(Aspekte der Thermodynamik in der Strukturbiologie)

Einführung in die Bioinformatik

Wintersemester 2012/13
16:00-16:45 Hörsaal N100 B3

Peter Güntert

Finding information:
Literature
Proteins
Alignments

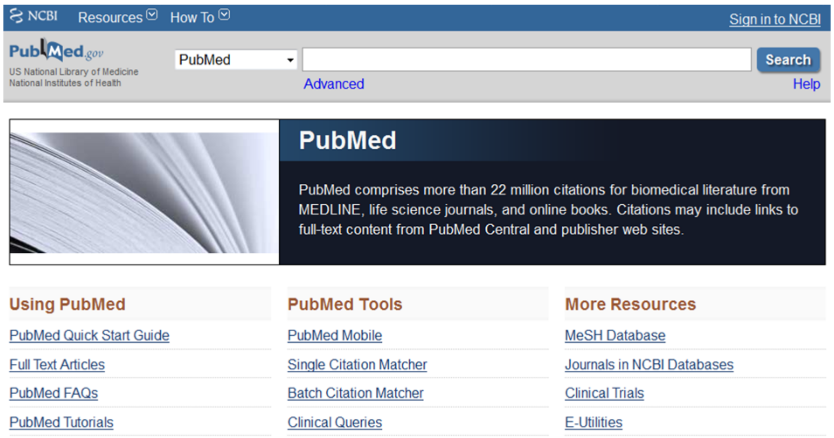
Outline

- Retrieving scientific information with PubMed
- Retrieving scientific information with Web of Knowledge

- Fetching the protein or DNA sequence you need
- Searching a database with BLAST
- Making a multiple sequence alignment

PubMed

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

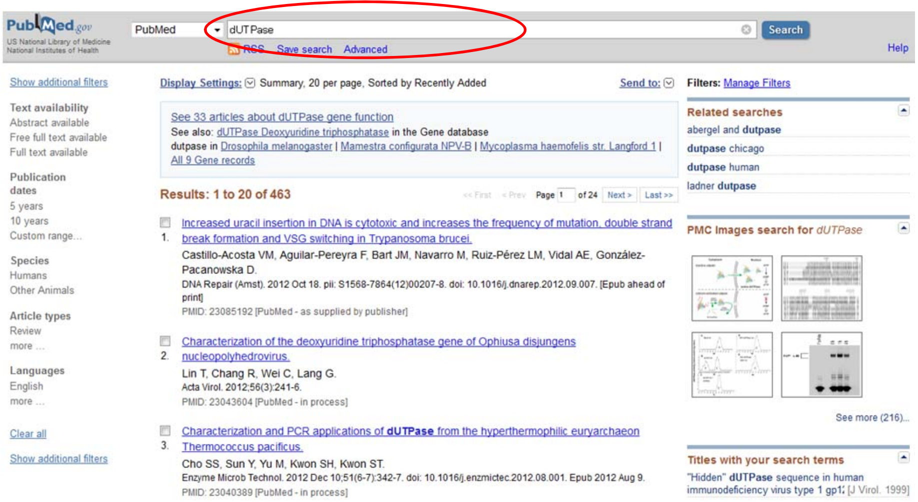


PubMed comprises more than 22 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

- PubMed is a database containing the recent scientific publications in biology and medicine.
- You can search PubMed using any keyword you are interested in.
- PubMed is free.

Searching PubMed Rapidly

- Open www.ncbi.nlm.nih.gov/pubmed
- Type your favorite keywords



Display Settings: Summary, 20 per page, Sorted by Recently Added

Send to:

Filters: Manage Filters

Show additional filters

Text availability
Abstract available
Free full text available
Full text available

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

Article types
Review
more ...

Languages
English
more ...

Clear all

Show additional filters

See 33 articles about dUTPase gene function
See also: dUTPase Deoxyuridine triphosphatase in the Gene database
dutpase in Drosophila melanogaster | Mamestra configurata NPV-B | Mycoplasma haemofelis str. Langford 1 |
All 9 Gene records

Results: 1 to 20 of 463

<< First < Prev Page 1 of 24 Next > Last >>

☐ 1. Increased uracil insertion in DNA is cytotoxic and increases the frequency of mutation, double strand break formation and VSG switching in Trypanosoma brucei.
Castillo-Acosta VM, Aguilar-Pereyra F, Bart JM, Navarro M, Ruiz-Pérez LM, Vidal AE, González-Pacanowska D.
DNA Repair (Amst). 2012 Oct 18. pii: S1568-7864(12)00207-8. doi: 10.1016/j.dnarep.2012.09.007. [Epub ahead of print]
PMID: 23085192 [PubMed - as supplied by publisher]

☐ 2. Characterization of the deoxyuridine triphosphatase gene of Ophiusa disjungens nucleopolyhedrovirus.
Lin T, Chang R, Wei C, Lang G.
Acta Virol. 2012;56(3):241-6.
PMID: 23043604 [PubMed - in process]

☐ 3. Characterization and PCR applications of **dUTPase** from the hyperthermophilic euryarchaeon Thermococcus pacificus.
Cho SS, Sun Y, Yu M, Kwon SH, Kwon ST.
Enzyme Microb Technol. 2012 Dec 10;51(6-7):342-7. doi: 10.1016/j.enzmictec.2012.08.001. Epub 2012 Aug 9.
PMID: 23040389 [PubMed - in process]

Related searches
abergel and **dutpase**
dutpase chicago
dutpase human
ladner **dutpase**

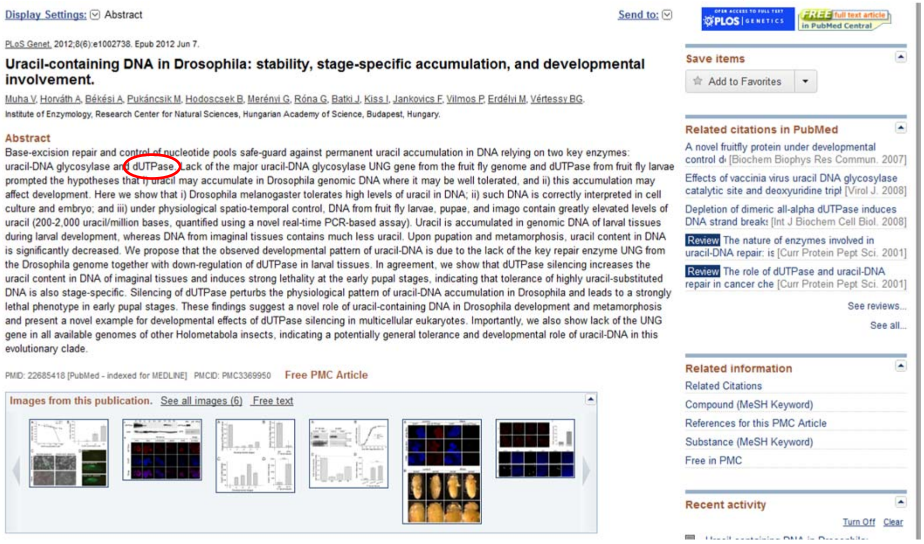
PMC Images search for *dUTPase*

See more (216)...

Titles with your search terms
"Hidden" **dUTPase** sequence in human immunodeficiency virus type 1 gp1: [J Virol. 1999]

PubMed entry

Links to full text (sometimes)



Display Settings: Abstract

Send to:

PLoS Genet. 2012;8(6):e1002738. Epub 2012 Jun 7.

Uracil-containing DNA in Drosophila: stability, stage-specific accumulation, and developmental involvement.

Muha V, Horváth A, Békési A, Pukáncsik M, Hodoscsek B, Merényi G, Róna G, Batki J, Kiss I, Jankovics F, Vilmos P, Erdélyi M, Vértessy BG.
Institute of Enzymology, Research Center for Natural Sciences, Hungarian Academy of Science, Budapest, Hungary.

