

Proteomika: adatbázisok



Gáspári Zoltán, 2019

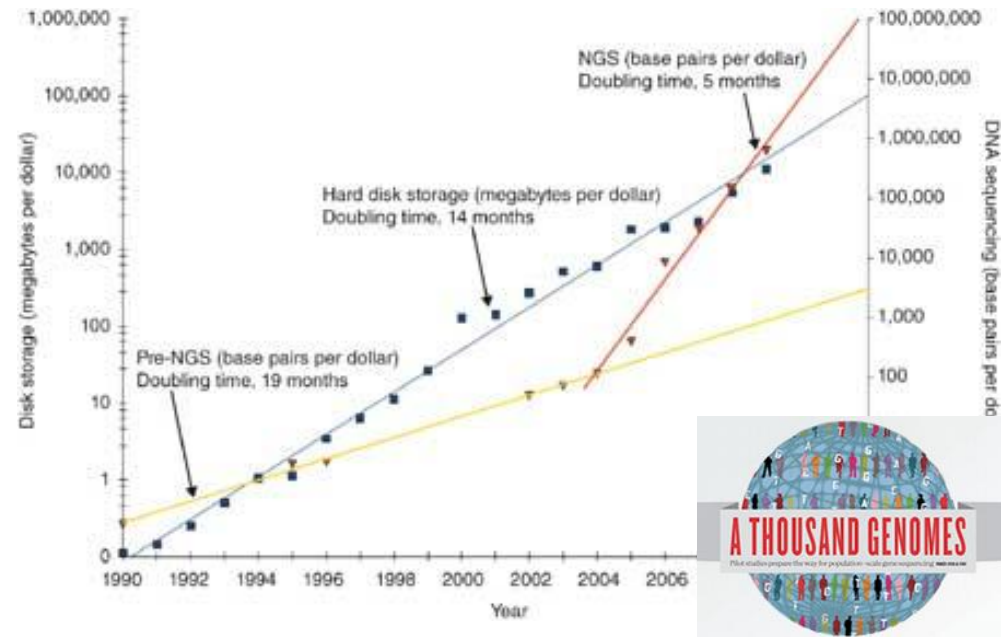
gaspari.zoltan@itk.ppke.hu

Miért kellene adatbázisok?

- Adatok forrása bioinformatikai kutatásokhoz:
 - adatbázisok
 - kollaborációk (majdnem mindig kiegészítve adatbázisokból származó adatokkal a tágabb kontextusba helyezés céljából)
- A tárolandó és rendszerezendő adatmennyiség hatalmas:

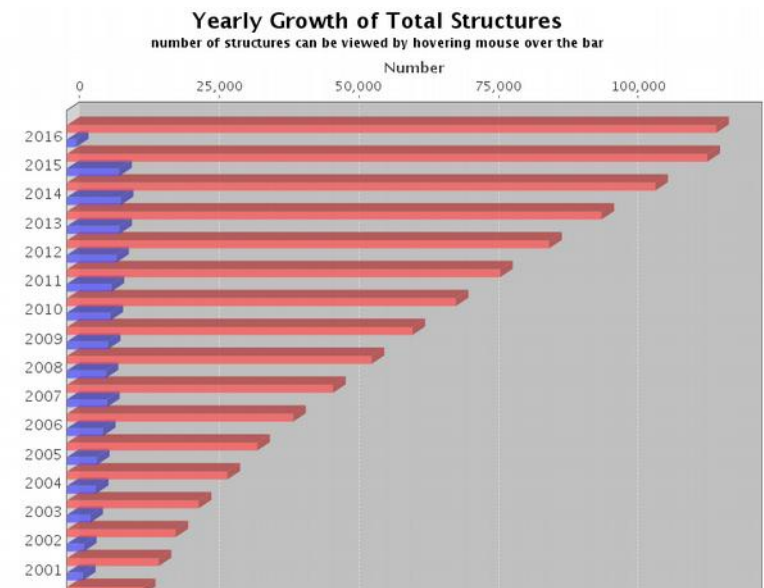
DNS szekvenciák:

- 1990: 25 ezer bp / hét
- 2000: 5 millió bp / hét
- 2010: 250 milliárd bp / hét (NGS)
- komplexitás: NGS eredmények többszörös illesztésként tárolva
- 1000 USD/emberi genom már realitás!



fehérjeszerkezetek:

- > 9 000 új szerkezet 2015-ben (összesen > 110 000)
- Növekvő méret és komplexitás (RNS-fehérje komplexek stb.)
- Dinamika mint új jelleg
- Fehérje-fehérje kölcsönhatások: adatbázisonként nagyságrendben 100 ezer
- > 23 millió kivonat tudományos közleményekből



Mi van az adatbázisokban?

- Amit a kutatók beleraknak
- Az adatok nem lesznek attól megbízhatóak, hogy bekerülnek az adatbázisba
- Az adatbázisok messze vannak a teljességtől:
 - Tudásunk hiányos
 - Nem minden publikált adat kerül be (lustaság, időhiány, szándékosság)
 - (És viszont: nem minden adatbázis rekordhoz van publikáció...)
- Az adatbázisok hibákat is tartalmaznak
 - Emberek vagyunk...
 - Kísérleti hibák (van, amit nehéz kiszűrni/észrevenni!)
 - Egy kísérlet komoly ellenőrzése csak a reprodukálásával lehetséges – erre nyilvánvalóan nincs kapacitás
 - Tudatos csalások (pl. H.M. Krishna Murthy)
- Ki teszi be az adatokat?
 - Bárki, aki produkálta őket: elsődleges adatbázisok
 - pl. DNS szekvenciák, fehérjeszerkezetek
 - Az adatok minősége leginkább a kutató hozzáállását tükrözi (természetesen vannak törekvések az ellenőrzésre stb.)
 - Az adatok nyilvános adatbázisban való elhelyezését a legtöbb tudományos folyóirat megköveteli, de a kutatók megtalálják a módját, hogy ha nem akarják, nem teszik be (vagy az igazán érdekes részeket elmismásolják)
 - Kurátorok: másodlagos adatbázisok, illetve egyes elsődlegesek esetében annotáció
 - Adatok elsődleges adatbázisokból/irodalomból
 - Annotálás, esetleg részletes ellenőrzés
 - Adott esetben komoly automatizálás is lehet, kézi beavatkozás csak szükség esetén

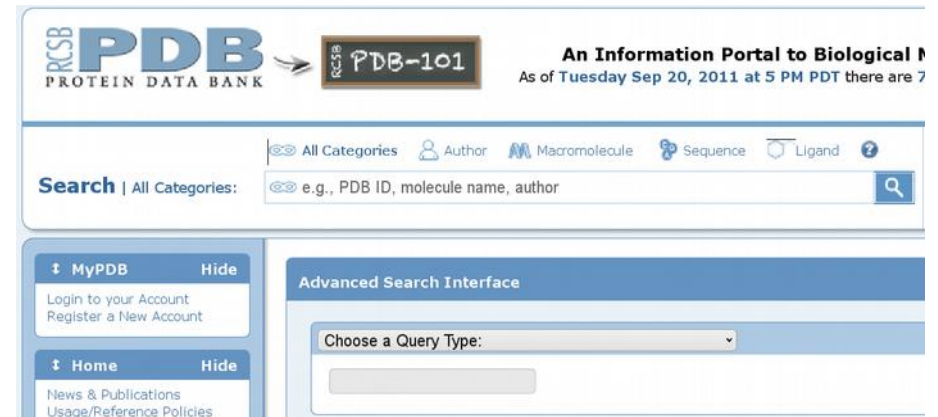
Adatbázisok működése és tudományetika

- Hozzáférés az adatokhoz
 - Teljesen szabad: nonprofit szervezetek, közösségi források segítségével fenntartott adatbázisok esetén
 - fizetős: céges adatbázisok.
 - vegyes: bizonyos részek ingyen, a teljes adatbázis pénzért – bevett gyakorlat
- Ki az adatok tulajdonosa?
 - Közösségi forrásokból finanszírozott kutatások esetén az adatokat általában kötelező szabad hozzáférésű adatbázisokba tenni
 - Céges adatbázisokból származó adatokat nem lehet akárkivel megosztani
- Az adatok felhasználása
 - Ha az adatok egyszer hozzáférhetőek lettek, a szerzőknek nincs beleszólásuk, ki és mire használja őket
 - Az adatok generálása általában nagyobb munka, mint az elemzésük (de az NGS esetében már ez fordítva igaz!), de az elemzés teszi az adatokat biológiailag értelmezhetővé: konfliktus a kísérleteket és a feldolgozást végző kutatók között
 - embargó: az adatokat elhelyezik, de adott ideig nem hozzáférhetőek ill. nem használhatóak fel tudományos közleményekben
 - Valós eset: 2009 augusztusában megjelent egy cikk, ami Laura Bierut csoportjának a dbGaP (database of genotypes and phenotypes) adatbázisban elhelyezett adatait használta, bár az adatok szeptemberig embargó alatt voltak (mivel a cikket márciusban küldték be, az embargót 6 hónappal sértették meg)



Néhány gyakorlati szempont

- Egy vagy néhány adatbázisbejegyzés (rekord) használata:
 - Érdeemes a webes keresőfelületet használni
 - Általában működik, a kívánt adat megtalálható, letölthető stb.
 - Részletes minőség-ellenőrzés lehetséges
- Nagy adatmennyiség elemzése
 - Meg lehet próbálni a webes keresést, de jó eséllyel nem teljes listát kapunk, elszalasztunk valamit, vagy csak túl sok lesz a találat
 - Ilyenkor sokszor szükséges a teljes adatbázis letöltése helyi elemzéshez, DE sokszor amit kapunk, távolról sem adatbázis, csak szöveges állományok
 - Akár magunknak kellhet belőle „igazi” adatbázist csinálni a hatékony munkához
 - Minőség-ellenőrzés nem könnyű, egyszerű automatizált megoldás kell
- Adatbázisok verziói
 - A legtöbb adatbázist többé-kevésbé rendszeresen frissítik
 - Ha túl ritkán, még nem lesz bent, amit keresünk
 - Ha túl gyakran, mire befejezzük az elemzést, új változatok lesznek kint, fontos, hogy legyen meg helyben az a verzió, amivel dolgoztunk, hogy ellenőrizni tudjuk az eredményeinket később is!



