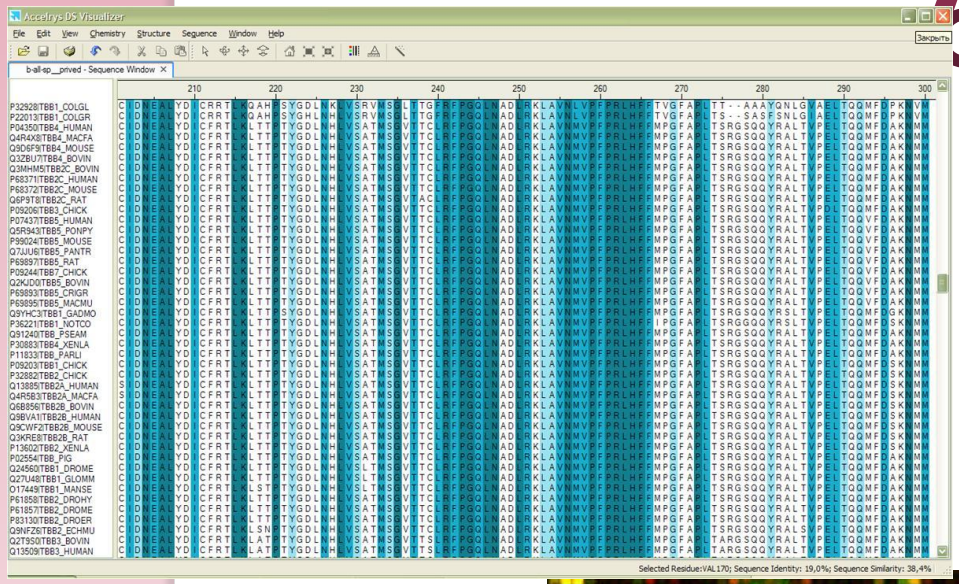
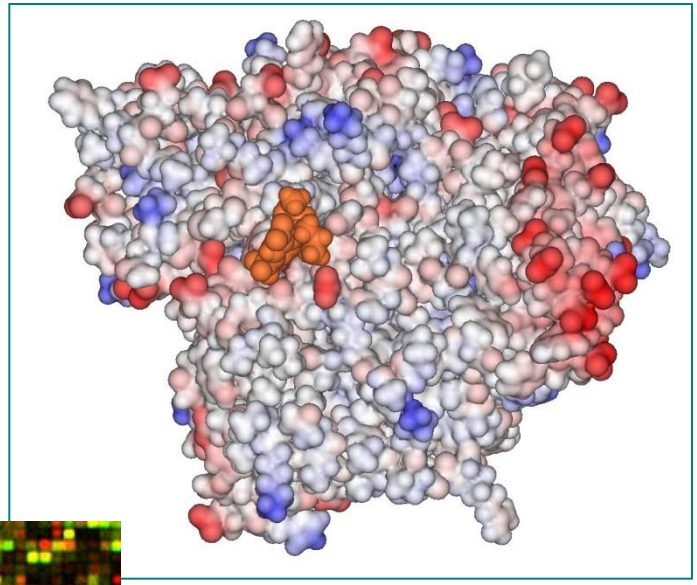
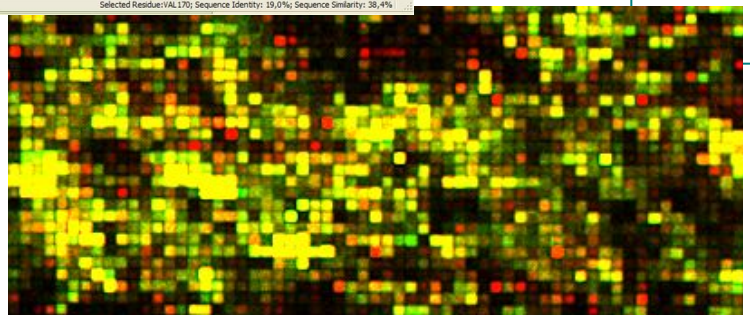
Київський національний університет ім. Тараса Шевченка
Інститут високих технологій

БІОІНФОРМАТИК

к.б.н. Нилорко О.А.







Банки (бази) данних – 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 данних обов’язково містять в себе також набір програмних інструментів, які забезпечують :

- доступ до банка 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 курованих) 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 різнорідну інформацію (архівну, куровану, 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 дрозофіли) 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 CATCH, 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, BIOCYC, 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

Taxonomy

Reference

Annotations (comments)

Cross-references

Keywords

```
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.
GN   EPO.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=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
RT   erythropoietin.";
RL   Nature 313:806-810(1985).
….
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).
…
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.
….
```

Sequence database: example (cont.)

Annotations (features)

Sequence

```
FT   SIGNAL          1     27
FT   CHAIN          28    193       ERYTHROPOIETIN.
FT   PROPEP         190   193       MAY BE REMOVED IN PROCESSED PROTEIN.
FT   DISULFID        34   188
FT   DISULFID        56    60
FT   CARBOHYD        51    51       N-LINKED (GLCNAC...).
FT   CARBOHYD        65     65      N-LINKED (GLCNAC...).
FT   CARBOHYD       110    110      N-LINKED (GLCNAC...).
FT   CARBOHYD       153    153      O-LINKED (GALNAC...).
FT   VARIANT        131    132      SL -> NF (IN AN HEPATOCELLULAR
FT                                  CARCINOMA).
FT                                  /FTId=VAR_009870.
FT   VARIANT        149    149       P -> Q (IN AN HEPATOCELLULAR CARCINOMA).
FT                                  /FTId=VAR_009871.
FT   CONFLICT      40     40         E -> Q (IN REF. 1; CAA26095).
FT   CONFLICT      85     85         Q -> QQ (IN REF. 5).
FT   CONFLICT     140    140         G -> R (IN REF. 1; CAA26095).
**
**   ##################    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 TADTFRKLFR VYSNFLRGKL
     KLYTGEACRT GDR
//
```

Sequence database: example

…The fasta format:

```
> My_Sequence_Name
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAE
NITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPD
AASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR
```

…The RAW format:

```
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAE
NITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPD
AASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR
```

Database 1: 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, tm RNA, uRNA, etc…)

- 3D structure (DNA and RNA) - PDB

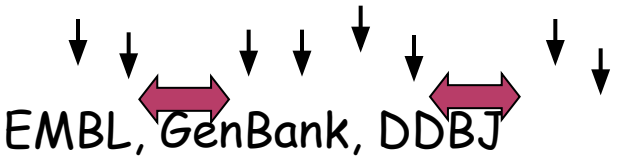
- Others: Aberrant splicing db; Eukaryotic promoter db (EPD); RNA editing sites, Multimedia Telomere Resource ……

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)

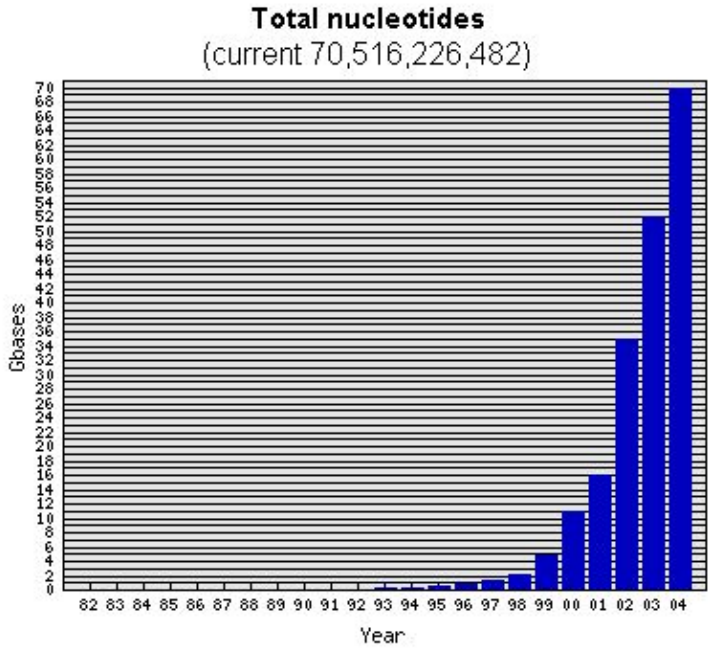
International Nucleotide Sequence Database Collaboration

(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)

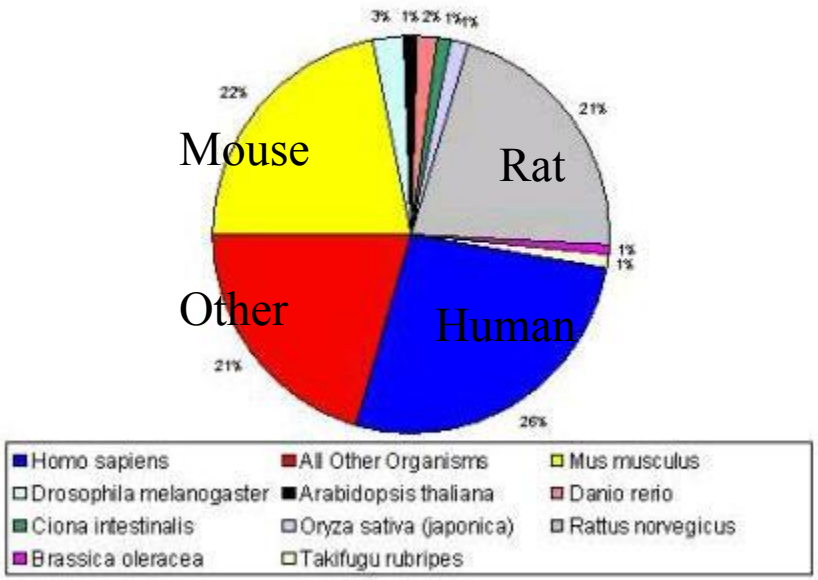
- Currently: 106,533,156,756 bases in 108,431,692 sequence records

The tremendous increase in nucleotide sequences (1980-2004)



1980: 80 genes fully sequenced !

