3	4	5	6	7
1234567890123456789012345678901234567890123456789012345678901234567890							
OBSLTE	31-JAN-94	1MBP	2MBP				

# Типи записів в заголовному розділі

● **TITLE** – описує експеримент та аналіз надходження

```
1 2 3 4 5 6 7
123456789012345678901234567890123456789012345678901234567890
TITLE      RHIZOPUSPEPSIN COMPLEXED WITH REDUCED PEPTIDE INHIBITOR

TITLE      BETA-GLUCOSYLTRANSFERASE, ALPHA CARBON COORDINATES ONLY

TITLE      NMR STUDY OF OXIDIZED THIOREDOXIN MUTANT (C62A,C69A,C73A)
TITLE      2 MINIMIZED AVERAGE STRUCTURE
```

**SAVEAT** – повідомляє про помилки хіральності

# Типи записів в заголовному розділі

## COMPND – описує макромолекулярний компонент надходження

```

1          2          3          4          5          6          7
1234567890123456789012345678901234567890123456789012345678901234567890
COMPND    MOL_ID: 1;
COMPND    2 MOLECULE: HEMOGLOBIN;
COMPND    3 CHAIN: A, B, C, D;
COMPND    4 ENGINEERED: YES;
COMPND    5 MUTATION: YES
COMPND    6 OTHER_DETAILS: DEOXY FORM

COMPND    MOL_ID: 1;
COMPND    2 MOLECULE: COWPEA CHLOROTIC MOTTLE VIRUS;
COMPND    3 CHAIN: A, B, C;
COMPND    4 SYNONYM: CCMV;
COMPND    5 MOL_ID: 2;
COMPND    6 MOLECULE: RNA (5'-( *AP*UP*AP*U)-3' );
COMPND    7 CHAIN: D, F;
COMPND    8 ENGINEERED: YES;
COMPND    9 MOL_ID: 3;
COMPND   10 MOLECULE: RNA (5'-( *AP*U)-3' );
COMPND   11 CHAIN: E;
COMPND   12 ENGINEERED: YES

COMPND    MOL_ID: 1;
COMPND    2 MOLECULE: HEVAMINE A;
COMPND    3 CHAIN: A;
COMPND    4 EC: 3.2.1.14, 3.2.1.17;
COMPND    5 OTHER_DETAILS: PLANT ENDOCHITINASE/LYSOZYME
```



# розшифровка деталей запису COMPND

TOKEN	VALUE DEFINITION
MOL_ID	Numbers each component; also used in SOURCE to associate the information
MOLECULE	Name of the macromolecule.
CHAIN	Comma-separated list of chain identifier(s).
FRAGMENT	Specifies a domain or region of the molecule.
SYNONYM	Comma-separated list of synonyms for the MOLECULE.
EC	The Enzyme Commission number associated with the molecule. If there is more than one EC number, they are presented as a comma-separated list.
ENGINEERED	Indicates that the molecule was produced using recombinant technology or by purely chemical synthesis.
MUTATION	Indicates if there is a mutation.
OTHER_DETAILS	Additional comments.

# Типи записів в заголовному розділі

**SOURCE** – описує біологічне та/або хімічне джерело кожної біологічної молекули в надходженні

```
          1           2           3           4           5           6           7
123456789012345678901234567890123456789012345678901234567890
SOURCE    MOL_ID: 1;
SOURCE    2 ORGANISM_SCIENTIFIC: AVIAN SARCOMA VIRUS;
SOURCE    3 STRAIN: SCHMIDT-RUPPIN B;
SOURCE    4 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE    5 EXPRESSION_SYSTEM_PLASMID: PRC23IN
```

```
SOURCE    MOL_ID: 1;
SOURCE    2 ORGANISM_SCIENTIFIC: GALLUS GALLUS;
SOURCE    3 ORGANISM_COMMON: CHICKEN;
SOURCE    4 ORGAN: HEART;
SOURCE    5 TISSUE: MUSCLE
```

```
SOURCE    MOL_ID: 1;
SOURCE    2 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE    3 EXPRESSION_SYSTEM_STRAIN: BE167;
SOURCE    4 FRAGMENT: RESIDUES 1-16;
SOURCE    5 ORGANISM_SCIENTIFIC: BACILLUS AMYLOLIQUEFACIENS;
SOURCE    6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE    7 FRAGMENT: RESIDUES 17-214;
SOURCE    8 ORGANISM_SCIENTIFIC: BACILLUS MACERANS
```



# Типи записів в заголовному розділі

## EXPDTA – містить інформацію щодо експерименту

```

      1           2           3           4           5           6           7
123456789012345678901234567890123456789012345678901234567890
EXPDTA      X-RAY DIFFRACTION

EXPDTA      NEUTRON DIFFRACTION; X-RAY DIFFRACTION

EXPDTA      NMR, 32 STRUCTURES

EXPDTA      NMR, REGULARIZED MEAN STRUCTURE

EXPDTA      FIBER DIFFRACTION
```

```

ELECTRON DIFFRACTION
ELECTRON MICROSCOPY
CRYO-ELECTRON MICROSCOPY
SOLUTION SCATTERING, THEORETICAL MODEL
FIBER DIFFRACTION
FLUORESCENCE TRANSFER
NEUTRON DIFFRACTION
NMR (may have a qualifier e.g. number of models see examples below)
SOLUTION SCATTERING
THEORETICAL MODEL*
X-RAY DIFFRACTION
```

# Типи записів в заголовному розділі

**KEYWDS – ключові слова, що стосуються надходження**

```
1           2           3           4           5           6           7
123456789012345678901234567890123456789012345678901234567890
KEYWDS      LYASE, TRICARBOXYLIC ACID CYCLE, MITOCHONDRION, OXIDATIVE
KEYWDS      2 METABOLISM
```

**AUTHOR – імена людей, що відповідають за надходження**

```
1           2           3           4           5           6           7
123456789012345678901234567890123456789012345678901234567890
AUTHOR      M.B.BERRY, B.MEADOR, T.BILDERBACK, P.LIANG, M.GLASER,
AUTHOR      2 G.N.PHILLIPS JUNIOR, T.L.ST. STEVENS
```

**REVDAT – історія внесення змін в**

```
1           2           3           4           5           6           7
123456789012345678901234567890123456789012345678901234567890
REVDAT      3      15-OCT-89 1PRC      1      REMARK
REVDAT      2      19-APR-89 1PRC      2      CONECT
REVDAT      1      09-JAN-89 1PRC      0
```

# Типи записів в заголовному розділі

**SPRSDE – які застарілі надходження  
замінені на поточне**

```
      1           2           3           4           5           6           7
1234567890123456789012345678901234567890123456789012345678901234567890
SPRSDE      17-JUL-84 4HHB           1HHB

SPRSDE      27-FEB-95 1GDJ           1LH4 2LH4
```

**JRNL – основне літературне джерело, яке  
описує результати, депоновані в**

```
      1           2           3           4           5           6           7
1234567890123456789012345678901234567890123456789012345678901234567890
JRNL      AUTH      G.FERMI,M.F.PERUTZ,B.SHAANAN,R.FOURME
JRNL      TITL      THE CRYSTAL STRUCTURE OF HUMAN DEOXYHAEMOGLOBIN AT
JRNL      TITL 2    1.74 A RESOLUTION
JRNL      REF       J.MOL.BIOL.           V. 175      159 1984
JRNL      REFN      ASTM JMOBAK   UK ISSN 0022-2836
```

**REMARK – різноманітна службова  
інформація.**

# PDB – міжнародний банк даних білкових структур

## <http://www.rcsb.org/>

Are you missing data updates? The PDB archive has moved to <ftp://ftp.wwpdb.org>. For more information click [here](#).

## Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data.

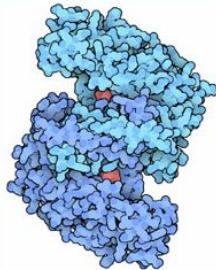
This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found [here](#).

A [narrated tutorial](#) illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? [info@rcsb.org](mailto:info@rcsb.org)

### Molecule of the Month: Citrate Synthase



Your body burns up a lot of food every day. However, cells don't burn food like a fireplace. Instead, food molecules are combined with oxygen molecules one-by-one, in many carefully controlled steps. In this way, the energy that is released can be captured in convenient forms, like ATP or NADH, which are then used elsewhere to power essential cellular functions. Our cells get most of their energy from a long series of reactions that combine oxygen and glucose, forming carbon dioxide and water, and creating lots of ATP and NADH in the process.

- More ...
- Previous Features

The RCSB PDB is managed by two members of the RCSB: Rutgers, The State University of New Jersey and the San Diego Supercomputer Center and Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California, San Diego. It is supported by funds from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of Science, Department of Energy (DOE), the National Library of Medicine (NLM), the National Cancer Institute (NCI), the National Center for Research Resources (NCRR), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institute of Neurological Disorders and Stroke (NINDS), and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

### News

- Complete News
- Newsletter
- Discussion Forum

18-September-2007

#### PDB Data Summaries

Various summaries of current data in the PDB archive are available through the /pub/pdb/derived\_data directory of the FTP site at <ftp://ftp.wwpdb.org>.

- Full article ...

11-September-2007

#### RCSB PDB Poster Prize Awarded, Art of Science shown at ISMB Meeting

- Full article ...

### Quick Tips:

Try the Web Services API for software developers using C/C++, Java, Python and Perl. [Click here.](#)

In citing the PDB please refer to: H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank, Nucleic Acids Research, 28 pp. 235-242 (2000).



# NDB - база даних просторових структур нуклеїнових кислот

<http://ndbserver.rutgers.edu/>



## WELCOME TO THE NUCLEIC ACID DATABASE

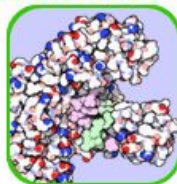
a repository of three-dimensional structural information about nucleic acids

- [Atlas](#)
- [Deposit Data](#)
- [Download Data](#)
- [Search](#)
- [Reports](#)
- [Education](#)
- [Standards](#)
- [Tools](#)
- [Links](#)

Number of Released Structures:  
**3557 Structures**  
Last Update: 19-June-2007

**Search the NDB by ID**  
Enter an NDB ID or PDB ID    
Search for Released Structures

### Nucleic Acids Highlight



### About NDB

#### NDB News

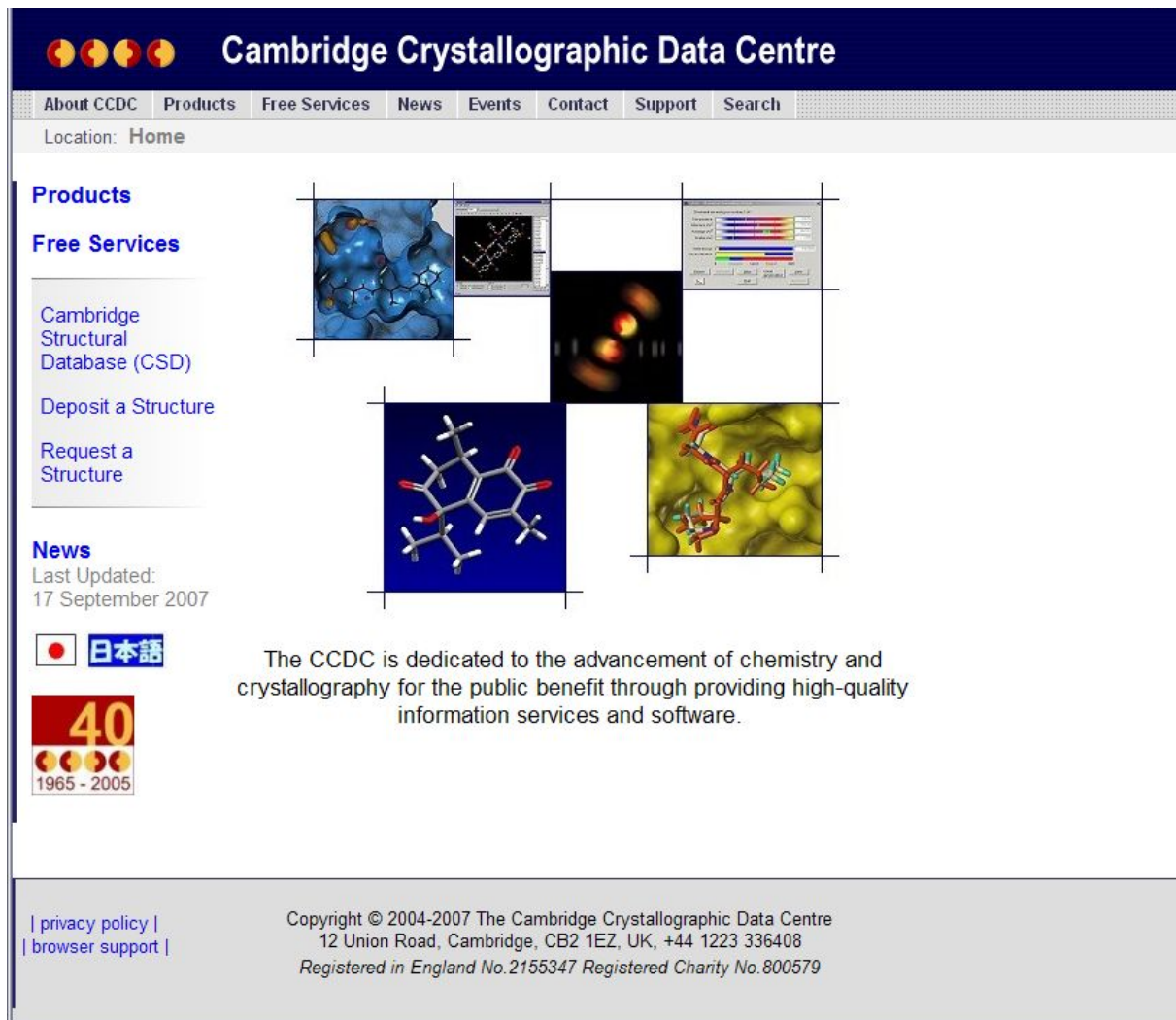
[Archive of NDB newsletters](#)

The NDB is supported by funds from the [National Science Foundation](#) and the [Department of Energy](#).