Index of ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/

[Up to higher level directory](#)

Name	Size	Last Modified
File: README	5 KB	07/27/2011 02:00:00 PM
File: README.gunzip	1 KB	07/27/2011 02:00:00 PM
File: README.reldate	1 KB	07/27/2011 02:00:00 PM
File: README.varsplc	11 KB	07/27/2011 02:00:00 PM
docs		07/29/2011 02:12:00 PM
reldate.txt	1 KB	07/27/2011 02:00:00 PM
File: uniprot.xsd	50 KB	07/27/2011 02:00:00 PM
uniprot_sprot.dat.gz	435772 KB	07/27/2011 02:00:00 PM
uniprot_sprot.fasta.gz	77273 KB	07/27/2011 02:00:00 PM
uniprot_sprot.xml.gz	751560 KB	07/27/2011 02:00:00 PM
uniprot_sprot_varsplc.fasta.gz	6436 KB	07/27/2011 02:00:00 PM
uniprot_trembl.dat.gz	6595994 KB	07/27/2011 02:00:00 PM
uniprot_trembl.fasta.gz	3183457 KB	07/27/2011 02:00:00 PM
uniprot_trembl.xml.gz	12917040 KB	07/27/2011 02:00:00 PM

Adatbázisok és fájlformátumok

- Sokféle adatbázis még többféle adatot tárol
- Sok adatbázisnak saját adatformátuma van (megadhatnak egyes adatokat gyakran használt formátumban is)
- Több formátumot az adatbázisokról neveztek el (GenBank, PDB, EMBL...)
- Formátumkonverzió fontossága:
 - Csak első ránézésre egyszerű
 - Adatvesztés: csak az marad(hat) meg, amit mindkét formátum képes reprezentálni, ténylegesen szerepel a bemenetben ÉS a konverter is kezeli!
 - A formátumok összetettsége miatt nem mindig könnyű saját konvertert írni – ha már létezik elérhető, használjuk azt, de körültekintéssel!
 - Ritkán használt szintaktikai megoldások, kulcsszavak gondot okozhatnak a saját programjainknak is, de akár a „hivatalosoknak” is.
- Valós eset: FT ismétlődések
 - Egy kolléga kérte, hogy nézzek utána egy adott szekvenciában periodikusan előforduló Phe-Thr (FT) aminosavpárok jelentőségének
 - Kiderült, hogy az FT jelenléte hibás konverzió eredménye, az EMBL formátum 'feature' sorait vezeti be

```
FT          /translation="EEKYTTMFDNVNLDEILANDRLLNNYVKCLLEDGEANCTADGKEL
FT          KKAVPDALSNECAKCNKQKEGTTKVLKHLINHKPDIWAQLKAKYDPDGTYSKKYEDKE
FT          KELHE"
```

Egy EMBL formátumú file 'feature' részének részlete az adott DNS-szekvencia fordításával

Adatbázisok fajtái

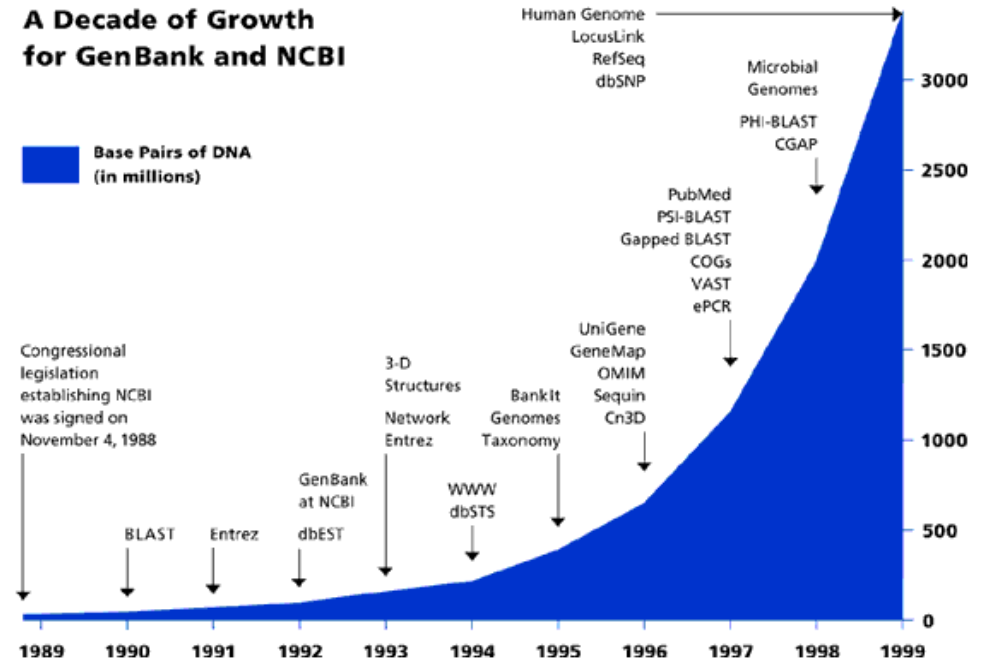
- Nyers adatok (pl. a TrEMBL automatikusan fordított fehérjeszekvenciái)
- Annotált szekvenciák (Swiss-Prot, Uniprot)
- A nyers és az annotált adatbázisok is elsődlegesek
- A másodlagos adatbázisok tipikusan elsődlegesekből vett adatokon alapulnak.
- Szekvenciák esetén: a PFAM fehérjedomének, HMM-ek és illesztések gyűjteménye, a COG BLAST hasonlóság alapján csoportosított bakteriális szekvenciáké, az SBASE fehérjedomének kollekciója BLAST alapokon. Ezek mindegyike Uniprot alapú.
- 3D szerkezetek esetén: a SCOP és a CATH 3D szerkezetek hierarchikus adatbázisa, a PDB-n alapulnak. A SCOP főleg manuálisan, a CATH főleg automatikusan annotált / készített adatbázis

Bioinformatikai portálok: integrált források (adatok, eszközök stb.)

- NCBI: National Center for Biotechnology Information



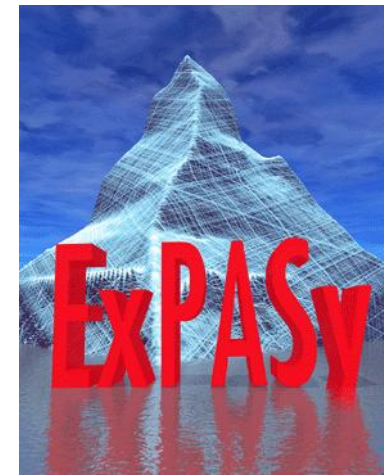
A Decade of Growth for GenBank and NCBI



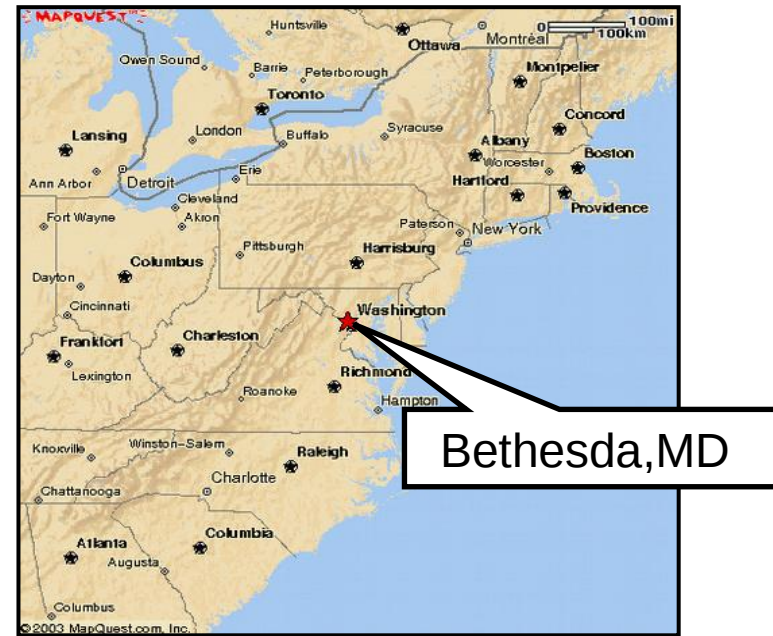
- EBI: European Bioinformatics Institute



- ExpASy: a Swiss Institute for Bioinformatics portálja
 - a UniProt konzorcium tagja



NCBI: The National Center for Biotechnology Information



***1988-ban hozták létre a
National Library of Medicine, NIH részeként***

- Publikus adatbázisok létrehozása
- Számítógépes biológiai kutatások
- Programfejlesztés szekvenciaelemzéshez
- Információmegosztás

NCBI Resources How To My NCBI | Sign In

National Center for Biotechnology Information Search All Databases for Search

Resources

- NCBI Home
- All Resources (A-Z)
- Literature
- DNA & RNA
- Proteins
- Sequence Analysis
- Genes & Expression
- Genomes
- Maps & Markers
- Domains & Structures
- Genetics & Medicine
- Taxonomy
- Data & Software
- Training & Tutorials
- Homology
- Small Molecules
- Variation


Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[More about the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS](#)

Genome

1000 prokaryotic genomes are now completed and available in the Genome database.



