Youth Bioinformatics Symposium 2016

Cecilia Arighi Hongzhan Huang
Karen Ross C.R. Vinayaka
Cathy Wu

Protein Information Resource
Georgetown University Medical Center
University of Delaware

Outline

I. Historical Background

II. Searching for Protein Information in “Free-Text” Resources

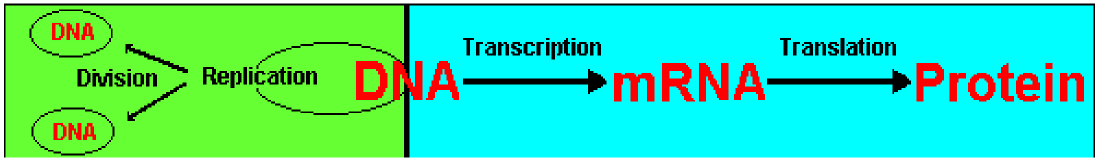
III. The UniProtKB database

IV. Protein Sequence Similarity Search

V. Multiple Sequence Alignment (MSA)

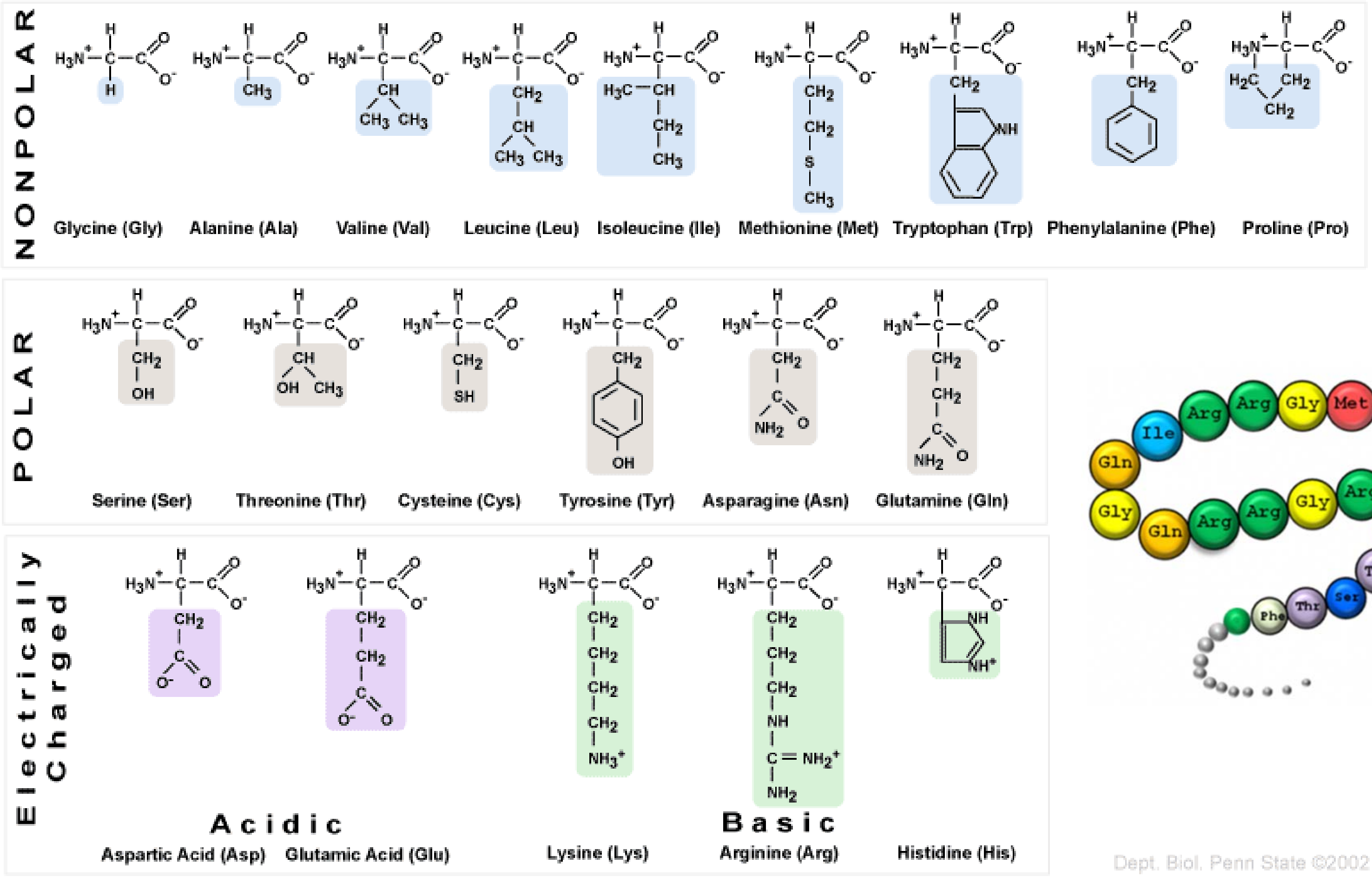
I. Historical Background

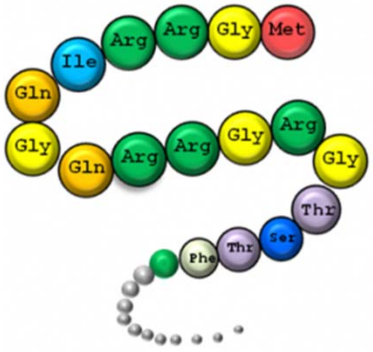
The Central Dogma of Modern Biology



www.worldofteaching.com

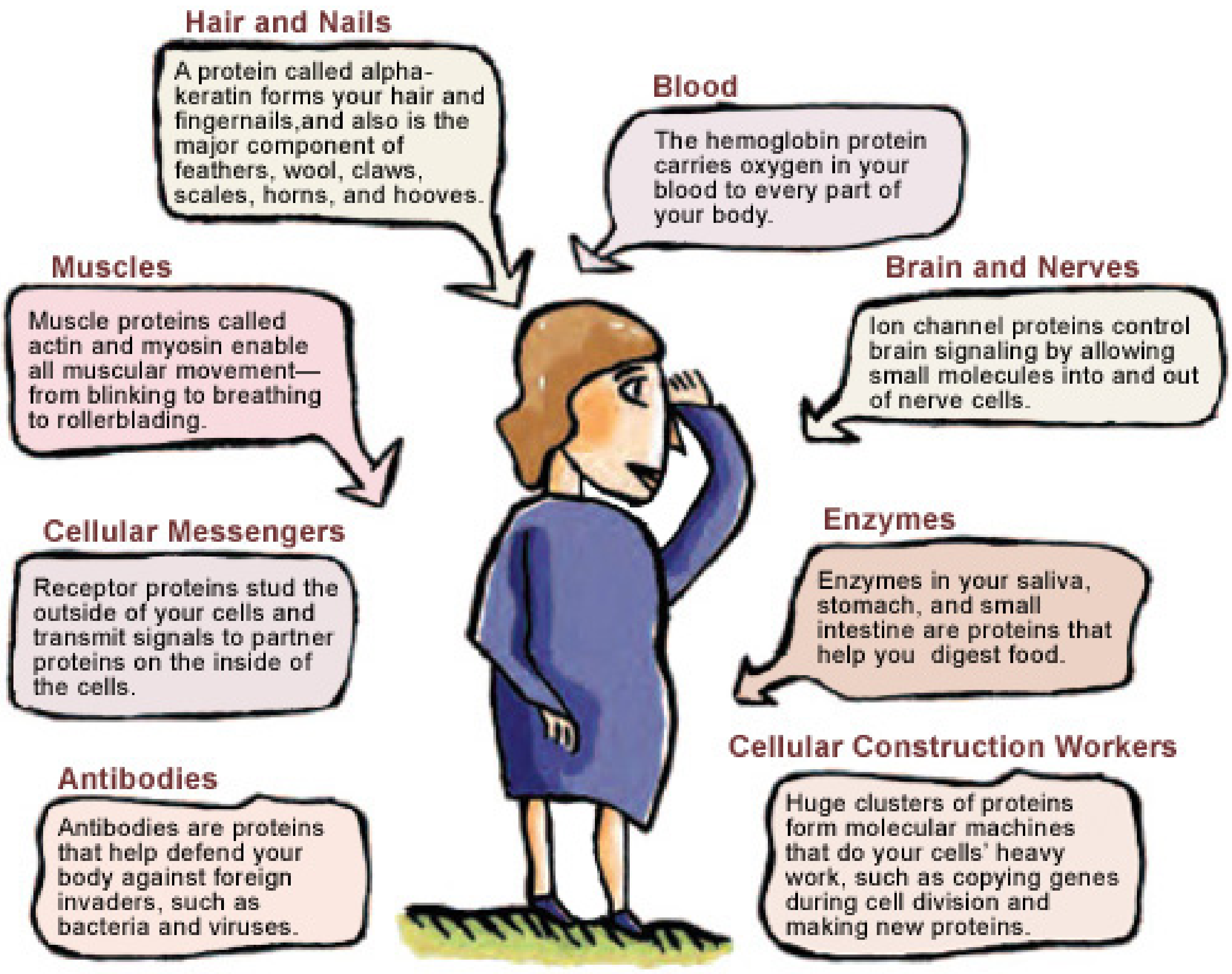
- Proteins are composed of chains of amino acids
- Size and chemical properties of amino acids vary





Dept. Biol. Penn State ©2002

What Do Proteins Do?



Hair and Nails

A protein called alpha-keratin forms your hair and fingernails,and also is the major component of feathers, wool, claws, scales, horns, and hooves.

Blood

The hemoglobin protein carries oxygen in your blood to every part of your body.

Muscles

Muscle proteins called actin and myosin enable all muscular movement—from blinking to breathing to rollerblading.

Brain and Nerves

Ion channel proteins control brain signaling by allowing small molecules into and out of nerve cells.

Cellular Messengers

Receptor proteins stud the outside of your cells and transmit signals to partner proteins on the inside of the cells.

Enzymes

Enzymes in your saliva, stomach, and small intestine are proteins that help you digest food.

