

wykład 3

# **Bazy danych**

dr Jacek Śmietański

[jacek.smietanski@ii.uj.edu.pl](mailto:jacek.smietanski@ii.uj.edu.pl)

<http://jaceksmietanski.net>

1. Źródła wiedzy biologicznej
2. GenBank / RefSeq
3. UniProt / SwissProt
4. PDB
5. Systemy powiązań
6. Inne popularne bazy
7. Poszukiwanie baz specjalistycznych

# Źródła wiedzy biologicznej

## **Dane eksperymentalne** (badania *in vitro*, *in vivo*)

- sekwencjonowanie
  - eksperymenty mikromacierzowe
  - rentgenografia (X-ray)
  - rezonans magnetyczny (NMR)
- itp.

## **Dane obliczeniowe** (badania *in silico*)

- analizy porównawcze (homologia)
  - eksploracja danych – odkrywanie wiedzy
  - przewidywanie ab initio
  - symulacje (dynamika molekularna, obliczenia kwantowo-mechaniczne)
- itp.



Bezpłatne bazy sekwencji nukleotydowych oraz struktur białkowych.



Ogromna liczba baz danych – w większości darmowych i dostępnych bez ograniczeń (on-line; FTP).



Narzędzia wyszukiwania dostępne on-line.



Liczne adnotacje i powiązania między różnymi bazami.



Wiele projektów *open source*.



Literatura naukowa coraz częściej udostępniana w trybie *open access* (za darmo dla wszystkich).

# Ile jest tych danych?

Liczba zdeponowanych danych rośnie bardzo szybko (choć w niektórych grupach natknęliśmy już na barierę).

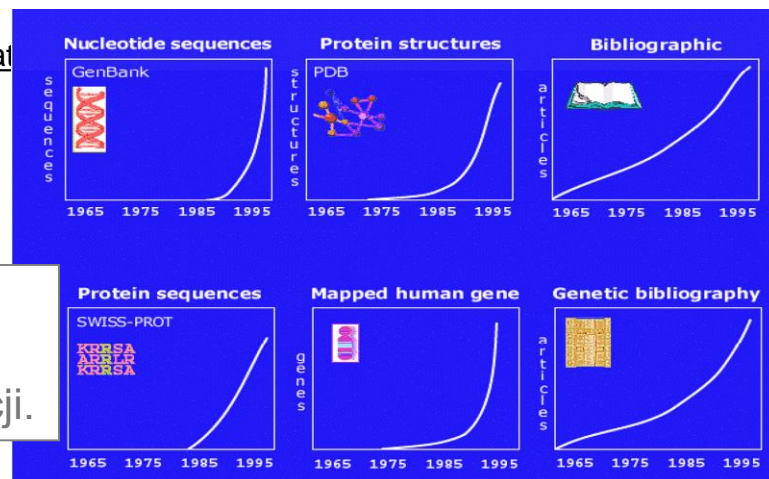
Np. GenBank – ponad **260 mld** nukleotydów;  
– ponad **208 mln** sekwencji genów;  
(sierpień 2018; <http://www.ncbi.nlm.nih.gov/genbank/statistics>)

Dane WGS: 3,2 bln nukleotydów – sierpień 2018; 2,2 bln nukleotydów – sierpień 2018.

UniProt/SwissProt – **558 tys** sekwencji białkowych (październik 2018; 552 tys – październik 2016),  
<http://www.uniprot.org/statistics/Swiss-Prot>)

UniProt/TrEMBL – **127 mln** sekwencji aminokwasowych  
(październik 2018; 67 mln - wrzesień 2015, <http://www.uniprot.org/statistics/TrEMBL>)

PDB – **145 tys** struktur (październik 2018,  
[http://www.rcsb.org/pdb/static.do?p=general\\_information/pdb\\_statistics](http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics))



Do pewnego momentu liczba deponowanych danych rosła w tempie wykładniczym (wykres obok z roku 1995).  
Dzisiaj w większości baz nie obserwujemy już takiej tendencji.

- Poszukiwanie dodatkowych informacji o badanej sekwencji (np. literatura, adnotacje).
- Poszukiwanie sekwencji homologicznych.
- Określenie, czy dana sekwencja jest już zdeponowana w bazie danych.
- Poszukiwanie sekwencji DNA homologicznej do niekodujących regionów DNA (np. sekwencje regulatorowe, elementy powtarzające się).
- Poszukiwanie sekwencji nadających się do wykorzystania w PCR.
- Poszukiwanie charakterystycznych motywów sekwencyjnych lub strukturalnych.
- Poszukiwanie / przewidywanie struktury, aktywności lub funkcji nieznanej sekwencji.
- itp.

## Wyszukiwanie tekstowe (np. ENTREZ)

LOCUS	SCU49845	5028 bp	DNA	PLN	21-JUN-1999
DEFINITION	Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.				
ACCESSION	U49845				
VERSION	U49845.1 GI:1293613				
KEYWORDS	.				
SOURCE	Saccharomyces cerevisiae (baker's yeast)				
ORGANISM	Saccharomyces cerevisiae Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.				
REFERENCE	1 (bases 1 to 5028)				
AUTHORS	Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.				
TITLE	Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in Saccharomyces cerevisiae				
JOURNAL	Yeast 10 (11), 1503-1509 (1994)				

## Wyszukiwanie na podstawie sekwencji (np. BLAST)

**Hemoglobin α** VLSPADKTNVKAAMGKVGAAHAGEYCAEALERMFLSFPTTKTYFPHF-----D  
**Myoglobin** GLSEG EWQLVLNVAGKVEADIPGHQCEVLIIRLFKGGHPETLEKFDKFKHLKSED

LSHGSAQVKGGHKKVADALTNAVAHVDDMPNALSA LSDLHAHKL R VDPVNKK L  
EMKASEDLKKHGATVLTALGGI LKKKGHHEAE I KPLAQSHATKH K I PVKYLE F

LSHCLLMTLAAHLPAEFTPAVHASLDKFLASVSTVLT SKYR  
I SEC I IQVLQSKHPGDFCADAQGAMNKA LELFRKDMASNYKELGFQG



## Przeszłość

Bezproblemowo



Mała ilość informacji,  
nieduża złożoność  
zadań => użytkownik  
daje sobie radę

## Teraźniejszość

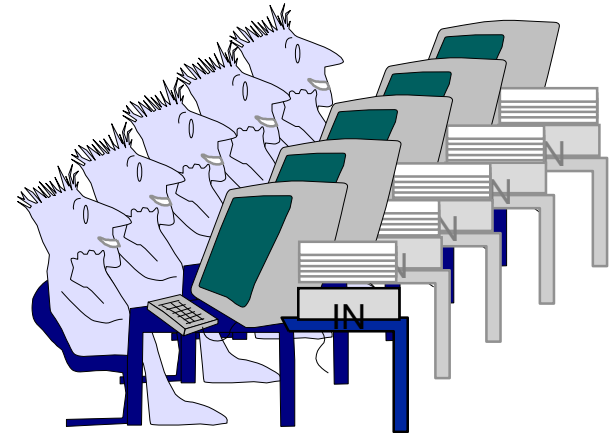
Potrzeba automatyzacji



Trudność w dostępie do  
dużej ilości informacji,  
rośnie trudność zadań  
=> użytkownik pod  
presją

## Przyszłość

Nowa jakość podziału zadań



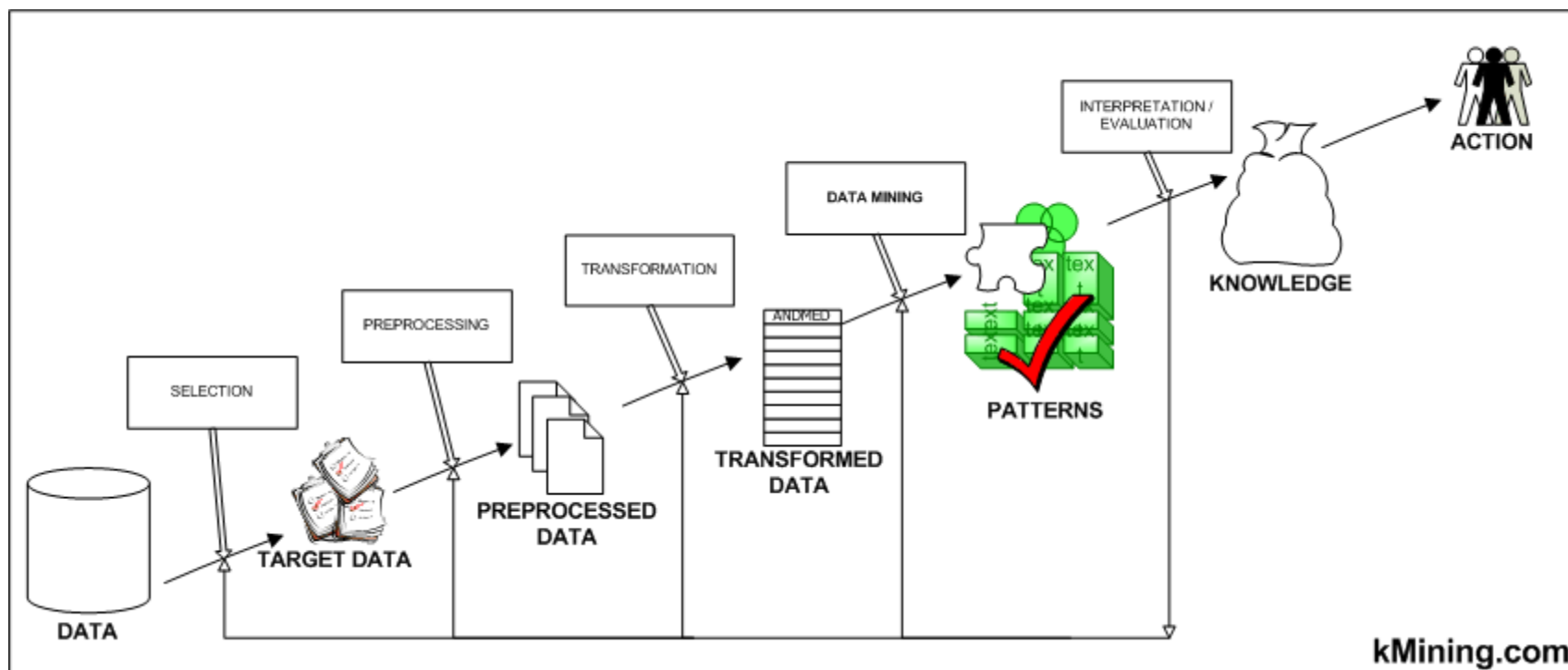
Duża ilość informacji,  
duża trudność zadań +  
wspierające systemy  
decyzyjne => użytkownik  
daje sobie radę

- błędne dane
- niekompletne dane
- powtarzające się dane (np. identyczne sekwencje)
- brak spójności
- niejednoznaczne nazewnictwo
- niejawne powiązania
- modyfikacja wpisu (wpływ na bazy zależne; dotychczasowe publikacje itp.)
- odnajdowanie i oznaczanie błędów
- itp.

## Knowledge discovery

poszukiwanie powiązań między informacjami, których nie znano w momencie wprowadzania danych do bazy.