Popular Resources

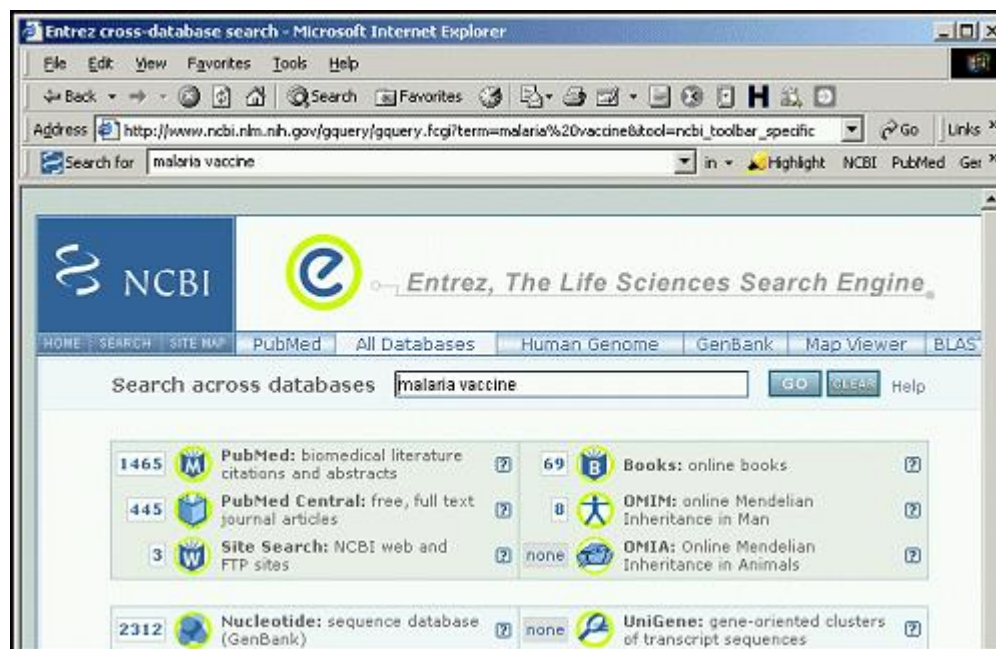
- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene
- Nucleotide
- Protein
- GEO
- Conserved Domains
- Structure
- PubChem

You are here: NCBI Help Desk

GETTING STARTED	RESOURCES	POPULAR	FEATURED	NCBI INFORMATION
Site Map	Literature	PubMed	GenBank	About NCBI
NCBI Help Manual	DNA & RNA	PubMed Central	Reference Sequences	Research at NCBI
NCBI Handbook	Proteins	Bookshelf	Map Viewer	NCBI Newsletter
Training & Tutorials	Sequence Analysis	BLAST	Genome Projects	NCBI FTP Site
	Genes & Expression	Gene	Human Genome	Contact Us
	Genomes	Nucleotide	Mouse Genome	
	Maps & Markers	Protein	Influenza Virus	
	Domains & Structures	GEO	Primer-BLAST	
	Genetics & Medicine	Conserved Domains	Short Read Archive	
	Taxonomy	Structure		
	Data & Software	PubChem		
	Training & Tutorials			
	Homology			
	Small Molecules			
	Variation			

Néhány NCBI adatbázis és szolgáltatás

- GenBank: elsődleges (DNS) szekvencia adatbázis (NGS adatok mennyiségével jelentősége valamelyest csökken, bár fontos referencia maradt!)
- Szabad hozzáférés biomedicinális irodalomhoz
 - PubMed: szabad Medline (3 millió keresés naponta)
 - PubMed Central: Ingyenes teljes szövegű hozzáférés (open access publikációk + a NIH által támogatott kutatásokból származó, kötelezően beteendő cikkek)
- Entrez: integrált molekuláris és irodalmi adatbázisok



- BLAST: leggyakrabban használt szekvenciahasonlóság-kereső szolgáltatás (100 – 200 ezer keresés naponta)

Entrez: az összes NCBI adatbázis egyidőjű keresése

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases Alzheimer's Disease GO Clear Help

- Result counts displayed in gray indicate one or more terms not found

66868	PubMed: biomedical literature citations and abstracts	892	Books: online books
12588	PubMed Central: free, full text journal articles	145	OMIM: online Mendelian Inheritance in Man
none	Site Search: NCBI web and FTP sites	165	OMIA: online Mendelian Inheritance in Animals

10785	Nucleotide: Core subset of nucleotide sequence records	180	dbGaP: genotype and phenotype
none	EST: Expressed Sequence Tag records	5	UniGene: gene-oriented clusters of transcript sequences
none	GSS: Genome Survey Sequence records	31	CDD: conserved protein domain database
4445	Protein: sequence database	435	3D Domains: domains from Entrez Structure
1	Genome: whole genome sequences	none	UniSTS: markers and mapping data
102	Structure: three-dimensional macromolecular structures	1	PopSet: population study data sets
none	Taxonomy: organisms in GenBank	48581	GEO Profiles: expression and molecular abundance profiles
none	SNP: single nucleotide polymorphism	58	GEO DataSets: experimental sets of GEO data
525	Gene: gene-centered information	none	Cancer Chromosomes: cytogenetic databases
none	SRA: Short Read Archive	82	PubChem BioAssay: bioactivity screens of chemical substances
15	BioSystems: Pathways and systems of interacting molecules	5	PubChem Compound: unique small molecule chemical structures
none	HomoloGene: eukaryotic homology groups	28	PubChem Substance: deposited chemical substance records
114	GENSAT: gene expression atlas of mouse central nervous system	none	Protein Clusters: a collection of related protein sequences
61	Probe: sequence-specific reagents	none	Peptidome: MS/MS proteomic experiments
1	Genome Project: genome project information		

3	Journals: detailed information about the journals indexed in PubMed and other Entrez databases	3	MeSH: detailed information about NLM's controlled vocabulary
1476	NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections		

- 40 (és növekvő számú) integrált adatbázis

PubMed: cikkek kivonatai (biomedicinális területen)

> 21 millió kivonat összetett kereshetősége

The screenshot shows a PubMed search results page for the query "neanderthal genome". The search bar at the top contains the text "neanderthal genome" and a "Search" button. Below the search bar, there are options for "Display Settings" (Summary, 20 per page, Sorted by Recently Added) and "Send to" options. The results are listed in a table with three entries:

- Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression.**
Currat M, Excoffier L.
Proc Natl Acad Sci U S A. 2011 Sep 13;108(37):15129-34. Epub 2011 Sep 12.
PMID: 21911389 [PubMed - in process]
[Related citations](#)
- Cross-comparison of the genome sequences from human, chimpanzee, Neanderthal and a Denisovan hominin identifies novel potentially compensated mutations.**
Zhang G, Pei Z, Ball EV, Mort M, Kehrer-Sawatzki H, Cooper DN.
Hum Genomics. 2011 Jul 1;5(5):453-84.
PMID: 21807602 [PubMed - in process]
[Related citations](#)
- On characterizing adaptive events unique to modern humans.**
Crisci JL, Wong A, Good JM, Jensen JD.
Genome Biol Evol. 2011;3:791-8. Epub 2011 Jul 29.
PMID: 21803765 [PubMed - in process] **Free PMC Article**
[Free full text](#) [Related citations](#)

On the right side of the page, there are sections for "Filter your results:" (All (33), Free Full Text (14), Review (3)), "Titles with your search terms" (Neanderthal genome sees first light, Neanderthal DNA yields to genome foray, Premises of the Neanderthal genome), and "11 free full-text articles in PubMed Central" (On characterizing adaptive events unique to modern humans).

Kulcsszó: "neanderthal genome" (más eredmény, mint a "neandertal genome" kulcsszóra!)