In citing the NDB please refer to:  
H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.-H. Hsieh, A. R. Srinivasan, and B. Schneider. (1992) The Nucleic Acid Database: A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys. J.*, 63, 751-759.

[ndbadmin@ndbserver.rutgers.edu](mailto:ndbadmin@ndbserver.rutgers.edu)  
©1995-2007 The Nucleic Acid Database Project. Rutgers, The State University of New Jersey

# CSD (Cambridge Crystallographic Data Centre) – банк кристалографічних даних низькомолекулярних сполук <http://www.ccdc.cam.ac.uk/>



The screenshot shows the homepage of the Cambridge Crystallographic Data Centre. At the top, there is a dark blue header with the logo (four colored circles) and the text "Cambridge Crystallographic Data Centre". Below the header is a navigation menu with links: "About CCDC", "Products", "Free Services", "News", "Events", "Contact", "Support", and "Search". The main content area is divided into several sections. On the left, there is a sidebar with "Products" and "Free Services" sections. The "Free Services" section includes links for "Cambridge Structural Database (CSD)", "Deposit a Structure", and "Request a Structure". Below this is a "News" section with the text "Last Updated: 17 September 2007". There is also a language selection button for Japanese (日本語) and a 40th anniversary logo (1965-2005). The main content area features a grid of images: a blue molecular model, a diffraction pattern, a 3D molecular model, and a yellow molecular model. At the bottom, there is a footer with a privacy policy link, browser support link, and copyright information: "Copyright © 2004-2007 The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK, +44 1223 336408. Registered in England No. 2155347 Registered Charity No. 800579".

**Cambridge Crystallographic Data Centre**

Location: Home

**Products**

**Free Services**

Cambridge Structural Database (CSD)

Deposit a Structure

Request a Structure

**News**

Last Updated: 17 September 2007

日本語

40  
1965 - 2005

The CCDC is dedicated to the advancement of chemistry and crystallography for the public benefit through providing high-quality information services and software.

[| privacy policy |](#)  
[| browser support |](#)

Copyright © 2004-2007 The Cambridge Crystallographic Data Centre  
12 Union Road, Cambridge, CB2 1EZ, UK, +44 1223 336408  
Registered in England No. 2155347 Registered Charity No. 800579

# VMRB - банк даних ЯМР-спектроскопії макромолекул

<http://www.bmrw.wisc.edu/>



Biological Magnetic Resonance Data Bank

Google Search

Search Archive

Deposit Data

NMR Statistics

Spectroscopists' Corner

Programmers' Corner

Home

Site Map

FTP Access

Structural Genomics  
and other "omics"

Metabolomics

Educational Outreach

NMR Data Formats

WWW Sites

Home



News

About BMRB

Feedback

FTP Access

BMRB List Server

BMRB Job Opportunity

BMRB  
BioMagResBank

A Repository for Data from NMR Spectroscopy  
on Proteins, Peptides, and Nucleic Acids

Department of Biochemistry  
University of Wisconsin-Madison

BMRB Data Listed By:

Macromolecular types

NMR spectral parameters

Kinetics

Thermodynamics

Restraints

Structure

Time-domain sets

Solid-state NMR

Search BMRB

Data Browser, FASTA Search of BMRB, NMR Restraints, Time-domain Data Sets

Deposit Data

ADIT NMR

BMRB Mirrors

Madison USA, Osaka Japan, Florence Italy

About BMRB

Mission Statement, Aims and Policies, Data Accepted, Distribution

Biomolecular

Highlight:

**Lysozyme**



BMRB is a member of the wwPDB



# PQS - база даних четвертинних структур білків

<http://pqs.ebi.ac.uk/>



Biological Magnetic Resonance Data Bank

Google Search

Search Archive

Deposit Data

NMR Statistics

Spectroscopists' Corner

Programmers' Corner

Home

Site Map

FTP Access

Structural Genomics  
and other "omics"

Metabolomics

Educational Outreach

NMR Data Formats

WWW Sites

Home



News

About BMRB

Feedback

FTP Access

BMRB List Server

BMRB Job Opportunity

BMRB  
BioMagResBank

A Repository for Data from NMR Spectroscopy  
on Proteins, Peptides, and Nucleic Acids

Department of Biochemistry  
University of Wisconsin-Madison

BMRB Data Listed By:

Macromolecular types

NMR spectral parameters

Kinetics

Thermodynamics

Restraints

Structure

Time-domain sets

Solid-state NMR

Search BMRB

Data Browser, FASTA Search of BMRB, NMR Restraints, Time-domain Data Sets

Deposit Data

ADIT NMR

BMRB Mirrors

Madison USA, Osaka Japan, Florence Italy

About BMRB

Mission Statement, Aims and Policies, Data Accepted, Distribution

Biomolecular

Highlight:

**Lysozyme**



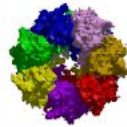
BMRB is a member of the wwPDB





# Profunc – аналіз структури для передбачення функцій

● <http://www.ebi.ac.uk/thornton-srv/databases/profunc/>



**ProFunc**

*Analysis of a protein's 3D structure to help identify its likely biochemical function*

The aim of the ProFunc server is to help identify the likely biochemical function of a protein from its three-dimensional structure. It uses a series of methods, including fold matching, residue conservation, surface cleft analysis, and functional 3D templates, to identify both the protein's likely active site and possible homologues in the PDB.

Some of the methods take minutes to run; others take hours. You will be notified by e-mail when the entire process is complete, but can check on preliminary results as they become available.

From this page you can submit your own structure, analyse an existing PDB entry, or retrieve the results of a previously submitted run.

Please note that if you have several structures to submit, try to limit the number of structures you upload to about 6 per hour to avoid overloading the server. If you wish to run a very large number of structures, please contact Roman Laskowski ([roman@ebi.ac.uk](mailto:roman@ebi.ac.uk)) about arranging some form of batch run.

Choose option A, B or C:

**A**

 Upload PDB-format file:

or

**B**

 Get existing PDB file\* PDB code:

\* The ProFunc analyses have already been computed for a number of existing PDB entries. In such cases you will be taken



# SCOP – структурна класифікація білків

<http://scop.mrc-lmb.cam.ac.uk/scop/>

Structural Classification of Proteins



Welcome to SCOP: Structural Classification of Proteins.

**1.71 release** (October 2006)

27599 PDB Entries. 1 Literature Reference. 75930 Domains. (excluding nucleic acids and theoretical models).

Folds, superfamilies, and families [statistics here](#).

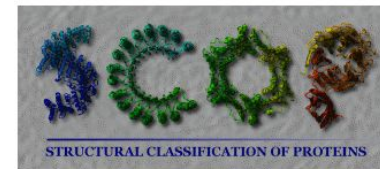
[New folds superfamilies families](#).

[List of obsolete entries and their replacements](#).

