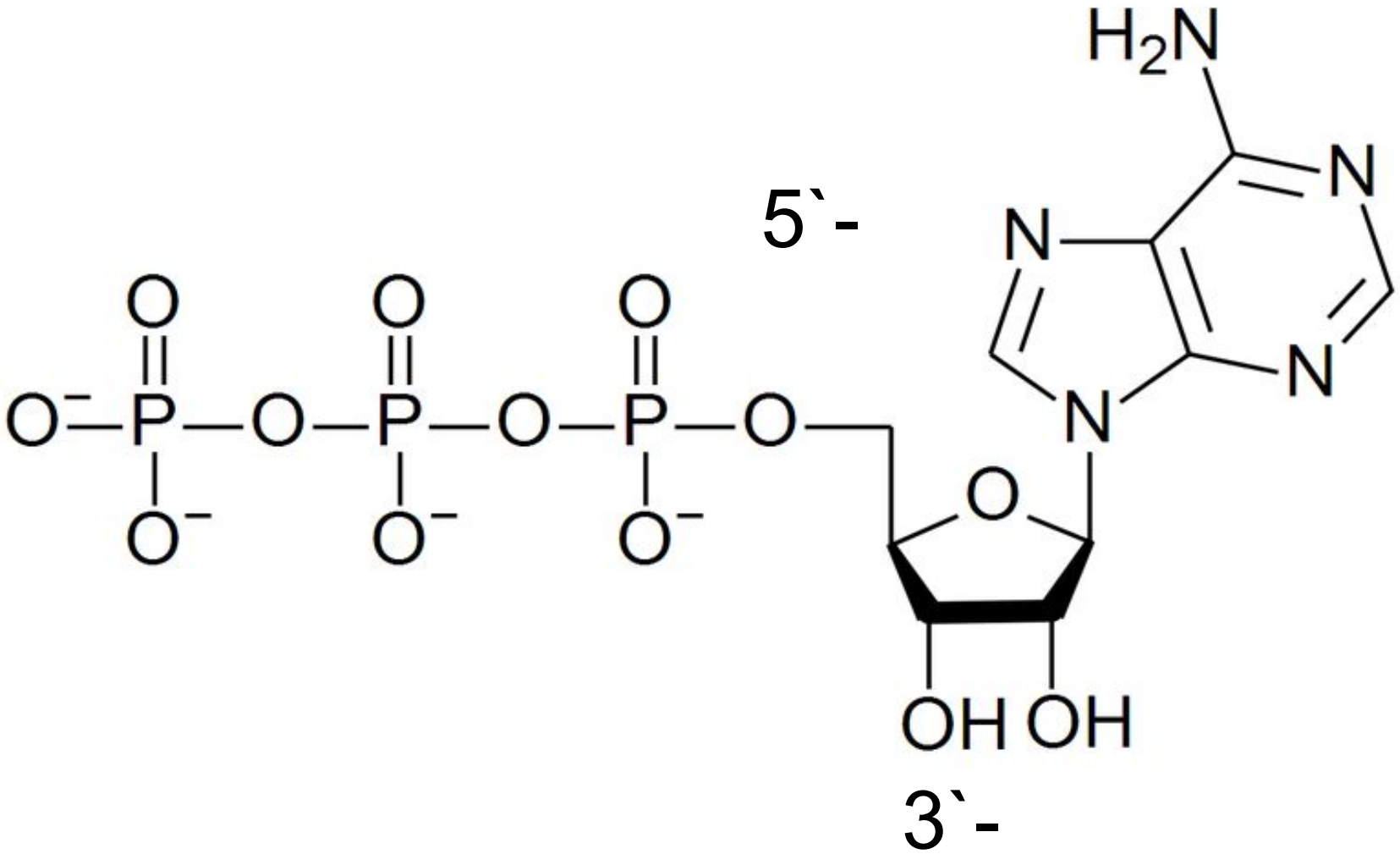
DNA vs. computer

1. Про 5’-3’ и всякую химию
2. Про банки данных (архивные vs. курируемые)
3. Святая троица EMBL – GenBank – DDBJ
4. Собственно EMBL, его разделы, классы даных и поля; CDS, кодирующие участки, ссылки из Swiss-Prot.

АТФ



Как записывают последовательности нуклеиновых кислот ?

1. Последовательность = последовательность однобуквенных символов.
   Никаких дефисов и обозначений фосфодиэфирных связей.

2. Одни и те же однобуквенные символы для последовательностей РНК и ДНК (при записи РНК обычно ‘U’ ⇒ ‘T’).
   Любая последовательность по умолчанию считается ДНК
   (т.е. полимером 2'-дезоксирибонуклеотидов).

3. Одни и те же символы используются для обозначения азотистых оснований, нуклеозидов и нуклеотидов
   Допустимы заглавные и строчные буквы, хотя рекомендованы заглавные.