Abstract
Base-excision repair and control of nucleotide pools safe-guard against permanent uracil accumulation in DNA relying on two key enzymes: uracil-DNA glycosylase and dUTPase. Lack of the major uracil-DNA glycosylase UNG gene from the fruit fly genome and dUTPase from fruit fly larvae prompted the hypotheses that i) uracil may accumulate in Drosophila genomic DNA where it may be well tolerated, and ii) this accumulation may affect development. Here we show that i) Drosophila melanogaster tolerates high levels of uracil in DNA; ii) such DNA is correctly interpreted in cell culture and embryo; and iii) under physiological spatio-temporal control, DNA from fruit fly larvae, pupae, and imago contain greatly elevated levels of uracil (200-2,000 uracil/million bases, quantified using a novel real-time PCR-based assay). Uracil is accumulated in genomic DNA of larval tissues during larval development, whereas DNA from imaginal tissues contains much less uracil. Upon pupation and metamorphosis, uracil content in DNA is significantly decreased. We propose that the observed developmental pattern of uracil-DNA is due to the lack of the key repair enzyme UNG from the Drosophila genome together with down-regulation of dUTPase in larval tissues. In agreement, we show that dUTPase silencing increases the uracil content in DNA of imaginal tissues and induces strong lethality at the early pupal stages, indicating that tolerance of highly uracil-substituted DNA is also stage-specific. Silencing of dUTPase perturbs the physiological pattern of uracil-DNA accumulation in Drosophila and leads to a strongly lethal phenotype in early pupal stages. These findings suggest a novel role of uracil-containing DNA in Drosophila development and metamorphosis and present a novel example for developmental effects of dUTPase silencing in multicellular eukaryotes. Importantly, we also show lack of the UNG gene in all available genomes of other Holometabola insects, indicating a potentially general tolerance and developmental role of uracil-DNA in this evolutionary clade.

PMID: 22685418 [PubMed - indexed for MEDLINE] PMCID: PMC3369950 **Free PMC Article**

Images from this publication. See all images (6) Free text

Save items
Add to Favorites

Related citations in PubMed
A novel fruitfly protein under developmental control d [Biochem Biophys Res Commun. 2007]
Effects of vaccinia virus uracil DNA glycosylase catalytic site and deoxyuridine tripl [Virol J. 2008]
Depletion of dimeric all-alpha dUTPase induces DNA strand break: [Int J Biochem Cell Biol. 2008]
Review The nature of enzymes involved in uracil-DNA repair: is [Curr Protein Pept Sci. 2001]
Review The role of dUTPase and uracil-DNA repair in cancer che [Curr Protein Pept Sci. 2001]
See reviews...
See all...

Related information
Related Citations
Compound (MeSH Keyword)
References for this PMC Article
Substance (MeSH Keyword)
Free in PMC

Recent activity
Turn Off Clear

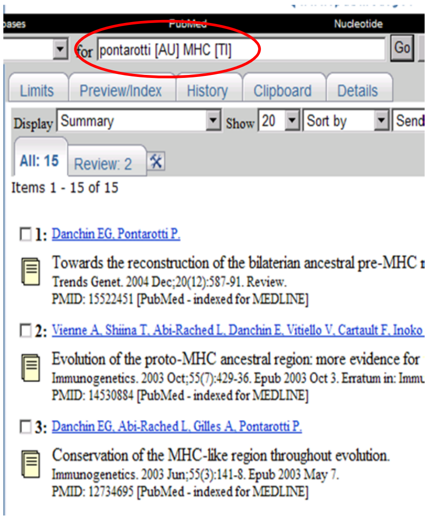
Searching PubMed Precisely

Restrict the search with fields

– [AU]	Author
– [SO]	Source (journal)
– [TI]	Title
– [AD]	Address
– [MH]	Keywords

Search Field Descriptions and Tags

- Affiliation [AD]
- Article Identifier [AID]
- All Fields [ALL]
- Author [AU]
- Book [book]
- Comment Corrections
- Corporate Author [CN]
- Create Date [CRDT]
- Completion Date [DCOM]
- EC/RN Number [RN]
- Editor [ED]
- Entrez Date [EDAT]
- Filter [FILTER]
- First Author Name [1AU]
- Full Author Name [FAU]
- Full Investigator Name [FIR]
- Grant Number [GR]
- Investigator [IR]
- ISBN [ISBN]
- Issue [IP]
- Journal [TA]
- Language [LA]
- Last Author [LASTAU]
- Location ID [LID]
- MeSH Date [MHDA]
- MeSH Major Topic [MAJR]
- MeSH Subheadings [SH]
- MeSH Terms [MH]
- Modification Date [LR]
- NLM Unique ID [JID]
- Other Term [OT]
- Owner
- Pagination [PG]
- Personal Name as Subject [PS]
- Pharmacological Action [PA]
- Place of Publication [PL]
- PMID [PMID]
- Publisher [PUBN]
- Publication Date [DP]
- Publication Type [PT]
- Secondary Source ID [SI]
- Subset [SB]
- Supplementary Concept[NM]
- Text Words [TW]
- Title [TI]
- Title/Abstract [TIAB]
- Transliterated Title [TT]
- UID [PMID]
- Version
- Volume [VI]



for pontarotti [AU] MHC [TI]

Items 1 - 15 of 15

☐ 1: Danchin EG, Pontarotti P.

Towards the reconstruction of the bilaterian ancestral pre-MHC r
Trends Genet. 2004 Dec;20(12):587-91. Review.
PMID: 15522451 [PubMed - indexed for MEDLINE]

☐ 2: Vienne A, Shiina T, Abi-Rached L, Danchin E, Vitiello V, Cartault F, Inoko

Evolution of the proto-MHC ancestral region: more evidence for
Immunogenetics. 2003 Oct;55(7):429-36. Epub 2003 Oct 3. Erratum in: Immu
PMID: 14530884 [PubMed - indexed for MEDLINE]

☐ 3: Danchin EG, Abi-Rached L, Gilles A, Pontarotti P.

Conservation of the MHC-like region throughout evolution.
Immunogenetics. 2003 Jun;55(3):141-8. Epub 2003 May 7.
PMID: 12734695 [PubMed - indexed for MEDLINE]

Tips for Searching PubMed

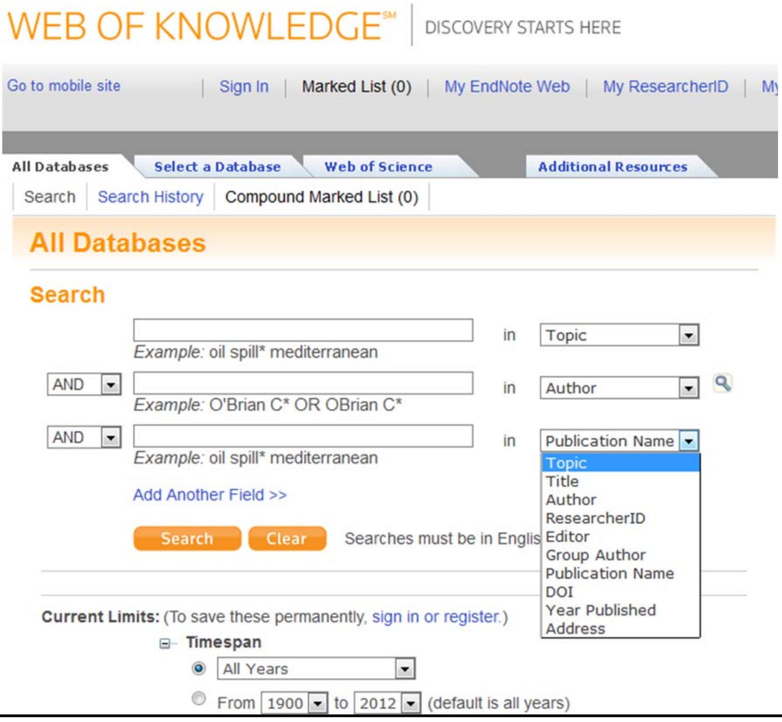
- Quoted queries like **“down syndrome”** behave as a single word, and can make the search more relevant.
- Use **OR** and **AND** to refine your queries
- Add the initials of the paper’s author, as in **Smith TF**
- Save the **PMID** of your papers
  - Very precise
  - Very short
- PubMed contains only papers published after 1965
- Use no more than 10 names for papers before 1995