**Authors:** Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. [scop@mrc-lmb.cam.ac.uk](mailto:scop@mrc-lmb.cam.ac.uk)

**Reference:** Murzin A. G., Brenner S. E., 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. [\[PDF\]](#)

**Recent changes** are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [\[PDF\]](#) and Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229. [\[PDF\]](#).



## Access methods

- Enter SCOP at the [top of the hierarchy](#)
- [Keyword search of SCOP entries](#)
- [SCOP parseable files](#)
- [All SCOP releases and reclassified entry history](#)
- SCOP domain sequences and pdb-style coordinate files ([ASTRAL](#))
- Hidden Markov Model library for SCOP superfamilies ([SUPERFAMILY](#))
- **NEW** Structural alignments for proteins with non-trivial relationships ([SISYPHUS](#))

### **NEW** pre-SCOP - preview of the next release

We provide limited access to the SCOP development version ([pre-SCOP](#)) that classifies a substantial part of PDB entries released between January 2005 and April 2006. Priority was given to representatives of new sequence families targeted by structural genomics initiatives.

- [Online resources](#) of potential interest to SCOP users

SCOP [mirrors](#) around the world may speed your access.

## News

- SCOP has been updated to include all PDB entries released up to 18 January 2005. See [folds, superfamilies, and families statistics](#).
- This release, corresponding to the PDB snapshot at the beginning of 2005, is the last one that classifies all PDB structures released before a certain date. The process of classification of new entries has been changed. For more information please visit [pre-SCOP](#) - a preview of the next release.
- This release is similar in appearance to the previous release, so the generic [release notes](#) from that release still apply. Please read the notes; they contain more detailed explanations and examples of SCOP features.



# CATH – структурна класифікація білків

<http://cathwww.biochem.ucl.ac.uk/latest/index.html>

The screenshot shows the homepage of the CATH Protein Structure Classification website. The header features the CATH logo and navigation links for CATH, DHS, Gene3D, and FTP. The main content area is titled 'CATH Protein Structure Classification' and includes a search bar, a 'Goto' section with links to SSAP Server, CATHEDRAL Server, DHS, and Gene3D, and a 'Navigation' section with links to Home and Top of hierarchy. The main text area contains the following sections:

- Version 3.1.0: Released Jan 2007**
- CATH Group**: Dr. Alison Cuff, Dr. Ian Sillitoe, Dr. Mark Dibley, Mr. Tony Lewis, Mr. Oliver Redfern, Dr. Frances M.G. Pearl
- Contributors to the CATH Version 3.1.0 Release**: Ms. Sarah Addou, Mr. Tim Dallman, Mr. Benoit Dessailly, Dr. Lesley Greene, Dr. David Lee, Dr. Jon Lees, Dr. Russell L. Marsden, Mr. Adam Reid, Mr. Stathis Sideris, Dr. Corin Yeats, Prof. Janet Thornton, Prof. Christine A. Orengo
- Links**:
  - Browse or search the classification
  - CATH statistics and release information
  - General information on CATH
  - CATH lists and FTP site
  - [NEW]** Raw data files for CATH (including CATH Domain PDB files)
    - Full HMM Library (right-click link and select "Save as...")
    - Concatenated file of 7794 models representing all sequence families in CATH v3.1.0 (gzipped HMMER2.0 format, 83MB)
  - [NEW]** CrossLinks between superfamilies in CATH
  - DHS - Dictionary of Homologous Superfamilies. Summary of structural and functional features for CATH Homologous Superfamilies
  - CATH File Formats (for FTP files)
- Introduction**: CATH is a hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H). Class, derived from secondary structure content, is assigned for more than 90% of protein structures automatically. Architecture, which describes the gross orientation of secondary structures, independent of connectivities, is currently assigned manually. The topology level clusters structures into fold groups according to their topological connections and numbers of secondary structures. The homologous superfamilies cluster proteins with highly similar structures and functions. The assignments of structures to fold groups and homologous superfamilies are made by sequence and structure comparisons.

On the right side, there is a 'CATH v3.1.0 Release statistics' table and a 'Technical notes' section.

	v3.0.0	v3.1.0	New
<b>Domains</b>	86151	93885	7734
<b>Chains</b>	57741	63453	5712
<b>PDBs</b>	27522	30028	2506

...more

**Technical notes**  
This release has incorporated a great deal of internal development including:

- Development of backend PostgreSQL database
- Development of the central code library
- New web interface for domain chopping (DomChop)
- Improved public pages to show very latest information
- Added numerous maintenance scripts and regression tests

...more

# SCOR – структурна класифікація РНК

<http://scor.lbl.gov/>

## SCOR

### Structural Classification of RNA

[SCOR 1.1](#) [SCOR 1.2](#) [Programs](#) [Help](#) [PDB/NDB list](#) [Links](#) [SCOR XML files](#) [Release notes](#) [Glossary](#) [MeRNA](#) [Survey](#)

We will be seeking NIH funding for continued research, development and maintenance of the Structural Classification of RNA (SCOR) database. We are surveying current and potential users to learn who uses SCOR, and how it is used. This survey is important to our future development of SCOR. We ask that you please [take our survey](#).

SCOR 2.0.3, 24 Oct. 2004: 579 PDB entries (15 May 2004).  
5350 internal loops, 2920 hairpin loops.

- [Structural Classification](#)
- [Functional Classification](#)
- [RNA Tertiary Interaction](#)

PDB/NDB ID or Keyword ([help](#)):   [Advanced Search](#)

**Primary reference:** Klosterman PS, Tamura M, Holbrook SR, Brenner SE. 2002. SCOR: a structural classification of RNA database. *Nucleic Acids Res.* **30**: 392-394. [[pdf](#)], [additional references](#)

#### Synopsis

The Structural Classification of RNA (SCOR) is a database designed to provide a comprehensive perspective and understanding of RNA motif structure, function, tertiary interactions and their relationships. SCOR 2.0.3 provides a survey of the three-dimensional motifs contained in 579 NMR and X-ray RNA structures available as of May 15, 2004. This includes 8,270 secondary structural elements, of which 2,920 are hairpin loops and 5,350 are internal loops. The structural elements are organized in a directed acyclic graph (DAG) architecture, allowing multiple parent classes for a motif. Users can browse the database or search by PDB or NDB identifier, keyword or sequence. Descriptions and cartoon representations of each of the classes are available. RNA motifs reported in the literature (e.g. Kink turns, S-turns, GNRA loops) are incorporated into the classification.

**Authors:** Makio Tamura, Donna K. Hendrix, Peter S. Klosterman, Nancy R. B. Schimmelman, Steven E. Brenner and Stephen R. Holbrook. Supported by Lawrence Berkeley National Laboratory and NIGMS of the NIH.

#### Who is using SCOR?

Contact us at [scor@compbio.berkeley.edu](mailto:scor@compbio.berkeley.edu)  
Generated from scor database 2.0.3 on October 24, 2004  
Copyright © 2001-2005

# The Protein Kinase Resource – структури кіназ

<http://www.kinasenet.org/pkr/Welcome.do>

The screenshot shows the homepage of the Protein Kinase Resource. At the top, there is a navigation bar with the PKR logo and the text "Protein Kinase Resource". Below this, there is a search bar with a "Go" button and a "Site Help" link. The main content area features a welcome message and a 3D protein structure visualization. On the left side, there is a sidebar with a search bar and a list of navigation links. At the bottom, there is a footer with contact information and logos for SDSC, UCSD, and the Protein Kinase Resource.