More than 50'000 species, but…



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          1..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                   AKEAENITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVLRG
FT                   QALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTITAD
FT                   TFRKLFRVYSNFLRGKLKLYTGEACRTGDR
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
```

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

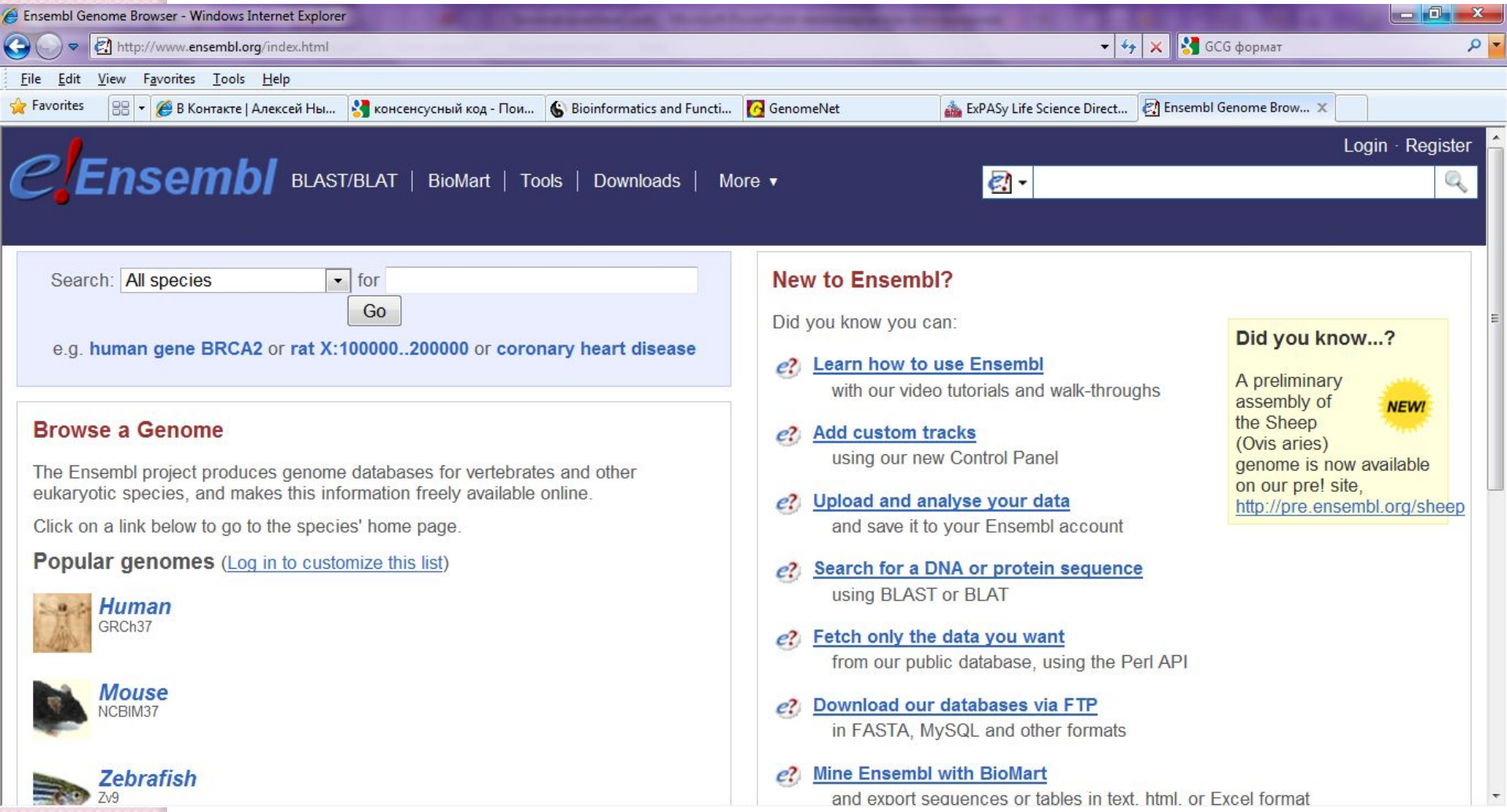
i.e. Ensembl, TIGR

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


Ensembl Genome Browser - Windows Internet Explorer
http://www.ensembl.org/index.html
GCG формат
File Edit View Favorites Tools Help
Favorites
В Контакте | Алексей Ны...
консенсусный код - Пои...
Bioinformatics and Functi...
GenomeNet
ExPASy Life Science Direct...
Ensembl Genome Brow...
Login · Register
e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | More
Search: All species for
Go
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)
Human
GRCh37
Mouse
NCBIM37
Zebrafish
Zv9
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
NEW!

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);

Since 1986

TrEMBL -> EBI (European Bioinformatics Institute)
(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: O75144

Printer-friendly view | Quick Blast search

[General] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

General information about the entry

Entry name	ICOL_HUMAN
Primary accession number	O75144
Secondary accession numbers	Q9NRO1 Q9HD18
Entered in SWISS-PROT in	Release 38, July 1999
Sequence was last modified in	Release 40, October 2001
Annotations were last modified in	Release 41, March 2002

Name and origin of the protein

Protein name	ICOS ligand [Precursor]
Synonyms	B7 homolog 2 B7-H2 [illegible] GL50 B7 related protein 1 B7RP-1
Gene name	ICOSL or B7H2 or B7RP1 or KIAA0653
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

names and taxonomy

References

[1] SEQUENCE FROM NUCLEIC ACID (ISOFORM 1).
TISSUE=Dendritic cell;
MEDLINE=20377846; PubMed=11023515; [NCBI, ExPASy, EBI, Israel, Japan]
Wang S., Zhu G., Chapoval A.I., Dong H., Tamada K., Ni J., Chen L.;
"Costimulation of T cells by B7-H2, a B7-like molecule that binds ICOS.";
Blood 96:2808-2813(2000).

[2] SEQUENCE FROM NUCLEIC ACID (ISOFORM 1), AND CHARACTERIZATION.
TISSUE=Peripheral blood lymphocyte;
MEDLINE=20465019; PubMed=11007763; [NCBI, ExPASy, EBI, Israel, Japan]
Yoshinaga S.K., Zhang M., Perillo T., Heran T., Zhang S.L., Moser J., Sonnenberg M., Boone T., Brankow D., Dy T., Delaney J., Han H., Hu A., Kohno T., Manoukian R., Whoriskey J.S., Coccia M.A.;
"Characterization of a new human B7-related protein: B7RP-1 is the ligand for the co-stimulatory molecule ICOS.";
Int. Immunol. 12:1459-1447(2000).

[3] SEQUENCE FROM NUCLEIC ACID (ISOFORM 2).
TISSUE=Leukocyte;
MEDLINE=21126021; PubMed=10657606; [NCBI, ExPASy, EBI, Israel, Japan]
Ling V., Wu P.W., Finnerty H.F., Bean K.M., Spaulding V., Fouser L.A., Leonard J.P., Hunter S.E., Zollner R., Thomas J.L., Miyashiro J.S., Jacobs K.A., Collins M.;
"Identification of GL50, a novel B7-like protein that functionally binds to ICOS receptor.";
J. Immunol. 164:1653-1657(2000).

[4] SEQUENCE FROM NUCLEIC ACID.
TISSUE=Brain;
MEDLINE=99403880; PubMed=9734811; [NCBI, ExPASy, EBI, Israel, Japan]
Ishikawa K.-I., Nagase T., Suyama M., Miyawaki N., Tanaka A., Kotani H., Nomura 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.";
DNA Res. 5:169-176(1998).

[5] SEQUENCE FROM NUCLEIC ACID (ISOFORM 2).
Ling V., Dumont-Ioannopoulos K.;
"GL50 molecules and uses therefor.";
Patent number WO0021796, 20-MAR-2000.

Comments

- *FUNCTION*: 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 INFLAMMATORY CONDITIONS, AS WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-STIMULATING MEMORY T-CELL FUNCTION (*BY SIMILARITY*).
- *SUBCELLULAR LOCATION*: Type I membrane protein (*By similarity*).
- *ALTERNATIVE PRODUCTS*: AT LEAST 2 ISOFORMS; 1 (SHOWN HERE) AND 2 ARE PRODUCED BY ALTERNATIVE SPLICING.
- *TISSUE SPECIFICITY*: ISOFORM 1 IS WIDELY EXPRESSED. HIGH LEVELS IN HEART, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE MARROW, COLON, [illegible] LYMPH NODES, LEUKOCYTES, SPLEEN, THYMUS AND TONSIL). WHILE ISOFORM 2 IS DETECTED ONLY IN LYMPH NODES, LEUKOCYTES AND SPLEEN.
- *INDUCTION*: CONSTITUTIVE EXPRESSION IS SUBSEQUENTLY ENHANCED BY TREATMENT WITH THE ALPHA IN PERIPHERAL BLOOD B-CELLS AND MONOCYTES. WHILE IT IS DECREASED IN DENDRITIC CELLS.
- *SIMILARITY*: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY, B7 FAMILY.
- *SIMILARITY*: CONTAINS 1 IMMUNOGLOBULIN-LIKE V-TYPE DOMAIN.
- *SIMILARITY*: CONTAINS 1 IMMUNOGLOBULIN-LIKE C2-TYPE DOMAIN.
- *CAUTION*: Ref.4 sequence differs from that shown in position 300 onward for an unknown reason.

comments

Copyright

Cross-references

EMBL	AF199108; AAF64739.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AF289028; AAG01176.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AF216749; AAR19241.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AB014553; BAA31528.1; ALT_SEQ [EMBL / GenBank / DDBJ] [CoDingSequence] AX103595; CAC36765.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
MIM	605717 [NCBI / EBI]
GeneCards	ICOSL
GeneCensus	ICOSL
Ensembl	O75144
HUGE	KIAA0653
InterPro	IPR003599; Ig IPR003006; Ig_MHC IPR003600; Ig_like Graphical view of domain structure.
Pfam	PF00047; ig; 3
SMART	SM00409; IG; 1 SM00410; IG_like; 1.
ProDom	[Domain structure / List of seq. sharing at least 1 domain]
BLOCKS	O75144
ProtoMap	O75144
PRESAGE	O75144
DIP	O75144
MaxBase	O75144
SWISS-2DPAGE	GET REGION ON 2D PAGE

cross-references

Keywords

B-cell activation; Immune response; Glycoprotein; Immunoglobulin domain; Signal; Transmembrane; Multigene family; Alternative splicing

keywords

Features

Key	From	To	Length	Description
SIGNAL	1	18	18	POTENTIAL.
CHAIN	19	302	284	ICOS L GAND.
DOMAIN	19	256	238	EXTRACELLULAR (POTENTIAL).
TRANSMEM	257	277	21	POTENTIAL.
DOMAIN	278	302	25	CYTOPLASMIC (POTENTIAL).
DOMAIN	30	120	91	IG-LIKE V-TYPE DOMAIN.
DOMAIN	131	223	73	IG-LIKE C2-TYPE DOMAIN.
DISULFID	37	113		POTENTIAL.
DISULFID	130	210		POTENTIAL.
CARBOHYD	70	70		N-LINKED (GLCNAC...) (POTENTIAL).
CARBOHYD	127	127		N-LINKED (GLCNAC...) (POTENTIAL).
CARBOHYD	173	173		N-LINKED (GLCNAC...) (POTENTIAL).
CARBOHYD	186	186		N-LINKED (GLCNAC...) (POTENTIAL).
CARBOHYD	225	225		N-LINKED (GLCNAC...) (POTENTIAL).
VARSPLIC	300	302		GHV -> ENQHLLELLS (IN ISOFORM 2).

features

Feature alignment

Feature table viewer

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] | CRC64: 647934E21B55E34A [This is a checksum on the sequence]