4. **Последовательность записывается в направлении 5'→3'**

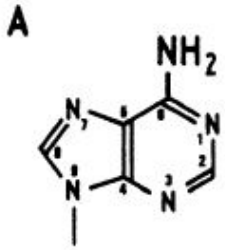
Пример:

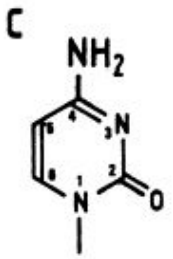
5'-CTCGAC-3'

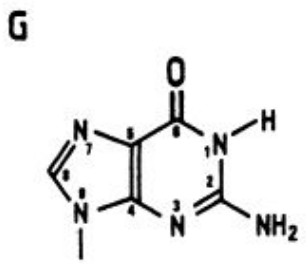
Nomenclature Committee of the International Union of Biochemistry (NC-IUB)
Nomenclature for incompletely specified bases in nucleic acid sequences
Recommendations 1984
Biochem. J. (1985) 229, 281-286

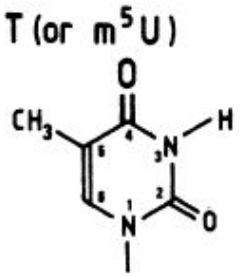
Общепринятые однобуквенные обозначения для стандартных азотистых оснований (остатков нуклеозидов и нуклеотидов) и вырожденных позиций в выравниваниях нуклеиновых кислот

Символ	Расшифровка	Происхождение обозначения
G	G, guanine (гуанин)	
A	A, adenine, (аденин)	
T	T/U, thymine/uracyl (тимин в ДНК и урацил в РНК)	
C	C, cytosine (цитозин)	
R	A или G, purine(пурины)	pu**R**ine
Y	C или T или U, pyrimidine (пиримидин)	p**Y**rimidine
M	A или C	a**M**ino
K	G или T	**K**eto
S	G или C	**S**trong interaction (3 H bonds)
W	A или T	**W**eak interaction (2 H bonds)
H	A или C или T , но не G	в алфавите 'H' следует за 'G'
B	G или T или C, но не A	'B' следует за 'A'
V	G или C или A , но не T или U	'V' следует за 'U'
D	G или A или T , но не C	'D' следует за 'C'
N	G или A или T или C	a**N**y



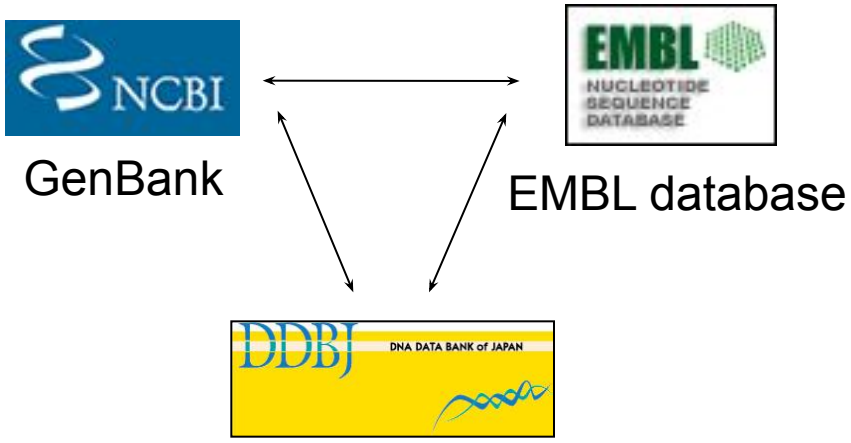






NCBI и EBI

- National Center for Biotechnology Information и European Bioinformatics Institute (подразделение EMBL – European Molecular Biology Laboratory)
- Три базы данных – GenBank, EMBL и DDBJ (японская) – по сути, одно и то же.



Что надо знать про банк EMBL

-что это архив (за содержание записи несёт ответственность её автор)

- поэтому разнобой в терминологии
- поэтому одно и то же по многу раз
- поэтому много неисправленных ошибок

- что у последовательности из записи часто нет естественных границ

- что это часть триединства (EMBL, GenBank, DDBJ)

- ежедневный обмен данными

- … ну и смысл основных полей, конечно (особенно структуру поля FT!)

GenBank EMBL DDBJ

компьютерный поиск гена, трансляция и компьютерная аннотация

Базы данных научной литературы

TrEMBL TRANSLATED EMBL ~2 500 000 последовательностей

Экспертиза

swissprot ~200 000 последовательностей

UniParc (UniProt Archive)

PIR-PSD pir.georgetown.edu

UniProt the universal protein resource

UniRef (UniProt non-redundant Reference databases)

Класс данных

3.1 Data Class

The data class of each entry, representing a methodological approach to the generation of the data or a type of data, is indicated on the first (ID) line of the entry. Each entry belongs to exactly one data class.

Class	Definition
CON	Entry constructed from segment entry sequences, drawing annotation from segment entries
ANN	Entry constructed from segment entry sequences with its own annotation
PAT	Patent
EST	Expressed Sequence Tag
GSS	Genome Survey Sequence
HTC	High Thoughput CDNA sequencing
HTG	High Thoughput Genome sequencing
MGA	Mass Genome Annotation
WGS	Whole Genome Shotgun
TPA	Third Party Annotation
STS	Sequence Tagged Site
STD	Standard (all entries not classified as above)

Таксономический раздел

3.2 Taxonomic Division

The entries which constitute the database are grouped into taxonomic divisions, the object being to create subsets of the database which reflect areas of interest for many users.