**Protein Kinase Resource**

SEQUENCE STRUCTURE PKR FUNCTION

Search Site Help

Go

- + Search PKR
- + Sequence
- + Structure
- + Kinome Tree
- + Applications
- + Resources
- + Community
- + About PKR

The Protein Kinase Resource has been redesigned and expanded with new content and cross-links with other online resources. In addition, we have introduced several new technologies to integrate information search, retrieval and visualization. For support setting up the browser environment please refer to site help.

Any comments or suggestions will be greatly appreciated and will help us design a better resource for the benefit of the kinase research community.

Structure Search

Advanced Search

Browse Structures

Resources

The Protein Kinase Resource © 2004 | SDSC | UC San Diego, MC 0505  
9500 Gilman Dr. La Jolla, CA 92093-0505 Tel. 858-534-5000 Fax. 858-534-5152 info@sdsc.edu

SDSC UCSD

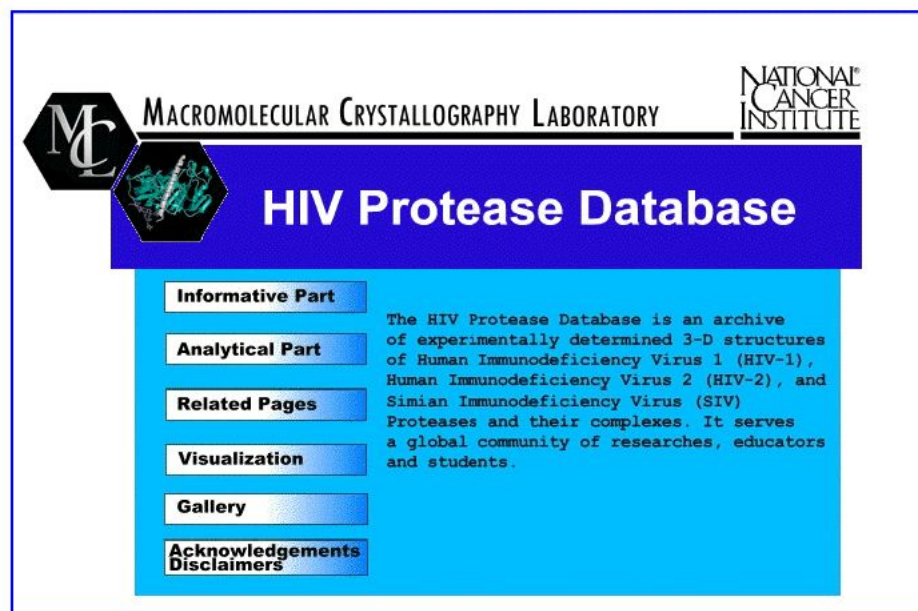


# HIV Protease Database

<http://mcll.ncifcrf.gov/hivdb/index.html>

The Database contains 272 structures.  
148 taken from PDB, 124 are unique entries for exclusive purpose of HIV PR Database. 10/23/2002

Check the last updated version of the HIVdb. 65 new structures added!



The screenshot shows the website's header with the Macromolecular Crystallography Laboratory (MCL) logo and the National Cancer Institute logo. The main title is "HIV Protease Database". Below the title is a navigation menu with buttons for "Informative Part", "Analytical Part", "Related Pages", "Visualization", "Gallery", and "Acknowledgements Disclaimers". A text block on the right describes the database as an archive of experimentally determined 3-D structures of HIV-1, HIV-2, and SIV proteases and their complexes.

This version of the database is for archival purposes only.  
The current version of database that offers better tools for its interrogation can be found at [NIST](#)

Questions about this site should be directed to [J. Vondrasek \(vondrasek@ncifcrf.gov\)](mailto:J.Vondrasek@ncifcrf.gov)  
Copyright © 2000 National Cancer Institute - Frederick Cancer Research and Development Center

You are visitor number

Click

Powered by [counter.bloke.com](http://counter.bloke.com)

# Databases 8: metabolic


- Contain informations that describe enzymes, biochemical reactions and metabolic pathways;
- ENZYME and BRENDA: nomenclature databases that store informations on enzyme names and reactions;
- Metabolic databases: EcoCyc (specialized on Escherichia coli), KEGG, EMP/WIT;

Usually these databases are tightly coupled with query software that allows the user to visualise reaction schemes.



BRENDA  
Useful to prepare  
lab's experiments !

<http://www.brenda.uni-koeln.de/>

 **BRENDA**  
The Comprehensive Enzyme Information System 

### Complete Entry of EC-Number 1.2.3.4

 [open printable version in a new window](#)

EC NUMBER	COFACTOR	LOCALIZATION
ORGANISM	SPECIFIC ACTIVITY [ $\mu\text{mol}/\text{min}/\text{mg}$ ]	PURIFICATION
SYSTEMATIC NAME	KM VALUE [mM]	MOLECULAR WEIGHT
RECOMMENDED NAME	PH OPTIMUM	SUBUNITS
SYNONYMS	PH RANGE	PH STABILITY
CAS REGISTRY NUMBER	TEMPERATURE OPTIMUM [°C]	TEMPERATURE STABILITY [°C]
REACTION	METALS, IONS	STORAGE STABILITY
REACTION TYPE	INHIBITORS	LINKS TO OTHER DATABASES
SUBSTRATES/PRODUCTS	SOURCE/TISSUE	REFERENCES

<\_>= reference; #\_#=organism

**EC NUMBER**  
1.2.3.4

**ORGANISM**  
#1# *Sorghum vulgare* (variety CSH-5 <1>; CSH-1 <3>) <1, 3>  
#2# *Pseudomonas* sp. (OX-53) <2>  
#3# *Hordeum vulgare* (barley) <4>  
#4# *Tilletia contraversa* (parasitic fungus) <5>  
#5# *Hylocomium splendens* <6>  
#6# *Rhytidiadelphus squarrosus* <6>  
#7# *Hylocomium loreum* <6>

**SYSTEMATIC NAME**  
Oxalate: oxygen oxidoreductase

**RECOMMENDED NAME**  
Oxalate oxidase

**SYNONYMS**  
Oxidase, oxalate  
Aero-oxalo dehydrogenase  
Oxalic acid oxidase #4# <5>

**CAS REGISTRY NUMBER**  
9031-79-2

**REACTION**  
Oxalate + O<sub>2</sub> = 2 CO<sub>2</sub> + H<sub>2</sub>O<sub>2</sub>

**REACTION TYPE**  
Redox reaction

**SUBSTRATES/PRODUCTS**  
S 1. Oxalate + O<sub>2</sub> #1-7# (almost specific for oxalate, #2# <2>; no other electron acceptor found, #3# <4>) <1-6>  
P 1. CO<sub>2</sub> + H<sub>2</sub>O<sub>2</sub>  
S 2. **Additional information:** #2# (oxidation at extreme low rate: glyoxylic acid, DL-malic acid, citric acid, not oxidized: succinic acid, formic acid, <2>  
P 2. ?