The screenshot shows the abstract page for the article "Neanderthal genomics and the evolution of modern humans" by Noonan JP. The abstract text is as follows:

Abstract
Humans possess unique physical and cognitive characteristics relative to other primates. Comparative analyses of the human and chimpanzee genomes are beginning to reveal sequence changes on the human lineage that may have contributed to the evolution of human traits. However, these studies cannot identify the genetic differences that distinguish modern humans from archaic human species. Here, I will discuss efforts to obtain genomic sequence from Neanderthal, the closest known relative of modern humans. Recent studies in this nascent field have focused on developing methods to recover nuclear DNA from Neanderthal remains. The success of these early studies has inspired a Neanderthal genome project, which promises to produce a reference Neanderthal genome sequence in the near future. Technical issues, such as the level of Neanderthal sequence coverage that can realistically be obtained from a single specimen and the presence of modern human contaminating sequences, reduce the detection of authentic human-Neanderthal sequence differences but may be remedied by methodological improvements. More critical for the utility of a Neanderthal genome sequence is the evolutionary relationship of humans and Neanderthals. Current evidence suggests that the modern human and Neanderthal lineages diverged before the emergence of contemporary humans. A fraction of biologically relevant human-chimpanzee sequence differences are thus likely to have arisen and become fixed exclusively on the modern human lineage. A reconstructed Neanderthal genome sequence could be integrated into human-primate genome comparisons to help reveal the evolutionary genetic events that produced modern humans.

PMID: 20439435 [PubMed - indexed for MEDLINE] PMID: PMC2860157 **Free PMC Article**

On the right side of the page, there are sections for "Related citations" (Sequencing and analysis of Neanderthal genomic DNA, Analysis of one million base pairs of Neanderthal DNA, Neanderthal DNA yields to genome foray) and "Related information" (Related Citations).

Az ExPASy szerver

<http://www.expasy.org>



- Az első molekuláris biológiai szerver (1993 augusztus), proteomikai fókusszal
 - Adatbázisok: **UniProtKB**, **PROSITE**, ENZYME, **Swiss-2DPAGE**, stb.;
 - Sokféle 2D/MS fehérjeazonosító/jellemző és szekvenciaelemző eszköz
- Szekvenciaelemzés (Blast, ScanProsite, ProtParam, ProtScale, RandSeq, Translate, etc.);
- Proteomika (AACompldent, FindMod, FindPept, Aldente, PeptideMass, TagIdent, etc.);
- 3D szerkezetelemzés és -megjelenítés (Swiss-Model, Swiss-PDBviewer)

Az ExpASy WWW szerver: genomikai és proteomikai eszközök és adatbázisok

Visual Guidance

Categories

proteomics

genomics

sequence alignment

similarity search

characterisation/annotation

structural bioinformatics

systems biology

phylogeny/evolution





population genetics










transcriptomics

biophysics

imaging












Databases

-  **STRING** • protein-protein interactions • [\[more\]](#)
-  **EPD** • collection of eukaryotic promoters • [\[more\]](#)
-  **SwissRegulon** • annotations of regulatory sites • [\[more\]](#)
-  **smirnaDB** • miRNA expression profiles analysis • [\[more\]](#)

-  **CLIPZ** • binding sites of RNA-binding proteins • [\[more\]](#)
-  **EIMMo** • miRNA target predictions • [\[more\]](#)
-  **GPSDB** • gene and protein synonyms • [\[more\]](#)
-  **ImmunoDB** • insect immune-related genes and gene families • [\[more\]](#)
-  **miOrtho** • catalogue of animal microRNA genes • [\[more\]](#)
-  **MyHits** • protein domains database and tools • [\[more\]](#)
-  **OMA** • orthology inference among complete genomes. • [\[more\]](#)
-  **OpenFlu** • Influenza genetic and epidemiological data • [\[more\]](#)
-  **OrthoDB** • Hierarchical catalog of eukaryotic orthologs • [\[more\]](#)

Tools

-  **EPD** • collection of eukaryotic promoters • [\[more\]](#)
-  **smirnaDB** • miRNA expression profiles analysis • [\[more\]](#)

-  **Association Viewer** • SNPs display in a genetic context • [\[more\]](#)
-  **BayeScan** • identify natural selection • [\[more\]](#)
-  **BLAST** • sequence similarity search • [\[more\]](#)
-  **boxshade** • MSA pretty printer • [\[more\]](#)
-  **ChIP-Seq** • ChIP-Seq data analysis tools • [\[more\]](#)
-  **CLIPZ** • binding sites of RNA-binding proteins • [\[more\]](#)
-  **Codon Suite** • codon-based sequence analysis • [\[more\]](#)
-  **Dotlet** • sequence similarity plots • [\[more\]](#)
-  **EIMMo** • miRNA target predictions • [\[more\]](#)
-  **EMBNet services** • bioinformatics tools and databases • [\[more\]](#)
-  **ESTscan** • coding region detection • [\[more\]](#)

Visual Guidance

Categories

proteomics

protein sequences and identification

mass spectrometry and 2-DE data

protein characterisation and function

families, patterns and profiles

post-translational modification

protein structure

protein-protein interaction

similarity search/alignment

genomics

structural bioinformatics

systems biology

phylogeny/evolution








population genetics


transcriptomics

biophysics

imaging












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 • [\[more\]](#)
-  **neXtProt** • human proteins • [\[more\]](#)
-  **PROSITE** • protein domains and families • [\[more\]](#)
-  **ViralZone** • portal to viral UniProtKB entries • [\[more\]](#)

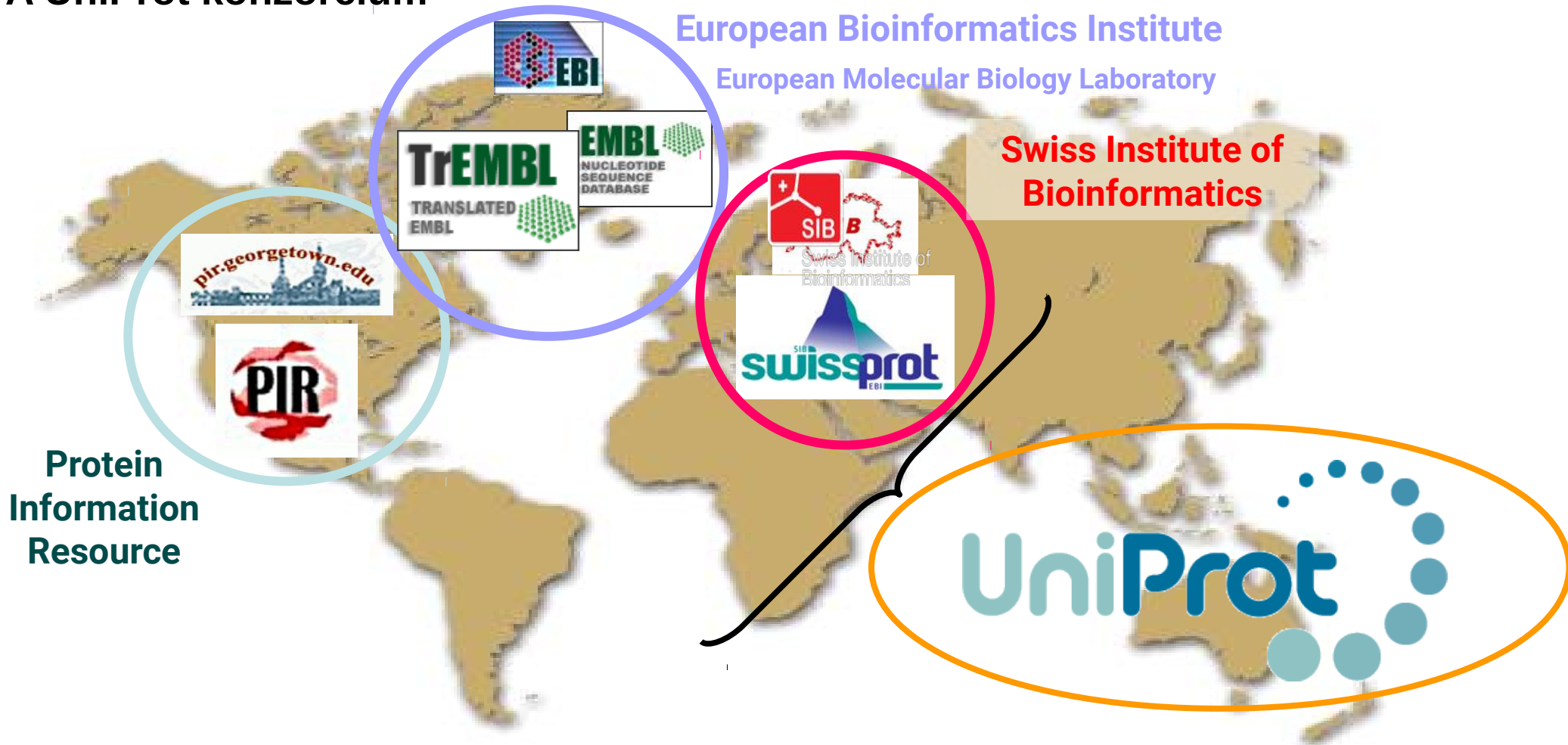
-  **ENZYME** • nomenclature of enzymes • [\[more\]](#)
-  **GlycoSuiteDB** • glycan database • [\[more\]](#)
-  **GPSDB** • gene and protein synonyms • [\[more\]](#)
-  **HAMAP** • manual protein annotation • [\[more\]](#)
-  **MIAPEGelDB** • MIAPE document edition • [\[more\]](#)
-  **MyHits** • protein domains database and tools • [\[more\]](#)
-  **PaxDb** • protein abundance database • [\[more\]](#)
-  **Prolune** • Popular science articles (in French) • [\[more\]](#)
-  **Protein Model Portal** • structural information for a protein • [\[more\]](#)
-  **Protein Spotlight** • Informally written reviews on proteins • [\[more\]](#)