```
        10         20         30         40         50         60
MELGSPGLLF LLFSHLRADT QEVEVRAMVG SDVELSCACT SSSRFELNDV YVYWQTSESK
        70         80         90        100        110        120
TVVTYRIPQN SSLENVDSRV PSRALWSPAS ELKGNFSLRL YNVTPDDSGK YRCLVSLSDL
       130        140        150        160        170        180
GFQEVLSVEV TLHVAANFSV PVVSAPHSPS QDELTFTCTS IGGYPEPHLS WWEKEDEQLP
       190        200        210        220        230        240
DQALFNDTVT LNNRGTFTVV SVLRIAPTPN VNIGCCIENV LLQQNLTVGS QTGNDIGERD
       250        260        270        280        290        300
KITENPVSTG EKNAATWSIL AVLCLLVVVA VAIGWVCRDR CLQHSYAGAW AVSPETELTG
       310
HV
```

Sequence

O75144 in FASTA format

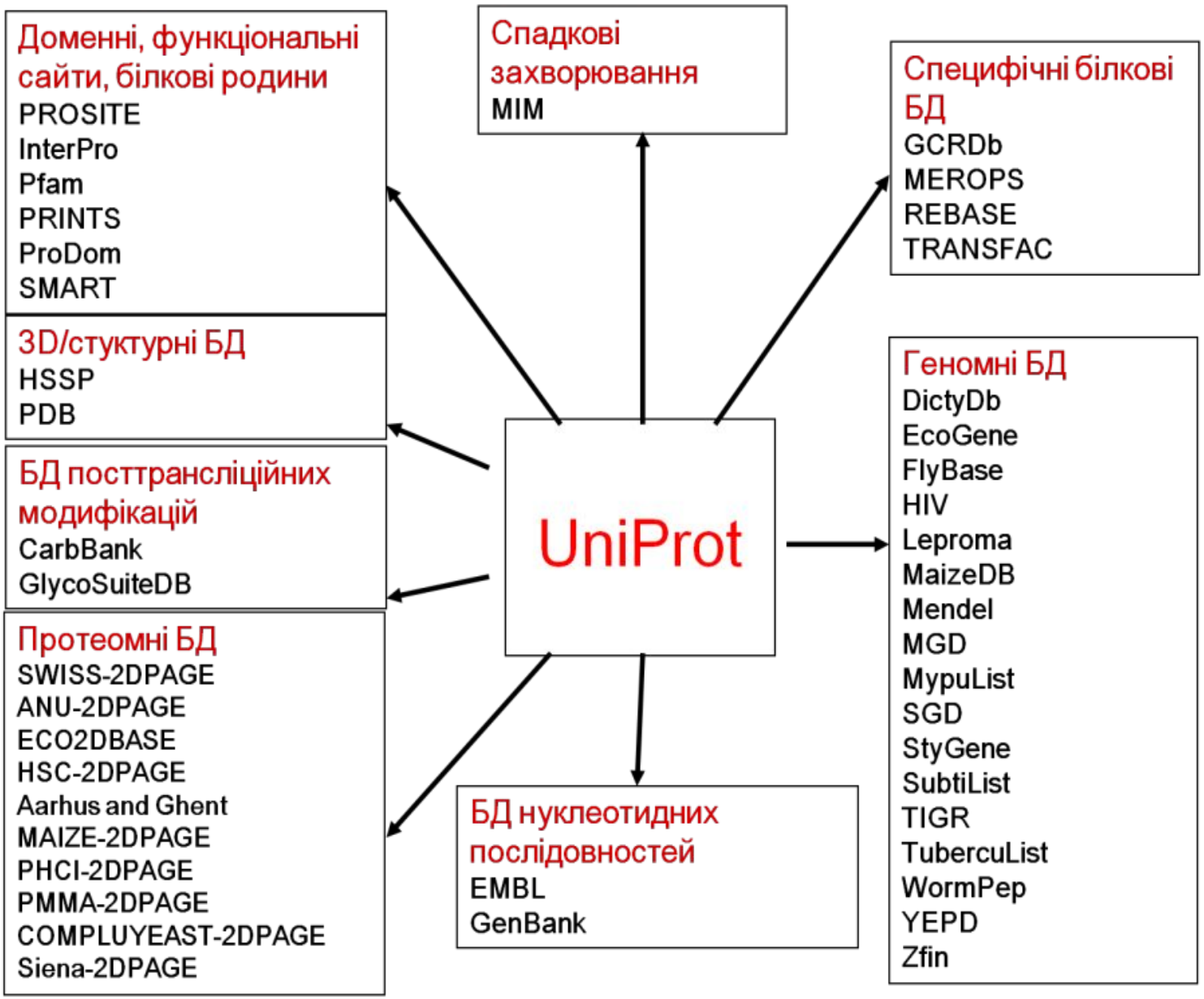
Annotation/Curation (Comment lines)

- 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 (Nothern 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…


Доменні, функціональні сайти, білкові родини
PROSITE
InterPro
Pfam
PRINTS
ProDom
SMART
Спадкові захворювання
MIM
Специфічні білкові БД
GCRDb
MEROPS
REBASE
TRANSFAC
3D/стуктурні БД
HSSP
PDB
Геномні БД
DictyDb
EcoGene
FlyBase
HIV
Leproma
MaizeDB
Mendel
MGD
MypuList
SGD
StyGene
SubtiList
TIGR
TubercuList
WormPep
YEPD
Zfin
БД посттрансліційних модифікацій
CarbBank
GlycoSuiteDB
UniProt
Протеомні БД
SWISS-2DPAGE
ANU-2DPAGE
ECO2DBASE
HSC-2DPAGE
Aarhus and Ghent
MAIZE-2DPAGE
PHCI-2DPAGE
PMMA-2DPAGE
COMPLUYEAST-2DPAGE
Siena-2DPAGE
БД нуклеотидних послідовностей
EMBL
GenBank

Cross-references

ADN
(Index of low redundancy)

3D

genomic

Cross-references	ICE8_HUMAN Q14790
EMBL	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] 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.
MIM	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

Sequence description:

Derived from:

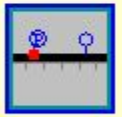
- Publications;
- Databases;
- Personal communication;
- Prediction.

Features — ICOL_HUMAN, O75144

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

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…

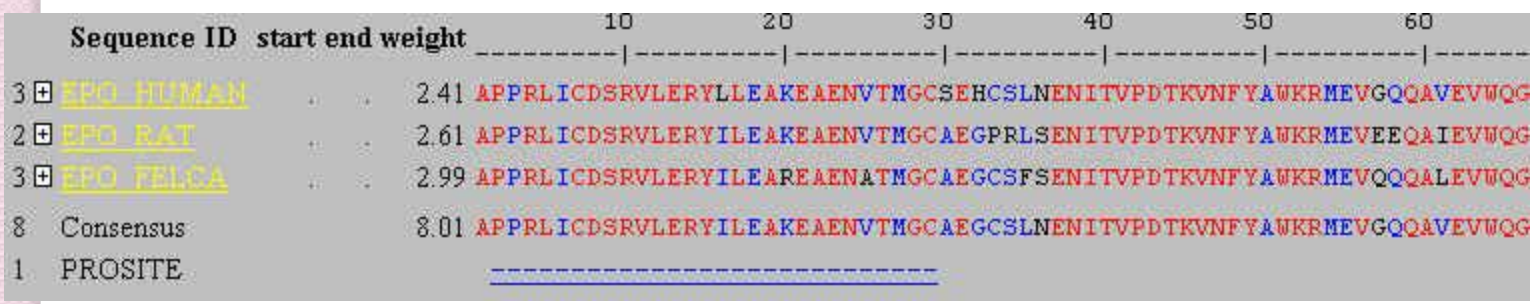
Database 5: protein domain/family

Protein domain/family: some definitions

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

Protein domain/family: some definitions

- Domains (conserved sequences or structures) are identified by multiple sequence alignments



- Domains can be defined by different methods:
  - **Pattern** (regular expression); used for very conserved domains
  - **Profiles** (weighted matrices): two-dimensional tables of position specific match-, gap-, and insertion-scores, derived from aligned sequence families; used for less conserved domains
  - **Hidden Markov Model** (HMM); probabilistic models; an other method to generate profiles.

Protein domain/family databases

- Contains biologically significant « pattern / profiles/ HMM » formulated in such a way that, with appropriate computional 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

Interpro

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

Wellcome Trust Sanger Institute

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

1buy Display pdb

Accession number: PF00758

Erythropoietin/thrombopoietin

Add Annotation

This family forms **structural complexes** with other Pfam families, to view them click here

INTERPRO description (entry IPR001323)

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.

Thrombopoeitin (TPO), a glycoprotein, is the mammalian0 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.

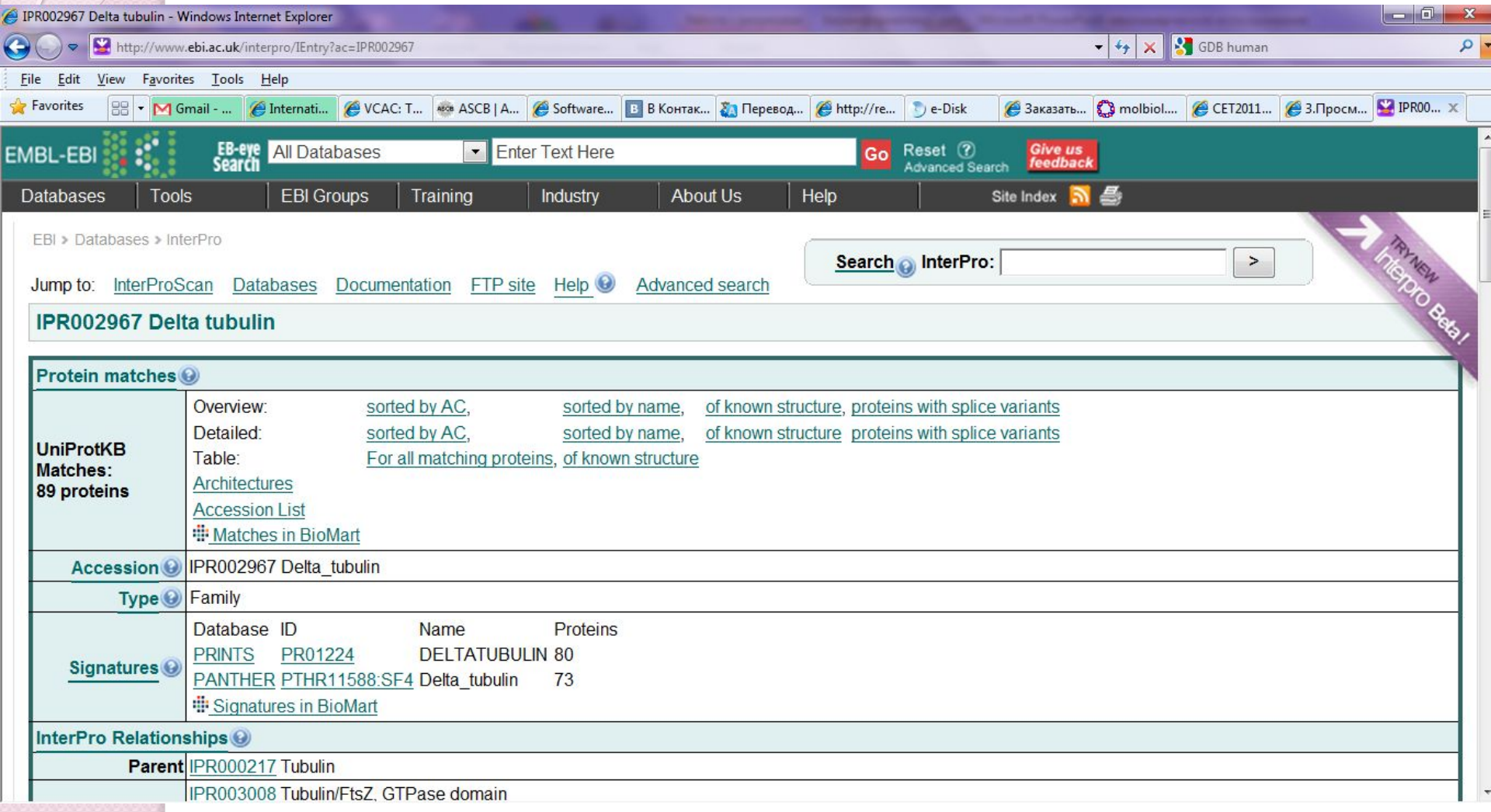
QuickGO	
FUNCTION :	hormone activity (GO:0005179)
COMPONENT :	extracellular (GO:0005576)

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

Alignment	Domain organisation
Seed (7) Full (32)	View 1 representative architecture View architectures for 32 proteins
Format Coloured alignment	Zoom 0.5 pixels/aa.
Get alignment View HMM logo	View Graphic
Further alignment options here Help relating to Pfam alignments here	

InterPro
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 - Windows Internet Explorer
http://www.ebi.ac.uk/interpro/IEntry?ac=IPR002967
GDB human
File Edit View Favorites Tools Help
Favorites
EMBL-EBI
EB-eye Search
All Databases
Enter Text Here
Go
Reset
Advanced Search
Give us feedback
Databases
Tools
EBI Groups
Training
Industry
About Us
Help
Site Index
EBI › Databases › InterPro
Search InterPro:
>
TRY NEW Interpro Beta!
Jump to: InterProScan Databases Documentation FTP site Help Advanced search
IPR002967 Delta tubulin
Protein matches
UniProtKB Matches: 89 proteins
Overview: sorted by AC, sorted by name, of known structure, proteins with splice variants
Detailed: sorted by AC, sorted by name, of known structure proteins with splice variants
Table: For all matching proteins, of known structure
Architectures
Accession List
Matches in BioMart
Accession
IPR002967 Delta_tubulin
Type
Family
Signatures
Database ID Name Proteins
PRINTS PR01224 DELTATUBULIN 80
PANTHER PTHR11588:SF4 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 ренгенівської/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 14α-Sterol Demethylase (Cyp51) From Mycobacterium Tuberculosis In Complex With Fluconazole**

Compound: **Mol_Id: 1; Molecule: Cytochrome P450 51-Like Rv0764C; Chain: A; Synonym: Cyp51, 14α-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*	*X*	*Y*	*Z*	*occ*	*B-factor*
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
12345678901234567890123456789012345678901234567890123456789012345678901234567890
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
12345678901234567890123456789012345678901234567890123456789012345678901234567890
OBSLTE     31-JAN-94 1MBP      2MBP
```

Типи записів в заголовному розділі

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

```
          1         2         3         4         5         6         7
12345678901234567890123456789012345678901234567890123456789012345678901234567890
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
```

CAVEAT – 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
1234567890123456789012345678901234567890123456789012345678901234567890
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
12345678901234567890123456789012345678901234567890123456789012345678901234567890
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
1234567890123456789012345678901234567890123456789012345678901234567890
KEYWDS    LYASE, TRICARBOXYLIC ACID CYCLE, MITOCHONDRION, OXIDATIVE
KEYWDS   2 METABOLISM
```

AUTHOR – імена людей, що відповідають за надходження

```
         1         2         3         4         5         6         7
1234567890123456789012345678901234567890123456789012345678901234567890
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
1234567890123456789012345678901234567890123456789012345678901234567890
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
```

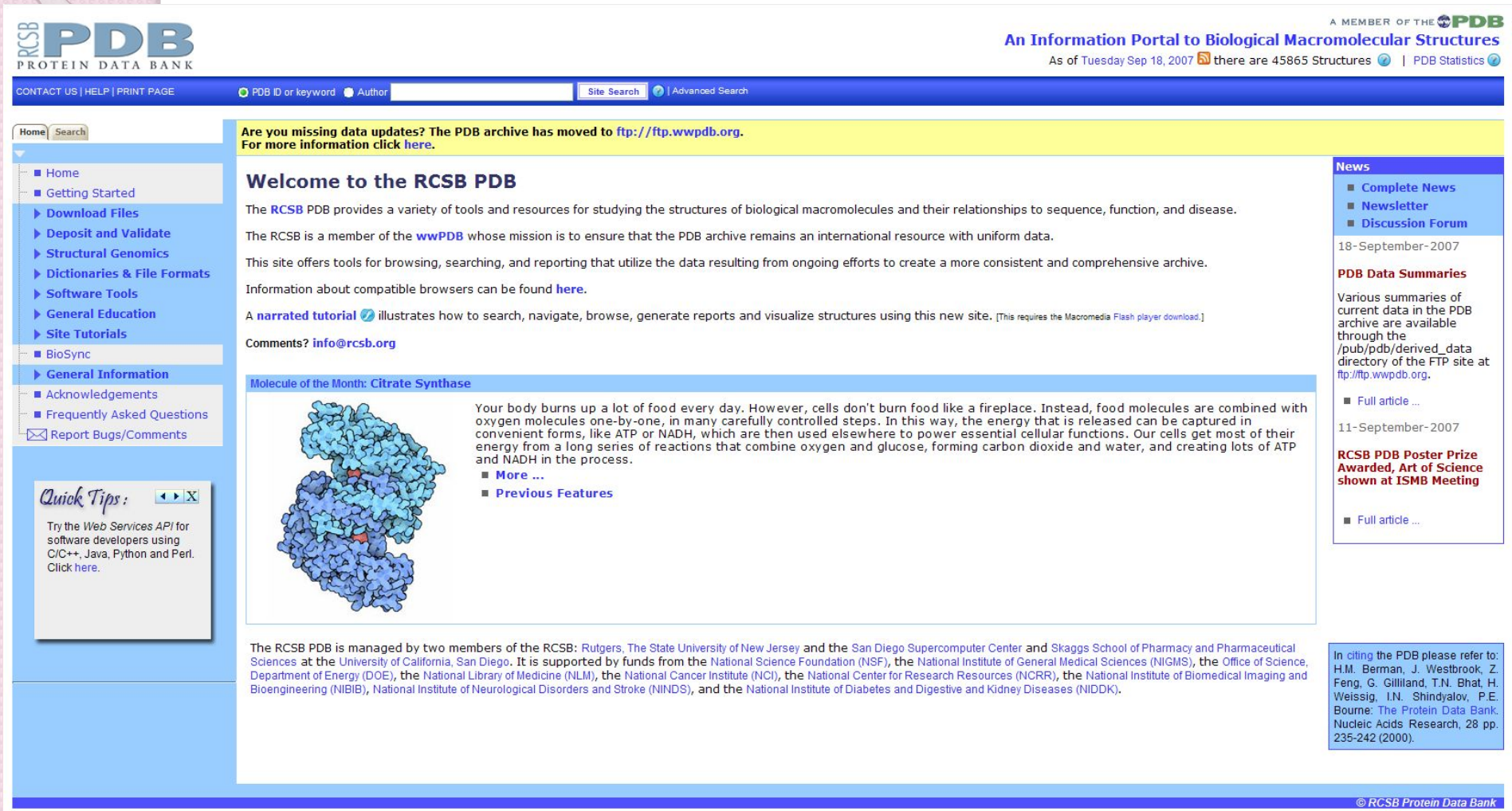
JRNL – основне літературне джерело, яке описує результати, депоновані в

```
         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.