Web of Knowledge

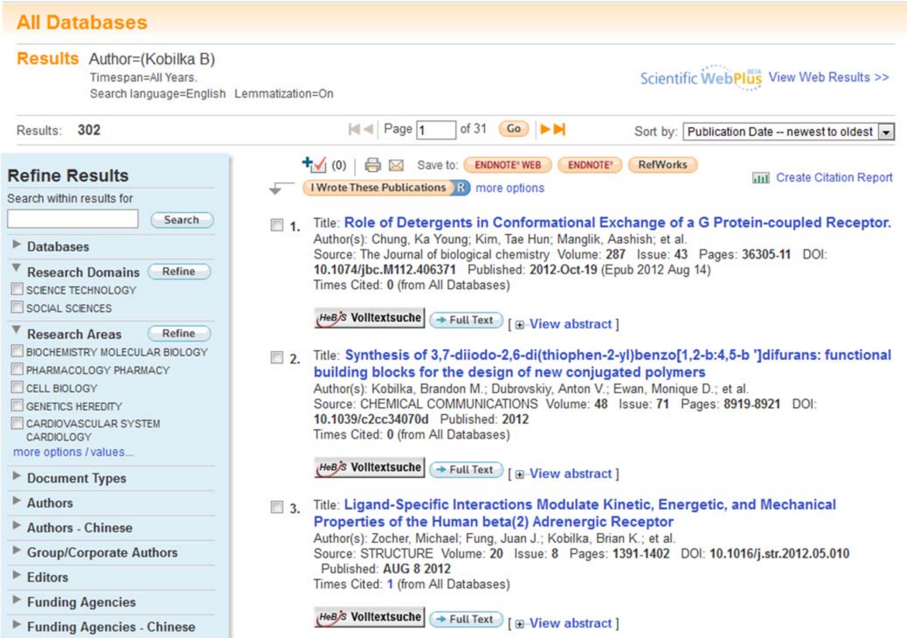
- Web of Knowledge is a database containing "all" scientific publications.
- Web of Knowledge includes citation information.
- Web of Knowledge includes links to the full text of publications.
- Web of Knowledge is not free but the Goethe University has a site license, and it is also accessible from home:
  http://apps.webofknowledge.com.ubproxy.ub.uni-frankfurt.de

---

http://apps.webofknowledge.com/



Web of Knowledge Search Results



Results Author=(Kobilka B)
Timespan=All Years.
Search language=English Lemmatization=On

Results: **302** Page 1 of 31 Sort by: Publication Date -- newest to oldest

Refine Results
Search within results for

- Databases
- Research Domains
  - ☐ SCIENCE TECHNOLOGY
  - ☐ SOCIAL SCIENCES
- Research Areas
  - ☐ BIOCHEMISTRY MOLECULAR BIOLOGY
  - ☐ PHARMACOLOGY PHARMACY
  - ☐ CELL BIOLOGY
  - ☐ GENETICS HEREDITY
  - ☐ CARDIOVASCULAR SYSTEM CARDIOLOGY
  - more options / values...
- Document Types
- Authors
- Authors - Chinese
- Group/Corporate Authors
- Editors
- Funding Agencies
- Funding Agencies - Chinese

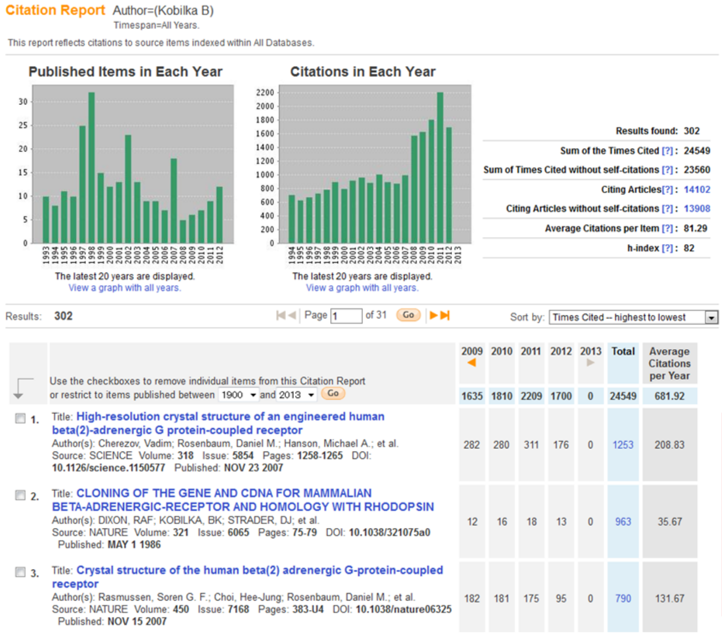
(0)

☐ 1. Title: **Role of Detergents in Conformational Exchange of a G Protein-coupled Receptor.**
Author(s): Chung, Ka Young; Kim, Tae Hun; Manglik, Aashish; et al.
Source: The Journal of biological chemistry Volume: **287** Issue: **43** Pages: **36305-11** DOI: **10.1074/jbc.M112.406371** Published: **2012-Oct-19** (Epub 2012 Aug 14)
Times Cited: **0** (from All Databases)
[View abstract]

☐ 2. Title: **Synthesis of 3,7-diiodo-2,6-di(thiophen-2-yl)benzo[1,2-b:4,5-b ']difurans: functional building blocks for the design of new conjugated polymers**
Author(s): Kobilka, Brandon M.; Dubrovskiy, Anton V.; Ewan, Monique D.; et al.
Source: CHEMICAL COMMUNICATIONS Volume: **48** Issue: **71** Pages: **8919-8921** DOI: **10.1039/c2cc34070d** Published: **2012**
Times Cited: **0** (from All Databases)
[View abstract]

☐ 3. Title: **Ligand-Specific Interactions Modulate Kinetic, Energetic, and Mechanical Properties of the Human beta(2) Adrenergic Receptor**
Author(s): Zocher, Michael; Fung, Juan J.; Kobilka, Brian K.; et al.
Source: STRUCTURE Volume: **20** Issue: **8** Pages: **1391-1402** DOI: **10.1016/j.str.2012.05.010**
Published: **AUG 8 2012**
Times Cited: **1** (from All Databases)
[View abstract]

Web of Knowledge Citation Report



Citation Report Author=(Kobilka B)
Timespan=All Years.
This report reflects citations to source items indexed within All Databases.

The latest 20 years are displayed.
View a graph with all years.

The latest 20 years are displayed.
View a graph with all years.

Results found: 302
Sum of the Times Cited [?]**:** 24549
Sum of Times Cited without self-citations [?]**:** 23560
Citing Articles[?]**:** 14102
Citing Articles without self-citations [?]**:** 13908
Average Citations per Item [?]**:** 81.29
h-index [?]**:** 82

Results: **302** Page 1 of 31 Sort by: Times Cited -- highest to lowest

	2009	2010	2011	2012	2013	Total	Average Citations per Year
Use the checkboxes to remove individual items from this Citation Report or restrict to items published between 1900 and 2013	1635	1810	2209	1700	0	24549	681.92
☐ 1. Title: **High-resolution crystal structure of an engineered human beta(2)-adrenergic G protein-coupled receptor** Author(s): Cherezov, Vadim; Rosenbaum, Daniel M.; Hanson, Michael A.; et al. Source: SCIENCE Volume: **318** Issue: **5854** Pages: **1258-1265** DOI: **10.1126/science.1150577** Published: **NOV 23 2007**	282	280	311	176	0	1253	208.83
☐ 2. Title: **CLONING OF THE GENE AND CDNA FOR MAMMALIAN BETA-ADRENERGIC-RECEPTOR AND HOMOLOGY WITH RHODOPSIN** Author(s): DIXON, RAF; KOBILKA, BK; STRADER, DJ; et al. Source: NATURE Volume: **321** Issue: **6065** Pages: **75-79** DOI: **10.1038/321075a0** Published: **MAY 1 1986**	12	16	18	13	0	963	35.67
☐ 3. Title: **Crystal structure of the human beta(2) adrenergic G-protein-coupled receptor** Author(s): Rasmussen, Soren G. F.; Choi, Hee-Jung; Rosenbaum, Daniel M.; et al. Source: NATURE Volume: **450** Issue: **7168** Pages: **383-U4** DOI: **10.1038/nature06325** Published: **NOV 15 2007**	182	181	175	95	0	790	131.67