Tools

-  **SWISS-MODEL Workspace** • structure homology-modeling • [\[more\]](#)
-  **SwissDock** • protein ligand docking server • [\[more\]](#)

-  **AACompSim** • amino acid composition comparison • [\[more\]](#)
-  **AllAll** • protein sequences comparisons • [\[more\]](#)
-  **BLAST** • sequence similarity search • [\[more\]](#)
-  **boxshade** • MSA pretty printer • [\[more\]](#)
-  **Compute pI/MW** • theoretical pI and Mw computation • [\[more\]](#)
-  **Dotlet** • sequence similarity plots • [\[more\]](#)
-  **EMBNet services** • bioinformatics tools and databases • [\[more\]](#)
-  **FindMod** • protein post-translational modifications • [\[more\]](#)
-  **FindPept** • peptide identification from unspecific cleavage • [\[more\]](#)
-  **GlycanMass** • oligosaccharide structure mass calculation • [\[more\]](#)
-  **GlycoMod** • oligosaccharide structure prediction • [\[more\]](#)
-  **HAMAP** • manual protein annotation • [\[more\]](#)
-  **HamapScan** • scan sequences against HAMAP • [\[more\]](#)
-  **HCD/CID spectra merger** • combine HCD and CID MS/MS spectra • [\[more\]](#)
-  **ImageMaster / Melanie** • software for 2-D PAGE analysis • [\[more\]](#)

A UniProt konzorcium



UniProt (Universal Protein Resource): a világon a legteljesebb fehérjekatalógus

<http://www.uniprot.org>

- UniProt Knowledgebase
- UniRef clusters (100/90/50% azonosság)
- UniParc (UniProt Archive)
- UniMES (Metagenomic and Environmental Sequences)

UniProt adatbázisok

UniProt Knowledgebase



UniRef100

UniRef90

UniRef50

UniParc

UniProtKB:



UniProtKB/TrEMBL
Automatikusan annotált
fehérjeszekvenciák
(>50 millió)



UniProtKB/Swiss-Prot
Kutatók által annotált
fehérjeszekvenciák
(kb. 500 ezer)

· Egy **UniRef100** bejegyzés = **Minden azonos szekvenca** (fragmensek is).

· Egy **UniRef90** bejegyzés = Legalább **90%-os azonosságot mutató szekvenciák**

·
Fajtól függetlenül

UniProt Archives:

Archivált nyers fehérjeszekvenciák, publikus adatbázisokból

Swiss-Prot, TrEMBL, PIR, EMBL, Ensembl, IPI, PDB, RefSeq, FlyBase, WormBase, Patent Offices.

Körültekintés szükséges:
pseudogének, inkorrekt CDS predikciók stb

+ UniMES metagenomikai és környezeti minták

A UniProt website

- www.uniprot.org
- Teljes adatkészletek letöltéséhez: [ftp.uniprot.org](ftp://ftp.uniprot.org)



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

Swiss-Prot (547,599)
Manually annotated and reviewed.

TrEMBL (90,860,905)
Automatically annotated and not reviewed.

UniRef

Sequence clusters

UniParc

Sequence archive

Proteomes

Supporting data

Literature citations

Taxonomy

Subcellular locations

Cross-ref. databases

Diseases

Keywords

News

Mosquitoes prefer human
[UniProt release 2015_02](#)

Thalidomide, the pharmacological version of yin and yang | Cross-references to DEPOD, MoonProt and Proteomes
[UniProt release 2015_01](#)

Higher and higher | New mouse and zebrafish variation files | Structuring of 'cofactor' annotations
[UniProt release 2015_01](#)

[News archive](#)

Getting started

- [Text search](#)
Our basic text search allows you to search all the resources available
- [BLAST](#)
Find regions of similarity between your sequences
- [Sequence alignments](#)
Align two or more protein sequences using the Clustal Omega program
- [Retrieve/ID mapping](#)
This tool merges the "Retrieve" and "ID Mapping" tools



UniProt data

- [Download latest release](#)
Get the UniProt data
- [Statistics](#)
View Swiss-Prot and TrEMBL statistics
- [Forthcoming changes](#)
Planned changes for the UniProt knowledgebase
- [Submit your data](#)
Submit your sequences and annotation updates

Protein spotlight



The Hidden Things

December 2014

Nature has its secret ways. During the course of the 19th

century, the Augustinian friar Gregor Mendel worked out the basics of genetic inheritance as he crossbred pea plants. About a century later, it has become obvious that the inheritance of a given trait is in fact not so straightforward...

Tools

BLAST
Align
Retrieve/ID mapping

Core data

Protein knowledgebase (UniProtKB)
Sequence clusters (UniRef)
Sequence archive (UniParc)
Proteomes

Supporting data

Literature citations
Taxonomy
Keywords
Subcellular locations
Cross-referenced databases
Diseases

Information

About UniProt
Help
FAQ
UniProtKB manual
Technical corner
Annotation programs

UniProtKB/Swiss-Prot annotáció



- Egy adott fehérjéről való kurrens tudásunk
- Ún. 'controlled vocabulary', azaz adott kifejezések, (kulcs)szavak következetes használata
Keywords, Tissues, Post-translational modifications, Strains, Species, Subcellular location, Extracellular domains, Journals...
- Megbízható annotációt ad, ami felhasználható:
 - Nem jellemzett genomokból származó szekvenciákra az annotáció átviteléhez (de legyünk óvatosak az ortológia/paralógia kapcsán!)
 - Programok, eszközök tesztelése (adatbányászat, predikciók stb.)
- UniProtKB/Swiss-Prot adatforrások:
 - publikációk (Pubmed)
 - adatbázisok
 - Nevezéktani bizottságok
 - predikciók
 - Szerzőkkel való kapcsolatfelvétel
- A manuális annotációra kiválasztott fehérjék:
 - Új, nagy hatású publikációkban leírt fehérjék (pl. cereblon, Q96SW2, http://www.expasy.org/spotlight/back_issues/sptlt117.shtml)
 - Adott metabolikus vagy jelátviteli útvonal (pl. az ubiquitin-szerű konjugációs útvonal)
 - Felhasználói kérések
 - 3D szerkezettel rendelkező fehérjék

Az annotáció eredetét/megbízhatóságát jellemző azonosítók

Coiled coil

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Coiled coil ⁱ	842 – 1939	1098	Sequence analysis			Add BLAST

Domainⁱ

The rodlike tail sequence is highly repetitive, showing cycles of a 28-residue repeat pattern composed of 4 heptapeptides, characteristic for alpha-helical coiled coils. Each myosin heavy chain can be split into 1 light meromyosin (LMM) and 1 heavy meromyosin (HMM). It can later be split further into 2 globular subfragments (S1) and 1 rod-shaped subfragment (S2).

Sequence similaritiesⁱ

Belongs to the TRAFAC class myosin-kinesin ATPase superfamily. Myosin family. Curated

Contains 1 IQ domain. PROSITE-ProRule annotation

Contains 1 myosin motor domain. Curated

Keywords - Domainⁱ

Coiled coil

Natural variant

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Natural variant ⁱ	56 – 56	1	G → R. 1 Publication Corresponds to variant rs28711516 [dbSNP Ensembl].		VAR_063550	
Natural variant ⁱ	88 – 88	1	E → Q. 2 Publications Corresponds to variant rs442275 [dbSNP Ensembl].		VAR_030203	
Natural variant ⁱ	275 – 275	1	I → N. 1 Publication		VAR_063551	
Natural variant ⁱ	721 – 721	1	R → W in SSS3; rare variant predisposing to sick sinus syndrome. 1 Publication		VAR_065561	

Manual assertion
inferred by curator

Manual assertion
inferred by rules

Manual assertion
based on experiment

Fehérje izoformák a SwissProt adatbázisban

- Egy reprezentatív izoforma az „alapértelmezett”
- A többi izoforma variánsként listázva, letölthetőek
- Teljes adatkészletek az összes variánssal letölthetőek (számos proteom)
- Információk további variánsokról, módosításokról