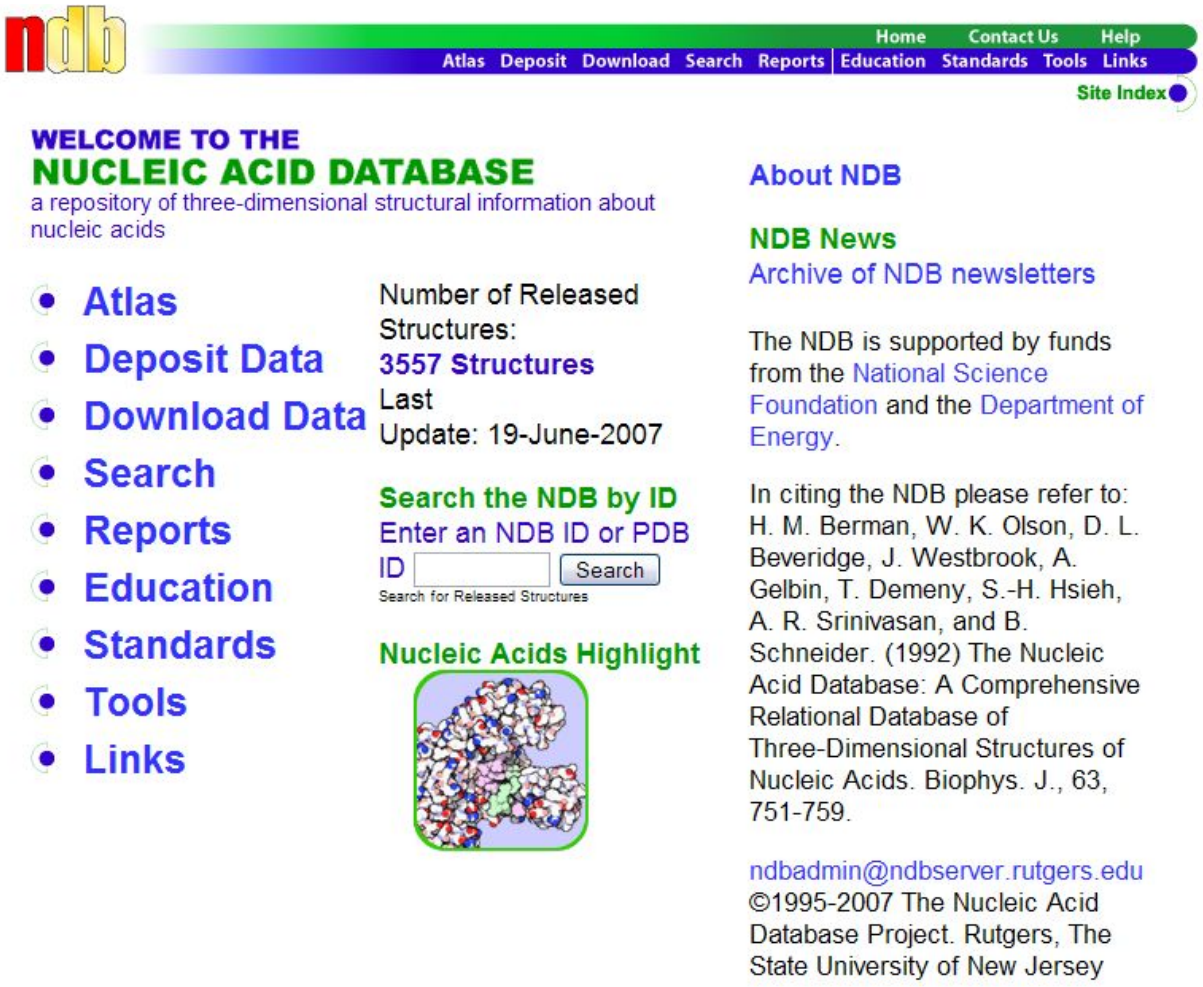
PDB – міжнародний банк даних білкових струкутр

http://www.rcsb.org/

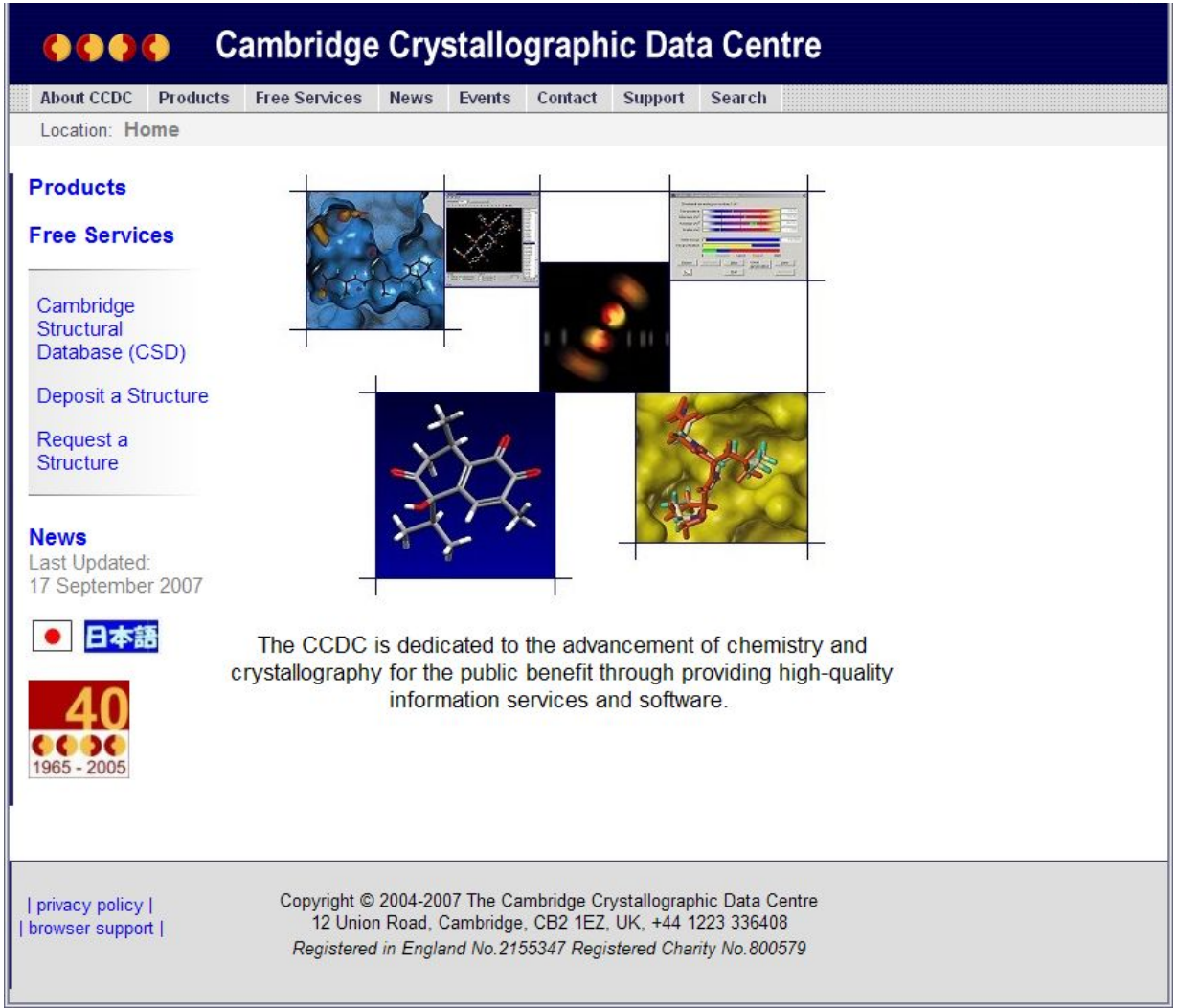


NDB - база даних просторових структур нуклеїнових кислот

http://ndbserver.rutgers.edu/

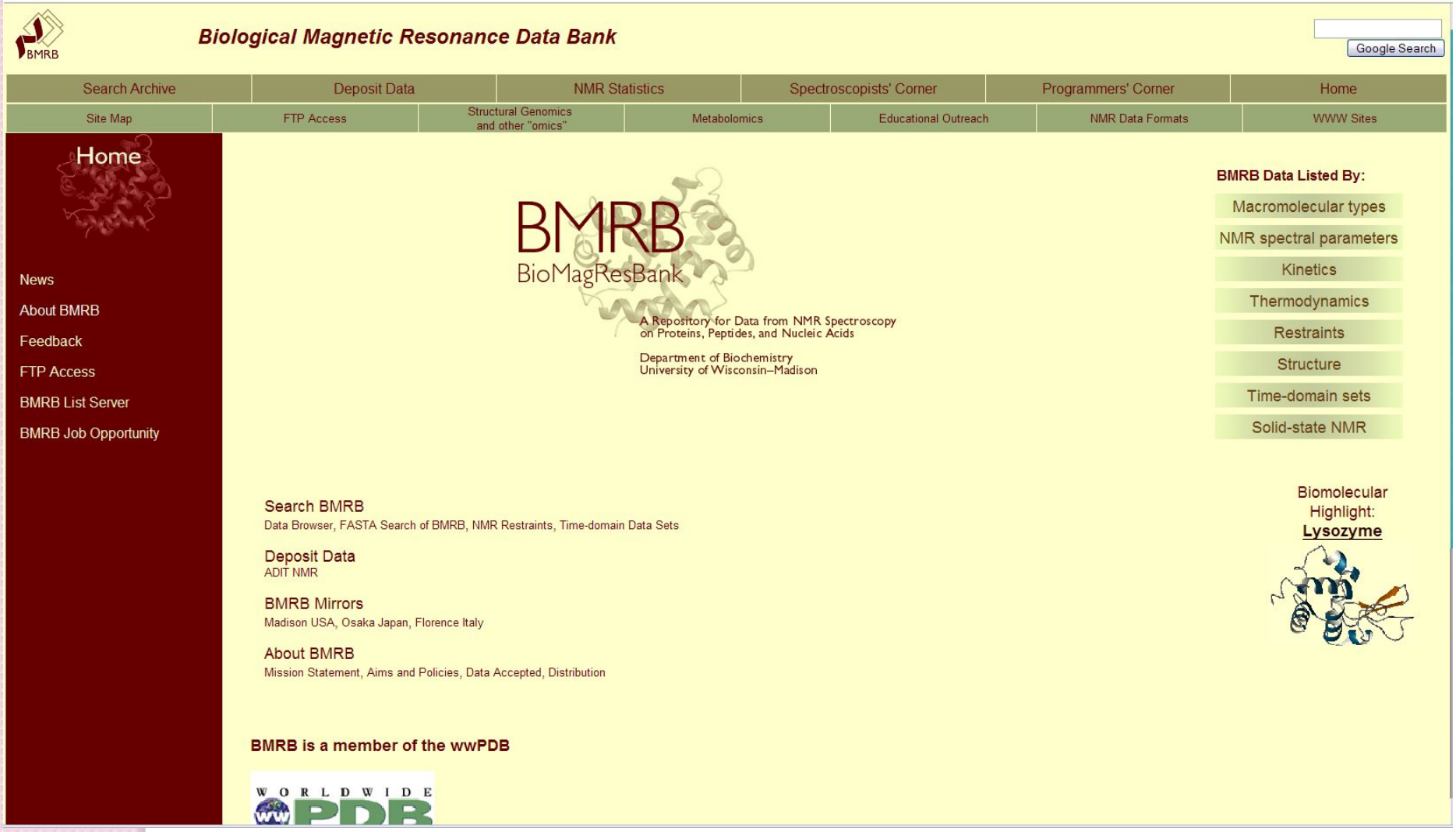


CSD (Cambridge Crystallographic Data Centre) – банк кристалографічних даних низькомолекулярних сполук http://www.ccdc.cam.ac.uk/



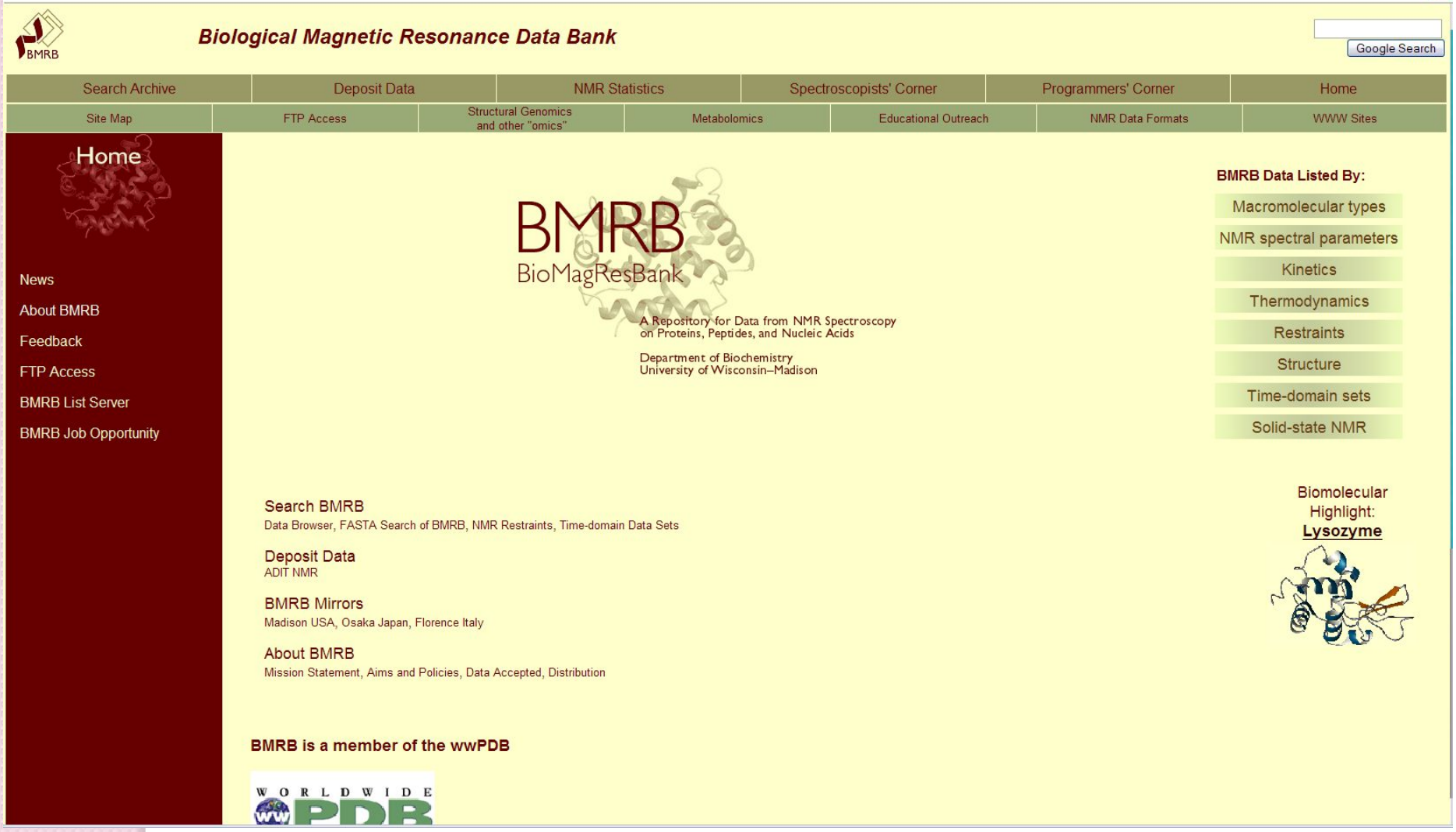
BMRB - банк даних ЯМР-спектроскопії макромолекул