Papers sorted by number of citations

h-Index

- Gegeben:
  n Publikationen, die $C_1 \ge C_2 \ge \ldots \ge C_n$ Mal zitiert werden
- Definition *h*-Index:
  h so, dass $C_h \ge h$ und $C_{h+1} \le h+1$
- Beispiel: $h = 20$ bedeutet: Es gibt 20 Publikationen, die 20 oder mehr Mal zitiert werden.
- Der *h*-Index wird hoch, wenn es viele Publikationen gibt, die viel zitiert werden.
- Der *h*-Index bleibt klein, wenn
  - viele Publikationen nur selten zitiert werden
  - nur einzelne Publikationen sehr oft zitiert werden

Web of Knowledge Publication Entry

Link to full text (via journal home page) →

Papers that cite this paper →

Papers cited by this paper →

High-resolution crystal structure of an engineered human beta(2)-adrenergic G protein-coupled receptor

Author(s): Cherezov, V (Cherezov, Vadim); Rosenbaum, DM (Rosenbaum, Daniel M.); Hanson, MA (Hanson, Michael A.); Rasmussen, SGF (Rasmussen, Soren G. F.); Thian, FS (Thian, Foon Sun); Kobilka, TS (Kobilka, Tong Sun); Choi, HJ (Choi, Hee-Jung); Kuhn, P (Kuhn, Peter); Weis, WI (Weis, William I.); Kobilka, BK (Kobilka, Brian K.); Stevens, RC (Stevens, Raymond C.)

Source: SCIENCE **Volume:** 318 **Issue:** 5854 **Pages:** 1258-1265 **DOI:** 10.1126/science.1150577 **Published:** NOV 23 2007

Times Cited: 1,229 (from Web of Science)

Cited References: 67 [view related records]

Abstract: Heterotrimeric guanine nucleotide-binding protein (G protein)-coupled receptors constitute the largest family of eukaryotic signal transduction proteins that communicate across the membrane. We report the crystal structure of a human beta(2)-adrenergic receptor-T4 lysozyme fusion protein bound to the partial inverse agonist carazolol at 2.4 angstrom resolution. The structure provides a high-resolution view of a human G protein-coupled receptor bound to a diffusible ligand. Ligand-binding site accessibility is enabled by the second extracellular loop, which is held out of the binding cavity by a pair of closely spaced disulfide bridges and a short helical segment within the loop. Cholesterol, a necessary component for crystallization, mediates an intriguing parallel association of receptor molecules in the crystal lattice. Although the location of carazolol in the beta(2)-adrenergic receptor is very similar to that of retinal in rhodopsin, structural differences in the ligand-binding site and other regions highlight the challenges in using rhodopsin as a template model for this large receptor family.

Accession Number: WOS:000251086600033

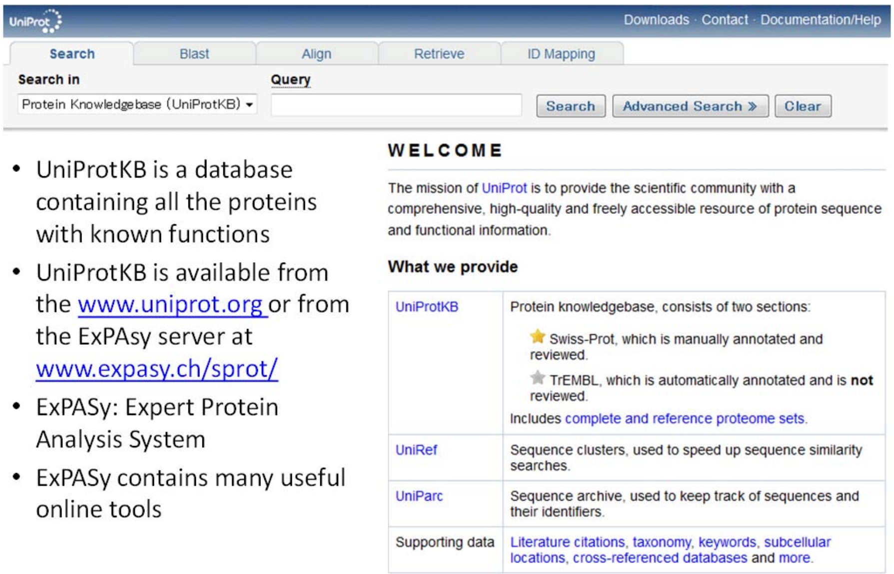
Document Type: Article

Language: English

KeyWords Plus: BETA-ADRENERGIC-RECEPTOR; HUMAN BETA-2-ADRENERGIC RECEPTOR; HUMAN GENOME; MOLECULAR-MECHANISM; LIGAND-BINDING; RHODOPSIN; ACTIVATION; CRYSTALLIZATION; MEMBRANE; AGONIST

Reprint Address: Stevens, RC (reprint author), Stanford Univ, Sch Med, Dept Cellular & Mol Physiol, Stanford, CA 94305 USA.

Retrieving Protein Sequences in UniProtKB



WELCOME

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.

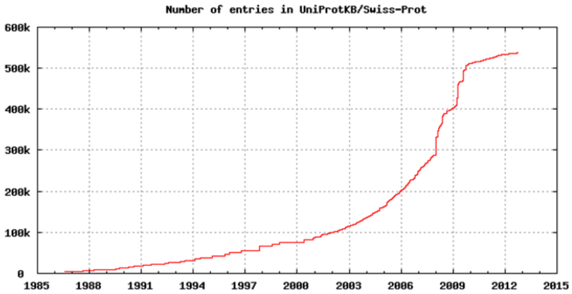
What we provide

UniProtKB	Protein knowledgebase, consists 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.

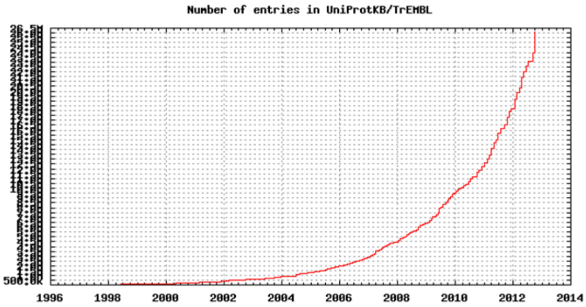
- UniProtKB is a database containing all the proteins with known functions
- UniProtKB is available from the www.uniprot.org or from the ExPAsy server at www.expasy.ch/sprot/
- ExPASy: Expert Protein Analysis System
- ExPASy contains many useful online tools

UniProtKB Statistics

- **UniProtKB/Swiss-Prot**
  - manually annotated and reviewed
  - 538,010 sequences
  - 190,998,508 a.a.



- **UniProtKB/TrEMBL**
  - automatically annotated and not reviewed
  - 26,079,526 sequences
  - 8,448,404,066 a.a.