Antibodies

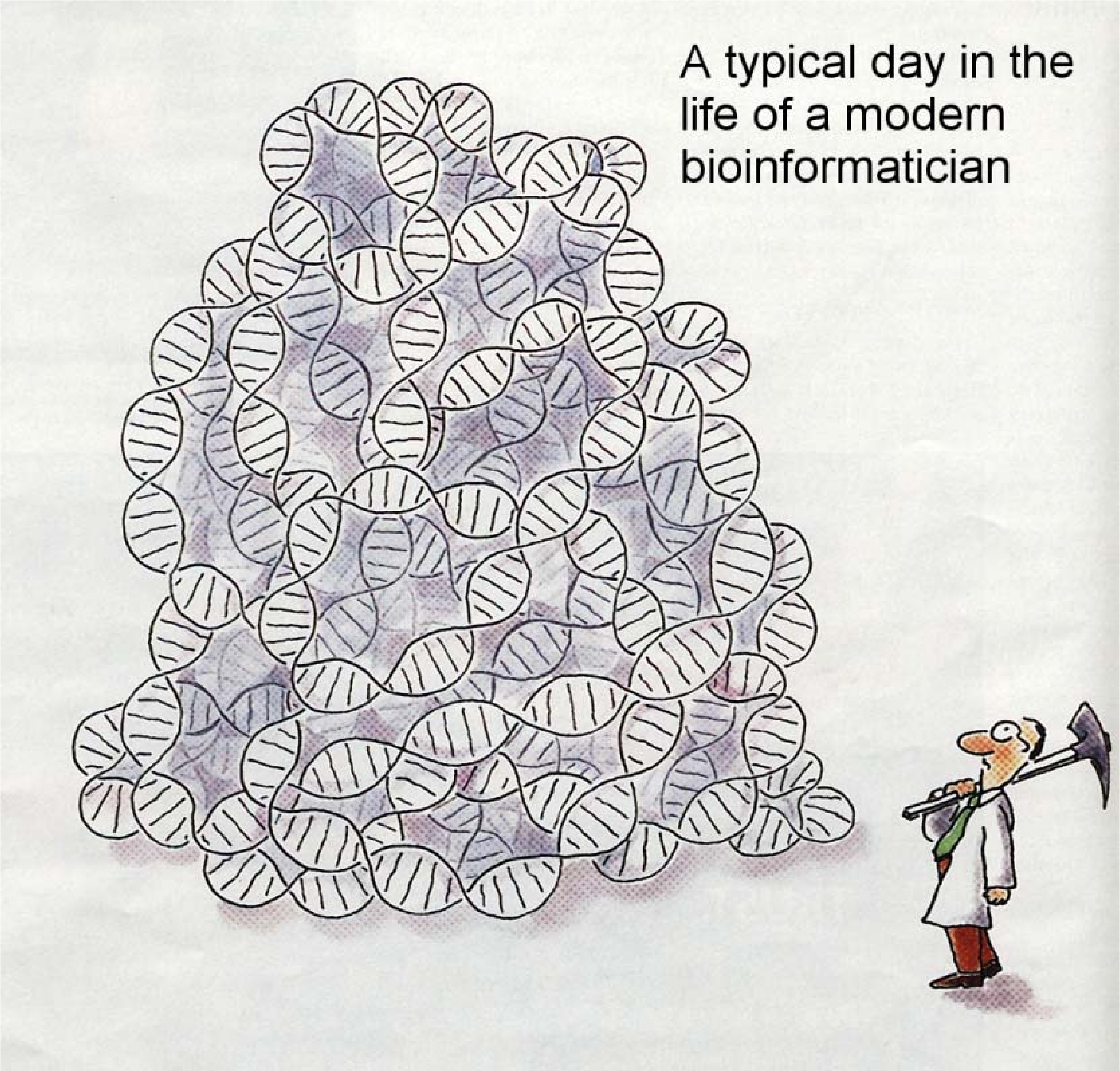
Antibodies are proteins that help defend your body against foreign invaders, such as bacteria and viruses.

Cellular Construction Workers

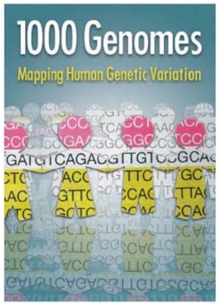
Huge clusters of proteins form molecular machines that do your cells' heavy work, such as copying genes during cell division and making new proteins.

https://publications.nigms.nih.gov/structlife/chapter1.html

Deluge of sequence data



Courtesy of the Swiss-Prot group (SIB Swiss Institute of Bioinformatics)



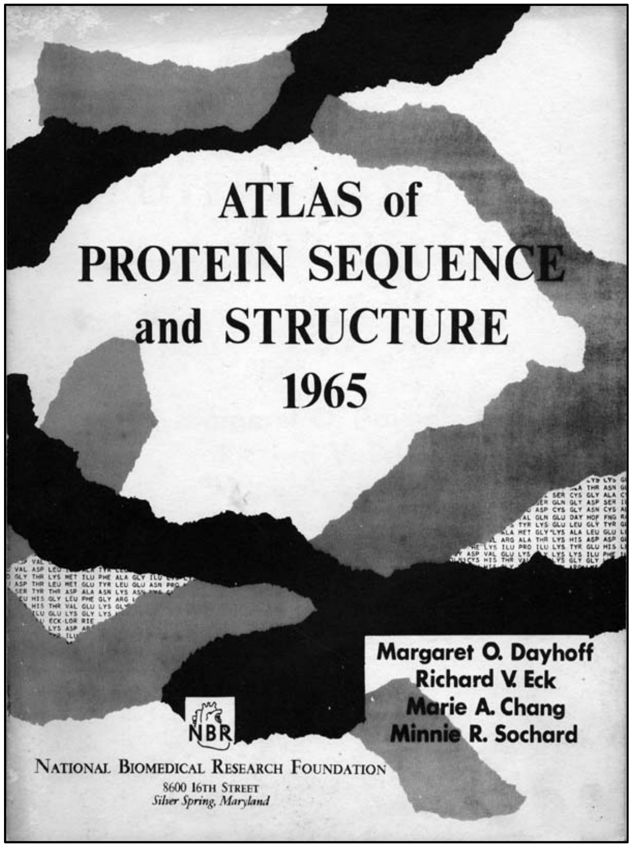
Source:http://1001genomes.org/

Source:http://www.scigenom.com/metagenomics

Source:https://www.broadinstitute.org/

Dr. Margaret Dayhoff

(1925 – 1983)