(wykorzystujemy metody eksploracji danych – *data mining*)



## Pierwszorzędowe (pierwotne)

surowe dane eksperymentalne

## Drugorzędowe (wtórne)

dane zawierające dodatkowe adnotacje, powiązania  
często nieredundantne

## Specjalistyczne

przetworzone, bądź np. dedykowane konkretnym  
organizmom, chorobom itp.

## **GenBank / RefSeq**

baza sekwencji nukleotydowych (baza genów)

## **GEO**

baza eksperymentów mikromacierzowych

## **UniProt / Swiss-Prot**

baza sekwencji aminokwasowych (baza białek)

## **PDB**

baza struktur przestrzennych

# GenBank / RefSeq

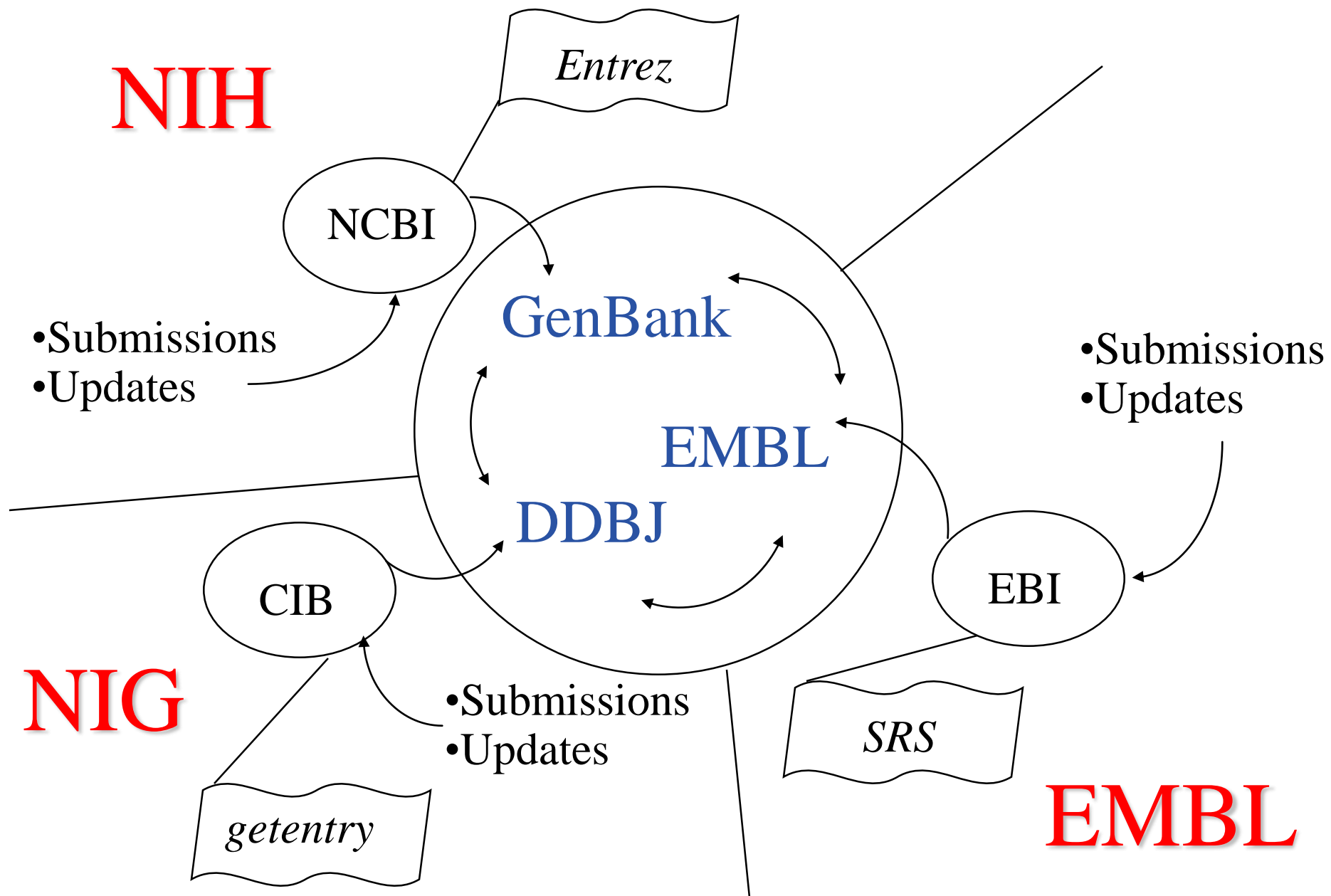
Baza sekwencji nukleotydowych, zarządzana przez NCBI (USA).

Analogiczne bazy funkcjonują również w Europie (EMBL) i Japonii (DDBJ).

Poszczególne bazy wymieniają informacje między sobą.

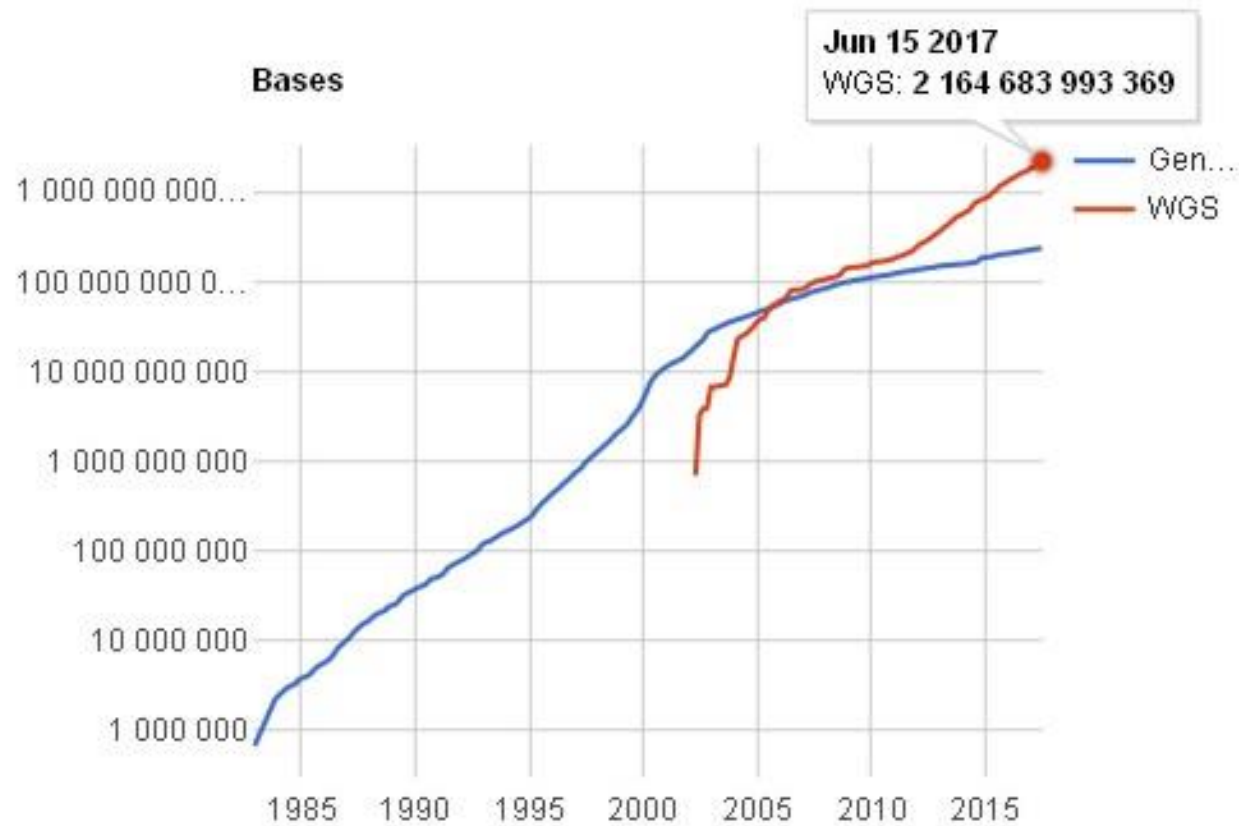
Dostępne on-line i przez FTP.

Autorzy samodzielnie wprowadzają nowe sekwencje (warunek publikacji).



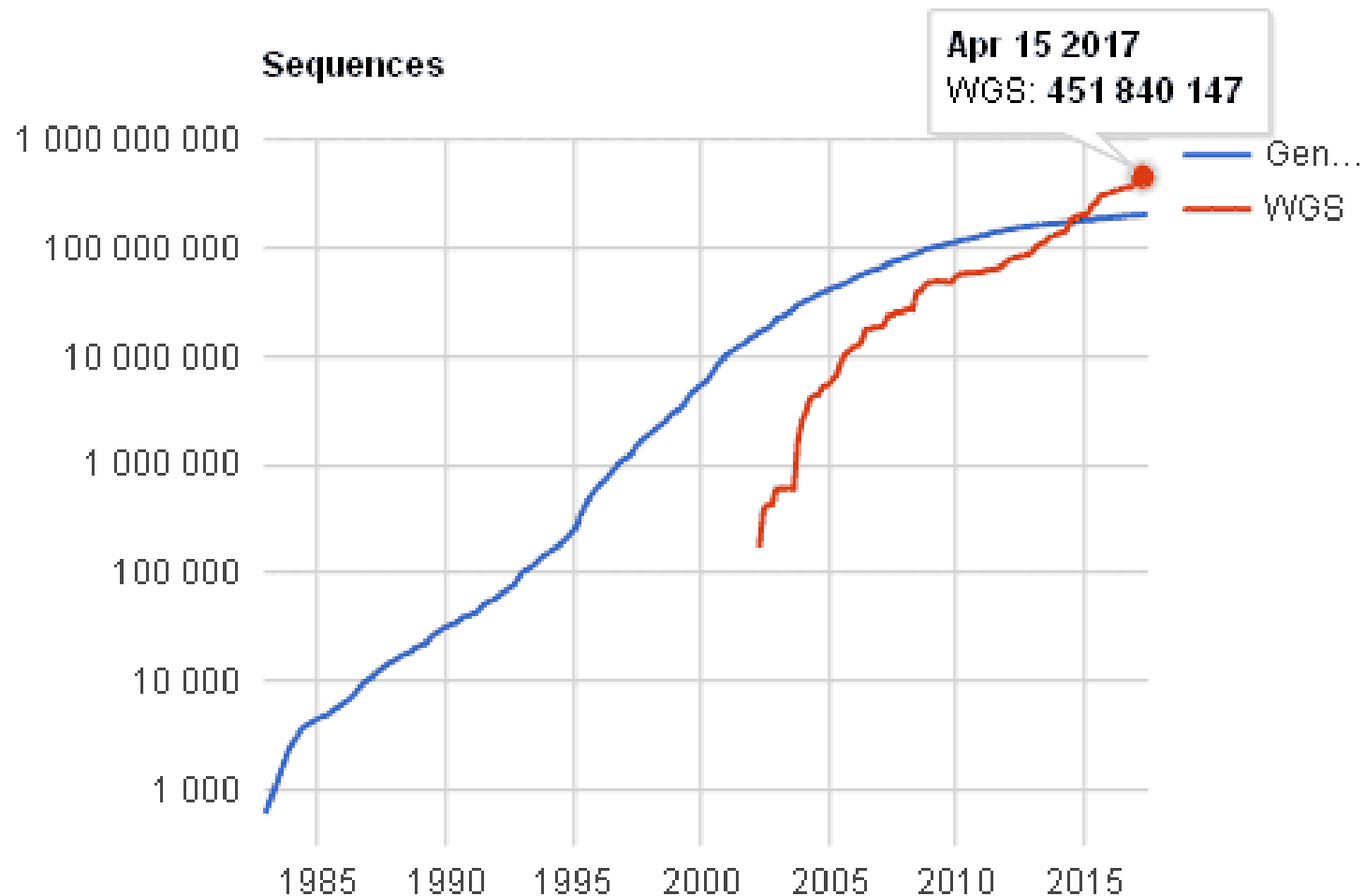


## GenBank and WGS Statistics



Skala wykładnicza...