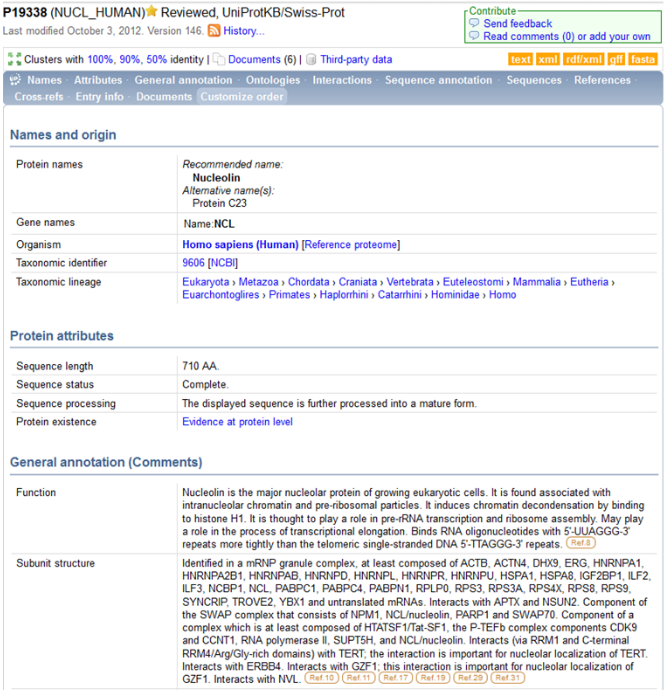
03.10.2012

The UniProtKB Entry

- Each UniProtKB entry is dedicated to a protein
- A UniProtKB entry summarizes everything that is known about a given protein
- The entry contains functional information and links to other databases mentioning this protein

Typical UniProtKB Entry

- Protein name
- Protein function
- Bibliography
- Links to other databases
  - Structure
  - Domains
  - Function



P19338 (NUCL_HUMAN) Reviewed, UniProtKB/Swiss-Prot

Last modified October 3, 2012. Version 146. History...

Clusters with 100%, 90%, 50% identity | Documents (6) | Third-party data

Names and origin

Protein names	*Recommended name:* **Nucleolin** *Alternative name(s):* Protein C23
Gene names	Name:**NCL**
Organism	**Homo sapiens (Human)** [Reference proteome]
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo

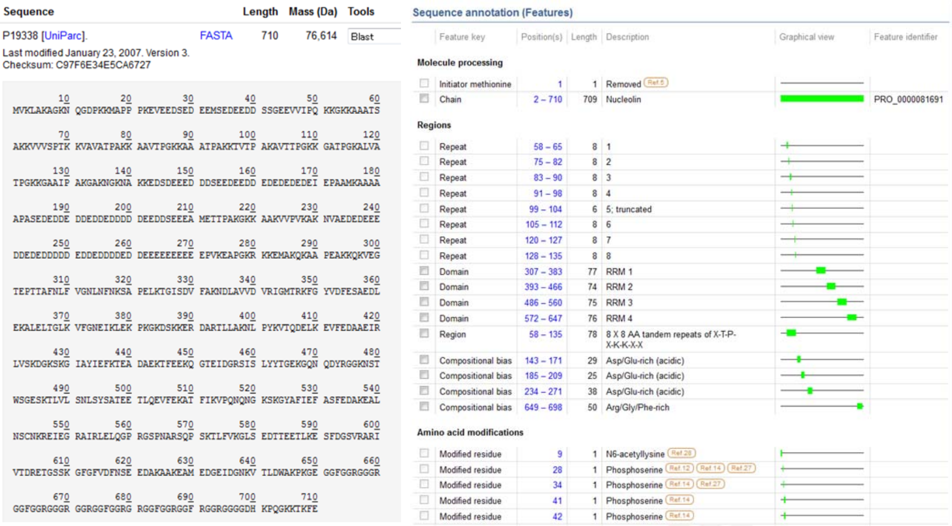
Protein attributes

Sequence length	710 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is further processed into a mature form.
Protein existence	Evidence at protein level

General annotation (Comments)

Function	Nucleolin is the major nucleolar protein of growing eukaryotic cells. It is found associated with intranucleolar chromatin and pre-ribosomal particles. It induces chromatin decondensation by binding to histone H1. It is thought to play a role in pre-rRNA transcription and ribosome assembly. May play a role in the process of transcriptional elongation. Binds RNA oligonucleotides with 5'-UUAGGG-3' repeats more tightly than the telomeric single-stranded DNA 5'-TTAGGG-3' repeats. Ref.8
Subunit structure	Identified in a mRNP granule complex, at least composed of ACTB, ACTN4, DHX9, ERG, HNRNPA1, HNRNPA2B1, HNRNPAB, HNRNPD, HNRNPL, HNRNPR, HNRNPU, HSPA1, HSPA8, IGF2BP1, ILF2, ILF3, NCBP1, NCL, PABPC1, PABPC4, PABPN1, RPLP0, RPS3, RPS3A, RPS4X, RPS8, RPS9, SYNCRIP, TROVE2, YBX1 and untranslated mRNAs. Interacts with APTX and NSUN2. Component of the SWAP complex that consists of NPM1, NCL/nucleolin, PARP1 and SWAP70. Component of a complex which is at least composed of HTATSF1/Tat-SF1, the P-TEFb complex components CDK9 and CCNT1, RNA polymerase II, SUPT5H, and NCL/nucleolin. Interacts (via RRM1 and C-terminal RRM4/Arg/Gly-rich domains) with TERT; the interaction is important for nucleolar localization of TERT. Interacts with ERBB4. Interacts with GZF1; this interaction is important for nucleolar localization of GZF1. Interacts with NVL. Ref.10 Ref.11 Ref.17 Ref.19 Ref.29 Ref.31

Sequence and Sequence annotation in a UniProtKB Entry



Sequence		Length	Mass (Da)	Tools
P19338 [UniParc].	FASTA	710	76,614	Blast

Last modified January 23, 2007. Version 3.
Checksum: C97F6E34E5CA6727

```
        10         20         30         40         50         60
MVKLAKAGKN QGDPKKMAPP PKEVEEDSED EEMSEDEEDD SSGEEVVIPQ KKGKKAAATS

        70         80         90        100        110        120
AKKVVVSPTK KVAVATPAKK AAVTPGKKAA ATPAKKTVTP AKAVTTPGKK GATPGKALVA

       130        140        150        160        170        180
TPGKKGAAIP AKGAKNGKNA KKEDSDEEED DDSEEDEEDD EDEDEDEDEI EPAAMKAAAA

       190        200        210        220        230        240
APASEDEDDE DDEDDEDDDD DEEDDSEEEA METTPAKGKK AAKVVPVKAK NVAEDEDEEE

       250        260        270        280        290        300
DDEDEDDDDD EDDEDDDDED DEEEEEEEEE EPVKEAPGKR KKEMAKQKAA PEAKKQKVEG

       310        320        330        340        350        360
TEPTTAFNLF VGNLNFNKSA PELKTGISDV FAKNDLAVVD VRIGMTRKFG YVDFESAEDL

       370        380        390        400        410        420
EKALELTGLK VFGNEIKLEK PKGKDSKKER DARTLLAKNL PYKVTQDELK EVFEDAAEIR

       430        440        450        460        470        480
LVSKDGKSKG IAYIEFKTEA DAEKTFEEKQ GTEIDGRSIS LYYTGEKGQN QDYRGGKNST

       490        500        510        520        530        540
WSGESKTLVL SNLSYSATEE TLQEVFEKAT FIKVPQNQNG KSKGYAFIEF ASFEDAKEAL

       550        560        570        580        590        600
NSCNKREIEG RAIRLELQGP RGSPNARSQP SKTLFVKGLS EDTTEETLKE SFDGSVRARI

       610        620        630        640        650        660
VTDRETGSSK GFGFVDFNSE EDAKAAKEAM EDGEIDGNKV TLDWAKPKGE GGFGGRGGGR

       670        680        690        700        710
GGFGGRGGGR GGRGGFGGRG RGGFGGRGGF RGGRGGGGDH KPQGKKTKFE
```

Sequence annotation (Features)

	Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Molecule processing						
☐	Initiator methionine	1	1	Removed Ref.5		
▨	Chain	2 – 710	709	Nucleolin		PRO_0000081691
Regions						
☐	Repeat	58 – 65	8	1		
☐	Repeat	75 – 82	8	2		
☐	Repeat	83 – 90	8	3		
☐	Repeat	91 – 98	8	4		
☐	Repeat	99 – 104	6	5; truncated		
☐	Repeat	105 – 112	8	6		
☐	Repeat	120 – 127	8	7		
☐	Repeat	128 – 135	8	8		
▨	Domain	307 – 383	77	RRM 1		
▨	Domain	393 – 466	74	RRM 2		
▨	Domain	486 – 560	75	RRM 3		
▨	Domain	572 – 647	76	RRM 4		
▨	Region	58 – 135	78	8 X 8 AA tandem repeats of X-T-P-X-K-K-X-X		
▨	Compositional bias	143 – 171	29	Asp/Glu-rich (acidic)		
▨	Compositional bias	185 – 209	25	Asp/Glu-rich (acidic)		
▨	Compositional bias	234 – 271	38	Asp/Glu-rich (acidic)		
▨	Compositional bias	649 – 698	50	Arg/Gly/Phe-rich		
Amino acid modifications						
☐	Modified residue	9	1	N6-acetyllysine Ref.28		
☐	Modified residue	28	1	Phosphoserine Ref.12 Ref.14 Ref.27		
☐	Modified residue	34	1	Phosphoserine Ref.14 Ref.27		
☐	Modified residue	41	1	Phosphoserine Ref.14		
☐	Modified residue	42	1	Phosphoserine Ref.14		

FASTA format for sequences

```
>sp|P19338|NUCL_HUMAN Nucleolin OS=Homo sapiens GN=NCL PE=1 SV=3
MVKLAKAGKNQGDPKKMAPPPKEVEEDSEDEEMSEDEEDDSSGEEVVIPQKKGKKAAATS
AKKVVVSPTKKVAVATPAKKAAVTPGKKAAATPAKKTVTPAKAVTTPGKKGATPGKALVA
TPGKKGAAIPAKGAKNGKNAKKEDSDEEEDDDSEEDEEDDEDEDEDEDEIEPAAMKAAAA
APASEDEDDEDDEDDEDDDDDEEDDSEEEAMETTPAKGKKAAKVVPVKAKNVAEDEDEEE
DDEDEDDDDDEDDEDDDDEDDEEEEEEEEEEPVKEAPGKRKKEMAKQKAAPEAKKQKVEG
TEPTTAFNLFVGNLNFNKSAPELKTGISDVFAKNDLAVVDVRIGMTRKFGYVDFESAEDL
EKALELTGLKVFGNEIKLEKPKGKDSKKERDARTLLAKNLPYKVTQDELKEVFEDAAEIR
LVSKDGKSKGIAYIEFKTEADAEKTFEEKQGTEIDGRSISLYYTGEKGQNQDYRGGKNST
WSGESKTLVLSNLSYSATEETLQEVFEKATFIKVPQNQNGKSKGYAFIEFASFEDAKEAL
NSCNKREIEGRAIRLELQGPRGSPNARSQPSKTLFVKGLSEDTTEETLKESFDGSVRARI
VTDRETGSSKGFGFVDFNSEEDAKAAKEAMEDGEIDGNKVTLDWAKPKGEGGFGGRGGGR
GGFGGRGGGRGGRGGFGGRGRGGFGGRGGFRGGRGGGGDHKPQGKKTKFE
```

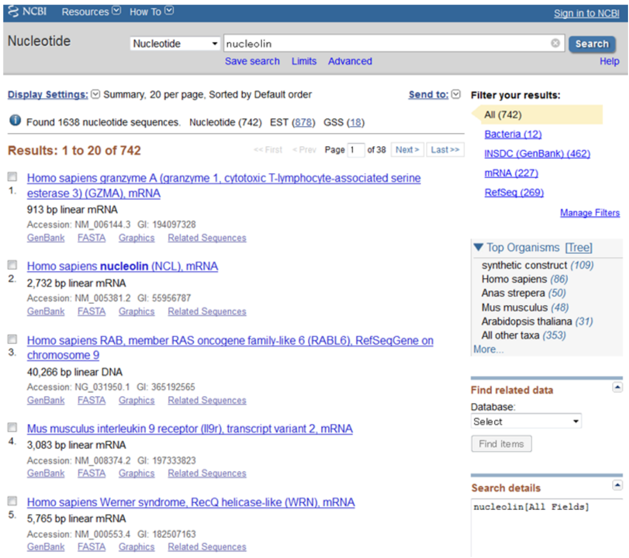
- First line starts with ‘>’ and is a comment
  (some programs need the sequence without comment line)
- Following lines are the sequence in one-letter amino acid codes

Looking for DNA Sequences

- There are many types of DNA sequences
- The most common are
  - Regulatory regions, often before genes
  - Untranslated regions, often around the genes
  - Protein-coding regions
  - Intergenic regions (between the genes)
- All these sequences can be found in GenBank

Fetching a DNA Sequence at the NCBI

- Navigate to www.ncbi.nlm.nih.gov/Genbank/
- Type in a keyword.
- You get a list of entries matching your keyword.
- Point, click, and explore…



Summary, 20 per page, Sorted by Default order

Found 1638 nucleotide sequences. Nucleotide (742) EST (878) GSS (18)

Results: 1 to 20 of 742

1. Homo sapiens granzyme A (granzyme 1, cytotoxic T-lymphocyte-associated serine esterase 3) (GZMA), mRNA
   913 bp linear mRNA
   Accession: NM_006144.3 GI: 194097328
2. Homo sapiens **nucleolin** (NCL), mRNA
   2,732 bp linear mRNA
   Accession: NM_005381.2 GI: 55956787
3. Homo sapiens RAB, member RAS oncogene family-like 6 (RABL6), RefSeqGene on chromosome 9
   40,266 bp linear DNA
   Accession: NG_031950.1 GI: 365192565
4. Mus musculus interleukin 9 receptor (Il9r), transcript variant 2, mRNA
   3,083 bp linear mRNA
   Accession: NM_008374.2 GI: 197333823
5. Homo sapiens Werner syndrome, RecQ helicase-like (WRN), mRNA
   5,765 bp linear mRNA
   Accession: NM_000553.4 GI: 182507163

Filter your results:
- All (742)
- Bacteria (12)
- INSDC (GenBank) (462)
- mRNA (227)
- RefSeq (269)

Top Organisms [Tree]
- synthetic construct *(109)*
- Homo sapiens *(86)*
- Anas strepera *(50)*
- Mus musculus *(48)*
- Arabidopsis thaliana *(31)*
- All other taxa *(353)*

More...

Find related data

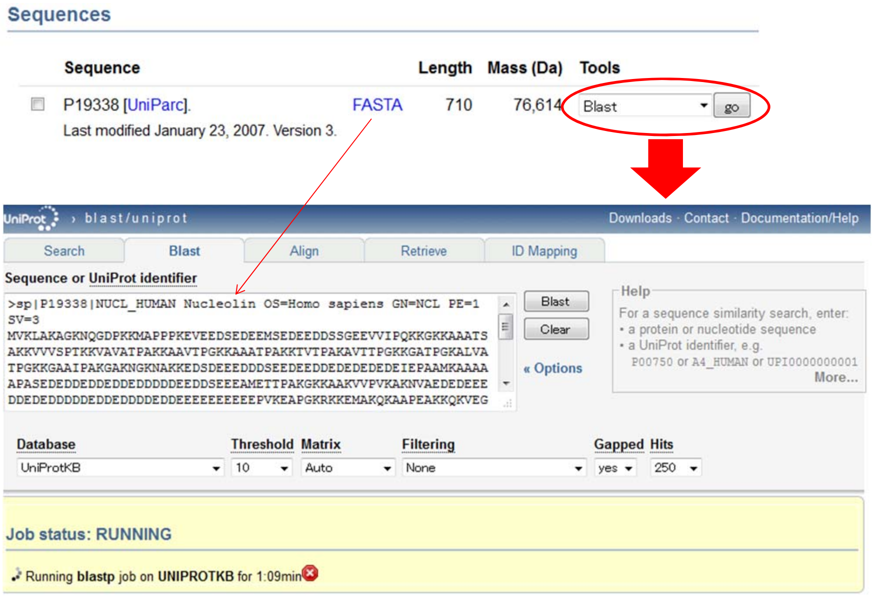
Search details

`nucleolin[All Fields]`

Searching for Your Sequence with BLAST

- BLAST: Basic Local Alignment Search Tools
- Compares your sequence with all other sequences in your favorite database
- Returns the most similar sequences