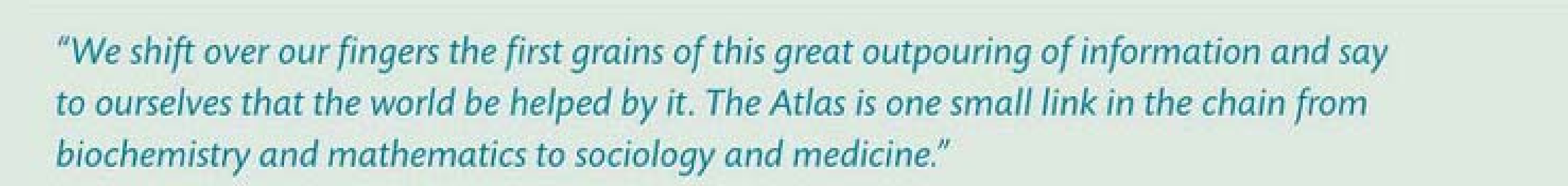
ATLAS of
PROTEIN SEQUENCE
and STRUCTURE
1965

Margaret O. Dayhoff
Richard V. Eck
Marie A. Chang
Minnie R. Sochard

NBR

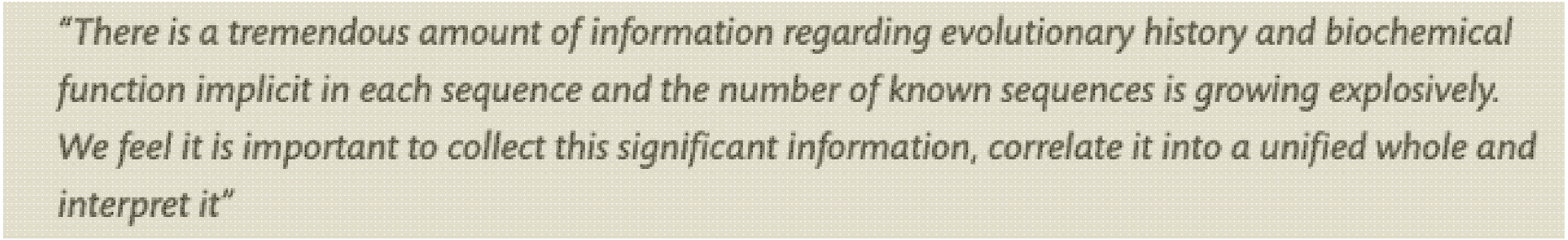
NATIONAL BIOMEDICAL RESEARCH FOUNDATION
8600 16TH STREET
Silver Spring, Maryland

- Interested in the possibility of deducing the evolutionary connections of the biological world from sequence evidence
- Formulated the first probability model of protein evolution - PAM substitution matrix
- The origin of the single-letter AA code
- Published the Atlas of Protein Sequence and Structure (1965-79), which became the Protein Information Resource Protein Sequence Database (PIR-PSD)



“We shift over our fingers the first grains of this great outpouring of information and say to ourselves that the world be helped by it. The Atlas is one small link in the chain from biochemistry and mathematics to sociology and medicine.”

Margaret O. Dayhoff to Susan Tideman, 18 October 1968, National Biomedical Research Foundation Archives



“There is a tremendous amount of information regarding evolutionary history and biochemical function implicit in each sequence and the number of known sequences is growing explosively. We feel it is important to collect this significant information, correlate it into a unified whole and interpret it”

Margaret O. Dayhoff to Carl Berkley, 27 February 1967, National Biomedical Research Foundation Archives