<http://www.ncbi.nlm.nih.gov/genbank/statistics>



Skala wykładnicza...

<http://www.ncbi.nlm.nih.gov/genbank/statistics>

# Whole Genome Shotgun Submissions

<https://www.ncbi.nlm.nih.gov/Traces/wgs/?view=WGS>

Found 133,610 projects											Download		Columns List		Page	1	2	3	4	5	6	7	8	...	2673	( 50	per page)
#	Prefix	Type	Targeted Locus Name	DIV	Organism	Bioproject	Biosample	Intraspecific Name	Other Source	Total length (Mbases)	Contigs			#	# Prot	Has Annot	#	# Prot									
											#	#	#														
1	LMTP01	WGS		PLN	Pinus lambertiana	PRJNA174450	SAMN03354659	isolate: USFS_5038	tissue_type: megagametophyte	26,323.9	4,253,097																
2	APFE02	WGS		PLN	Pinus taeda	PRJNA174450	SAMN02981512		tissue_type: needle	21,039.8	7,082,509																
3	JZKD01	WGS		PLN	Picea glauca	PRJNA242552	SAMN02736787		tissue_type: needles	20,925.9	5,261,503						3,353,683										
4	ALWZ04	WGS		PLN	Picea glauca	PRJNA83435	SAMN01120252		tissue_type: Flushing bud	19,229.5	6,221,640			Y			3,033,285										
5	LPNX01	WGS		PLN	Pseudotsuga menziesii	PRJNA174450	SAMN03333061	isolate: Weyco1	dev_stage: megagametophyte tissue_type: megagametophyte	13,993.5	1,236,665																
6	AUXO01	WGS		ENV	gut metagenome	PRJNA202380	SAMN02715735		host: Ovis aries isolation_source: sheep rumen	5,515.9	8,786,927			Y													
7	AVCP01	WGS		INV	Locusta migratoria	PRJNA185471	SAMN02261463		dev_stage: adult tissue_type: whole body	5,493	1,397,492			Y													
8	AZMS01	WGS		INV	Acanthoscurria geniculata	PRJNA222716	SAMN02720822			4,865.8	12,478,692			Y			4,986,575										
9	AACV02	WGS		ENV	marine	PRJNA12694	SAMN02954245		isolation_source:	4,261.5	4,124,495						2,087,206 6,122,306										

# Organismal Divisions

		Used in which database?
BCT	Bacterial	DDBJ - GenBank
FUN	Fungal	EMBL
HUM	Homo sapiens	DDBJ - EMBL
INV	Invertebrate	all
MAM	Other mammalian	all
ORG	Organelle	EMBL
PHG	Phage	all
PLN	Plant	all
PRI	Primate (also see HUM)	all (not same data in all)
PRO	Prokaryotic	EMBL
ROD	Rodent	all
SYN	Synthetic and chimeric	all
VRL	Viral	all
VRT	Other vertebrate	all

# Functional Divisions

**PAT** Patent

**EST** Expressed Sequence Tags

**STS** Sequence Tagged Site

**GSS** Genome Survey Sequence

**HTG** High Throughput Genome (unfinished)

**HTC** High throughput cDNA (unfinished)

**CON** Contig assembly instructions

Organismal divisions:

**BCT**    **FUN**    **INV**    **MAM**    **PHG**    **PLN**

**PRI**    **ROD**    **SYN**    **VRL**    **VRT**

## Strona domowa:

<http://www.ncbi.nlm.nih.gov/genbank/>

## Przykładowy rekord, opis formatu:

<http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>

## Wyszukiwanie:

Entrez – nukleotydy:

<http://www.ncbi.nlm.nih.gov/nucleotide/>

Główna kolekcja GenBank (bez sekwencji EST):

<http://www.ncbi.nlm.nih.gov/nuccore/>

BLAST:

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

# GenBank – format pliku GBFF [GenBank Flat File] (1): nagłówek

```
LOCUS      SCU49845      5028 bp      DNA      PLN      21-JUN-1999
DEFINITION Saccharomyces cerevisiae TSP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION  U49845
VERSION   U49845.1  GI:129361
SOURCE     Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE  1  (bases 1 to 5028)
AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Torpey,C.W.
TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
JOURNAL   Yeast 10 (11), 1503-1509 (1994)
PUBMED    7871890
REFERENCE  2  (bases 1 to 5028)
AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
JOURNAL   Genes Dev. 10 (7), 777-792 (1996)
```

identyfikator sekwencji

typ cząsteczki

długość sekwencji

identyfikator działu

## Identyfikatory:

- niezmiennie w czasie
- zawsze odnoszące się do określonych sekwencji
- umożliwiające przegląd historii aktualizacji

University, New

# GenBank – format pliku GBFF (2): właściwości

```
FEATURES                      Location/Qualifiers
     source                    1..5028
                                /organism="Saccharomyces cerevisiae"
                                /db_xref="taxon:4932"
                                /chromosome="IX"
                                /map="9"
     CDS                       <1..206
                                /codon_start=3
                                /product="TCP1-beta"
                                /protein_id="AAA98665.1"
                                /db_xref="GI:1293614"
                                /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLRKRAVVSSASEA
AEVLLRVDNIIRARPRTANRQHM"
     gene                      687..3158
                                /gene="AXL2"
     CDS                       687..3158
                                /gene="AXL2"
                                /note="plasma membrane glycoprotein"
                                /codon_start=1
                                /function="required for axial budding pattern of S.
cerevisiae"
                                /product="Axl2p"
                                /protein_id="AAA98666.1"
                                /db_xref="GI:1293615"
                                /translation="MTQLQISLLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN"
```



## ORIGIN

```
1  gatcctccat  atacaacggt  atctccacct  caggttttaga  tctcaacaac  ggaaccattg
61  ccgacatgag  acagtttaggt  atcgtcgaga  gttacaagct  aaaacgagca  gtagtcagct
121  ctgcatctga  agccgctgaa  gttctactaa  ggggtggataa  catcatccgt  gcaagaccaa
181  gaaccgccaa  tagacaacat  atgtaacata  tttaggatat  acctcgaaaa  taataaaccg
241  ccacactgtc  attattataa  ttagaaacag  aacgcaaaaa  ttatccacta  tataattcaa
301  agacgcgaaa  aaaaaagaac  aacgcgtcat  agaacttttg  gcaattcgcg  tcacaaataa
361  attttggcaa  cttatgtttc  ctcttcgagc  agtactcgag  ccctgtctca  agaatgtaat
421  aatacccatc  gtaggtatgg  ttaaagatag  catctccaca  acctcaaagc  tccttgccga
481  gagtcgccct  cctttgtcga  gtaattttca  cttttcatat  gagaacttat  tttcttattc
541  ttactctca  catcctgtag  tgattgacac  tgcaacagcc  accatcacta  gaagaacaga
601  acaattactt  aatagaaaaa  ttatatcttc  ctcgaaacga  tttcctgctt  ccaacatcta
661  cgtatatcaa  gaagcattca  cttaccatga  cacagcttca  gatttcatta  ttgctgacag
721  ctactatata  actactccat  ctagtagtgg  ccacgcccta  tgaggcatat  cctatcggaa
781  aacaataccc  cccagtggca  agagtcaatg  aatcgtttac  atttcaaatt  tccaatgata
841  cctataaatc  gtctgtagac  aagacagctc  aaataacata  caattgcttc  gacttaccga
901  gctggctttc  gtttgactct  agttctagaa  cgttctcagg  tgaaccttct  tctgacttac
961  tatctgatgc  gaacaccacg  ttgtattttc  atgtaatact  cgaggggtacg  gactctgccg
1021  acagcacgtc  tttgaacaat  acataccaat  ttgttggttac  aaaccgtcca  tccatctcgc
1081  tatcgtcaga  tttcaatcta  ttggcggtgt  taaaaaacta  tgggtatact  aacggcaaaa
1141  acgctctgaa  actagatcct  aatgaagtct  tcaacgtgac  ttttgaccgt  tcaatgttca
1201  ctaacgaaga  atccattgtg  tcgtattacg  gacgttctca  gttgtataat  gcgccgttac
1261  ccaattggct  gttcttcgat  tctggcgagt  tgaagtttac  tgggacggca  ccggtgataa
1321  actcggcgat  tgctccagaa  acaagctaca  gttttgtcat  catcgctaca  gacattgaag
1381  qattttctgc  cgttgaagta  qaattcqaat  taqtcatcqa  qactcaccag  ttaactacct
```

<http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>

- brak możliwości ograniczenia zapytania do pewnych pól (bez przeglądania całych plików);
- powolne zapytania, powolne dołączanie nowych wpisów (ponownie konieczność przeglądania całych plików);
- jednoczesność (co będzie jak kilka osób zmodyfikuje jednocześnie ten sam wpis?);
- spójność (jak sprawdzać czy wprowadzane wartości są prawidłowe – np. czy powiązania wskazują na istniejące zapisy?)

```
>gi|37993870|gb|CF805616.1|CF805616 TaRGA.C2 [...]  
ACAATTGGTTTATGCCCATGAGGAGAAAGACAAGAAAGACAACAAGGAAGGTCAC  
TTCATGTGGGTCCATGTCTCTCAGAGTTTTAGTGTGGGCGACATCT  
TCAAGGAGCTTATGAGGCAGCTTCAGAGCCTAAGGTTGCATGCCCTCAATTTCA  
TAACCTGAATGCCTTGGAAGGAATTGGAGAGGAACTAGATGGAAAAAGATTC  
CTTCTAGTACTAGATGATGTCTGGTGCAACAAGGATGTCGGTAACGAGGAGCTAC  
CAAAGTTACTTACTCCACTGAAGAAAGGAAAGAGAGGAAGCAAGATCCTAGTGAC  
AACTCGAAGTAAATTTCCATTGTCGGATCAAGGTCCCGGTGTGCGGCATACTGCA  
ATGCCAATAAATGAGGTTAATGATACTGCCTTCTTCGAGCTATTCATGCACTATG  
CCCTCGAAGAAGGCCAAGACTGGAGCCTGTTCAAGACCATTGGTGAGGAGATTGC  
AGAAAAGCTG
```

Numer GI

Bardzo prosty format – przydatny, gdy interesuje nas tylko sekwencja.  
Pierwszy wiersz (zaczynający się znakiem większości „>”) – nagłówek  
(ID, nazwa itp.);  
kolejne wiersze - sekwencja

Nieredundantna (nadzorowana, drugorzędowa) baza danych sekwencji.

Ograniczona tylko do najlepiej poznanych genetycznie organizmów.

(sekwencje z ok. 17tys. gatunków – w GenBanku 250tys.)