http://www.bmrb.wisc.edu/



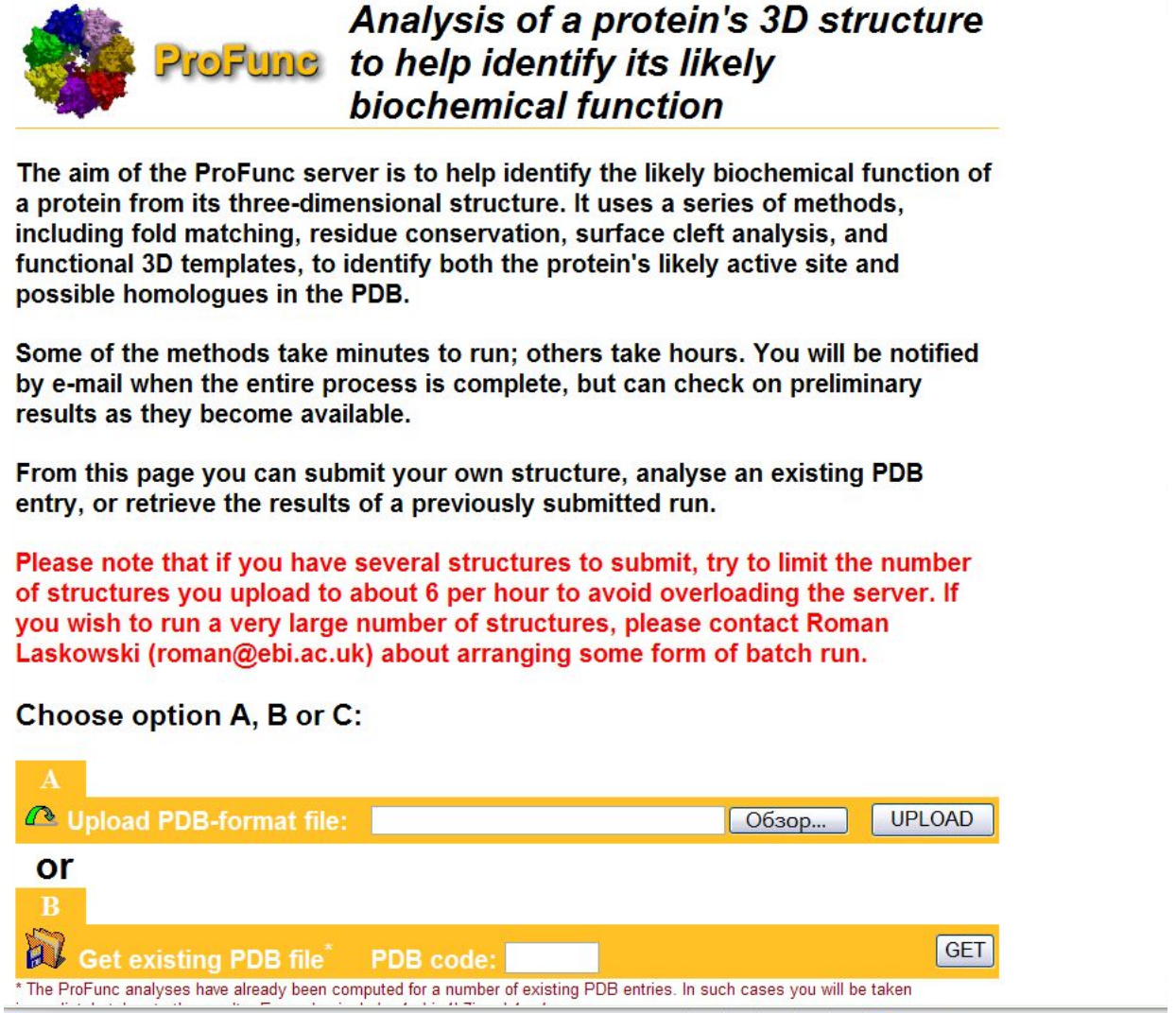
PQS - база даних четвертинних структур білків

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



Profunc – аналіз структури для пердбачення 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) about arranging some form of batch run.

Choose option A, B or C:

A

Upload PDB-format file: Обзор... UPLOAD

or

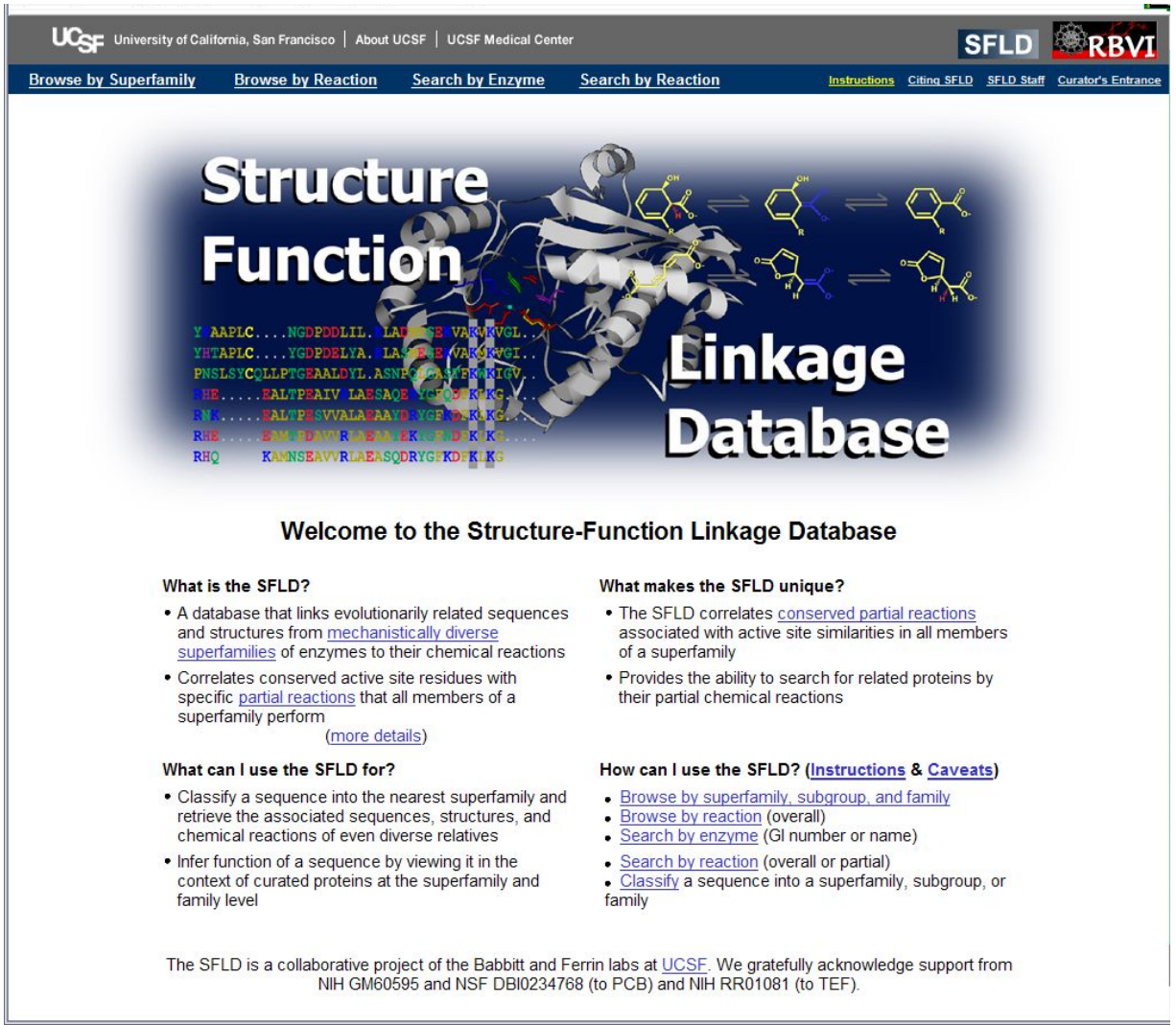
B

Get existing PDB file* PDB code: GET

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

SFLDB – база даних “структура-функція”

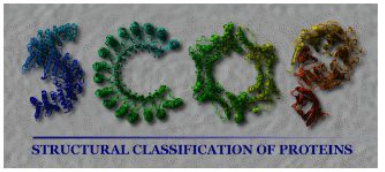
http://sfld.rbvi.ucsf.edu/



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
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

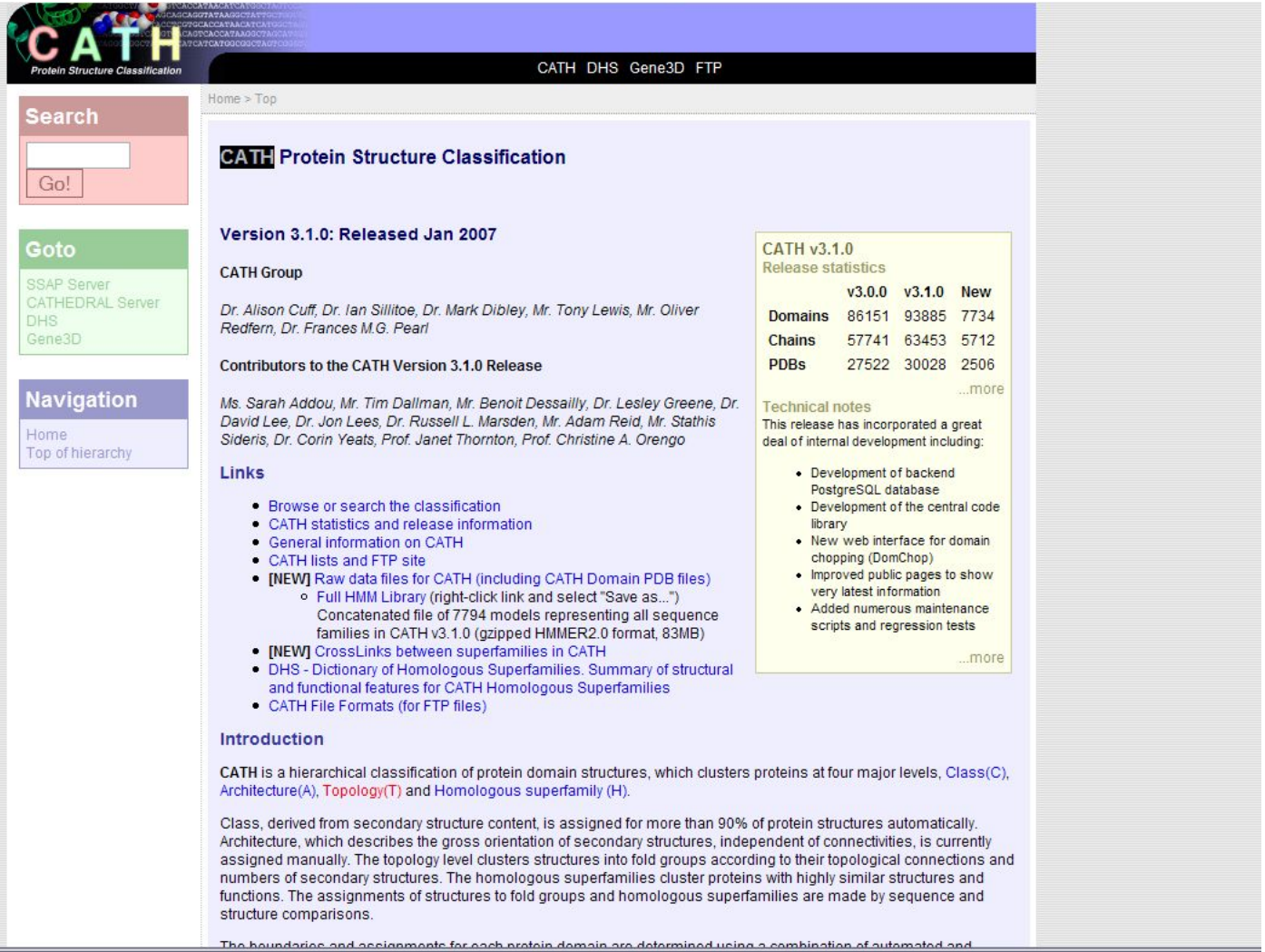
SCOP **mirrors** around the world may speed your access.

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.

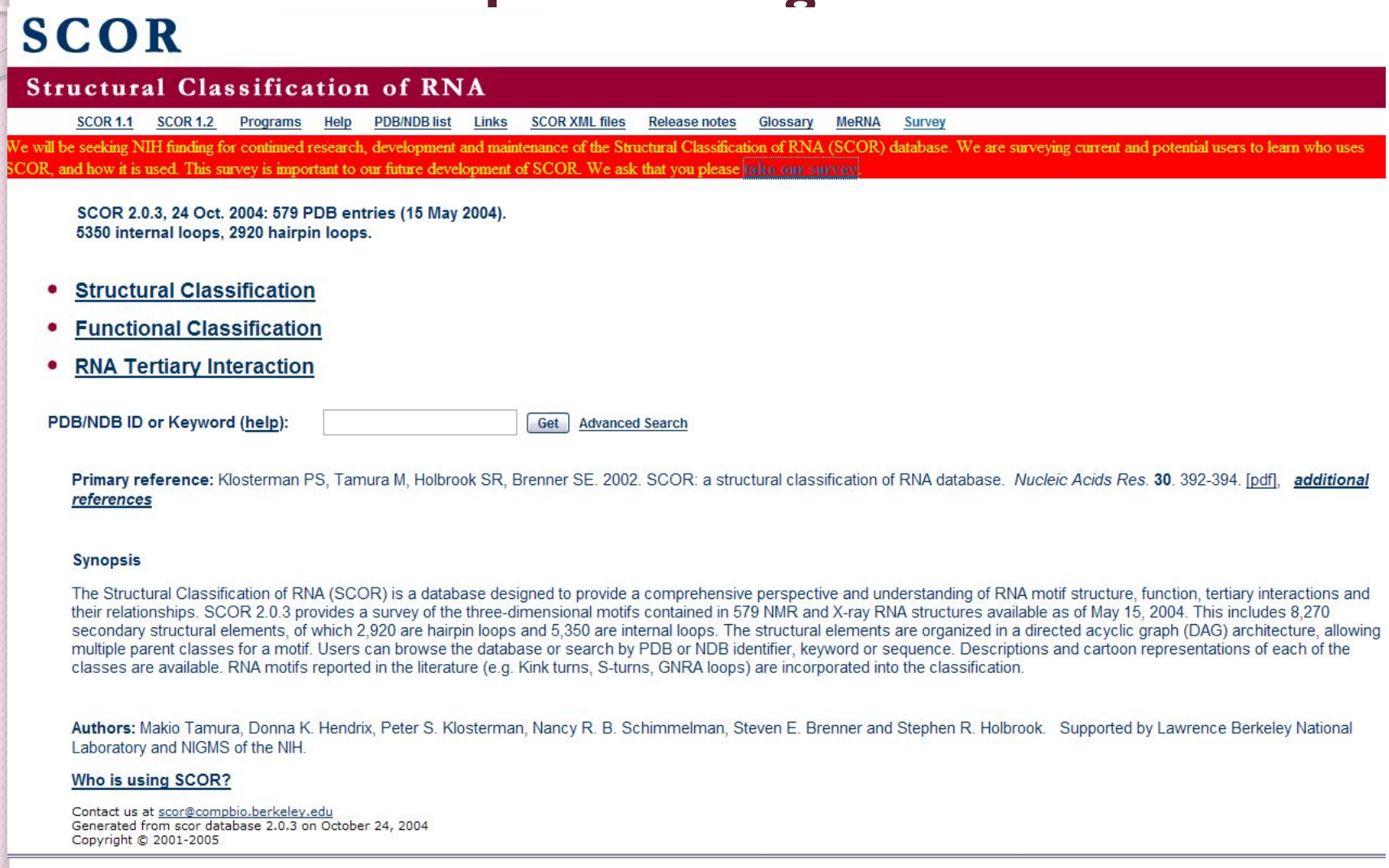
CATH – структурна класифікація білків

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



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): Get 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.