Start BLAST search from UniProtKB entry



Sequences

	Sequence		Length	Mass (Da)	Tools
☐	P19338 [UniParc]. Last modified January 23, 2007. Version 3.	FASTA	710	76,614	Blast ▾ go

Sequence or UniProt identifier

```
>sp|P19338|NUCL_HUMAN Nucleolin OS=Homo sapiens GN=NCL PE=1
SV=3
MVKLAKAGKNQGDPKKMAPPPKEVEEDSEDEEMSEDEEDDSSGEEVVIPQKKGKKAAATS
AKKVVVSPTKKVAVATPAKKAAVTPGKKAAATPAKKTVTPAKAVTTPGKKGATPGKALVA
TPGKKGAAIPAKGAKNGKNAKKEDSDEEEDDDSEEDEEDDEDEDEDEDEIEPAAMKAAAA
APASEDEDDEDDEDDEDDDDDEEDDSEEEAMETTPAKGKKAAKVVPVKAKNVAEDEDEEE
DDEDEDDDDDEDDEDDDDEDDEEEEEEEEEEEPVKEAPGKRKKEMAKQKAAPEAKKQKVEG
```

Help
For a sequence similarity search, enter:
- a protein or nucleotide sequence
- a UniProt identifier, e.g. `P00750` or `A4_HUMAN` or `UPI0000000001`

Database	Threshold	Matrix	Filtering	Gapped	Hits
UniProtKB	10	Auto	None	yes	250

Job status: RUNNING

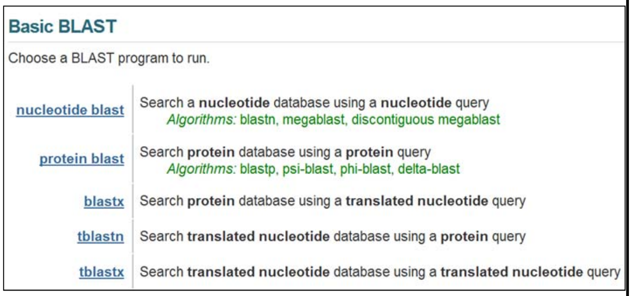
Running **blastp** job on **UNIPROTKB** for 1:09min

- Wait for ~4 minutes…

Choosing Your BLAST Program

- Navigate to the NCBI BLAST site at www.ncbi.nlm.nih.gov/blast

- Select your BLAST program:
  - Blastn DNA
  - Blastp Protein



Basic BLAST

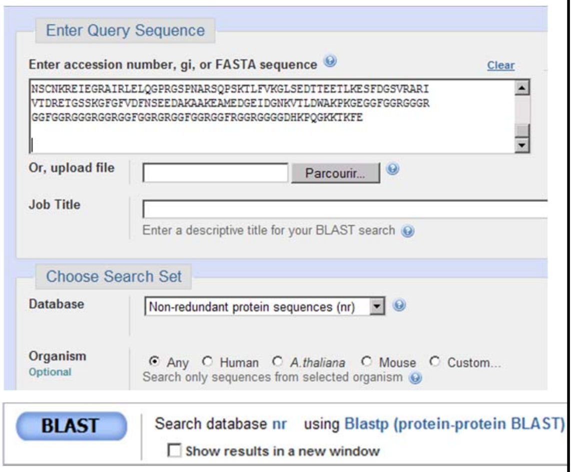
Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

---

Blasting a Protein Sequence

- Cut and paste your sequence
- Click the BLAST button at the bottom of the screen
- Wait



Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
NSCNKREIEGRAIRLELQGPRGSPNARSQPSKTLFVKGLSEDTTEETLKESFDGSVRARI
VTDRETGSSKGFGFVDFNSEEDAKAAKEAMEDGEIDGNKVTLDWAKPKGEGGFGGRGGGR
GGFGGRGGGRGGRGGFGGRGRGGFGGRGGFRGGRGGGGDHKPQGKKTKFE
```

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set

Database: Non-redundant protein sequences (nr)

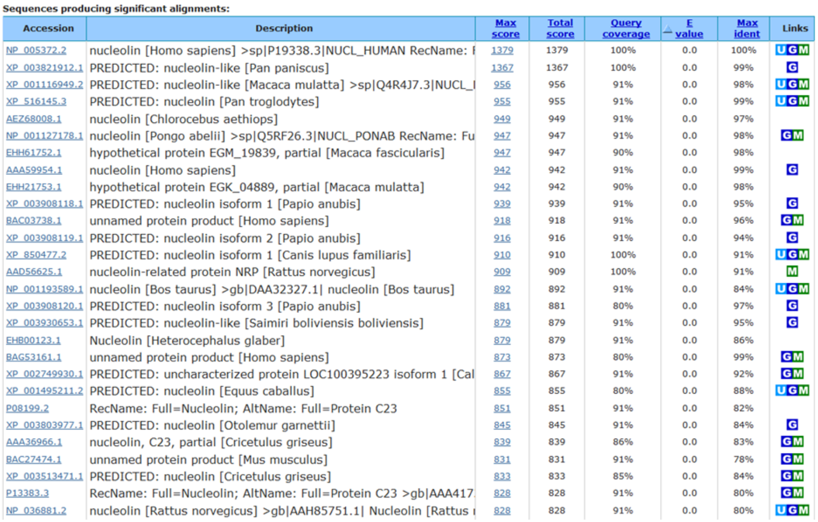
Organism (Optional): ● Any ○ Human ○ A.thaliana ○ Mouse ○ Custom...

Search only sequences from selected organism

Search database nr using Blastp (protein-protein BLAST)

☐ Show results in a new window

Reading a BLAST Output



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NP_005372.2	nucleolin [Homo sapiens] >sp|P19338.3|NUCL_HUMAN RecName: F	1379	1379	100%	0.0	100%	UGM
XP_003821912.1	PREDICTED: nucleolin-like [Pan paniscus]	1367	1367	100%	0.0	99%	G
XP_001116949.2	PREDICTED: nucleolin-like [Macaca mulatta] >sp|Q4R4J7.3|NUCL_I	956	956	91%	0.0	98%	UGM
XP_516145.3	PREDICTED: nucleolin [Pan troglodytes]	955	955	91%	0.0	99%	UGM
AEZ68008.1	nucleolin [Chlorocebus aethiops]	949	949	91%	0.0	97%	
NP_001127178.1	nucleolin [Pongo abelii] >sp|Q5RF26.3|NUCL_PONAB RecName: Fu	947	947	91%	0.0	98%	GM
EHH61752.1	hypothetical protein EGM_19839, partial [Macaca fascicularis]	947	947	90%	0.0	98%	
AAA59954.1	nucleolin [Homo sapiens]	942	942	91%	0.0	99%	G
EHH21753.1	hypothetical protein EGK_04889, partial [Macaca mulatta]	942	942	90%	0.0	98%	
XP_003908118.1	PREDICTED: nucleolin isoform 1 [Papio anubis]	939	939	91%	0.0	95%	G
BAC03738.1	unnamed protein product [Homo sapiens]	918	918	91%	0.0	96%	GM
XP_003908119.1	PREDICTED: nucleolin isoform 2 [Papio anubis]	916	916	91%	0.0	94%	G
XP_850477.2	PREDICTED: nucleolin isoform 1 [Canis lupus familiaris]	910	910	100%	0.0	91%	UGM
AAD56625.1	nucleolin-related protein NRP [Rattus norvegicus]	909	909	100%	0.0	91%	M
NP_001193589.1	nucleolin [Bos taurus] >gb|DAA32327.1| nucleolin [Bos taurus]	892	892	91%	0.0	84%	UGM
XP_003908120.1	PREDICTED: nucleolin isoform 3 [Papio anubis]	881	881	80%	0.0	97%	G
XP_003930653.1	PREDICTED: nucleolin-like [Saimiri boliviensis boliviensis]	879	879	91%	0.0	95%	G
EHB00123.1	Nucleolin [Heterocephalus glaber]	879	879	91%	0.0	86%	
BAG53161.1	unnamed protein product [Homo sapiens]	873	873	80%	0.0	99%	GM
XP_002749930.1	PREDICTED: uncharacterized protein LOC100395223 isoform 1 [Cal	867	867	91%	0.0	92%	GM
XP_001495211.2	PREDICTED: nucleolin [Equus caballus]	855	855	80%	0.0	88%	UGM
P08199.2	RecName: Full=Nucleolin; AltName: Full=Protein C23	851	851	91%	0.0	82%	
XP_003803977.1	PREDICTED: nucleolin [Otolemur garnettii]	845	845	91%	0.0	84%	G
AAA36966.1	nucleolin, C23, partial [Cricetulus griseus]	839	839	86%	0.0	83%	GM
BAC27474.1	unnamed protein product [Mus musculus]	831	831	91%	0.0	78%	GM
XP_003513471.1	PREDICTED: nucleolin [Cricetulus griseus]	833	833	85%	0.0	84%	GM
P13383.3	RecName: Full=Nucleolin; AltName: Full=Protein C23 >gb|AAA417	828	828	91%	0.0	80%	GM
NP_036881.2	nucleolin [Rattus norvegicus] >gb|AAH85751.1| Nucleolin [Rattus r	828	828	91%	0.0	80%	UGM

- Every line is a **hit**
- Best hits come first
- **Low E-Value = good hit**
- E-value >1 = bad hit
- **G** means a link onto a complete genome
- **U** means a link to UniGene, the transcript database

Multiple-Sequence Alignment (MSA)

- Multiple alignments reveal common features between sequences
- Multiple alignments are useful for
  - Comparing very different sequences
  - Making phylogenetic trees
  - Making structure predictions
- Multiple-sequence alignment is abbreviated as MSA

Making a Multiple-Sequence Alignment

- Identify a set of related sequences
- Do a BLAST of your favorite sequence
- Choose a method:

– www.ebi.ac.uk/clustalw	ClustalW	Popular
– www.tcoffee.org	T-Coffee	Accurate
– www.drive5.com/muscle/	Muscle	Fast
– www.tcoffee.org	M-Coffee	Consensus

Making an MSA with M-Coffee

- Open www.tcoffee.org
- Click MCoffee
- Cut and paste your sequences
- Submit your MSA

M-Coffee

Aligns DNA, RNA or Proteins by combining the output of popular aligners

Sequences input

Paste or upload your set of sequences in FASTA format

Sequences to align

Click here to use the sample file

- OR - Click here to upload a file

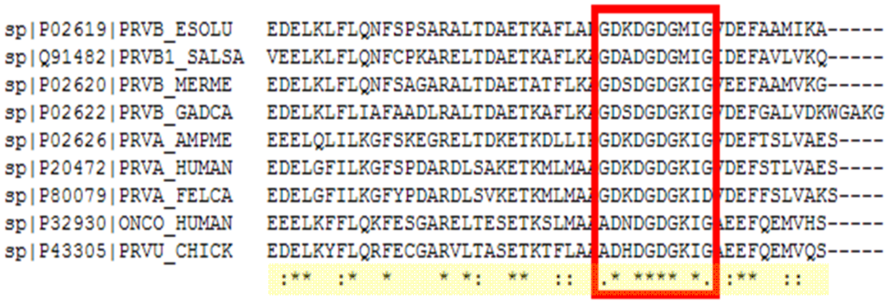
Show more options

Your email address

T-Coffee Server is hosted by the **Centre for Genomic Regulation** (CRG) of Barcelona

Making Sense of Your MSA

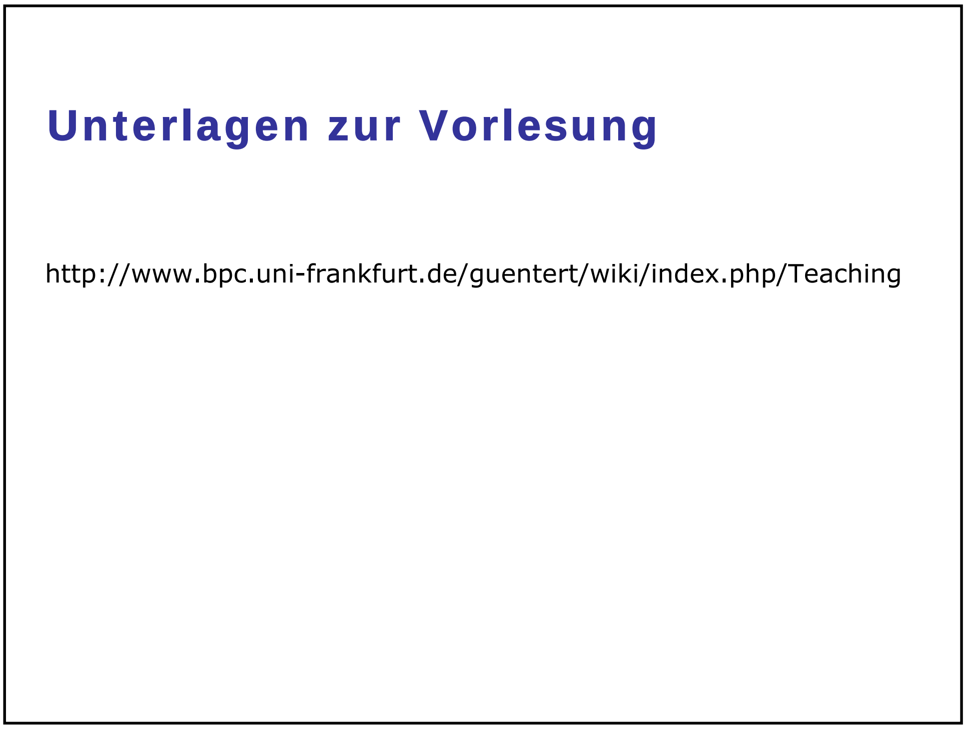
- Positions are marked:
  - Completely conserved = asterisk (*)
  - Highly conserved = colon (:)
  - Conserved = period (.)
- Look for highly conserved blocks:
  - The red box on this slide shows a highly conserved block.
  - These blocks are often functionally important positions.



```
sp|P02619|PRVB_ESOLU   EDELKLFLQNFSPSARALTDAETKAFLAIGDKDGDGMIGVDEFAAMIKA-----
sp|Q91482|PRVB1_SALSA  VEELKLFLQNFCPKARELTDAETKAFLKAGDADGDGMIGIDEFAVLVKQ-----
sp|P02620|PRVB_MERME   EDELKLFLQNFSAGARALTDAETATFLKAGDSDGDGKIGVEEFAAMVKG-----
sp|P02622|PRVB_GADCA   EDELKLFLIAFAADLRALTDAETKAFLKAGDSDGDGKIGVDEFGALVDKWGAKG
sp|P02626|PRVA_AMPME   EEELQLILKGFSKEGRELTDKETKDLLIIGDKDGDGKIGVDEFTSLVAES----
sp|P20472|PRVA_HUMAN   EDELGFILKGFSPDARDLSAKETKMLMAAGDKDGDGKIGVDEFSTLVAES----
sp|P80079|PRVA_FELCA   EDELGFILKGFYPDARDLSVKETKMLMAAGDKDGDGKIDVDEFFSLVAKS----
sp|P32930|ONCO_HUMAN   EEELKFFLQKFESGARELTESETKSLMAAADNDGDGKIGAEEFQEMVHS-----
sp|P43305|PRVU_CHICK   EDELKYFLQRFECGARVLTASETKTFLAAADHDGDGKIGAEEFQEMVQS-----
                        :**  :*  *    * *:  **  ::  .* **** *. :**  ::
```

Going Farther

- Bioinformatics is all about getting knowledge without having to make real-world experiments.
- More details in later chapters:
  - Databases: Chapters 3 and 4
  - BLAST: Chapter 7
  - Multiple-sequence alignments: Chapters 9 and 10



Unterlagen zur Vorlesung

http://www.bpc.uni-frankfurt.de/guentert/wiki/index.php/Teaching