Sequences Hide | Top

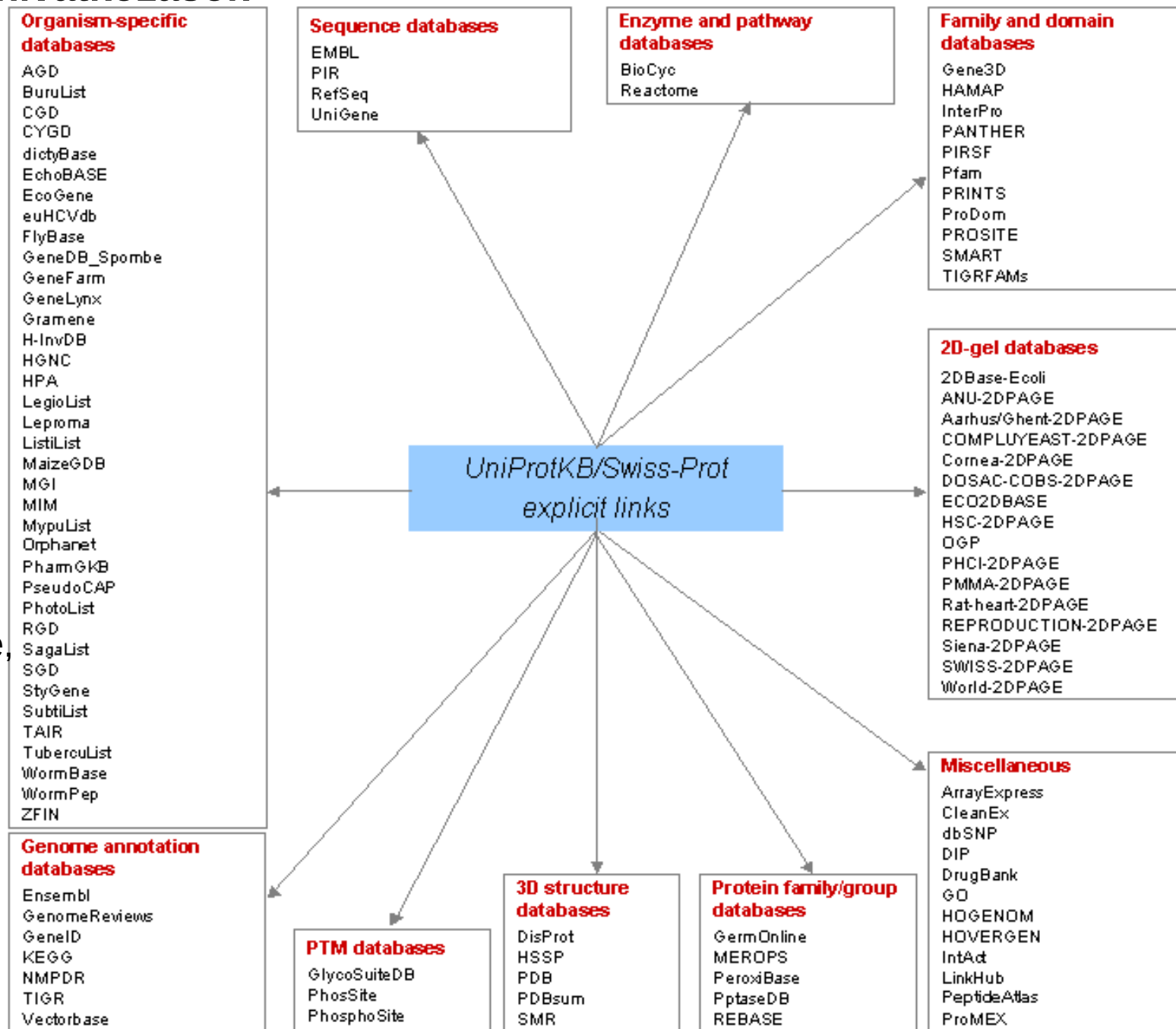
Sequence	Length	Mass (Da)	Tools
<input type="checkbox"/> Isoform Alpha (Alpha-A) [UniParc]. FASTA Last modified November 1, 1986, Version 1. Checksum: C5C90C9A5DD16AAB	777	85,659	Blast <input type="text"/> <input type="button" value="go"/>
<pre> 10 20 30 40 50 60 MDSKESLTPG REENPSSVLA QERGDVDMFY KTLRGGATVK VSASSPPLAV ASQSDSKQRR 70 80 90 100 110 120 LLVDFPKGSV SNAQQPDLK AVLSMGLYH GETETKVNGN DLGFPQQQI SLSSGETDLK 130 140 150 160 170 180 LLEESIANLN RSTSVPENPK SSASTAVSAA PTEKEFPKTH SDVSSEQQHL KGQTGTNGN 190 200 210 220 230 240 VKLYTTDQST FDILODLEFS SGSPGKETNE SPWRSLLID ENCLLSPLAG EDDSFLLEGN 250 260 270 280 290 300 SNEDCKPLIL PDTKPKIKDN GDLVLSSPSN VTLPOVKTEK EDFIELCTPG VIKQEKLGTV 310 320 330 340 350 360 YCAASFPGAN IIGNKMSAIS VHGVTSSGGQ MYHYDMNTAS LSQQQDQKI FNVIPPPIVG 370 380 390 400 410 420 SENWNRQGS GDDNLTSLGT LNFPGRIVFS NGYSSPSMRP DVSSPPSSS TATTGPPPKL 430 440 450 460 470 480 CLVCSDEASG CHYGLTCGS CKVFFKRAVE GQHNYLCAGR NDCIIDKIR KNCPCACRYK 490 500 510 520 530 540 CLQAGHLEA RKTKKKIKGI QQATTGVSQE TSENPGRKI VPATLPQLTP TLVSLLEVE 550 560 570 580 590 600 PEVLYAGYDS SVPDSTWRIM TTLNMLGGRQ VIAAVKWAKA IPGFRNLHLD DQHTLLQYSW 610 620 630 640 650 660 MFLMAFALGW RSYROSSANL LCFAPDLIIN EQRHTLPCHY DQCKHMLYVS SELHRLQVSY 670 680 690 700 710 720 EELYCHKTL LLSVFPKDLG KQELFDEIR HTYIKELGKA IVKREGNSSQ NWQRFYQLTK 730 740 750 760 770 LLDSRHVVVE NLLNYCFQTF LDKTMSIEFF EMLAEITNQ IPKYSNGNIK KLLFHQK </pre>			
← Hide			
<input type="checkbox"/> Isoform Beta (Beta-A) [UniParc]. Checksum: E2D1F6EA0EE14704 Show >	742	81,509	Blast <input type="text"/> <input type="button" value="go"/>
<input type="checkbox"/> Isoform Alpha-2 (Gamma) [UniParc]. Checksum: 23048D99B60B169C Show >	778	85,815	Blast <input type="text"/> <input type="button" value="go"/>
<input type="checkbox"/> Isoform Beta-2 [UniParc]. Checksum: 329BE98BCC1DC0E5 Show >	743	81,666	Blast <input type="text"/> <input type="button" value="go"/>
<input type="checkbox"/> Isoform GR-A alpha [UniParc]. Checksum: 8CB72DE6B0593EFD Show >	593	64,752	Blast <input type="text"/> <input type="button" value="go"/>
<input type="checkbox"/> Isoform GR-A beta [UniParc]. Checksum: 7C1C30CD84689FBE Show >	558	60,602	Blast <input type="text"/> <input type="button" value="go"/>
Isoform GR-P (Sequence not available). - -			
<input type="checkbox"/> Isoform Alpha-B (Beta-B) [UniParc]. Checksum: C6D7A2D88B4025C1 Show >	751	82,845	Blast <input type="text"/> <input type="button" value="go"/>
<input type="checkbox"/> Isoform Beta-B [UniParc]. Checksum: BCF1D97EFD06AB74 Show >	716	78,695	Blast <input type="text"/> <input type="button" value="go"/>

References Hide | Top

[← Hide Target scale references](#)

SwissProt: keresztivatkozások

- 125 adatbázishoz:
- EMBL/
GenBank/DDBJ,
RefSeq,
- PDB
- InterPro,
- PROSITE,
Pfam, Prints, etc.
- Organizmus-
specifikusak:
MIM, MGI, FlyBase,
SGD, GenoList,
- SWISS-2DPAGE
- PubMed



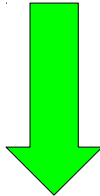
SwissProt és TrEMBL

- Definíciók:
 - Egy TrEMBL bejegyzés egy automatikusan generált bejegyzés egy EMBL CDS sor alapján, automatikus annotációval ('*Unreviewed*')
 - Egy Swiss-Prot bejegyzés egy manuálisan annotált rekord (*Reviewed*)
- A SwissProtba való integráció során az ID megváltozik, de az AC (accession number) megmarad

Serine/threonine protein phosphatase 2A 55 kDa regulatory subunit B beta isoform
Oryza sativa

```
ID   A2X2K3_ORYSI           Unreviewed;           524 AA.
AC   A2X2K3;
DT   20-MAR-2007, integrated into UniProtKB/TrEMBL.
DT   20-MAR-2007, sequence version 1.
DT   24-JUL-2007, entry version 5.

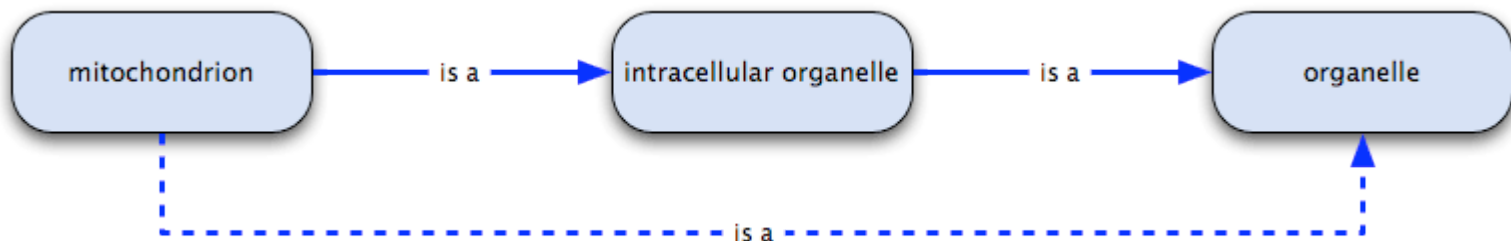
ID   2ABB_ORYSI           Reviewed;              525 AA.
AC   A2X2K3; O82774; Q6Z8B7;
DT   11-SEP-2007, integrated into UniProtKB/Swiss-Prot.
DT   11-SEP-2007, sequence version 2.
DT   11-SEP-2007, entry version 6.
```