On-line:

za pośrednictwem Entrez

FTP:

<ftp://ftp.ncbi.nih.gov/refseq/release/>

## Format danych podobnie jak w GenBanku. Dodatkowy prefiks przed identyfikatorem.

Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGSa
NW_	Genomic	Contig or scaffold, primarily WGSa
NS_	Genomic	Environmental sequence
NZ_b	Genomic	Unfinished WGS
NM_	mRNA	
NR_	RNA	
XM_c	mRNA	Predicted model
XR_c	RNA	Predicted model
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_c	Protein	
XP_c	Protein	Predicted model, associated with an XM_ accession
ZP_c	Protein	Predicted model, annotated on NZ_ genomic records



# UniProt / Swiss-Prot

# Protein Information Resource

<http://pir.georgetown.edu/>

PIR-PSD – historycznie pierwsza adnotowana baza sekwencji aminokwasowych, bezpośredni następca atlasu białek (1965-1978) Margaret Dayhoff.

Obecnie włączona przez EBI i SIB do bazy UniProt.

**PIR** Protein Information Resource

Integrated Protein Informatics Resource for Genomic, Proteomic and Systems Biology Research

The Universal Protein Resource (UniProt) provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information.

**PRO** Protein Ontology

- Representation of protein objects with descriptions and relationships
- Browse PRO
- Annotate with RACE-PRO
- Sample PRO report

**ProClass** Integrated Protein Knowledgebase

- Value-added reports for UniProt and unique UniProt proteins
- Functional analysis and protein ID mapping
- Sample protein report

**ProLINK** Literature Information & Knowledge

- Source for text mining and ontology development
- BioThesaurus text mining tool
- Bibliography mapping
- Sample Biblio. report

**OTHER RESOURCE**

- Representative Proteomes
- PIR Gold-Standard Data sets on NCBI's caZoo

**PEPTIDE SEARCH** DATABASE: UniProtKB

**TEXT SEARCH** DATABASE: ProClass

PIR is hosting the Fifth International Biocuration Conference April 2-4, 2012

**Bioinformatics & Computational Biology Graduate Programs:**

- MS program at Georgetown University
- MS, PhD and Graduate Certificate programs at University of Delaware
- Now recruiting new students for Fall of 2012
- Scholarships available

Home | About PIR | Databases | Search/Analysis | Download | Support | SITE MAP | TERMS OF USE

© 2008 Protein Information Resource

University of Delaware  
15 Innovation Way, Suite 005  
Newark, DE 19713, USA

Georgetown University Medical Center  
3300 Whitehaven Street, NW, Suite 1200  
Washington, DC 20007, USA

Adnotowana baza sekwencji białkowych:

- obszerny opis;
- minimalna redundancja
- integracja z innymi bazami

Przykładowe dane zawarte w adnotacjach:

- funkcja białka
- modyfikacje potranslacyjne np. fosforylacja, acetylacja, glikozylacja
- domeny i miejsca wiążące, motywy (palec cynkowy itp.)
- struktura drugorzędowa
- struktura czwartorzędowa
- podobieństwo do innych białek
- choroby związane z funkcją biologiczną białka
- sprzeczności w wyznaczeniu sekwencji, odmiany



Baza poddanych translacji sekwencji nukleotydowych

Zasoby bazy są uzupełniane i adnotowane automatycznie;

Część rekordów – po opracowaniu przez kuratora – jest dodawana do bazy Swiss-Prot

Podział:

*Swiss-Prot TrEMBL* – rekordy oczekujące na opracowanie i włączenie do Swiss-Prot

*REM-TrEMBL* – rekordy, których włączenie do Swiss-Prot nie jest planowane



<http://www.uniprot.org/>

Meta-baza powstała z połączenia zasobów  
Swiss-Prot, TrEMBL i PIR

The screenshot shows the UniProt website interface. At the top, there's a navigation bar with links: Search, Blast, Align, Retrieve, ID Mapping, Downloads, Contact, Documentation/Help. Below this is a search bar with a dropdown menu set to 'Protein Knowledgebase (UniProtKB)' and a 'Query' input field. To the right of the search bar are buttons for 'Search', 'Advanced Search', and 'Clear'.

The main content area is divided into several sections:

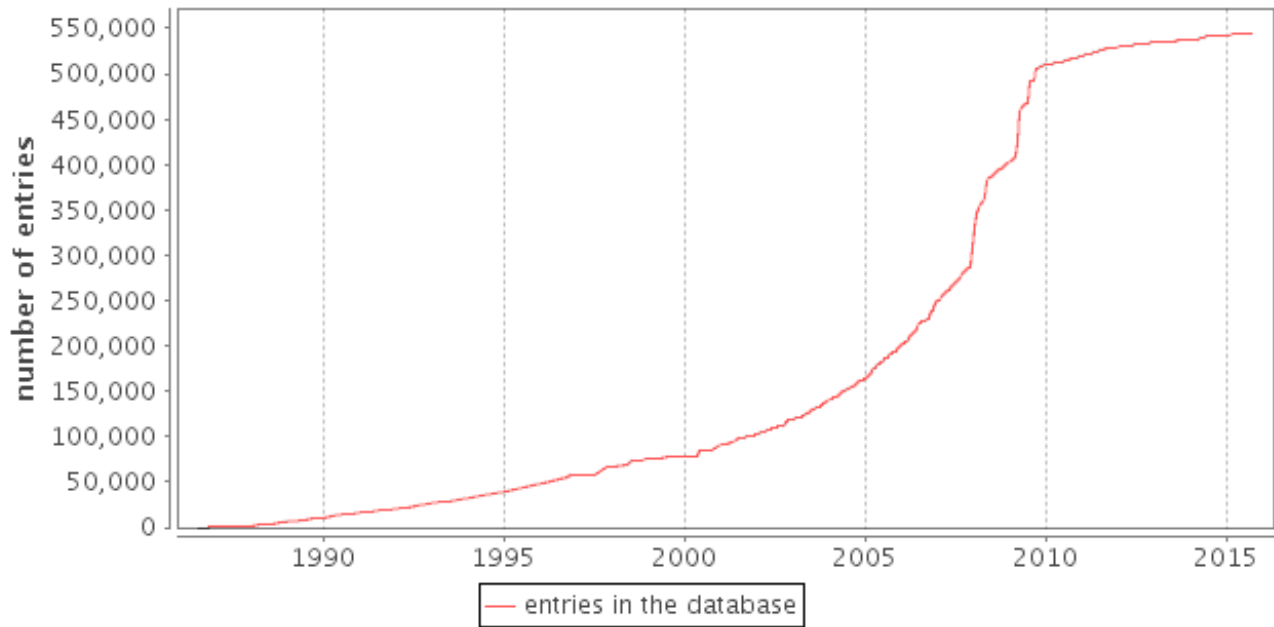
- WELCOME**: A brief mission statement about providing protein sequence and functional information.
- What we provide**: A table listing the services:
  - UniProtKB**: Protein knowledgebase, consisting of two sections:
    - ★ **Swiss-Prot**, which is manually annotated and reviewed.
    - ✳ **TrEMBL**, which is automatically annotated and is **not** reviewed.Includes complete and reference proteome sets.
  - UniRef**: Sequence clusters, used to speed up sequence similarity searches.
  - UniParc**: Sequence archive, used to keep track of sequences and their identifiers.
  - Supporting data**: Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more.
- Getting started**: A list of links for new users:
  - Text search
  - Sequence similarity searches (BLAST)
  - Sequence alignments
  - Batch retrieval
  - Database identifier mapping (ID Mapping)
- NEWS**: A section titled 'UniProt release 2012\_02 - Feb 22, 2012' with details about updates to the human proteome, GeneDB, Spombe to PomBase, and removal of the cross-reference NMPOR. It also includes links for 'Statistics for UniProtKB', 'Swiss-Prot', 'TrEMBL', 'Forthcoming changes', and 'News archives'. A 'Follow @uniprot' button shows 243 followers.
- SITE TOUR**: A section with a video player and the text 'Learn how to make best use of the tools and data on this site.'
- PROTEIN SPOTLIGHT**: A section titled 'get a grip January 2012' with a short story about a drapnipe that crawled up the front of a house to prevent cats from climbing up it.

At the bottom of the page is the UniProt logo, which consists of the word 'UniProt' in a blue sans-serif font, followed by a circular graphic made of blue dots of varying sizes.



Total	549,215
Entries with updated sequences	56
With a fragmented AA sequence	9,149
With known <a href="#">alternative products</a>	24,322

Number of entries in UniProtKB/Swiss-Prot over time

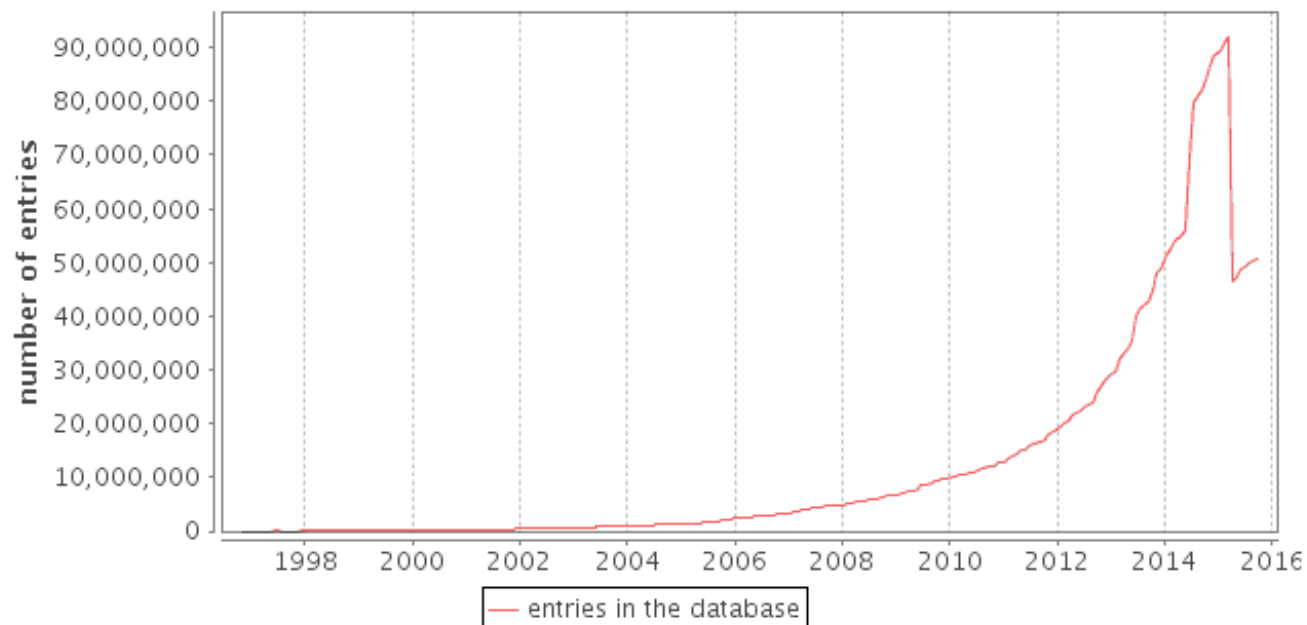


	Protein Existence (PE)	Number of entries
1	Evidence at protein level	90,456
2	Evidence at transcript level	57,714
3	Inferred from homology	387,606
4	Predicted	11,484
5	Uncertain	1,955



<b>Total</b>	50,825,784
Entries with updated sequences	857
With a fragmented AA sequence	6,573,313
With known <a href="#">alternative products</a>	0

**Number of entries in UniProtKB/TrEMBL over time**



	Protein Existence (PE)	Number of entries
1	Evidence at protein level	117,527
2	Evidence at transcript level	967,807
3	Inferred from homology	10,858,591
4	Predicted	38,881,859
5	Uncertain	0

# UniProt - wyszukiwanie

human antigen	All entries containing both terms.
human AND antigen	
human && antigen	
"human antigen"	All entries containing both terms in the exact order.
human -antigen	All entries containing the term human but not antigen.
human NOT antigen	
human ! antigen	
human OR mouse	All entries containing either term.
human    mouse	
antigen AND (human OR mouse)	Using parentheses to override boolean precedence rules.
anti*	All entries containing terms starting with anti. Asterisks can also be used at the beginning and within terms. <b>Note:</b> Terms starting with an asterisk or a single letter followed by an asterisk can slow down queries considerably.
author:Tiger*	Citations that have an author whose name starts with Tiger. To search in a specific field of a dataset, you must prefix your search term with the field name and a colon. To discover what fields can be queried explicitly, observe the query hints that are shown after submitting a query or use the query builder (see below).
length:[100 TO *]	All entries with a sequence of at least 100 amino acids.
citation:(author:Arai author:Chung)	All entries with a publication that was coauthored by two specific authors.