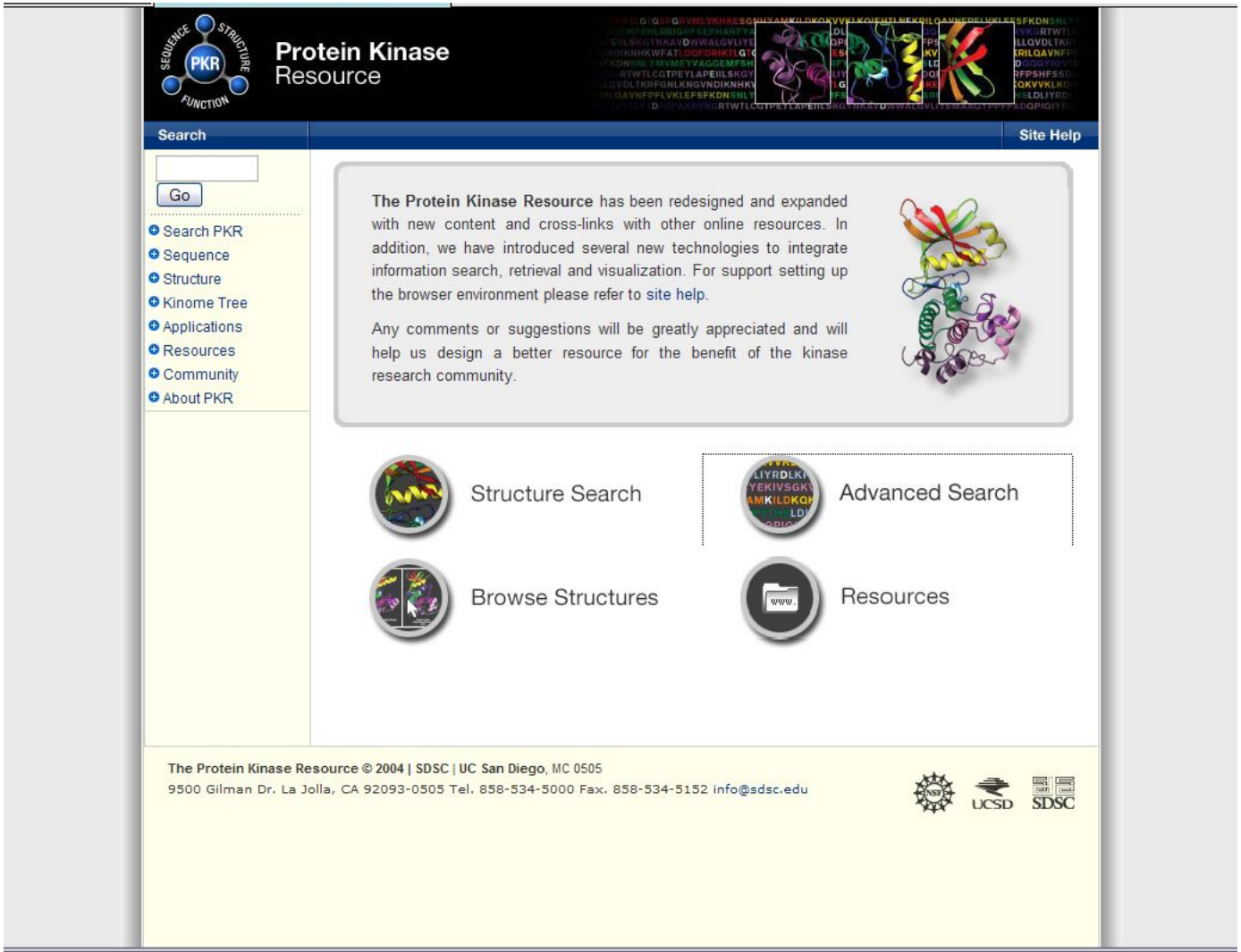
Authors: Makio Tamura, Donna K. Hendrix, Peter S. Klosterman, Nancy R. B. Schimmelman, Steven E. Brenner and Stephen R. Holbrook. Supported by Lawrence Berkeley National Laboratory and NIGMS of the NIH.

Who is using SCOR?

Contact us at scor@compbio.berkeley.edu
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

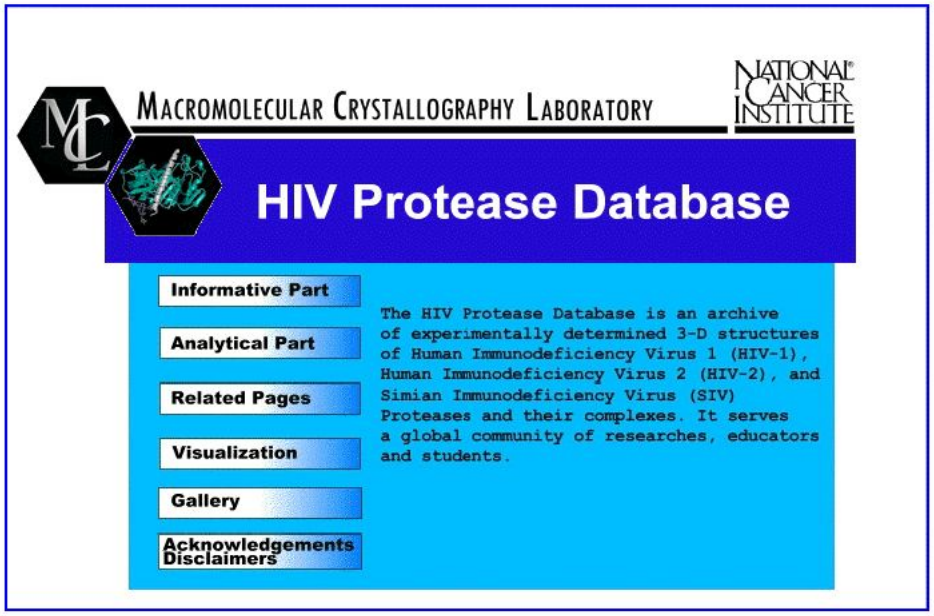


HIV Protease Database

http://mcl1.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!



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 (vondrasekj@ncifcrf.gov)

You are visitor number
Click
Powered by counter.bloke.com

Databases 8: metabolic

- Contain informations that describe enzymes, biochemical reactions and metabolic pathways;

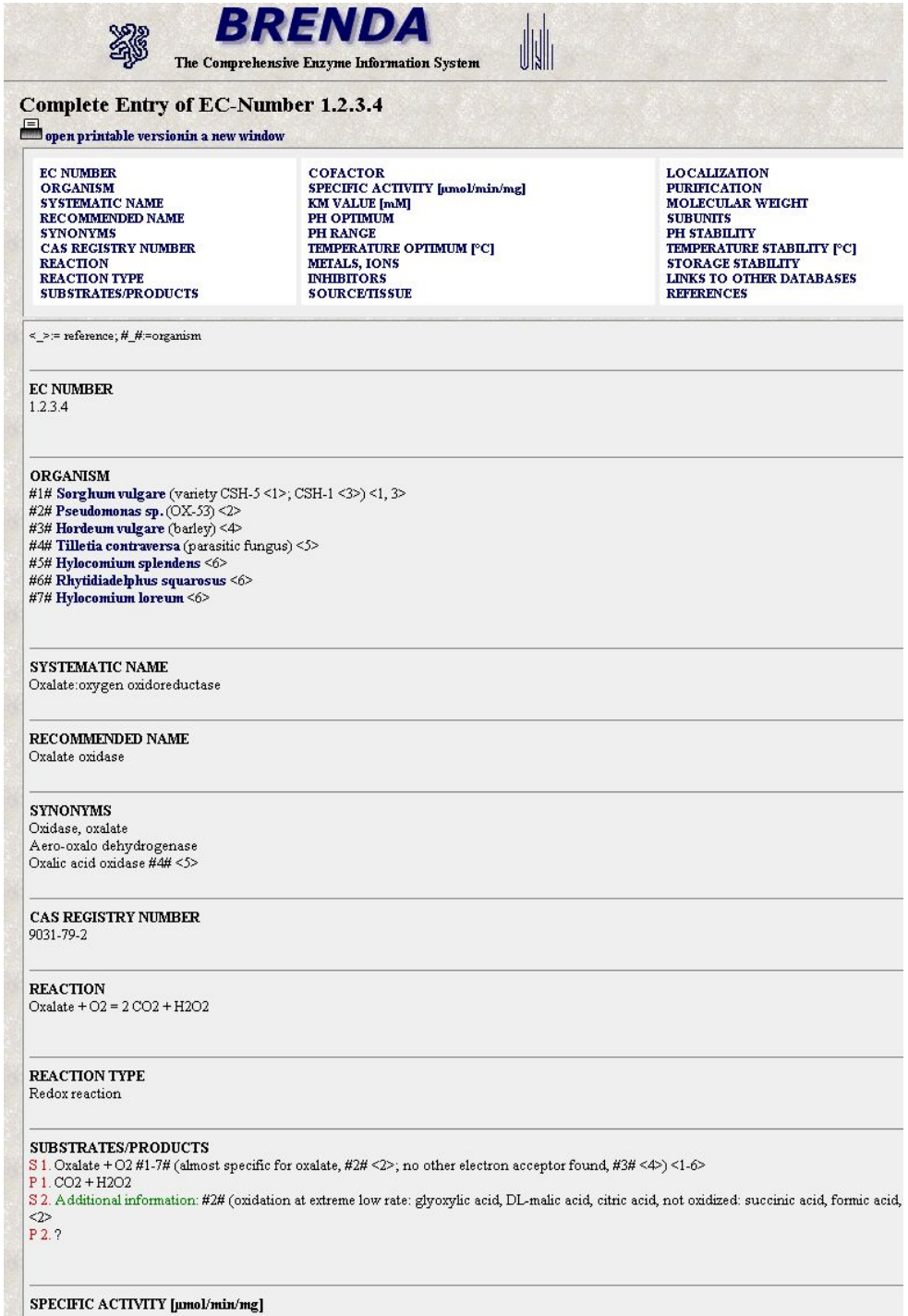
- ENZYME and BRENDA: nomenclature databases that store informations on enzyme names and reactions;

- 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 versionin a new window

EC NUMBER	COFACTOR	LOCALIZATION
ORGANISM	SPECIFIC ACTIVITY [µmol/min/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 squarosus** <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 + O2 = 2 CO2 + H2O2

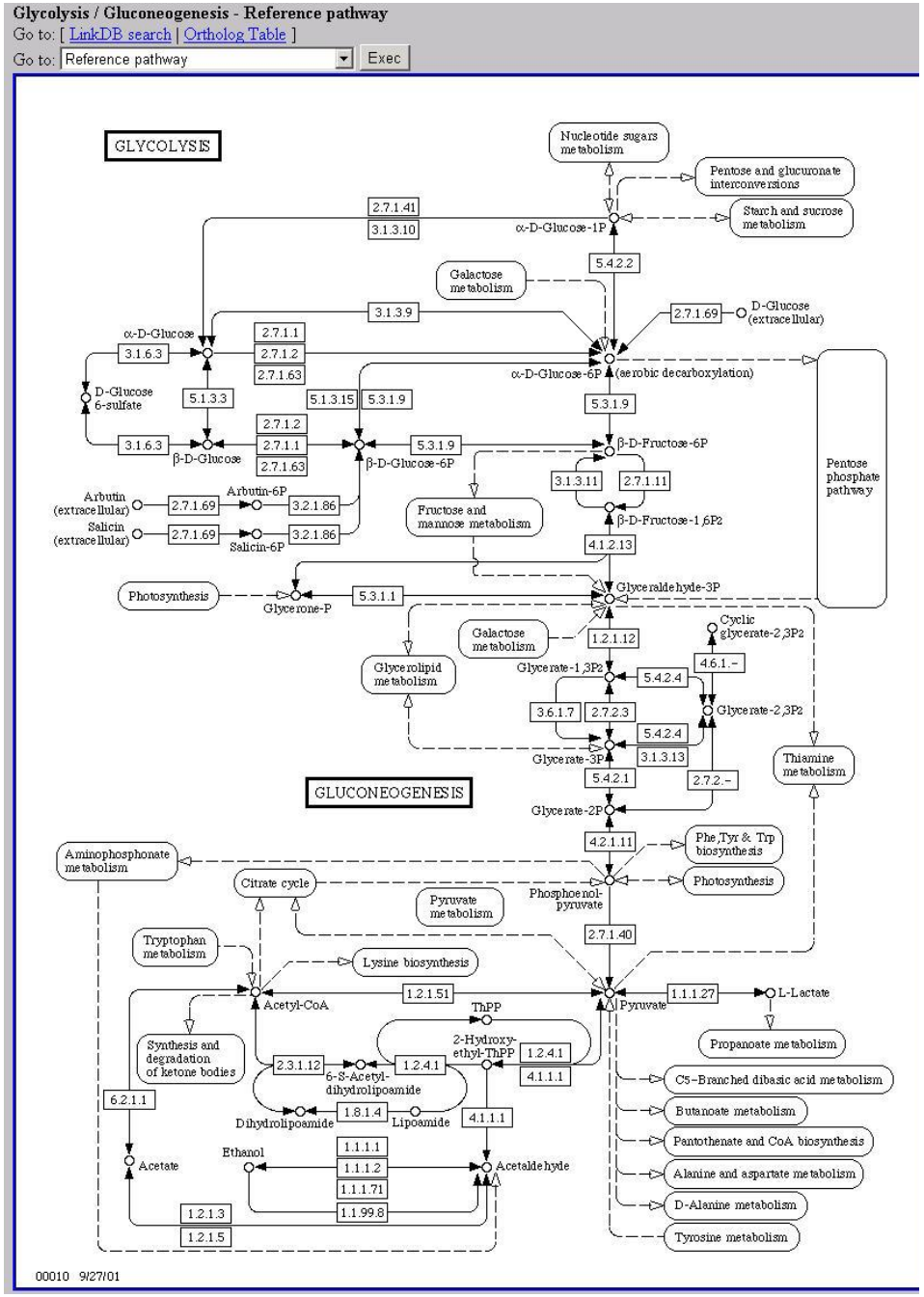
REACTION TYPE
Redox reaction

SUBSTRATES/PRODUCTS
S 1. Oxalate + O2 #1-7# (almost specific for oxalate, #2# <2>; no other electron acceptor found, #3# <4>) <1-6>
P 1. CO2 + H2O2
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 [µmol/min/mg]