**SPECIFIC ACTIVITY [ $\mu\text{mol}/\text{min}/\text{mg}$ ]**



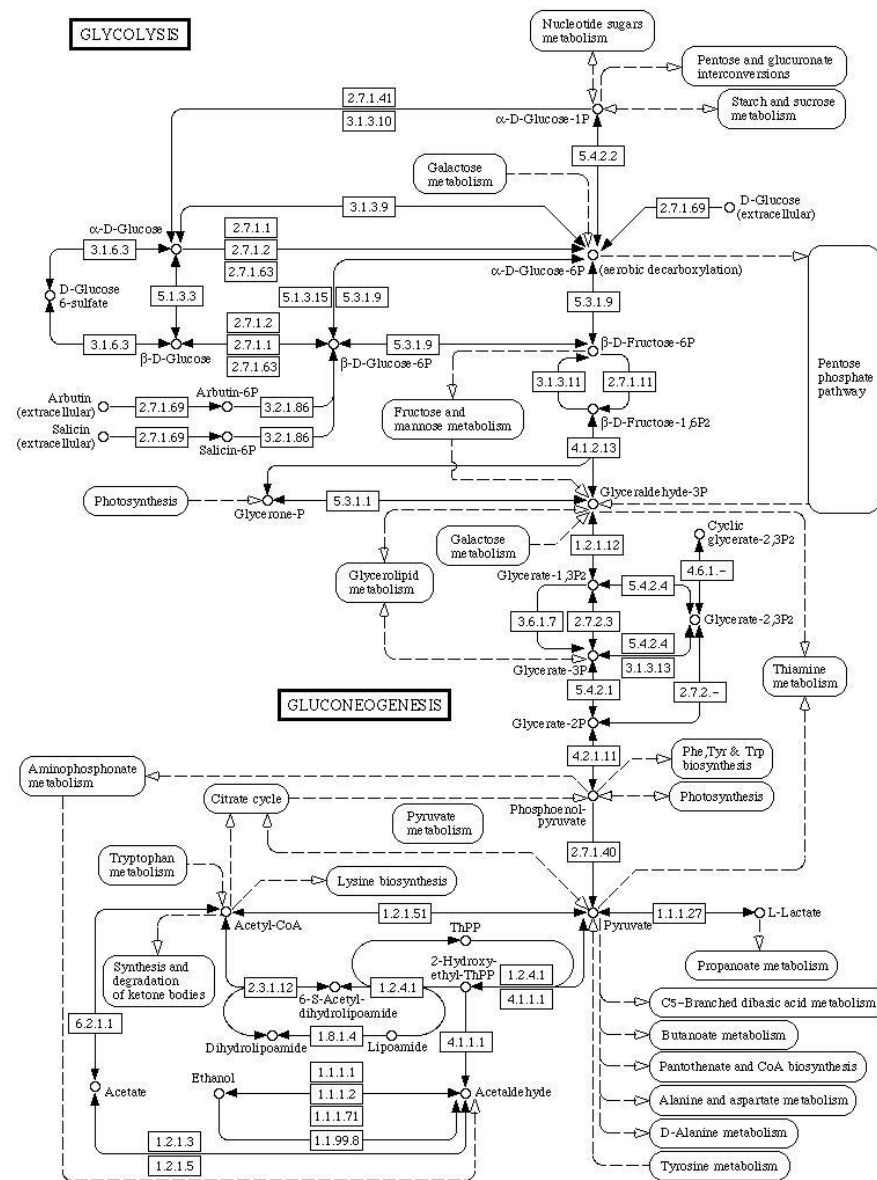
KEGG: Kyoto Encyclopedia of Genes and Genomes

<http://www.genome.ad.jp/kegg>

Glycolysis / Gluconeogenesis - Reference pathway

Go to: [ [LinkDB search](#) | [Ortholog Table](#) ]

Go to: Reference pathway [ Exec ]



# Databases 9: bibliographic

- Bibliographic reference databases contain citations and abstract informations of published life science articles;
- Example: Medline
- Other more specialized databases also exist (i.e. Agricola <http://agricola.nal.usda.gov/>).

# Databases 10: others

- There are many databases that cannot be classified in the categories listed previously;
- Examples: ReBase (restriction enzymes), TRANSFAC (transcription factors), CarbBank, GlycoSuiteDB (linked sugars), Protein-protein interactions db (Intact, BIND), Protease db (MEROPS), biotechnology patents db, etc.;
- As well as many other resources concerning any and new aspects of macromolecules and molecular biology (Ex: Microarrays).

# Proliferation of databases

- What is the best db for sequence analysis ?
- Which does contain the highest quality data ?
- Which is the more comprehensive ?
- Which is the more up-to-date ?
- Which is the less redundant ?
- Which is the more indexed (allows complex queries) ?
- Which Web server does respond most quickly ?
- .....??????

# Some important practical remarks

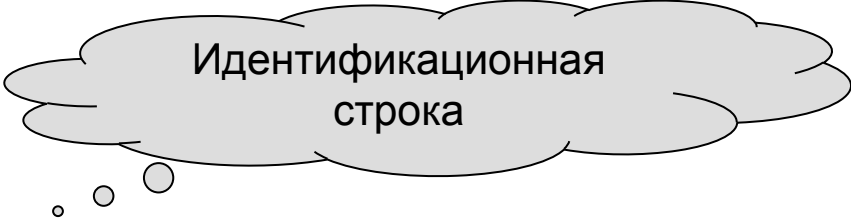
- Databases: many errors (automated annotation) !
- Not all db are available on all servers
- The update frequency is not the same for all servers; creation of db\_new between releases
- Some servers add automatically useful cross-references to an entry (implicit links) in addition to already existing links (explicit links)

# Представление аминокислотной последовательности в Raw формате:

```
MSEPQRLFFAIDLPAEIREQIIHWRATHFPPEAGRPVAADNLHLT  
LAFLGEVSAEKEKALSLLAGRIRQPGFTLTLDDAGQWLRSRVWVWL  
GMRQPPRGLIQLANMLRSQAARSGCFQSNRPFHPHITLLRDASEA  
VTIPPPGFNWSYAVTEFTLYASSFARGRTRYTPLKRWALTQ
```

# FASTA - формат

FASTA - популярная программа предназначенная для выравнивания последовательностей и сканирования баз данных, созданная W.R. Peerson и D.J. Lipman в 1988 году.