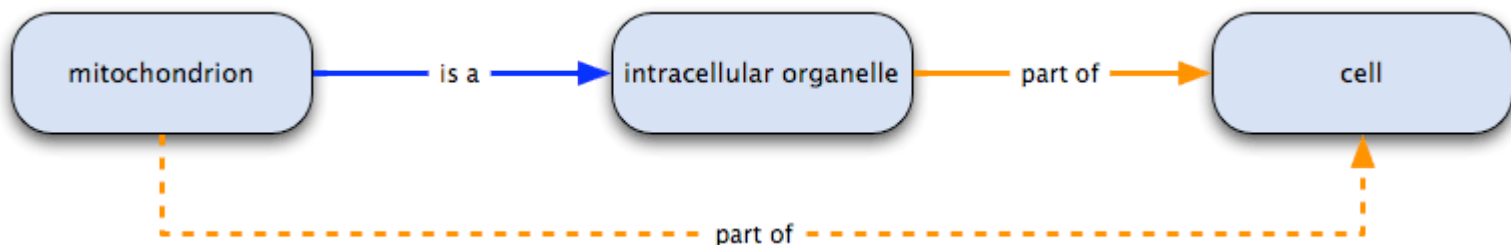
A GO (Gene Ontology) osztályozás

- Relációk

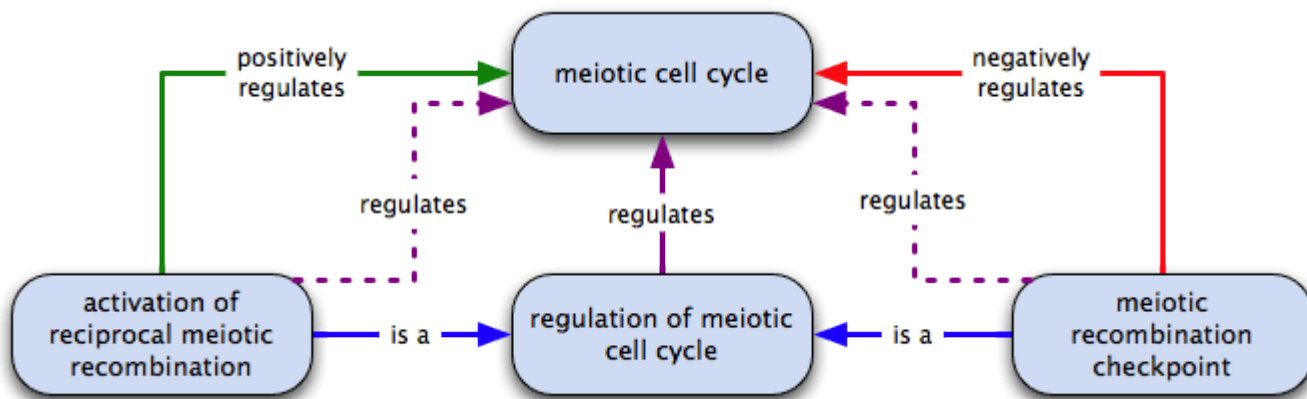
- Vannak alesetek (lásd a regulációnál)
- Egyes relációk implikálnak másokat (logikai összefüggések)



„is a”
= típusba sorolás



„part of”
= rész-egész
viszony



„regulates”
= szabályozás

A GO (Gene Ontology) osztályozás

- Jelentőség proteomikai vizsgálatokban:
 - Adott fehérjecsoportok összevetése: van-e köztük funkcionális eltérés (pl. azonosítunk n darab fehérjét, amik egy betegségben fontosak stb. - van-e statisztikailag kimutatható különbség a GO kulcsszavak eloszlásában)

BIOINFORMATICS

REVIEW

Vol. 23 no. 4 2007, pages 401–407
doi:10.1093/bioinformatics/btl633

Databases and ontologies

Enrichment or depletion of a GO category within a class of genes: which test?

Isabelle Rivals*, Léon Personnaz, Lieng Taing¹ and Marie-Claude Potier¹

Équipe de Statistique Appliquée and ¹Laboratoire de Neurobiologie et Diversité Cellulaire, École Supérieure de Physique et de Chimie Industrielles (ESPCI), 10 rue Vauquelin, 75005 Paris, France

Received on June 20, 2006; revised and accepted December 11, 2006

Advance Access publication December 20, 2006

Associate Editor: Jonathan Wren

Table 2. Classification of the genes expressed in a microarray experiment

	Category 1 (∈GO category)	Category 2 (∉GO category)	Total
Class 1 (DE)	n_{11}	n_{12}	n_{1+}
Class 2 (not DE)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Motívum és domén adatbázisok

- Másodlagos adatbázisok
- Motívum: egy szekvenciális szegmens, ami különböző fehérjékben előfordul, és funkcionális jelentősége van
- Domén: szerkezeti/feltekeredési/funkcionális/evolúciós egység (lásd még később)
- A motívumok és a domének szekvenciális hasonlóság alapján detektálhatók



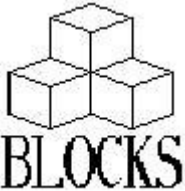
Doménygyűjtemény, cél a ritka variánsok felismerése/lefedése



Fehérjedomének/családok illesztésekkel és HMM-ekkel reprezentálva



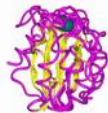
Illesztett fehérjeszekvenciák és domének



Illesztések konzervált „blokkjai”



Néhány száz domén manuálisan karbantartott modellje



Minden fehérje és genom funkcionális annotációja

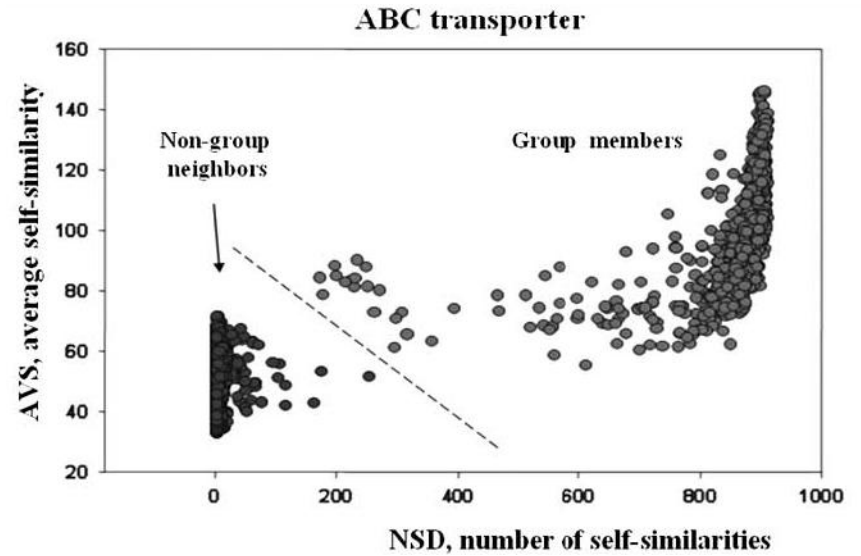
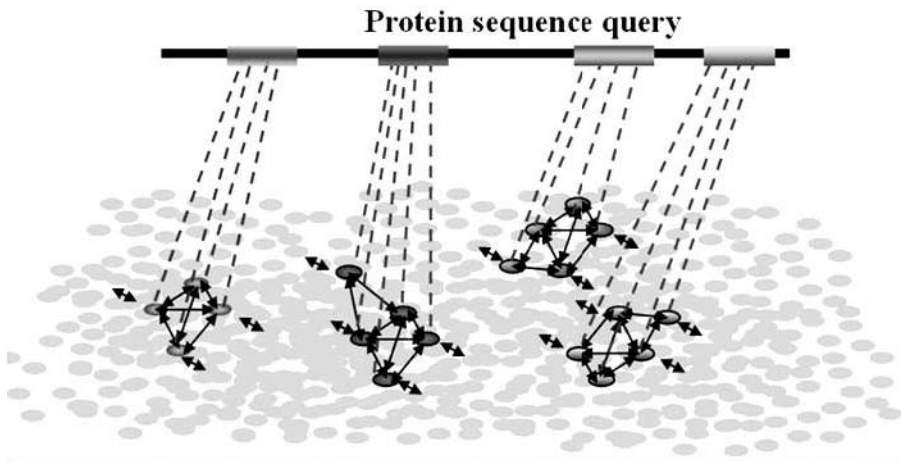