To use characters that have a special meaning in the query syntax literally in your query, you must escape them with a backslash, e.g. use `gene:L\ (1\ ) 2CB` to search for the gene name L(1)2CB. The current list of special characters is:

+ - && || ! ( ) { } [ ] ^ " ~ \* ? : \

# Format pliku UniProt / Swiss-Prot

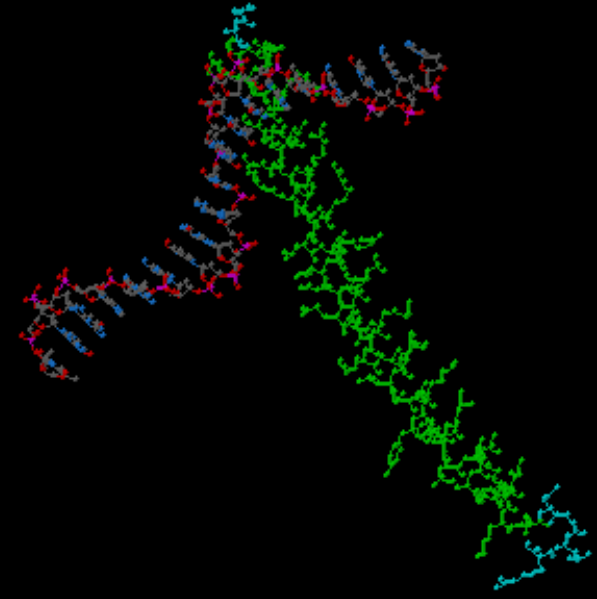
```
ID      HBB_HUMAN                      Reviewed;              147 AA.
AC      P68871; A4GX73; B2ZUE0; P02023; Q13852; Q14481; Q14510; Q45KT0;
AC      Q549N7; Q6FI08; Q6R7N2; Q8IZI1; Q9BX96; Q9UCD6; Q9UCP8; Q9UCP9;
DT      21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT      23-JAN-2007, sequence version 2.
DT      22-FEB-2015, entry version 104.
DE      RecName: Full=Hemoglobin subunit beta;
DE      AltName: Full=Beta-globin;
DE      AltName: Full=Hemoglobin beta chain;
DE      Contains:
DE          RecName: Full=LVV-hemorphin-7;
GN      Name=HBB;
OS      Homo sapiens (Human).
OC      Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC      Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC      Catarrhini; Hominidae; Homo.
OX      NCBI_TaxID=9606;
RN      [1]
RP      NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX      MEDLINE=77126403; PubMed=1019344;
RA      Marotta C., Forget B., Cohen-Solal M., Weissman S.M.;
RT      "Nucleotide sequence analysis of coding and noncoding regions of human
RT      beta-globin mRNA.";
RL      Prog. Nucleic Acid Res. Mol. Biol. 19:165-175(1976).
RN      [2]
RP      NUCLEOTIDE SEQUENCE [GENOMIC DNA].
...
```

# PDB

## Baza struktur molekularnych <http://www.pdb.org>



The screenshot shows the PDB website interface. At the top, there's a header with the PDB logo and a tagline: "An Information Portal to Biological Macromolecular Structures". Below this, a search bar is present with a dropdown menu for "PDB ID or Text" and a "Search" button. The left sidebar contains several sections: "MyPDB" (Login, Register), "Home" (News, Publications, etc.), "Deposition" (All Deposit Services, etc.), "Search" (Advanced Search, etc.), and "Tools" (Download, etc.). The main content area is titled "A Resource for Studying Biological Macromolecules" and includes a "Featured Molecules" section with a "Molecule of the Month: Integrase" and a "Protein Structure Initiative Featured Molecule: Nitrile Reductase QutF". There's also a "Latest Structures" section at the bottom. On the right, there's a "Customize This Page" section with "New Features" and "RCSB PDB News".

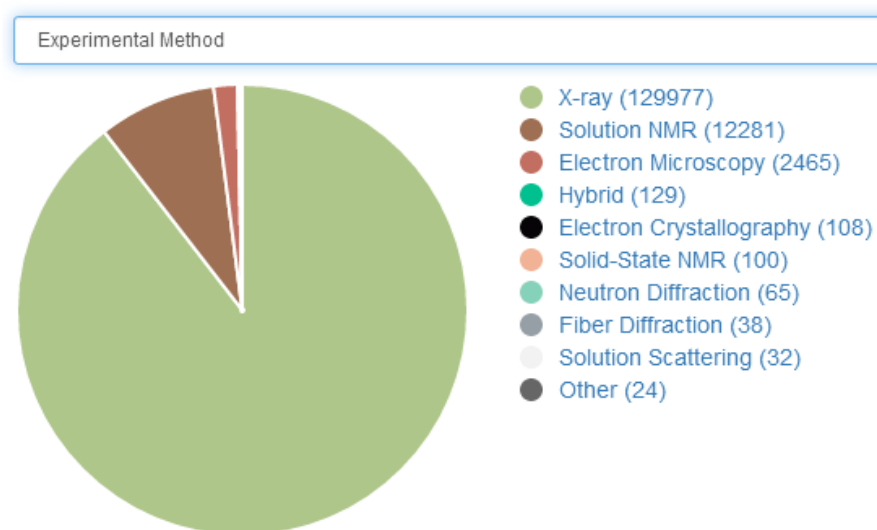
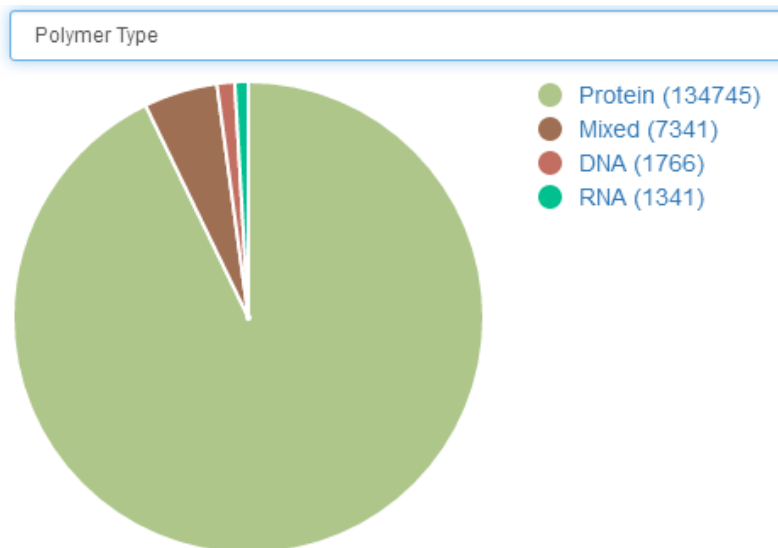


Dokumentacja formatu danych:  
<http://www.wwpdb.org/docs.html#format>

FTP:  
<ftp://ftp.wwpdb.org/>



Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	121728	1962	6283	4	129977
NMR	10864	1259	250	8	12381
ELECTRON MICROSCOPY	1790	31	644	0	2465
HYBRID	121	5	2	1	129
other	244	4	6	13	267
Total	134747	3261	7185	26	145219



<http://www.rcsb.org/pdb/statistics/holdings.do>

```
data_1EJ9
#
_entry.id      1EJ9
#
_audit_conform.dict_name      mmcif_pdbx.dic
_audit_conform.dict_version    4.007
_audit_conform.dict_location
http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
loop_
_database_2.database_id
_database_2.database_code
PDB      1EJ9
NDB      PD0125
RCSB     RCSB010631
#
loop_
_database_PDB_rev.num
_database_PDB_rev.date
_database_PDB_rev.date_original
_database_PDB_rev.status
_database_PDB_rev.replaces
_database_PDB_rev.mod_type
1 2000-08-03 2000-03-01 ? 1EJ9 0
2 2009-02-24 ?           ? 1EJ9 1
#
_database_PDB_rev_record.rev_num      2
_database_PDB_rev_record.type          VERSN
_database_PDB_rev_record.details      ?
```

```
#
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.Cartn_x_esd
_atom_site.Cartn_y_esd
_atom_site.Cartn_z_esd
_atom_site.occupancy_esd
_atom_site.B_iso_or_equiv_esd
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM 1 N N . PRO A 1 5 ? -3.218 23.313 19.768 1.00 65.32 ? ? ? ? ? 4 PRO A N 1
ATOM 2 C CA . PRO A 1 5 ? -2.926 24.681 19.350 1.00 62.03 ? ? ? ? ? 4 PRO A CA 1
ATOM 3 C C . PRO A 1 5 ? -3.532 24.954 17.967 1.00 52.41 ? ? ? ? ? 4 PRO A C 1
ATOM 4 O O . PRO A 1 5 ? -4.356 24.167 17.505 1.00 67.03 ? ? ? ? ? 4 PRO A O 1
ATOM 5 C CB . PRO A 1 5 ? -1.419 24.648 19.202 1.00 43.04 ? ? ? ? ? 4 PRO A CB 1
ATOM 6 C CG . PRO A 1 5 ? -1.192 23.263 18.562 1.00 39.23 ? ? ? ? ? 4 PRO A CG 1
ATOM 7 C CD . PRO A 1 5 ? -2.288 22.354 19.126 1.00 51.55 ? ? ? ? ? 4 PRO A CD 1
ATOM 8 N N . ALA A 1 6 ? -3.090 26.021 17.294 1.00 52.98 ? ? ? ? ? 5 ALA A N 1
```