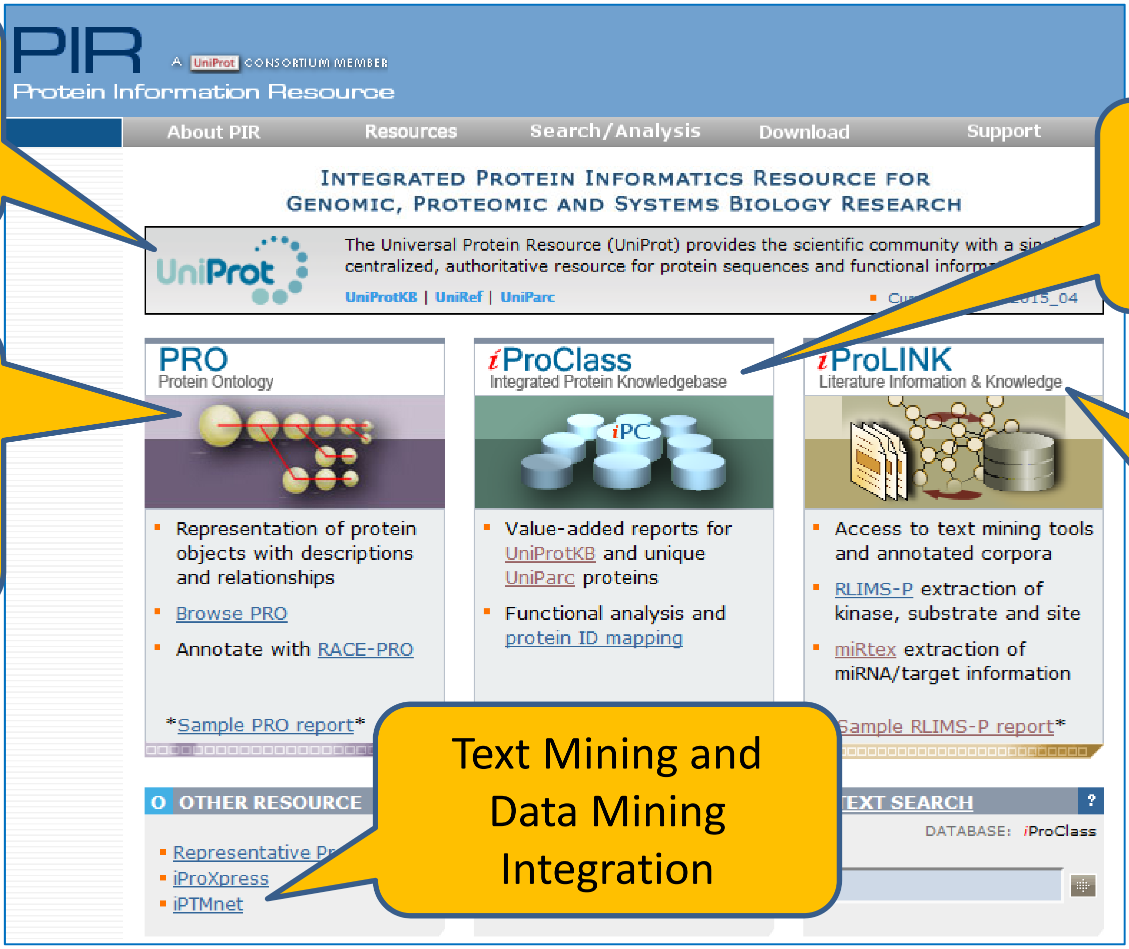
Quoted in: An Introduction to Molecular Evolution and Phylogenetics by Lindell Bromham

Protein Information Resource

http://proteininformationresource.org

Hub for protein functional information

Ontological representation of proteoforms and protein complexes



PIR A UniProt CONSORTIUM MEMBER
Protein Information Resource

About PIR | Resources | Search/Analysis | Download | Support

INTEGRATED PROTEIN INFORMATICS RESOURCE FOR GENOMIC, PROTEOMIC AND SYSTEMS BIOLOGY RESEARCH

UniProt

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

UniProtKB | UniRef | UniParc

- Cu 2015_04

PRO
Protein Ontology

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

Sample PRO report

***i*ProClass**
Integrated Protein Knowledgebase

- Value-added reports for UniProtKB and unique UniParc proteins
- Functional analysis and protein ID mapping

***i*ProLINK**
Literature Information & Knowledge

- Access to text mining tools and annotated corpora
- RLIMS-P extraction of kinase, substrate and site
- miRtex extraction of miRNA/target information

Sample RLIMS-P report*

O OTHER RESOURCE

- Representative Pr
- iProXpress
- iPTMnet

EXT SEARCH ?

DATABASE: *i*ProClass

Data warehouse

Access to text mining tools

Text Mining and Data Mining Integration

- National and international collaborative networks

EMBL-EBI | SIB | New York State Center of Excellence in Bioinformatics & Life Sciences | The Jackson Laboratory | MITRE | cnio | NCBI

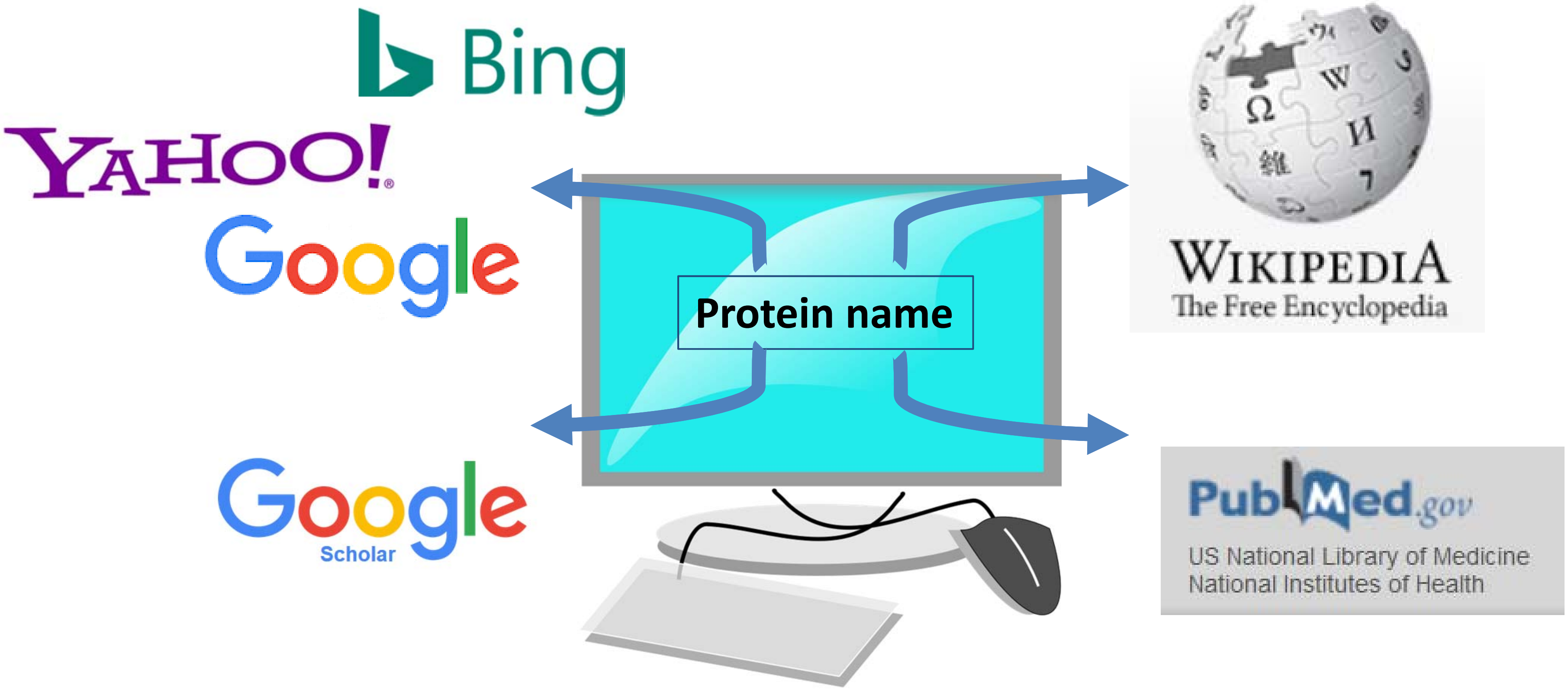
II. Searching for Protein Information in "Free-Text" Resources