KEGG: Kyoto Encyclopedia of Genes and Genomes

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



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
LAFLGEVSAEKEKALSLLAGRIRQPGFTLTLDDAGQWLRSRVVWL
GMRQPPRGLIQLANMLRSQAARSGCFQSNRPFHPHITLLRDASEA
VTIPPPGFNWSYAVTEFTLYASSFARGRTRYTPLKRWALTQ
```

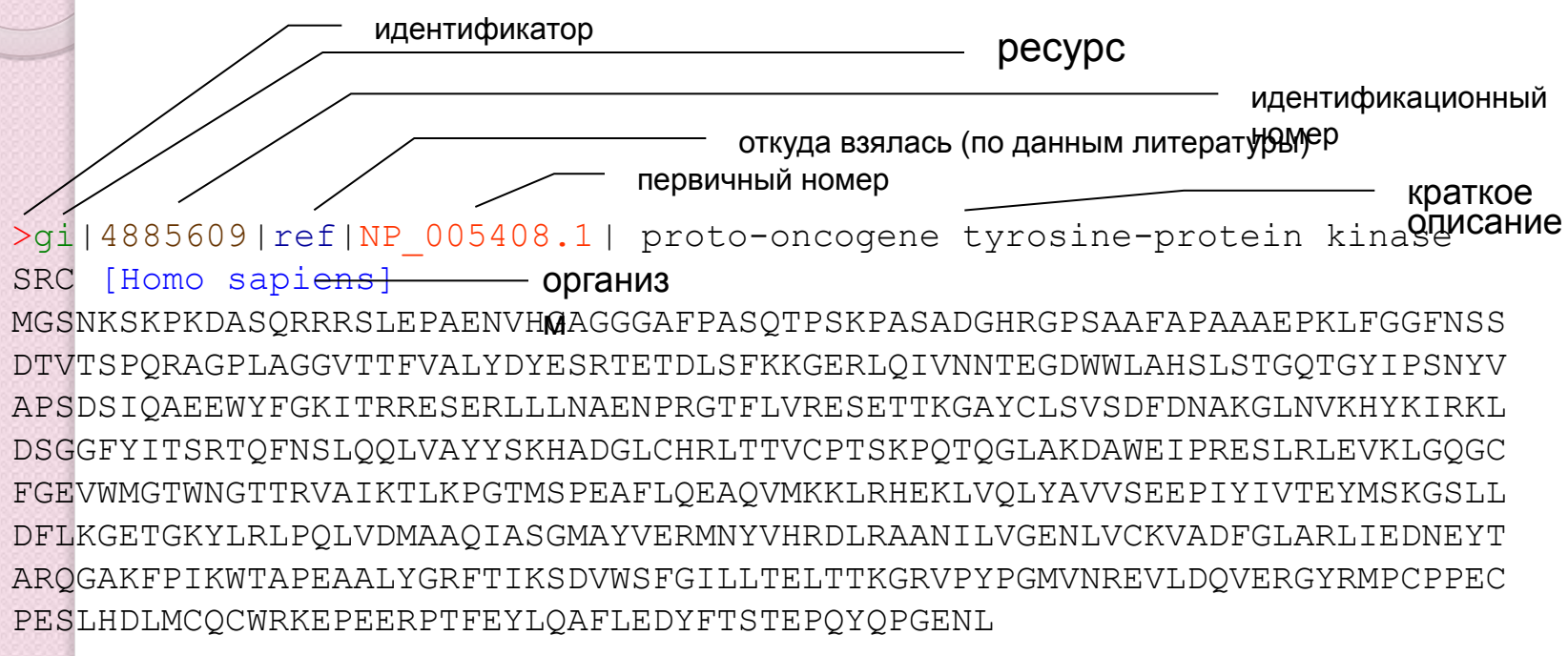
FASTA -формат

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