## Przykładowe pola:

- HEADER
- TITLE
- COMPND
- SOURCE
- AUTHOR
- DATE
- JRNL
- REMARK
- SEQRES
- ATOM COORDINATES

```
HEADER      SIGNAL TRANSDUCTION                      23-APR-97    1MPH
TITLE       PLECKSTRIN HOMOLOGY DOMAIN FROM MOUSE BETA-SPECTRIN, NMR,
TITLE       2 50 STRUCTURES
COMPND      MOL_ID: 1;
COMPND      2 MOLECULE: BETA SPECTRIN;
COMPND      3 CHAIN: NULL;
COMPND      4 FRAGMENT: PLECKSTRIN HOMOLOGY;
COMPND      5 SYNONYM: PH DOMAIN;
COMPND      6 ENGINEERED: YES
SOURCE      MOL_ID: 1;
SOURCE      2 ORGANISM_SCIENTIFIC: MUS MUSCULUS;
SOURCE      3 ORGANISM_COMMON: MOUSE;
SOURCE      4 ORGAN: BRAIN;
SOURCE      5 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE      6 EXPRESSION_SYSTEM_STRAIN: BL21 (DE3);
SOURCE      7 EXPRESSION_SYSTEM_PLASMID: PET21D
KEYWDS      SIGNAL TRANSDUCTION, INOSITOL PHOSPHATES
EXPDTA      NMR, 50 STRUCTURES
AUTHOR      M.NILGES,M.J.MACIAS,S.I.O'DONOGHUE,H.OSCHKINAT
REVDAT      1 16-JUN-97 1MPH 0
JRNL        AUTH  M.NILGES,M.J.MACIAS,S.I.O'DONOGHUE,H.OSCHKINAT
JRNL        TITL  AUTOMATED NOESY INTERPRETATION WITH AMBIGUOUS
JRNL        TITL 2 DISTANCE RESTRAINTS: THE REFINED NMR SOLUTION
JRNL        TITL 3 STRUCTURE OF THE PLECKSTRIN HOMOLOGY DOMAIN FROM
JRNL        TITL 4 BETA SPECTRIN
JRNL        REF   TO BE PUBLISHED
JRNL        REFN                                     0353
```

## Przykładowe pola:

- HEADER
- TITLE
- COMPND
- SOURCE
- AUTHOR
- DATE
- JRNL
- REMARK
- SEQRES
- ATOM COORDINATES

```
REMARK 999 1MPH          SWS      Q62261          1 - 2198 NOT IN ATOMS LIST
REMARK 999 1MPH          SWS      Q62261      2305 - 2363 NOT IN ATOMS LIST
DBREF  1MPH              1    106  SWS      Q62261      SPCO_MOUSE      2199    2304
SEQRES  1    106  MET GLU GLY PHE LEU ASN ARG LYS HIS GLU TRP GLU ALA
SEQRES  2    106  HIS ASN LYS LYS ALA SER SER ARG SER TRP HIS ASN VAL
SEQRES  3    106  TYR CYS VAL ILE ASN ASN GLN GLU MET GLY PHE TYR LYS
SEQRES  4    106  ASP ALA LYS SER ALA ALA SER GLY ILE PRO TYR HIS SER
SEQRES  5    106  GLU VAL PRO VAL SER LEU LYS GLU ALA ILE CYS GLU VAL
SEQRES  6    106  ALA LEU ASP TYR LYS LYS LYS LYS HIS VAL PHE LYS LEU
SEQRES  7    106  ARG LEU SER ASP GLY ASN GLU TYR LEU PHE GLN ALA LYS
SEQRES  8    106  ASP ASP GLU GLU MET ASN THR TRP ILE GLN ALA ILE SER
SEQRES  9    106  SER ALA
HELIX   1    1  ALA      41  SER      46  1
HELIX   2    2  ASP      93  SER     104  1
SHEET   1    A  7  PRO      55  SER      57  0
SHEET   2    A  7  GLU      34  TYR      38 -1  N  MET      35  O  VAL      56
SHEET   3    A  7  TRP      23  ASN      31 -1  N  ASN      31  O  GLU      34
SHEET   4    A  7  GLU       2  TRP      11 -1  N  ARG       7  O  HIS      24
SHEET   5    A  7  GLU      85  GLN      89 -1  N  GLN      89  O  ASN       6
SHEET   6    A  7  VAL      75  ARG      79 -1  N  LEU      78  O  TYR      86
SHEET   7    A  7  ILE      62  ALA      66 -1  N  ALA      66  O  VAL      75
```

## Przykładowe pola:

- HEADER
- TITLE
- COMPND
- SOURCE
- AUTHOR
- DATE
- JRNL
- REMARK
- SEQRES
- ATOM COORDINATES

				X	Y	Z				
MODEL	1									
ATOM	1	N	MET	1	-7.678	-13.900	-15.824	1.00	1.74	N
ATOM	2	CA	MET	1	-8.660	-12.928	-16.368	1.00	1.02	C
ATOM	3	C	MET	1	-8.995	-11.893	-15.292	1.00	0.87	C
ATOM	4	O	MET	1	-10.035	-11.971	-14.631	1.00	1.29	O
ATOM	5	CB	MET	1	-9.933	-13.643	-16.861	1.00	1.60	C
ATOM	6	CG	MET	1	-9.956	-13.743	-18.387	1.00	2.33	C
ATOM	7	SD	MET	1	-8.573	-14.694	-19.049	1.00	3.31	S
ATOM	8	CE	MET	1	-9.264	-16.356	-19.025	1.00	4.33	C
ATOM	9	1H	MET	1	-7.505	-13.691	-14.815	1.00	2.19	H
ATOM	10	2H	MET	1	-8.056	-14.870	-15.918	1.00	2.25	H
ATOM	11	3H	MET	1	-6.780	-13.823	-16.346	1.00	2.20	H
ATOM	12	HA	MET	1	-8.205	-12.416	-17.205	1.00	1.43	H
ATOM	13	1HB	MET	1	-9.969	-14.638	-16.440	1.00	1.91	H
ATOM	14	2HB	MET	1	-10.805	-13.092	-16.540	1.00	2.17	H
ATOM	15	1HG	MET	1	-10.881	-14.212	-18.688	1.00	2.75	H
ATOM	16	2HG	MET	1	-9.921	-12.744	-18.796	1.00	2.71	H
ATOM	17	1HE	MET	1	-10.194	-16.366	-19.578	1.00	4.65	H
ATOM	18	2HE	MET	1	-8.564	-17.040	-19.484	1.00	4.69	H
ATOM	19	3HE	MET	1	-9.451	-16.650	-18.002	1.00	4.75	H
ATOM	20	N	GLU	2	-8.100	-10.919	-15.120	1.00	0.51	N
ATOM	21	CA	GLU	2	-8.269	-9.846	-14.133	1.00	0.37	C
ATOM	22	C	GLU	2	-7.885	-8.506	-14.763	1.00	0.34	C
ATOM	23	O	GLU	2	-7.123	-8.463	-15.734	1.00	0.48	O
ATOM	24	CB	GLU	2	-7.386	-10.094	-12.897	1.00	0.53	C
ATOM	25	CG	GLU	2	-7.554	-11.510	-12.325	1.00	0.96	C
ATOM	26	CD	GLU	2	-6.450	-12.435	-12.822	1.00	1.61	C
ATOM	27	OE1	GLU	2	-5.265	-12.079	-12.684	1.00	2.09	O
ATOM	28	OE2	GLU	2	-6.763	-13.524	-13.344	1.00	2.16	O
ATOM	29	H	GLU	2	-7.288	-10.919	-15.684	1.00	0.67	H
ATOM	30	HA	GLU	2	-9.304	-9.804	-13.824	1.00	0.41	H
ATOM	31	1HB	GLU	2	-6.351	-9.936	-13.169	1.00	1.25	H
ATOM	32	2HB	GLU	2	-7.657	-9.381	-12.132	1.00	1.17	H

# Systemy powiązań

Systemy integrujące informacje pochodzące z różnych baz;  
często z dodatkowymi narzędziami.

Przykłady:

- **ENTREZ**
- **ExPASy**
- **RNA Central**
- **NDB**

**Entrez** (zarządzany przez NCBI) jest zintegrowanym systemem wyszukiwania informacji w bazach danych.

<https://www.ncbi.nlm.nih.gov/search/>



# Entrez – bazy wchodzące w skład sieci

## Literature

<b>Books</b>	books and reports
<b>MeSH</b>	ontology used for PubMed indexing
<b>NLM Catalog</b>	books, journals and more in the NLM Collections
<b>PubMed</b>	scientific & medical abstracts/citations
<b>PubMed Central</b>	full-text journal articles

## Health

<b>ClinVar</b>	human variations of clinical significance
<b>dbGaP</b>	genotype/phenotype interaction studies
<b>GTR</b>	genetic testing registry
<b>MedGen</b>	medical genetics literature and links
<b>OMIM</b>	online mendelian inheritance in man
<b>PubMed Health</b>	clinical effectiveness, disease and drug reports

## Genomes

<b>Assembly</b>	genome assembly information
<b>BioCollections</b>	museum, herbaria, and other biorepository collections
<b>BioProject</b>	biological projects providing data to NCBI
<b>BioSample</b>	descriptions of biological source materials
<b>Clone</b>	genomic and cDNA clones
<b>dbVar</b>	genome structural variation studies
<b>Genome</b>	genome sequencing projects by organism
<b>GSS</b>	genome survey sequences
<b>Nucleotide</b>	DNA and RNA sequences
<b>Probe</b>	sequence-based probes and primers
<b>SNP</b>	short genetic variations
<b>SRA</b>	high-throughput DNA and RNA sequence read archive
<b>Taxonomy</b>	taxonomic classification and nomenclature catalog

## Genes

<b>EST</b>	expressed sequence tag sequences
<b>Gene</b>	collected information about gene loci
<b>GEO DataSets</b>	functional genomics studies
<b>GEO Profiles</b>	gene expression and molecular abundance profiles
<b>HomoloGene</b>	homologous gene sets for selected organisms
<b>PopSet</b>	sequence sets from phylogenetic and population studies
<b>UniGene</b>	clusters of expressed transcripts