Question

If you wanted to find information about a protein and the diseases it was associated with, where would you look? What would search for?

Some Possible Answers

Web Search Engines



Protein name

Biomedical Literature Resource: Medline/PubMed

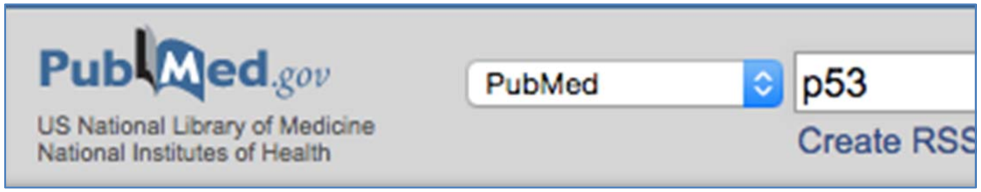
- Medline
  - ~26 million references to journal articles in life sciences with a concentration on biomedicine.
  - Indexed with NLM’s Medical Subject Headings (MeSH®)

- PubMed (http://www.ncbi.nlm.nih.gov/pubmed)
  - Provides free access to MEDLINE
  - Links to full-text articles found in PubMed Central or at publisher web sites, and other related resources.
  - Provides Advanced search, and special filters.

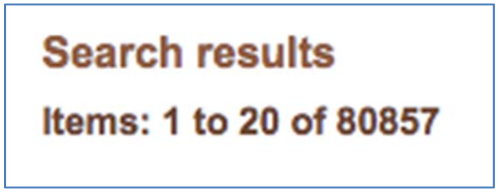
Source:http://www.nlm.nih.gov/pubs/factsheets/medline.html

Common issues

-Too many articles , e.g keyword p53


PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Create RSS

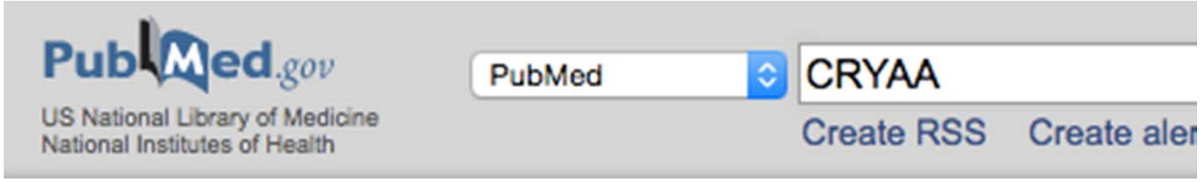
p53


Search results

Items: 1 to 20 of 80857

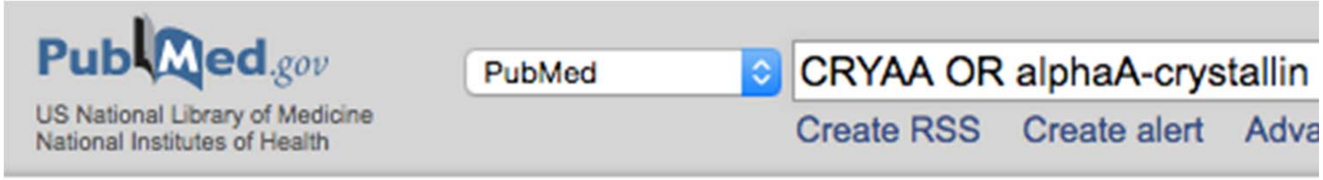
Possible solutions: Use Filters, MeSH terms, or Boolean operators to scope results

-Retrieval of a subset of articles due to narrow search (not including synonyms), e.g compare CRYAA alone and including alphaA-crystallin


PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Create RSS
Create aler

CRYAA

110 Articles


PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Create RSS
Create alert
Adva