In addition to the division, each entry contains a full taxonomic classification of the organism that was the source of the stored sequence, from kingdom down to genus and species (see below).

Each entry belongs to exactly one taxonomic division. The ID line of each entry indicates its taxonomic division, using the three letter codes shown below:

Division	Code
Bacteriophage	PHG
Environmental Sample	ENV
Fungal	FUN
Human	HUM
Invertebrate	INV
Other Mammal	MAM
Other Vertebrate	VRT
Mus musculus	MUS
Plant	PLN
Prokaryote	PRO
Other Rodent	ROD
Synthetic	SYN
Transgenic	TGN
Unclassified	UNC
Viral	VRL

Поле

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

FT

FT Key Location/Qualifiers=value

FT CDS 1..1000
/codon=(seq:"cug",aa:Ser)
/codon=(seq:"tga",aa:Trp)

http://www.ebi.ac.uk/embl/WebFeat/index.html

CDS и exons

CDS – кодирующая последовательность, то есть ровно те нуклеотиды, что соответствуют белку (по крайней мере его основной форме).

Кодирующие участки – те фрагменты ДНК, из которых составлен CDS.

Exons – экзоны, то из чего будет составлена зрелая матричная РНК, они включают в себя 5` и 3` - нетранслируемые области – те части РНК, которые отвечают за регуляцию трансляции.

Ссылки из записи Swiss-Prot на EMBL

	(2001) *Proteins* **44**:270-281
	Position: X-RAY CRYSTALLOGRAPHY (1.89 ANGSTROMS).
	Medline: 21348730
	DOI: 10.1002/prot.1092;
	PubMed: 11455600 CiteXplore

Comments	
FUNCTION	Decomposes hydrogen peroxide into water and oxygen; serves to protect cells from the toxic effects of hydrogen peroxide.
CATALYTIC ACTIVITY	2 H(2)O(2) = O(2) + 2 H(2)O.
COFACTOR	Heme group.
SUBUNIT	Homotetramer.
SUBCELLULAR LOCATION	Cytoplasm (Probable).
INDUCTION	By entry into stationary phase.
SIMILARITY	Belongs to the catalase family. HPII subfamily.

Copyright

Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms Distributed under the Creative Commons Attribution-NoDerivs License

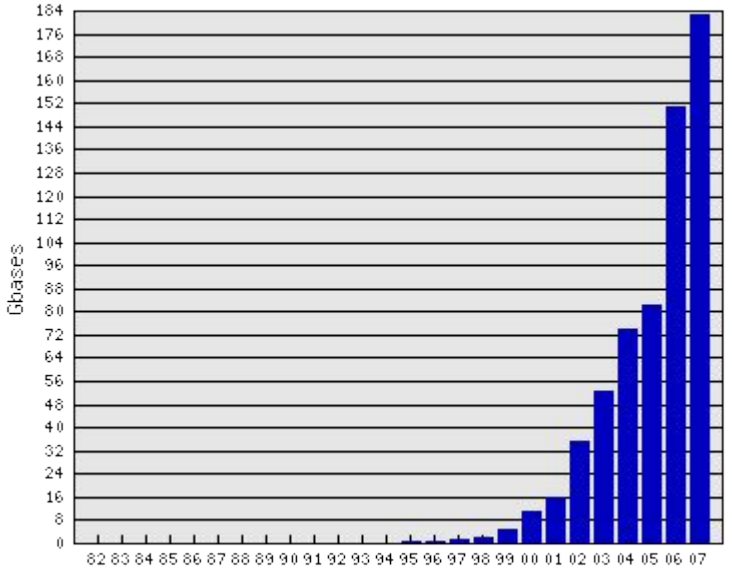
Database cross-references

EMBL	M55161; AAA24039.1; -; Genomic_DNA. U00096; AAT48137.1; -; Genomic_DNA. AP009048; BAA15513.1; -; Genomic_DNA.
PIR	A39129; A39129.
RefSeq	AP_002351.1; -. YP_025308.1; -.
PDB	1CF9; X-ray; 1.80 A; A/B/C/D=27-753. 1GG9; X-ray; 1.89 A; A/B/C/D=1-753. 1GGE; X-ray; 1.89 A; A/B/C/D=1-753. 1GGF; X-ray; 2.28 A; A/B/C/D=1-753. 1GGH; X-ray; 2.15 A; A/B/C/D=1-753. 1GGJ; X-ray; 1.92 A; A/B/C/D=1-753. 1GGK; X-ray; 2.26 A; A/B/C/D=1-753. 1IPH; X-ray; 2.80 A; A/B/C/D=1-753. 1P7Y; X-ray; 2.40 A; A/B/C/D=1-753.

Каждая строка – отдельный сиквенс; первая ссылка в строке – запись в EMBL, вторая – CDS внутри этой записи (здесь идентификатор, например, AAA24039.1 – это идентификатор CDS в специальном дочернем банке данных EMBL-Coding sequences).

Статистика EMBL

Total nucleotides



Number of entries