```
>My_Sequence_Name
MSEPQRLFFAIDLPAEIREQIIHWRATHFPPEAGRPVAADNLHLT
LAFLGEVSAEKEKALSLLAGRIRQPGFTLTLDDAGQWLRSRVVWL
GMRQPPRGLIQLANMLRSQAARSGCFQSNRPFHPHITLLRDASEA
VTIPPPGFNWSYAVTEFTLYASSFARGRTRYTPLKRWALTQ
```

> - уникальный идентификатор

ПРИМЕР:


идентификатор
ресурс
идентификационный номер
откуда взялась (по данным литературы)
первичный номер
краткое описание
организ
>gi|4885609|ref|NP_005408.1| proto-oncogene tyrosine-protein kinase SRC [Homo sapiens]
MGSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSAAFAPAAAEPKLFGGFNSS
DTVTSPQRAGPLAGGVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYV
APSDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVKHYKIRKL
DSGGFYITSRTQFNSLQQLVAYYSKHADGLCHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGC
FGEVWMGTWNGTTRVAIKTLKPGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLL
DFLKGETGKYLRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARLIEDNEYT
ARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLDQVERGYRMPCPPEC
PESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQYQPGENL

Внимание!!!

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

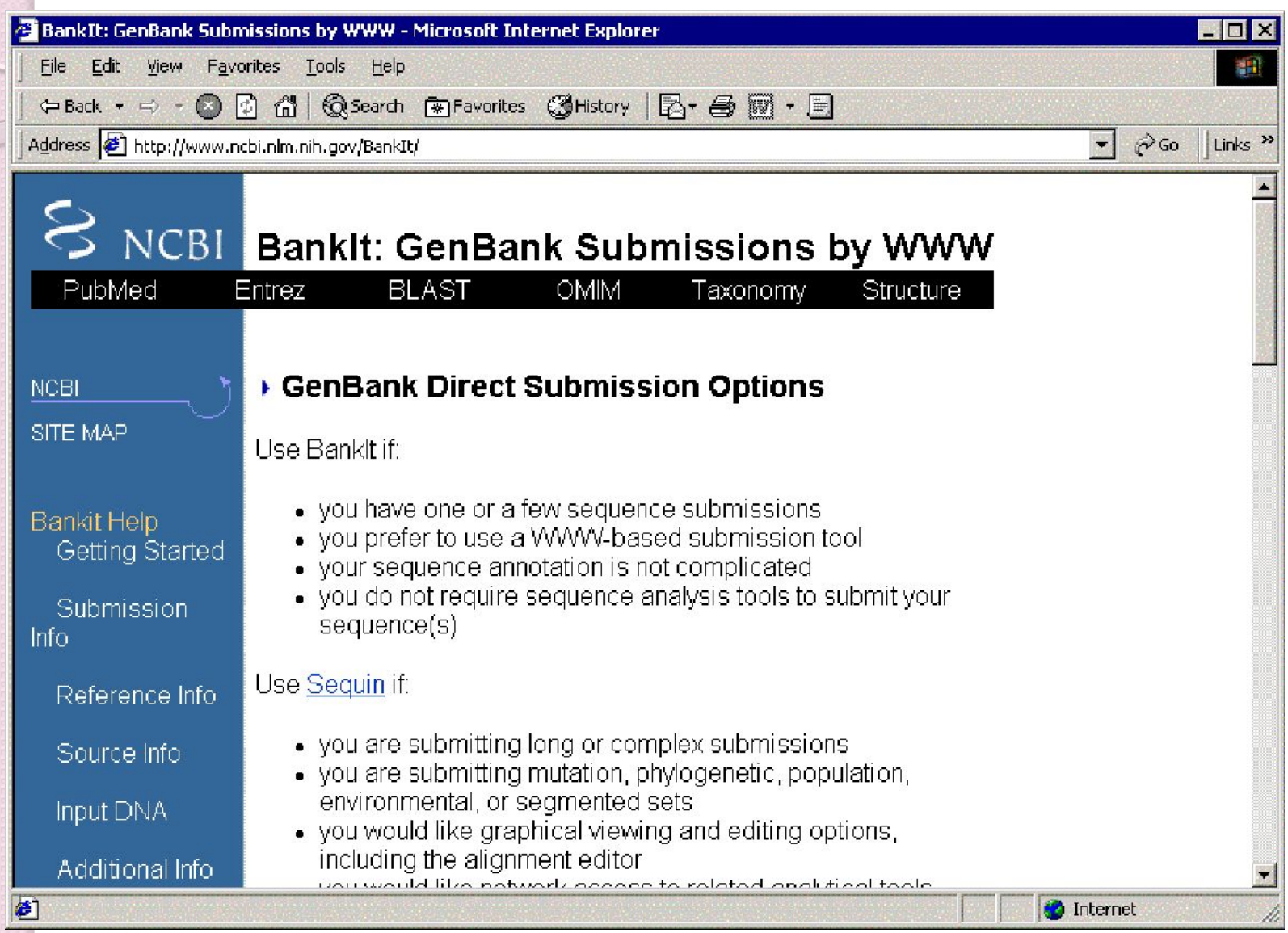
- При написании однобуквенного кода всегда используйте заглавные буквы;
- При работе с FASTA-последовательностями на ПК всегда используйте опцию TEXT;
- При работе с FASTA-форматом в текстовом процессоре Word, всегда используйте исключительно ASCII символы;
- Для правильного отображения этих последовательностей 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

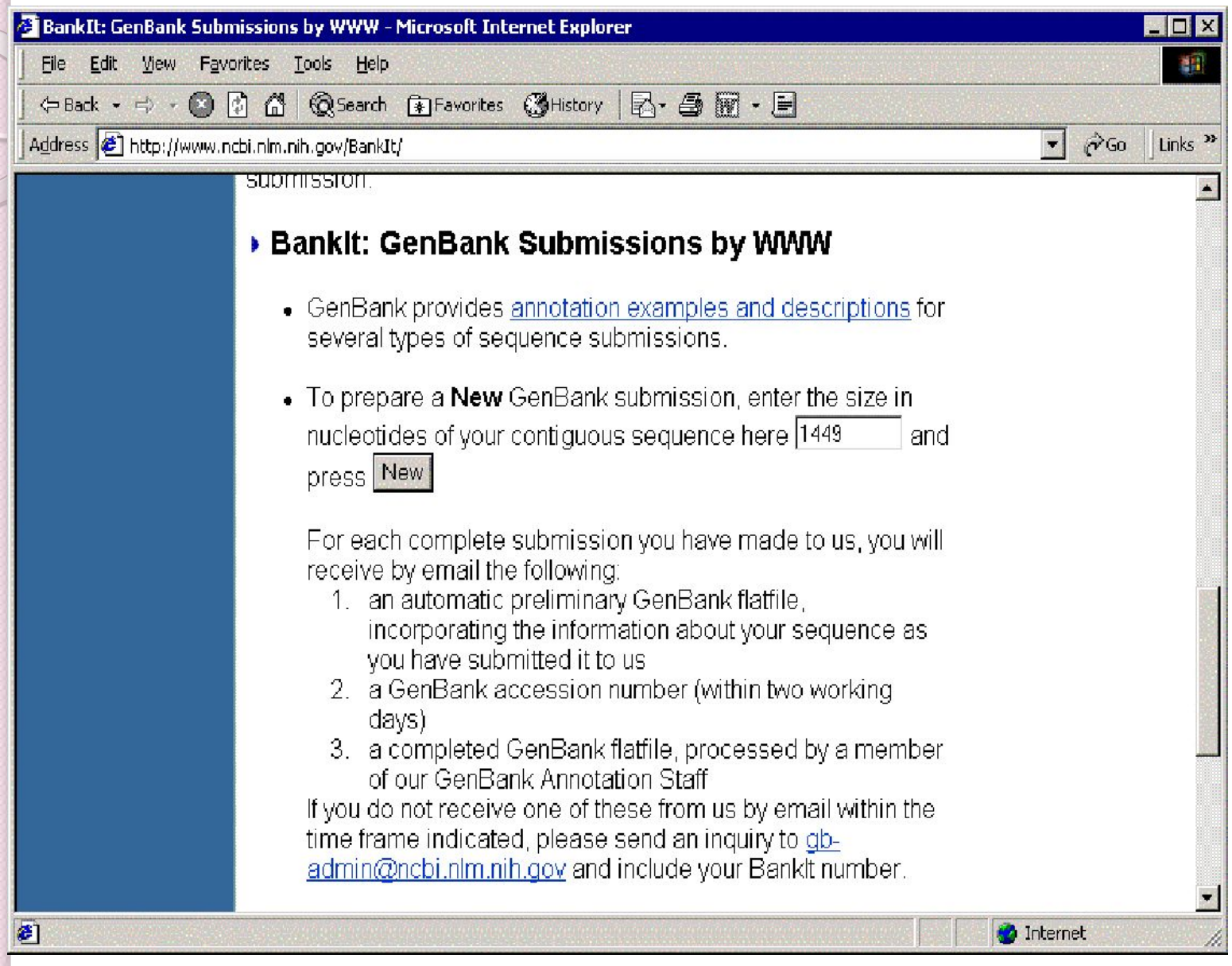
Isolate P876, 16S rRNA gene sequence. Length: 1449 bp

TGCAAGTCGA ACGGTAGCAG GAAGAAAGCT TGCTTTCTTT GCTGACGAGT GGCGGACGGG TGAGTAATGC TTGGGAATCT GGCTTATGGA GGGGGATAAC
TGTGGGAAAC TGCAGCTAAT ACCGCGTAAT CTCTGAGGAG TAAAGGGTGG GACyTTAGGG CCACCTGCCA TAAGATGAGC CCAAGTGGGA TTAGGTAGTT
GGTGGGGTAA AGGCCTACCA AGCCTGCGAT CTCTAGCTGG TCTGAGAGGA TGACCAGCCA CACTGGAACT GAGACACGGT CCAGACTCCT ACGGGAGGCA
GCAGTGGGGA ATATTGCGCA ATGGGGGGAA CCCTGACGCA GCCATGCCGC GTGAATGAAG AAGGCCTTCG GGTTGTAAAG TTCTTTCGGT AATGAGGAAG
GGGTGTTrTT kAATAGATAG CATCATTGAC GTTAATTACA GAAGAAGCAC CGGCTAACTC CGTGCCAGCA GCCGCGGTAA TACGGAGGGT GCGAGCGTTA
ATCGGAATAA CTGGGCGTAA AGGGCACGCA GGCGGACTTT TAAGTGAGAT GTGAAATCCC CGAGCTTAAC TTGGGAATTG CATTTCAGAC 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 AAAACTCAAA TGAATTGACG GGGGCCCGCA CAAGCGGTGG AGCATGTGGT
TTAATTCGAT GCAACGCGAA GAACCTTACC TACTCTTGAC ATCCTAAGAA GAGCTCAGAG ATGAGCTTGT GCCTTCGGGA ACTTAGAGAC AGGTGCTGCA
TGGCTGTCGT CAGCTCGTGT TGTGAAATGT TGGGTTAAGT CCCGCAACGA GCGCAACCCT TATCCTTTGT TGCCAGCGAT TTGGTCGGGA ACTCAAAGGA
GACTGCCAGT GACAAACTGG AGGAAGGTGG GGATGACGTC AAGTCATCAT GGCCCTTACG AGTAGGGCTA CACACGTGCT ACAATGGTGC ATACAGAGGG
CAGCGAGAGT GCGAGCTTAA GCGAATCTCA GAAAGTGCAT CTAAGTCCGG ATTGGAGTCT GCAACTCGAC TCCATGAAGT CGGAATCGCT AGTAATCGCA
AATCAGAATG TTGCGGTGAA TACGTTCCCG GGCCTTGTAC ACACCGCCCG TCACACCATG GGAGTGGGTT GTACCAGAAG TAGATAGCTT AACCTTCGGG
AGGGCGTTTA CCACGGTATG ATTCATGACT GGGGTGAAGT CGTAACAGA

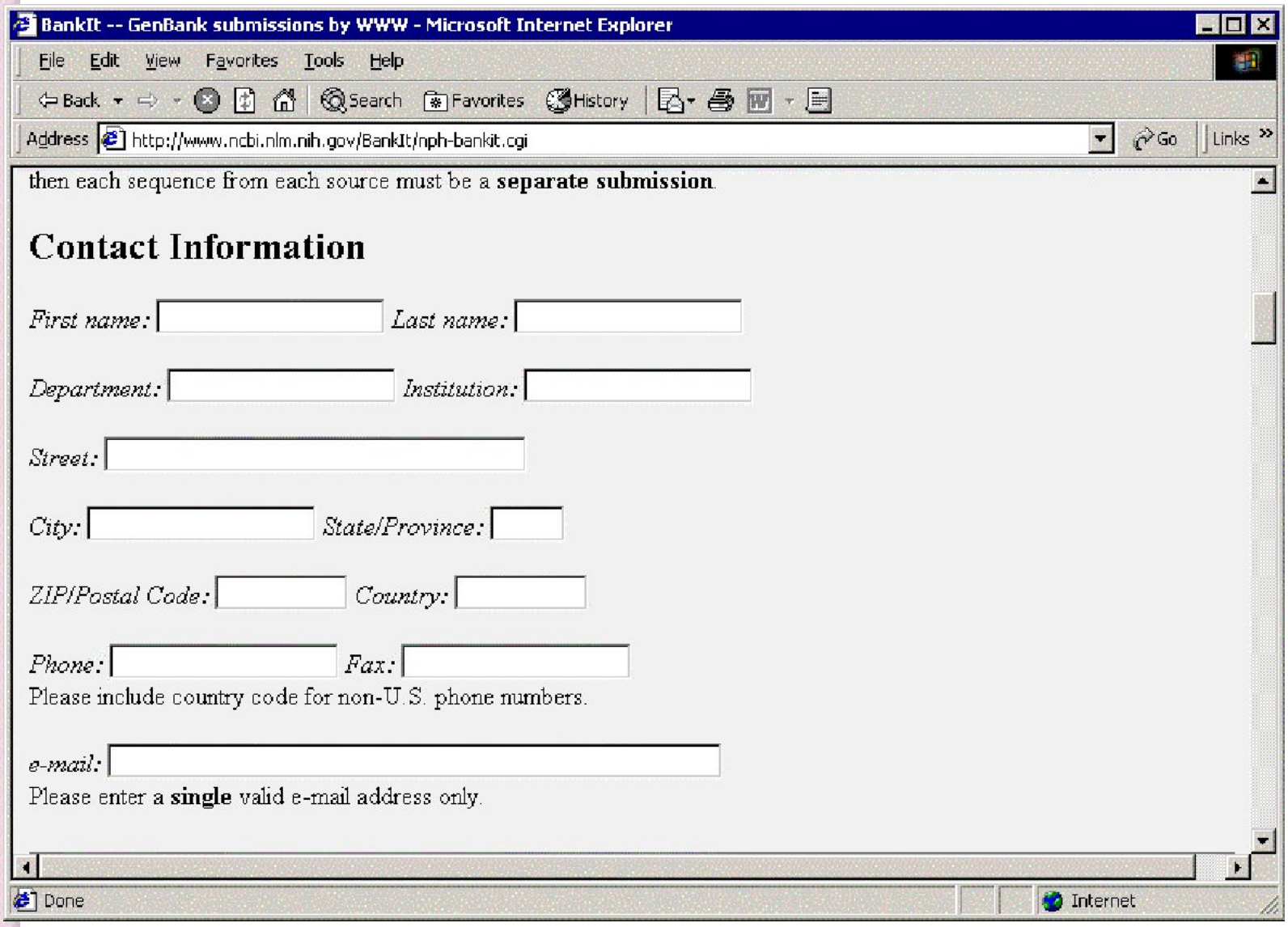
Подача в GenBank при помощи инструмента BankIt



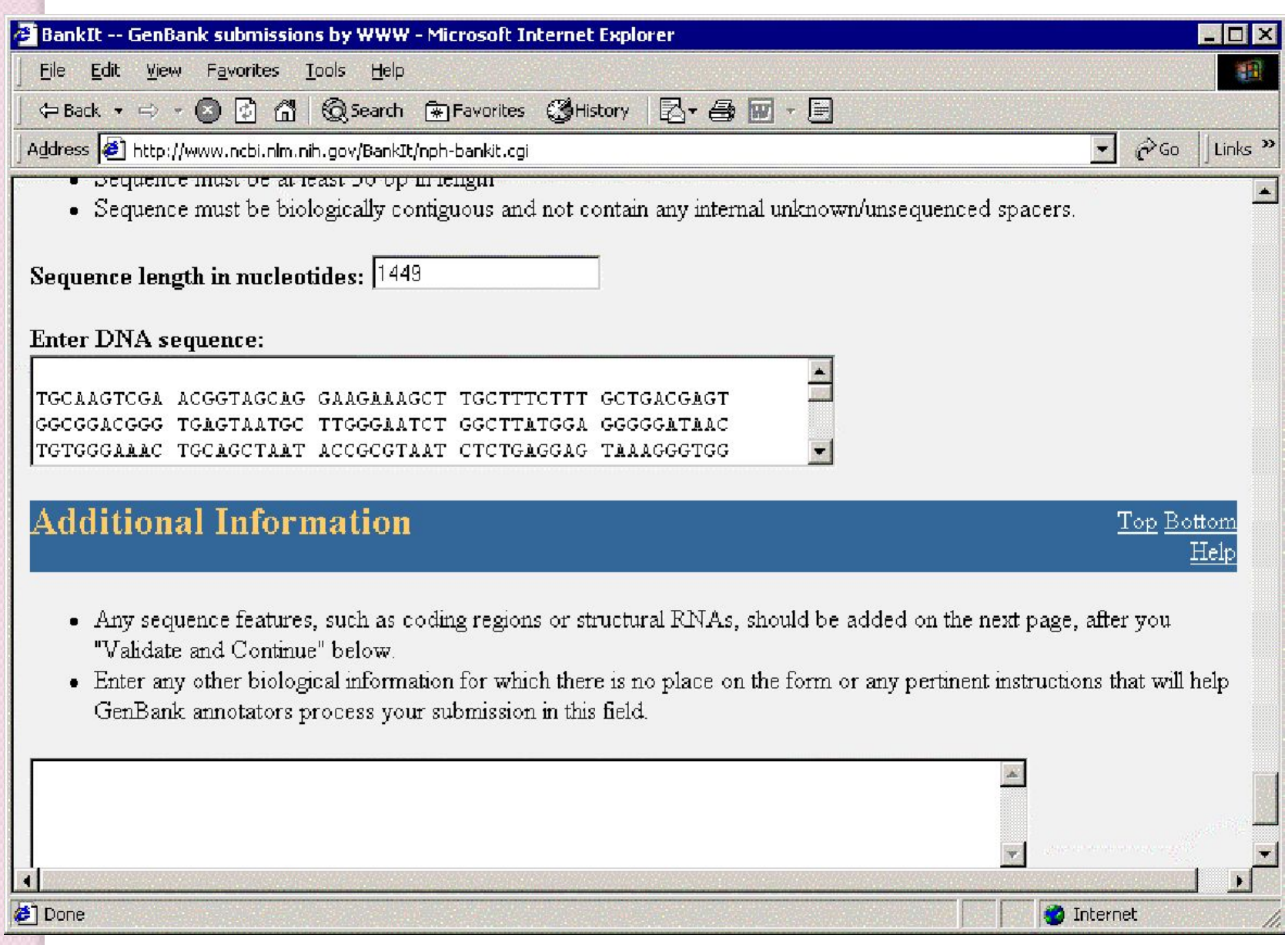
ШАГ 1.
Резервирование места в базе данных



ШАГ 2.
Контактная информация

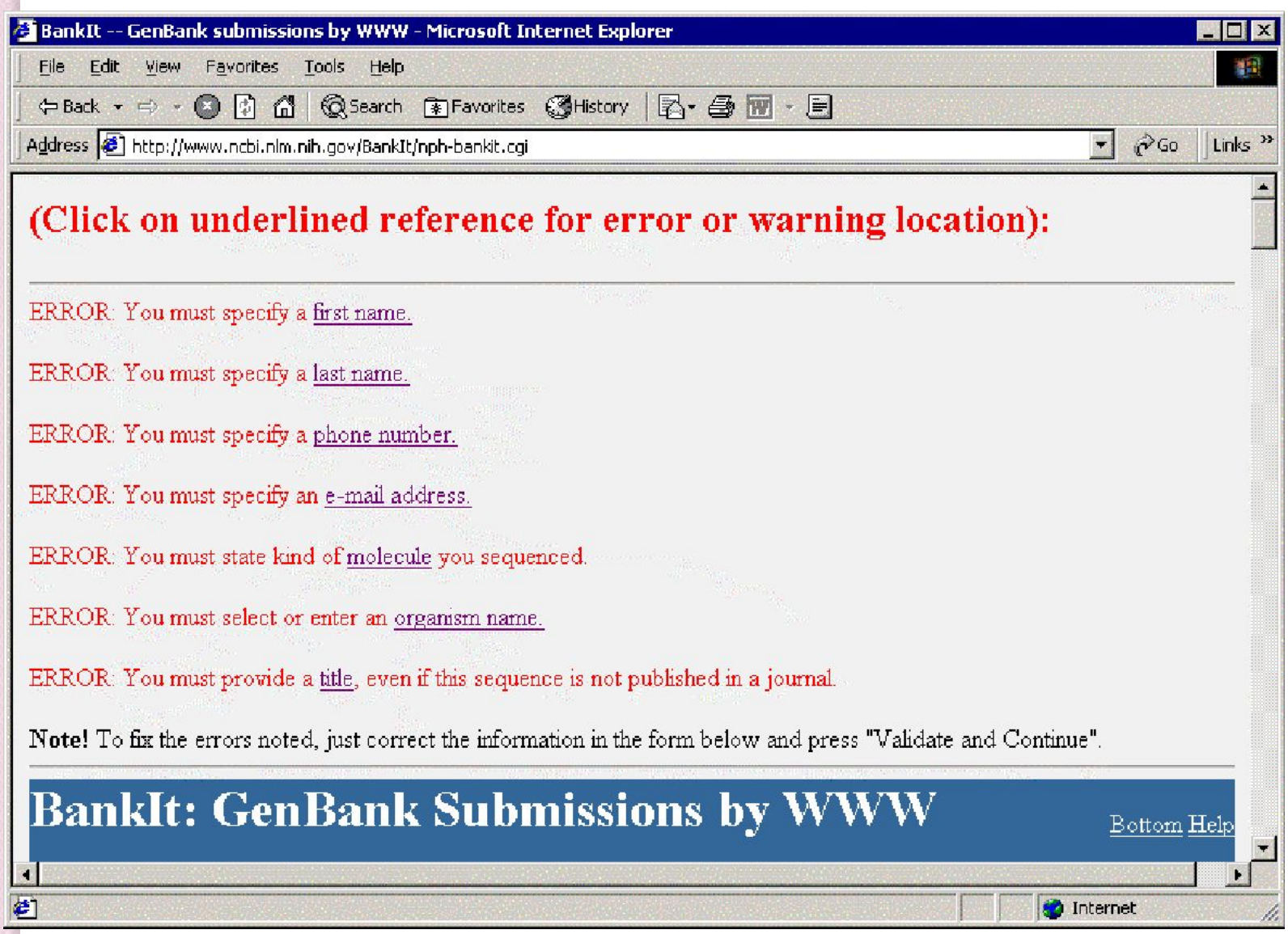


ШАГ 3.
Внесение текста последовательности.



ШАГ 4.

Подтверждение заявки и возможные ошибки.



Дякую за увагу
Благодарю за внимание
Thank you for your attention