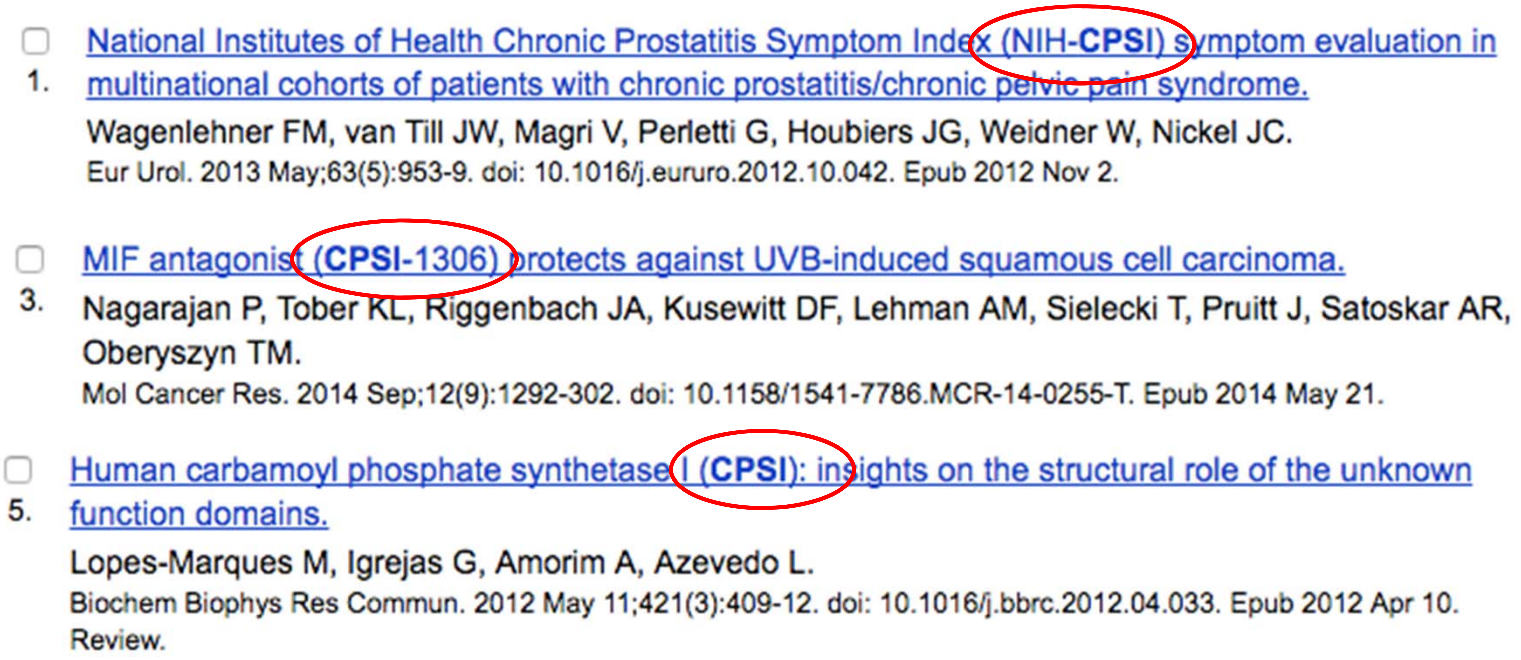
CRYAA OR alphaA-crystallin

510 Articles

Possible solutions: Include relevant synonyms

Common issues

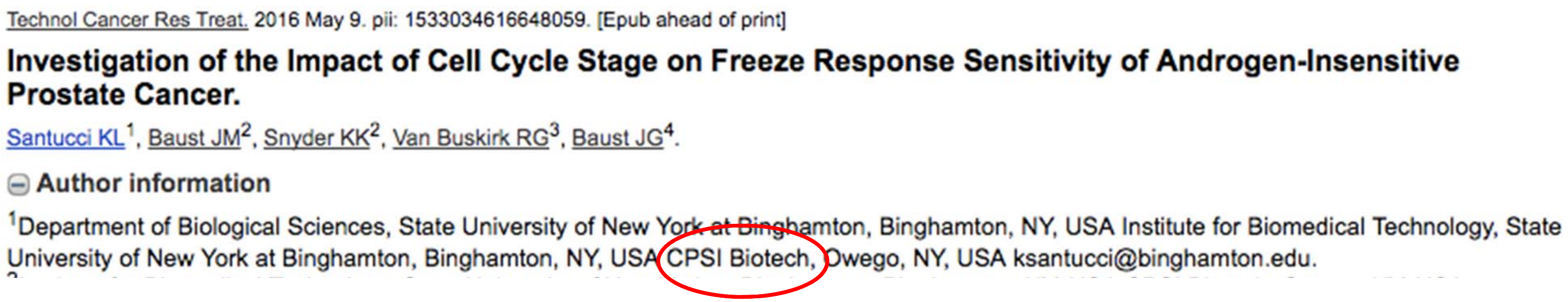
-Articles not relevant to the query due to many possible meanings for the word, e.g. CPSI



□ 1. National Institutes of Health Chronic Prostatitis Symptom Index (NIH-**CPSI**) symptom evaluation in multinational cohorts of patients with chronic prostatitis/chronic pelvic pain syndrome.
Wagenlehner FM, van Till JW, Magri V, Perletti G, Houbiers JG, Weidner W, Nickel JC.
Eur Urol. 2013 May;63(5):953-9. doi: 10.1016/j.eururo.2012.10.042. Epub 2012 Nov 2.

□ 3. MIF antagonist (**CPSI**-1306) protects against UVB-induced squamous cell carcinoma.
Nagarajan P, Tober KL, Riggenbach JA, Kusewitt DF, Lehman AM, Sielecki T, Pruitt J, Satoskar AR, Oberyszyn TM.
Mol Cancer Res. 2014 Sep;12(9):1292-302. doi: 10.1158/1541-7786.MCR-14-0255-T. Epub 2014 May 21.