## Proteins

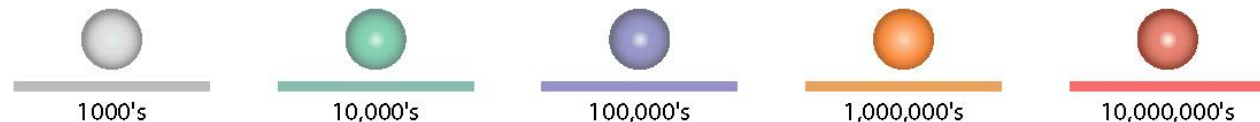
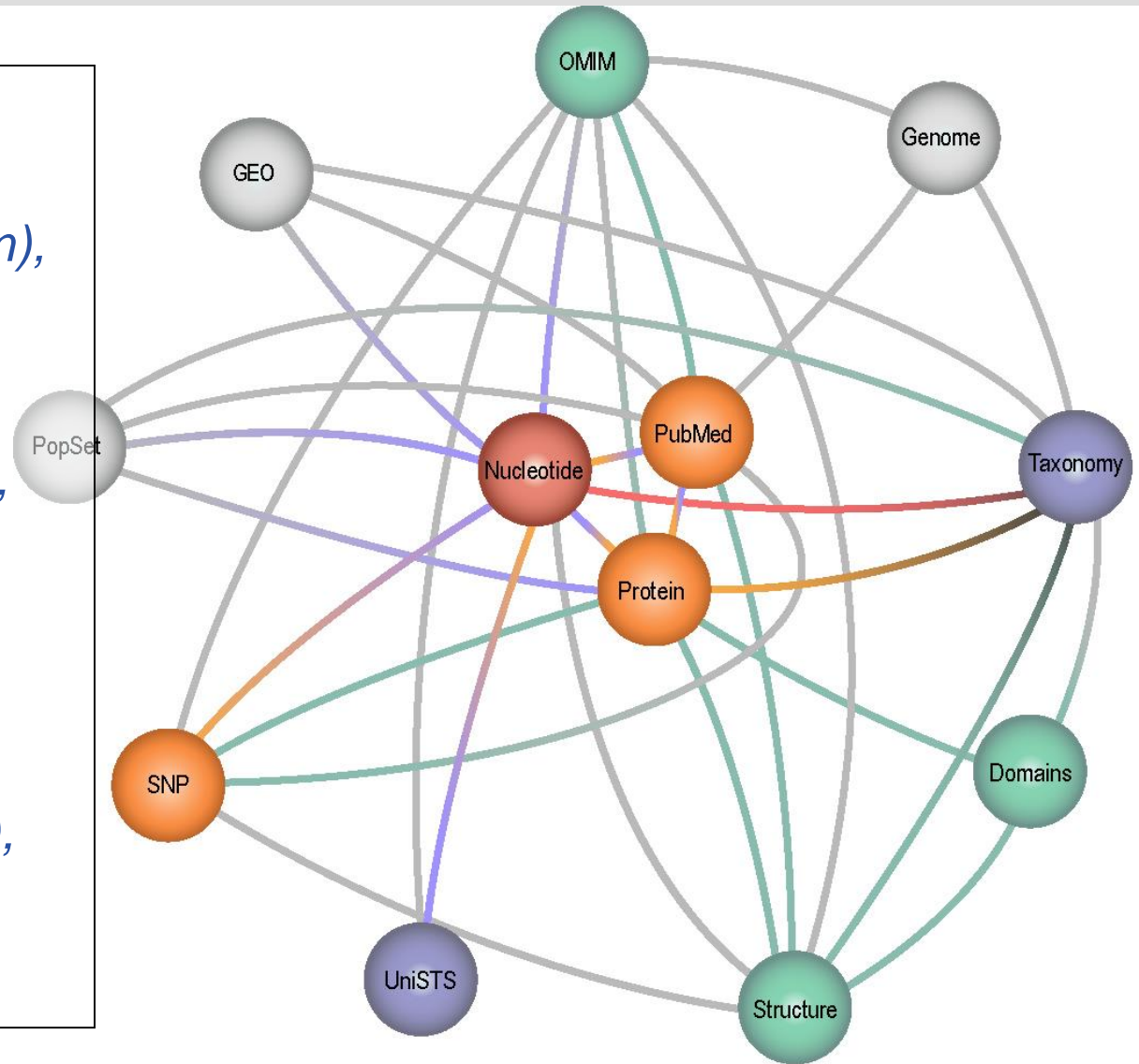
<b>Conserved Domains</b>	conserved protein domains
<b>Protein</b>	protein sequences
<b>Protein Clusters</b>	sequence similarity-based protein clusters
<b>Structure</b>	experimentally-determined biomolecular structures

## Chemicals

<b>BioSystems</b>	molecular pathways with links to genes, proteins and chemicals
<b>PubChem BioAssay</b>	bioactivity screening studies
<b>PubChem Compound</b>	chemical information with structures, information and links
<b>PubChem Substance</b>	deposited substance and chemical information

# Entrez – sieć powiązań

Bazy sekwencji DNA (*nucleotide*),  
sekwencje białek (*protein*),  
literatura (*PubMed*),  
polimorfizmy (*SNP*),  
systematyka (*taxonomy*),  
mutacje (*OMIM*),  
domeny białkowe (*domains*),  
eksperymenty mikromacierzowe (*GEO*),  
genomy (*genome*),  
itp.



## Link do tabeli z pełnym opisem:

[http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary\\_Matrices.html#Search\\_Fields\\_and\\_Qualifiers](http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Search_Fields_and_Qualifiers)

## Wybrane znaczniki:

[ACCN]	numer dostępu (accession number)
[ALL]	wszystkie pola (all fields)
[AUTH]	nazwisko autora (author name)
[FKEY]	najważniejsze cechy (feature key)
[ORGN]	organizm (organism)
[PROP]	właściwości (properties)
[SLEN]	długość sekwencji (sequence length)

## Przykłady zastosowania:

2:100[SLEN] – sekwencje o długości co najwyżej 100 nukleotydów



*Saccharomyces cerevisiae*[ORGN] – sekwencje pochodzące od wskazanego gatunku drożdży

1999/07/25:1999/07/31[MDAT] – sekwencje zmodyfikowane w podanym przedziale czasowym



# Expert Protein Analysis System

<http://expasy.org/>



**ExPASy Proteomics Server**

Search  for

Databases Tools Services Mirrors About Contact

You are here: ExPASy CH

The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE ([Disclaimer](#) / [References](#) / [Linking to ExPASy](#)).

## Databases

UniProtKB, PROSITE, HAMAP, SwissVar, ViralZone, SWISS-MODEL Repository, neXtProt, SWISS-2DPAGE, World-2DPAGE Repository, MIAPEGelDB, ENZYME, GlycoSuiteDB, UniPathway  
[details] [full list]

## Education & services

Downloads, Protein Spotlight, Protéines à la «Une», e-proxemis, Click2Drug, Bioinformatics core facility for Proteomics  
[full list]

## Tools & Software

Proteomics tools  
Make2D-DB, SwissDock, SwissProt  
[full list]

## Documentation

What's New?, E-How to link to ExPASy  
[full list]

Swiss Institute of Bioinformatics | [Disclaimer](#) | [Sitemap](#) | [Documentation](#) | [Contact](#)

Last modified



query

[Home](#) [About](#) [Contact](#)

## Visual Guidance

### Categories

proteomics  
genomics  
structural bioinformatics  
systems biology  
phylogeny/evolution  
population genetics  
transcriptomics  
biophysics  
imaging  
IT infrastructure  
drug design

### Resources A..Z

### Links/Documentation

ExPASy is the new SIB Bioinformatics Resource Portal which provides access to scientific databases and software tools in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see [Categories](#) in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

### Featuring today

#### Newick Utilities

Shell filters for high-throughput phylogenetic tree processing  
[\[details\]](#)



### How to use this portal?

- New features
- New to ExPASy
- Experienced ExPASy users: what is different

## Popular resources

☐ UniProtKB  
☒ SWISS-MODEL  
☒ STRING  
☒ PROSITE

## Latest News

### UniProt Knowledgebase release 2012\_02 - 2012-02-28

UniProtKB/SwissProt Release of 22-Feb-2012 contains 534,696 sequence entries. [More](#)  
UniProtKB/TrEMBL Release of 22-Feb-2012 contains 20,127,441 sequence entries. [More](#)

### Protein Spotlight: get a grip - 2012-02-03

Someone once told me that they had spread grease all over the drainpipe that crawled up the front of their house, to prevent cats from climbing up it. It's a very simple and pretty harmless way of keeping the enemy away...[More](#)

[\[More news\]](#) [\[SIB news\]](#)

RNAcentral provides unified access to the ncRNA sequence data supplied by the Expert Databases below [Learn more](#)

ENA



provides a comprehensive record of the world's nucleotide sequencing information

7.5 million sequences | Example

Updated

Rfam



is a collection of non-coding RNA families represented by manually curated sequence alignments, consensus secondary structures, and predicted homologues

2.5 million sequences | Example

RefSeq



is a comprehensive, integrated, non-redundant, well-annotated set of reference sequences

57,115 sequences | Example

Updated

Vega



is a repository for high-quality gene models produced by the manual annotation of vertebrate genomes. Human and mouse data from Vega are merged into GENCODE

32,429 sequences | Example

Updated

WormBase



curates, stores and displays genomic and genetic data about nematodes with primary emphasis on *C. elegans* and related nematodes

24,907 sequences | Example

New

gtRNAdb



contains tRNA gene predictions on complete or nearly complete genomes

10,625 sequences | Example

miRBase



is a database of published miRNA sequences and annotations that provides a centralised system for assigning names to miRNA genes

8,795 sequences | Example

RDP



provides quality-controlled, aligned and annotated rRNA sequences and a suite of analysis tools

4,779 sequences | Example

## Statistics

- > Release 3 (20 May 2015)
- > 8,607,919 distinct sequences
- > 15 Expert Databases
- > [Number of sequences over time](#)

## News

- > [RNAcentral release 3](#)
- > New training course: Online resources for ncRNA
- > RNAcentral release 2
- > New RNAcentral paper is online
- > RNAcentral release 1.0

Blog

RSS feed

Follow

437 followers

## Citing RNAcentral

If you use RNAcentral, please cite the following paper:

RNAcentral: an international database of ncRNA sequences

The RNAcentral Consortium, 2014 ([NAR](#))





A Portal for Three-dimensional Structural Information about Nucleic Acids  
As of 14-Oct-2015 number of released structures: 7796

[Search DNA](#)[Search RNA](#)[Advanced Search](#)

Search for released structures

## Welcome to the NDB

The NDB contains information about experimentally-determined nucleic acids and complex assemblies.

Use the NDB to perform searches based on annotations relating to sequence, structure and function, and to download, analyze, and learn about nucleic acids.

### Search Structures

[Search DNA](#)[Search DNA and its complexes](#)[Search RNA](#)[Search for RNA structures in the NDB archive or in the Non-Redundant list](#)[Advanced Search](#)[Search for structures based on structural features, chemical features, binding modes, citation and experimental information](#)

### Featured Tools

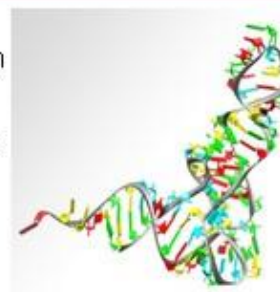
[RNA 3D Motif Atlas](#), a representative collection of RNA 3D internal and hairpin loop motifs

[Non-redundant Lists](#) of RNA-containing 3D structures

[RNA Base Triple Atlas](#), a collection of motifs consisting of two RNA basepairs

[WebFR3D](#), a webserver for symbolic and geometric searching of RNA 3D structures

[R3D Align](#), an application for detailed nucleotide to nucleotide alignments of RNA 3D structures



# Inne popularne bazy

OMIM – baza ludzkich genów i genotypów

SNP – baza mutacji punktowych

EST – baza sekwencji z technik wysokoprzepustowych

CATH – hierarchiczna klasyfikacja białek

PFAM – baza rodzin białkowych



## PubMed

<http://www.ncbi.nlm.nih.gov/pubmed/>

baza cytowań artykułów i książek naukowych z obszaru *life science*  
ponad 21 mln rekordów.