Идентификационная строка

>My\_Sequence\_Name

```
MSEPQRLFFAIDLPAEIREQIIHWRATHFPPEAGRPVAADNLHLT  
LAFLGEVSAEKEKALSLLAGRIRQPGFTLTLDDAGQWLRSRVWVW  
GMRQPPRGLIQLANMLRSQAARSGCFQSNRPFHPHITLLRDASEA  
VTIPPPGFNWSYAVTEFTLYASSFARGRTRYTPLKRWALTQ
```



- уникальный идентификатор



# ПРИМЕР:

идентификатор ресурс идентификационный номер  
откуда взялась (по данным литературы)  
первичный номер краткое описание

```
>gi|4885609|ref|NP_005408.1| proto-oncogene tyrosine-protein kinase  
SRC [Homo sapiens] организ
```

MGSNKSKPKDASQRRRSLEPAENVHMAAGGGAFFPASQTPSKPASADGHRGPSAAFAPAAAEPKLFGGFNSS  
DTVTSPQRAGPLAGGVTTFFVALYDYESRTE TDLSFKKGERLQIVNNTTEGDWLLAHSLSSTGQTGYIPSNYV  
APSDSIQAEWYFGKITRRESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVKHYKIRKL  
DSGGFYITSRTQFNSLQQLVAYYSKHADGLCHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGC  
FGEVVMGTWNGTTRVAIKTLKPGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPYIYIVTEYMSKGSLL  
DFLKGETGKYLRPLQQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARLIEDNEYT  
ARQGAKFPIKWTAPAAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLDQVERGYRMPCPPEC  
PESLHDLMCQCWRKEPEERPTFEYLLQAFLEDYFTSTEPQYQGENL

# Внимание!!!

Некоторые программы могут быть чувствительны к формату записи в FASTA-формате:

- При написании однобуквенного кода всегда используйте заглавные буквы;
- При работе с FASTA-последовательностями на ПК всегда используйте опцию TEXT;
- При работе с FASTA-форматом в текстовом процессоре Word, всегда используйте исключительно ASCII символы;
- Для правильного отображения этих последовательностей в текстовом процессоре Word используйте исключительно шрифт Courier;
- Применение FASTA-формата в тех случаях, когда требуется RAW-формат, может вызвать ошибки или привести к тому, что часть текста идентификационной линии будет воспринята программой как часть последовательности.

# Пример подачи последовательности в первичную базу данных

Isolate P876, 16S rRNA gene sequence. Length: 1449 bp

```
TGCAAGTCGA ACGGTAGCAG GAAGAAAGCT TGCTTTCTTT GCTGACGAGT GGC GGACGGG TGAGTAATGC TTGGGAATCT GGCTTATGGA GGGGGATAAC
TGTTGGGAAAC TGCAGСТААТ ACCGCGТААТ СТСТGAGGAG ТАААGGGTGG GACyTTAGGG CCACCTGCCA TAAGATGAGC CCAAGTGGGA TTAGGTAGTT
GGTGGGGTAA AGGCCTACCA AGCCTGCGAT СТСТAGCTGG TCTGAGAGGA TGACCAGCCA CACTGGAACT GAGACACGGT CCAGACTCCT ACGGGAGGCA
GCAGTGGGGА АТАТТGCGCA ATGGGGGGAA CCCTGACGCA GCCATGCCGC GTGAATGAAG AAGGCCTTCG GGTTGTAAAG TTCTTTTCGGT AATGAGGAAG
GGGTGTTTrTT kAATAGATAG CATCATTGAC GTTAATTACA GAAGAAGCAC CGGCTAACTC CGTGCCAGCA GCCGCGGTAA TACGGAGGGT GCGAGCGTTA
ATCGGAATAA CTGGGCGTAA AGGGCACGCA GGC GGACTTT TAAGTGAGAT GTGAAATCCC CGAGCTTAAC TTGGGAATTG CATTTAGAC TGGGAGTCTA
GAGTACTTTA GGGAGGGGTA GAATTCCACG TGTAGCGGTG AAATGCGTAG AGATGTGGAG GAATACCGAA GGCGAAGGCA GCCCCTTGGG AATGTACTGA
CGCTCATGTG CGAAAGCGTG GGGAGCAAAC AGGATTAGAT ACCCTGGTAG TCCACGCTGT AAACGCTGTC GATTTGGGGA TTGGGCTTTA AGCTTGGTGC
CCGAAGCTAA CGTGATAAAT CGACCGCCTG GGGAGTACGG CCGCAAGGTT AAAACTCAA TGAATTGACG GGGGCCCGCA CAAGCGGTGG AGCATGTGGT
TTAATTTCGAT GCAACGCGAA GAACCTTACC TACTCTTGAC ATCCTAAGAA GAGCTCAGAG ATGAGCTTGT GCCTTCGGGA ACTTAGAGAC AGGTGCTGCA
TGGCTGTCTG CAGCTCGTGT TGTGAAATGT TGGGTТААGT CCCGCAACGA GCGCAACCCT TATCCTTTGT TGCCAGCGAT TTGGTCGGGA ACTCAAAGGA
GACTGCCAGT GACAAACTGG AGGAAGGTGG GGATGACGTC AAGTCATCAT GGCCCTTACG AGTAGGGCTA CACACGTGCT ACAATGGTGC ATACAGAGGG
CAGCGAGAGT GCGAGCTTAA GCGAATCTCA GAAAGTGCAT СТААGTCCGG ATTGGAGTCT GCAACTCGAC TCCATGAAGT CGGAATCGCT AGTAATCGCA
AATCAGAATG TTGCGGTGAA TACGTTCCCG GGCCTTGТAC ACACCGCCCG TCACACCATG GGAGTGGGTT GTACCAGAAG TAGATAGCTT AACCTTCGGG
AGGGCGTTTTA CCACGGTATG ATTCATGACT GGGGTGAAGT CGTAAACAGA
```

# Подача в GenBank при помощи инструмента BankIt

The screenshot shows a Microsoft Internet Explorer browser window displaying the NCBI BankIt website. The address bar shows the URL <http://www.ncbi.nlm.nih.gov/BankIt/>. The page title is "BankIt: GenBank Submissions by WWW". The navigation menu includes PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The main content area is titled "GenBank Direct Submission Options" and provides instructions on when to use BankIt versus Sequin.

**BankIt: GenBank Submissions by WWW**

PubMed Entrez BLAST OMIM Taxonomy Structure

NCBI  
SITE MAP  
BankIt Help  
Getting Started  
Submission Info  
Reference Info  
Source Info  
Input DNA  
Additional Info

► **GenBank Direct Submission Options**

Use BankIt if:

- you have one or a few sequence submissions
- you prefer to use a WWW-based submission tool
- your sequence annotation is not complicated
- you do not require sequence analysis tools to submit your sequence(s)