18, főleg másodlagos szervert integráló ún. metaszerver

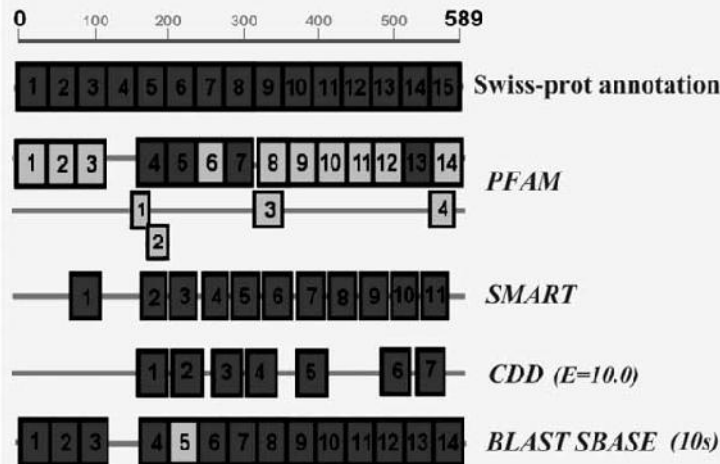
Példa: SBASE áttekintés



- Egyszerű szekvenciális hasonlóságokat keres
- Saját/nem saját megkülönböztetése hasonlósági csoportok alapján
- Képes ritka doménvariánsokat is detektálni



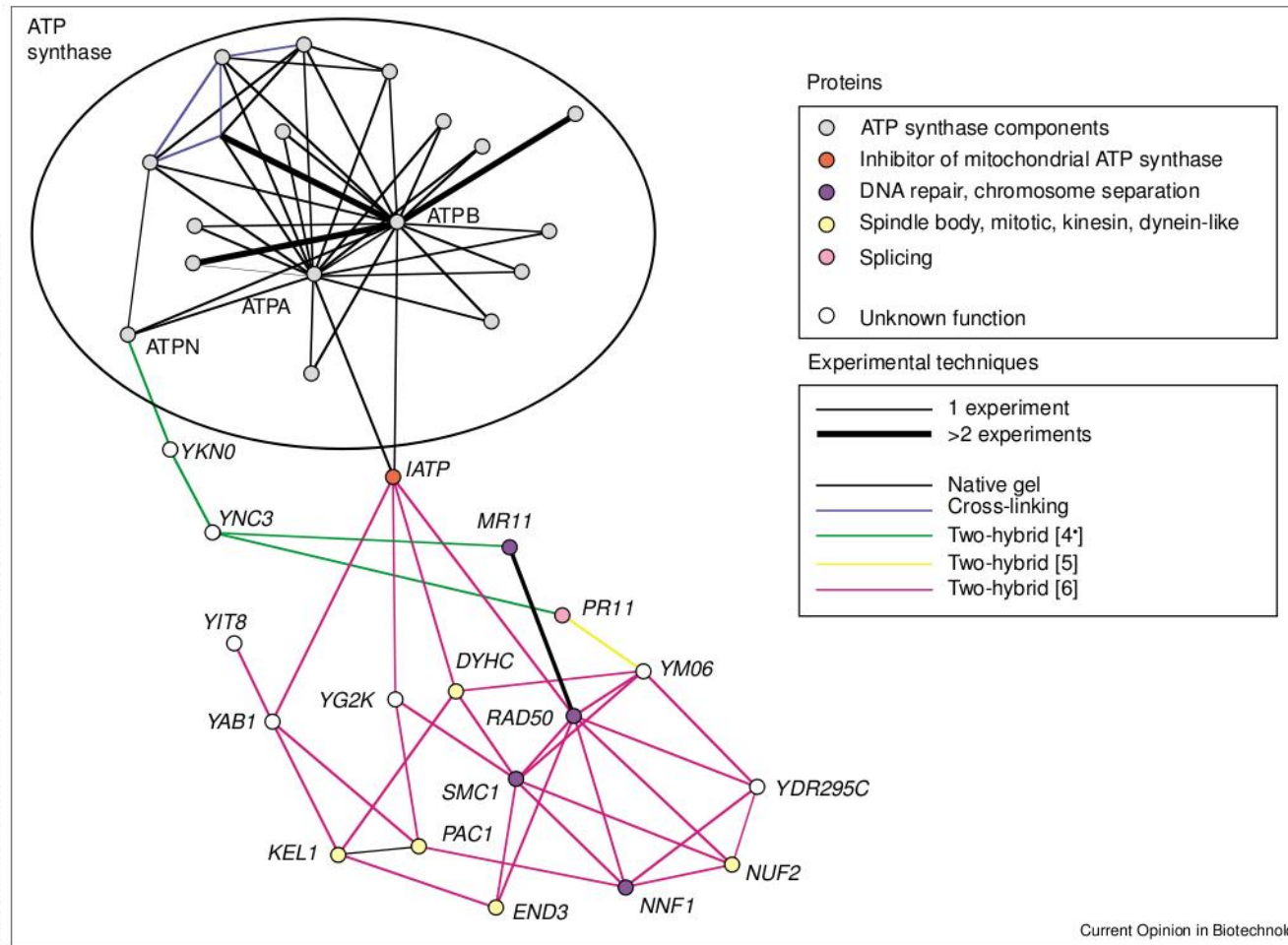
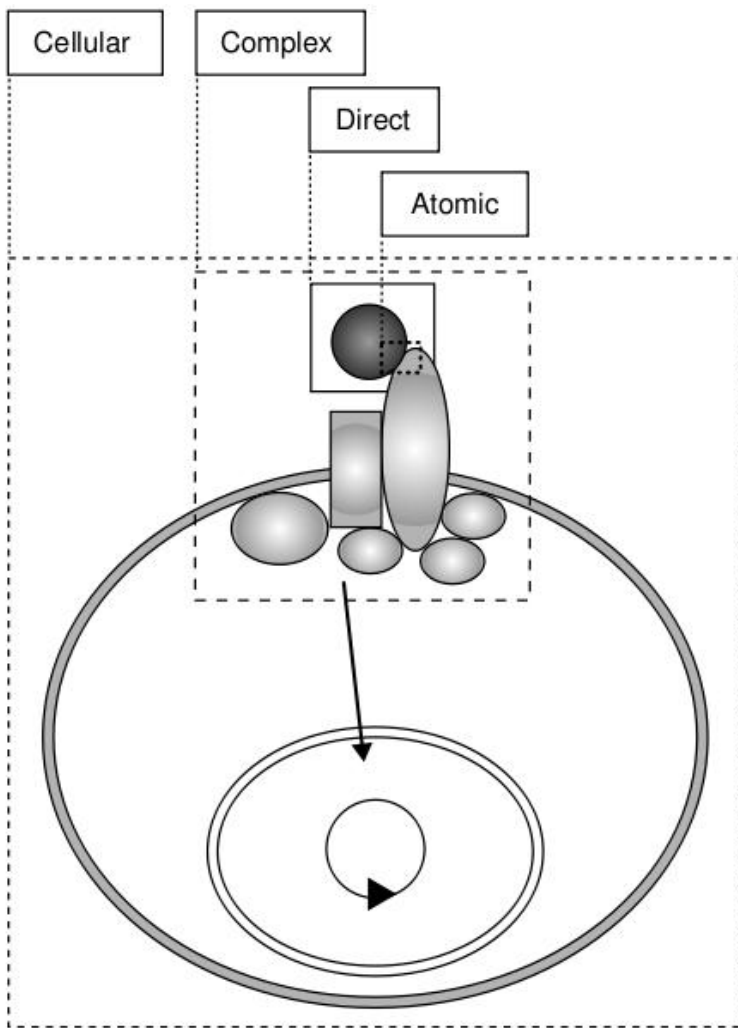
Prediction of HEAT repeat for 2AAA_HUMAN



Prediction: **Strong** Medium Weak

Fehérje-fehérje kölcsönhatási adatbázisok

- Probléma: honnan származik a kölcsönhatásról az információ
 - Rengeteg különféle kísérleti módszer, eltérő felbontással és megbízhatósággal
 - Egyedi vizsgálatok vs. nagy adatkészleten végzett kutatások (large scale)
 - A kölcsönhatások jellemző paraméterei is több nagyságrendet ölelnek fel
 - De szeretjük igen-nem alapon megjeleníteni a kölcsönhatásokat
- Sokféle adatbázis



Fehérje-fehérje kölcsönhatási adatbázisok



EMBL-EBI [Help](#) | [Feedback](#)

Databases Tools Research Training Industry About Us Help [Site Index](#)

EBI > Databases > Pathways & Networks > IntAct > View

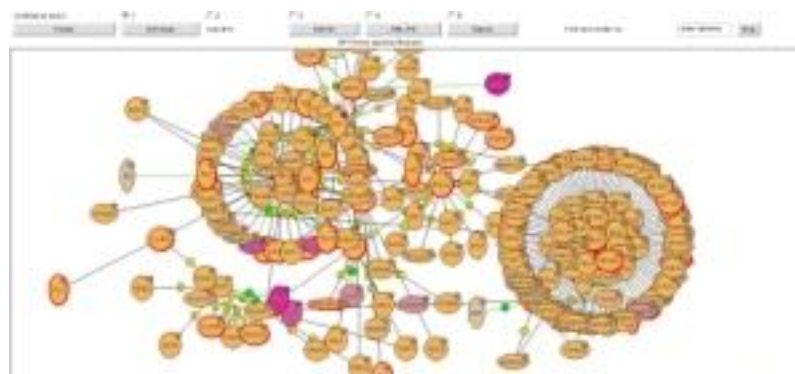
Search: [Show Advanced Fields >](#)

Home Search Interactions (7) Browse Lists Interaction Details Molecule View Graph

Network visualisation

Q81UX8 emrR Q81UJ6
Q81UM1 C1S Q7CKM5 Q81KW4
eutG yrbF

Navigation controls: pan, zoom, reset



MINT

STRING