## PubMed Central

<http://www.ncbi.nlm.nih.gov/pmc/>

pełne teksty artykułów  
ponad 2 mln rekordów

## BookShelf

<http://www.ncbi.nlm.nih.gov/books/>

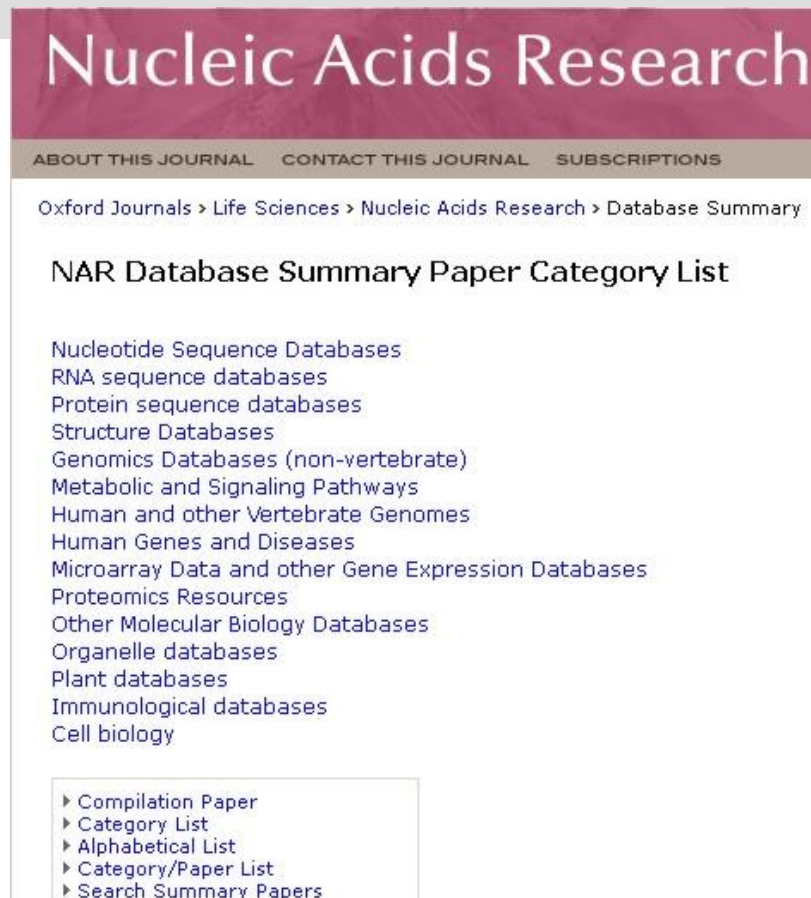
# Poszukiwanie baz specjalistycznych

Czasopismo Nucleic Acids Research corocznie wydaje specjalny numer opisujący nowe lub ulepszone niedawno bazy danych (w pojedynczym numerze opisane jest ok 100 baz).

Opisane bazy gromadzone są w kolekcji dostępnej dla każdego przez stronę czasopisma. Bazy można przeszukiwać alfabetycznie lub wg kategorii.

Kolekcja zawiera ponad 1500 baz danych.

UWAGA: w kolekcji są także bazy przestarzałe, zawierające nieaktualne lub niekompletne dane.



The screenshot shows the 'Nucleic Acids Research' journal header in a maroon box. Below it is a navigation bar with links: 'ABOUT THIS JOURNAL', 'CONTACT THIS JOURNAL', and 'SUBSCRIPTIONS'. The main content area shows the breadcrumb 'Oxford Journals > Life Sciences > Nucleic Acids Research > Database Summary'. The title 'NAR Database Summary Paper Category List' is followed by a list of database categories: Nucleotide Sequence Databases, RNA sequence databases, Protein sequence databases, Structure Databases, Genomics Databases (non-vertebrate), Metabolic and Signaling Pathways, Human and other Vertebrate Genomes, Human Genes and Diseases, Microarray Data and other Gene Expression Databases, Proteomics Resources, Other Molecular Biology Databases, Organelle databases, Plant databases, Immunological databases, and Cell biology. A separate box contains a list of paper types: Compilation Paper, Category List, Alphabetical List, Category/Paper List, and Search Summary Papers.

Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS

Oxford Journals > Life Sciences > Nucleic Acids Research > Database Summary

NAR Database Summary Paper Category List

Nucleotide Sequence Databases  
RNA sequence databases  
Protein sequence databases  
Structure Databases  
Genomics Databases (non-vertebrate)  
Metabolic and Signaling Pathways  
Human and other Vertebrate Genomes  
Human Genes and Diseases  
Microarray Data and other Gene Expression Databases  
Proteomics Resources  
Other Molecular Biology Databases  
Organelle databases  
Plant databases  
Immunological databases  
Cell biology

► Compilation Paper  
► Category List  
► Alphabetical List  
► Category/Paper List  
► Search Summary Papers


Lista baz wg kategorii: [http://www.oxfordjournals.org/our\\_journals/nar/database/c/](http://www.oxfordjournals.org/our_journals/nar/database/c/)


Lista alfabetyczna: [http://www.oxfordjournals.org/our\\_journals/nar/database/a/](http://www.oxfordjournals.org/our_journals/nar/database/a/)

<https://links.bioinformatics.ca/>

The Bioinformatics Links Directory features curated links to molecular resources, tools and databases. The links listed in this directory are selected on the basis of recommendations from bioinformatics experts in the field. We also rely on input from our community of bioinformatics users for suggestions. Starting in 2003, we have also started listing all links contained in the NAR Webserver issue.

 Hide Resources (176)

 Hide Databases (621)

 Hide Tools (1548)

## Computer Related (85)

This category contains links to resources relating to programming languages often used in bioinformatics. Other tools of the trade, such as web development and database resources, are also included here.

## DNA (604)

This category contains links to useful resources for DNA sequence analyses such as tools for comparative sequence analysis and sequence assembly. Links to programs for sequence manipulation, primer design, and sequence retrieval and submission are also listed here.

## Education (75)

Links to information about the techniques, materials, people, places, and events of the greater bioinformatics community. Included are current news headlines, literature sources, educational material and links to bioinformatics courses and workshops.

## Expression (396)

Links to tools for predicting the expression, alternative splicing, and regulation of a gene sequence are found here. This section also contains links to databases, methods, and analysis tools for protein expression, SAGE, EST, and microarray data.

## Human Genome (240)

This section contains links to draft annotations of the

## Literature (87)

Links to resources related to published literature,

<https://www.expasy.org>



[Home](#) [About](#) [Contact](#)

Query all databases



search

[help](#)

## Visual Guidance

## Categories

### proteomics

- [protein sequences and identification](#)
- [proteomics experiment](#)
- [function analysis](#)
- [sequence sites, features and motifs](#)
- [protein modifications](#)
- [protein structure](#)
- [protein interactions](#)
- [similarity search/alignment](#)

### genomics

- [structure analysis](#)

- [systems biology](#)

- [evolutionary biology](#)

- [population genetics](#)

- [transcriptomics](#)

- [biophysics](#)

- [imaging](#)

SIB resources

External resources - *(No support from the ExPASy Team)*

## Databases

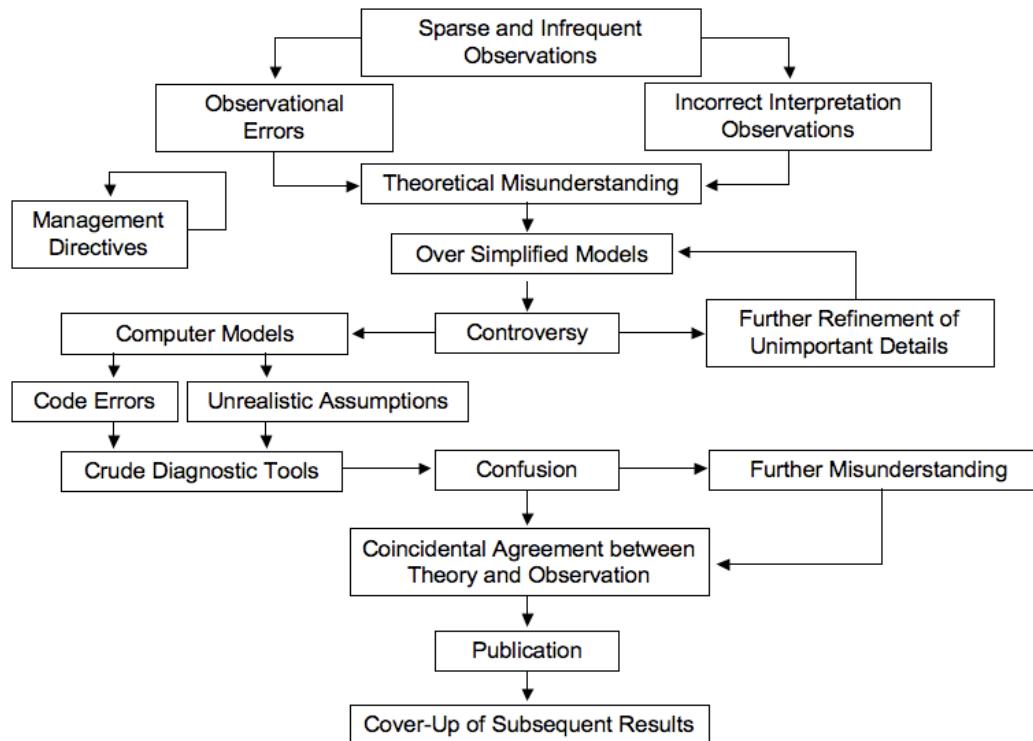
- UniProtKB** • functional information on proteins • [\[more\]](#)
- UniProtKB/Swiss-Prot** • protein sequence database • [\[more\]](#)
- STRING** • protein-protein interactions • [\[more\]](#)
- SWISS-MODEL Repository** • protein structure homology models • [\[more\]](#)
- PROSITE** • protein domains and families • [\[more\]](#)
- ViralZone** • portal to viral UniProtKB entries • [\[more\]](#)
- neXtProt** • human proteins • [\[more\]](#)
- EMBNET services** • bioinformatics tools, databases and courses • [\[more\]](#)
- ENZYME** • enzyme nomenclature • [\[more\]](#)
- GlyTouCan** • international glycan structure repository • [\[more\]](#)
- GPSDB** • gene and protein synonyms • [\[more\]](#)
- HAMAP** • UniProtKB family classification and annotation • [\[more\]](#)
- MatrixDB** • protein-glycosaminoglycan interactions • [\[more\]](#)
- MetaNetX** • Metabolic Network Repository & Analysis • [\[more\]](#)
- MIAPE** • MIAPE document edition • [\[more\]](#)

## Tools

- SWISS-MODEL Workspace** • structure homology-modeling • [\[more\]](#)
- SwissDock** • protein ligand docking server • [\[more\]](#)
- 2ZIP** • Prediction of leucine zipper domains • [\[more\]](#)
- 3of5** • find user-defined patterns in protein sequences • [\[more\]](#)
- AACompldent** • protein identification by aa composition • [\[more\]](#)
- AACompSim** • amino acid composition comparison • [\[more\]](#)
- Agadir** • Prediction of the helical content of peptides • [\[more\]](#)
- ALF** • simulation of genome evolution • [\[more\]](#)
- Alignment tools** • Four tools for multiple alignments • [\[more\]](#)
- AlIAl** • protein sequences comparisons • [\[more\]](#)
- APSSP** • Advanced Protein Secondary Structure Prediction • [\[more\]](#)
- Ascalaph** • Molecular modeling software • [\[more\]](#)
- big-PI** • predict GPI modification sites • [\[more\]](#)
- Biochemical Pathways** • Biochemical Pathways • [\[more\]](#)
- BLAST** • sequence similarity search • [\[more\]](#)
- BLAST (UniProt)** • BLAST search on the UniProt web site • [\[more\]](#)

# Koniec 😊

## The Course of Science



Kolejny wykład: Dopasowanie par sekwencji