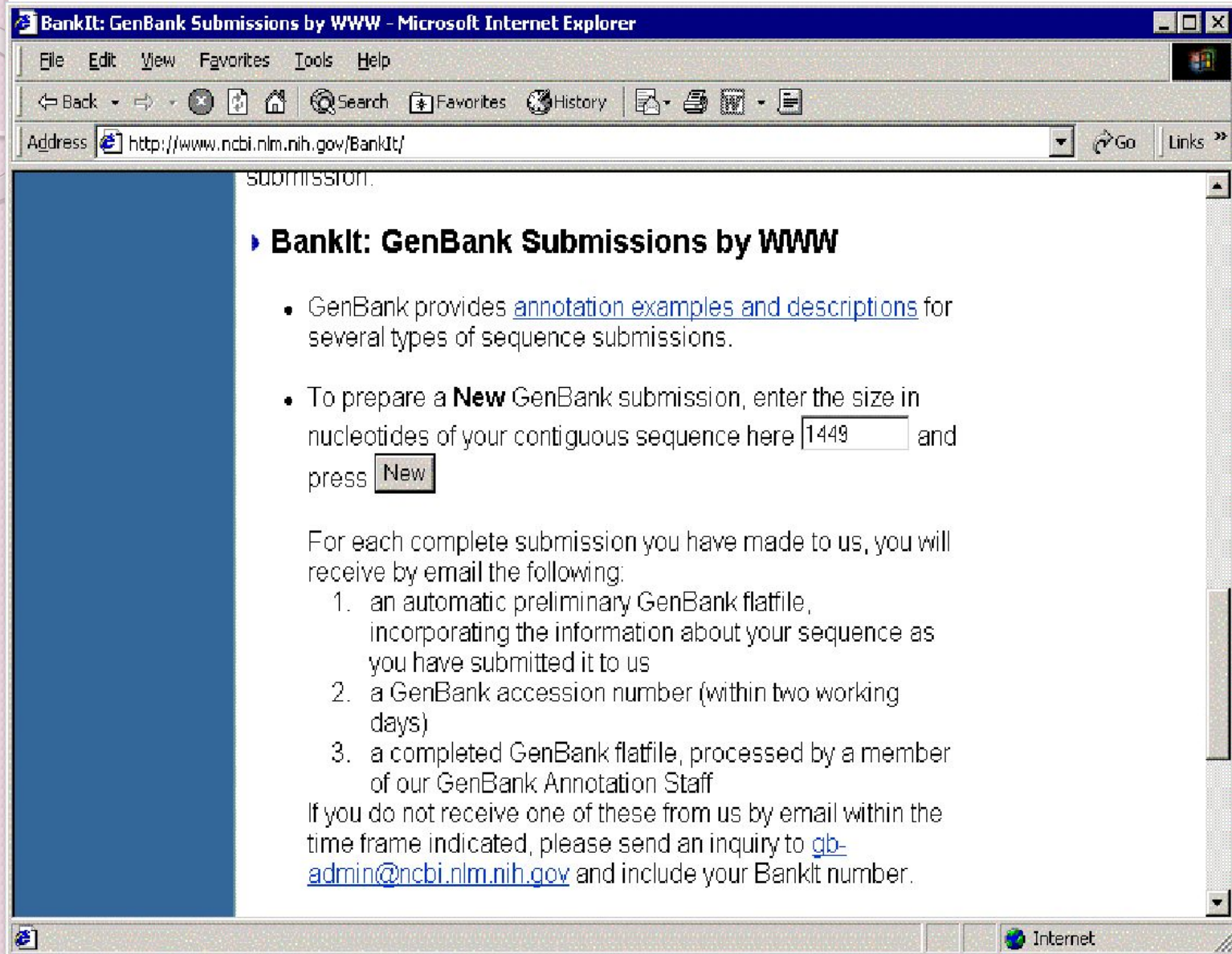
Use [Sequin](#) if:

- you are submitting long or complex submissions
- you are submitting mutation, phylogenetic, population, environmental, or segmented sets
- you would like graphical viewing and editing options, including the alignment editor
- you would like network access to related analytical tools

Internet

# ШАГ 1.

## Резервирование места в базе данных



The screenshot shows a Microsoft Internet Explorer browser window. The title bar reads "BankIt: GenBank Submissions by WWW - Microsoft Internet Explorer". The address bar contains "http://www.ncbi.nlm.nih.gov/BankIt/". The main content area has a blue sidebar on the left and a white main area. The main area contains the following text:

SUBMISSION.

► **BankIt: GenBank Submissions by WWW**

- GenBank provides [annotation examples and descriptions](#) for several types of sequence submissions.
- To prepare a **New** GenBank submission, enter the size in nucleotides of your contiguous sequence here  and press

For each complete submission you have made to us, you will receive by email the following:

1. an automatic preliminary GenBank flatfile, incorporating the information about your sequence as you have submitted it to us
2. a GenBank accession number (within two working days)
3. a completed GenBank flatfile, processed by a member of our GenBank Annotation Staff

If you do not receive one of these from us by email within the time frame indicated, please send an inquiry to [gb-admin@ncbi.nlm.nih.gov](mailto:gb-admin@ncbi.nlm.nih.gov) and include your BankIt number.

The browser's status bar at the bottom shows "Internet".



## ШАГ 2. Контактная информация

BankIt -- GenBank submissions by WWW - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print Mail

Address <http://www.ncbi.nlm.nih.gov/BankIt/nph-bankit.cgi> Go Links >>

then each sequence from each source must be a **separate submission**.

### Contact Information

*First name:*  *Last name:*

*Department:*  *Institution:*

*Street:*

*City:*  *State/Province:*

*ZIP/Postal Code:*  *Country:*

*Phone:*  *Fax:*

Please include country code for non-U.S. phone numbers.

*e-mail:*

Please enter a **single** valid e-mail address only.

Done Internet

## ШАГ 3.

### Внесение текста последовательности.

The screenshot shows a Microsoft Internet Explorer browser window with the title "BankIt -- GenBank submissions by WWW - Microsoft Internet Explorer". The address bar contains "http://www.ncbi.nlm.nih.gov/BankIt/nph-bankit.cgi". The main content area displays instructions for sequence submission:

- Sequence must be at least 50 bp in length
- Sequence must be biologically contiguous and not contain any internal unknown/unsequenced spacers.

Below the instructions, there is a text input field labeled "Sequence length in nucleotides:" with the value "1449" entered.

The "Enter DNA sequence:" section contains a large text area with the following sequence:

```
TGCAAGTCGA ACGGTAGCAG GAAGAAAGCT TGCTTTCTTT GCTGACGAGT
GGCGGACGGG TGAGTAATGC TTGGGAATCT GGCTTATGGA GGGGGATAAC
TGTGGGAAAC TGCAGCTAAT ACCGCGTAAT CTCTGAGGAG TAAAGGGTGG
```

### Additional Information

[Top](#) [Bottom](#)  
[Help](#)

- Any sequence features, such as coding regions or structural RNAs, should be added on the next page, after you "Validate and Continue" below.
- Enter any other biological information for which there is no place on the form or any pertinent instructions that will help GenBank annotators process your submission in this field.



## ШАГ 4.

### Подтверждение заявки и возможные ошибки.

BankIt -- GenBank submissions by WWW - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print View Source

Address <http://www.ncbi.nlm.nih.gov/BankIt/nph-bankit.cgi> Go Links

**(Click on underlined reference for error or warning location):**

---

ERROR: You must specify a [first name](#).

ERROR: You must specify a [last name](#).

ERROR: You must specify a [phone number](#).

ERROR: You must specify an [e-mail address](#).

ERROR: You must state kind of [molecule](#) you sequenced.

ERROR: You must select or enter an [organism name](#).

ERROR: You must provide a [title](#), even if this sequence is not published in a journal.


**Note!** To fix the errors noted, just correct the information in the form below and press "Validate and Continue".

---

**BankIt: GenBank Submissions by WWW** [Bottom Help](#)

Internet





**Дякую за увагу**  
**Благодарю за внимание**  
**Thank you for your attention**