□ 5. Human carbamoyl phosphate synthetase I (**CPSI**): insights on the structural role of the unknown function domains.
Lopes-Marques M, Igrejas G, Amorim A, Azevedo L.
Biochem Biophys Res Commun. 2012 May 11;421(3):409-12. doi: 10.1016/j.bbrc.2012.04.033. Epub 2012 Apr 10. Review.



Technol Cancer Res Treat. 2016 May 9. pii: 1533034616648059. [Epub ahead of print]

Investigation of the Impact of Cell Cycle Stage on Freeze Response Sensitivity of Androgen-Insensitive Prostate Cancer.

Santucci KL[1], Baust JM[2], Snyder KK[2], Van Buskirk RG[3], Baust JG[4].

⊖ **Author information**

[1]Department of Biological Sciences, State University of New York at Binghamton, Binghamton, NY, USA Institute for Biomedical Technology, State University of New York at Binghamton, Binghamton, NY, USA CPSI Biotech, Owego, NY, USA ksantucci@binghamton.edu.

Some Tips for PubMed Searches

- Use Filters to narrow your search

  e.g. selecting species human
- Use Advanced Search to narrow your search:

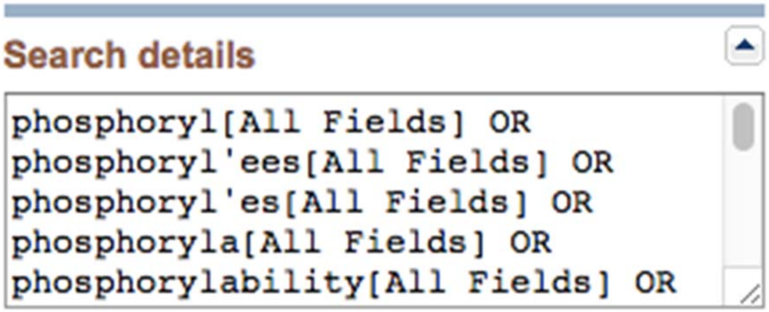
  Boolean operators AND, OR, NOT

  e.g. to retrieve articles on CPSI that are less likely to be about prostatitis score,

  search for CPSI NOT prostatitis
- Search for phrases in quotes

  e.g. "breast cancer"
- Use Wildcards * to expand your query

Phosphoryl*



Search details

```
phosphoryl[All Fields] OR
phosphoryl'ees[All Fields] OR
phosphoryl'es[All Fields] OR
phosphoryla[All Fields] OR
phosphorylability[All Fields] OR
```

Expanded Query

*Results can be saved locally, so you can review them at a later time

Hands-On #1: Protein Search

- Search for information about HEXA

  -What is its biochemical function?

  -What disease is it associated with?

  -Other information?

Divide into groups and each group use a different searching method (Google (not Wikipedia), Google Scholar, Wikipedia, PubMed)

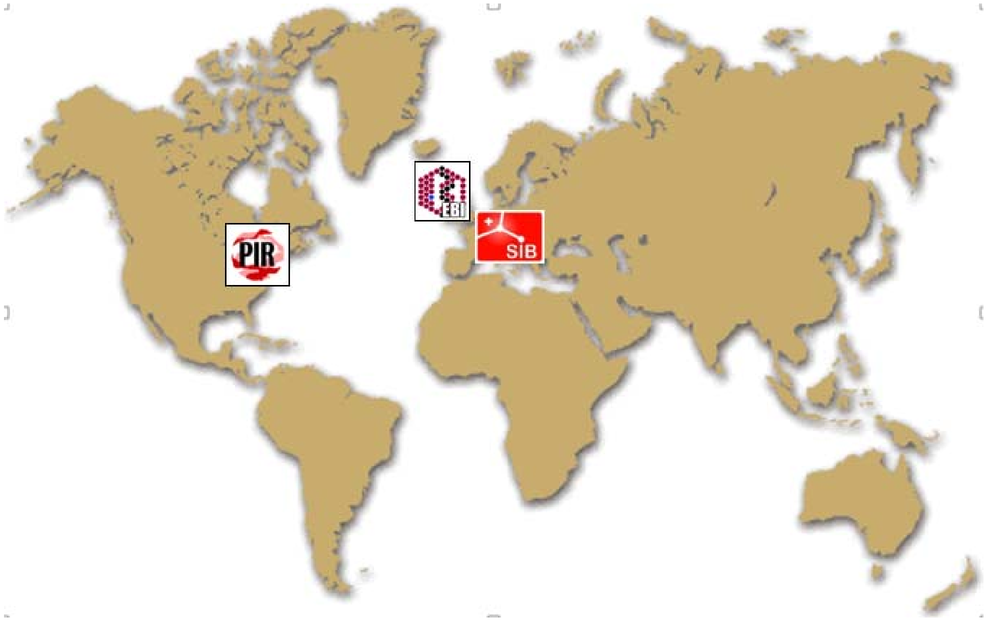
Discussion

- Did you find the information you were looking for?
- What were some pros and cons of the search method/resource you used?

III. Protein Resources UniProtKB Database

The Universal Protein Resource

www.uniprot.org



- ✓ **comprehensive**
- ✓ **high quality**
- ✓ **freely accessible**

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.

Modified from Michel Schneider, UniProt

UniProtKB, the Knowledge base component of UniProt

Where **data** becomes **structured knowledge**

Central hub for the collection of functional information on proteins. The core data mandatory for each UniProtKB entry:

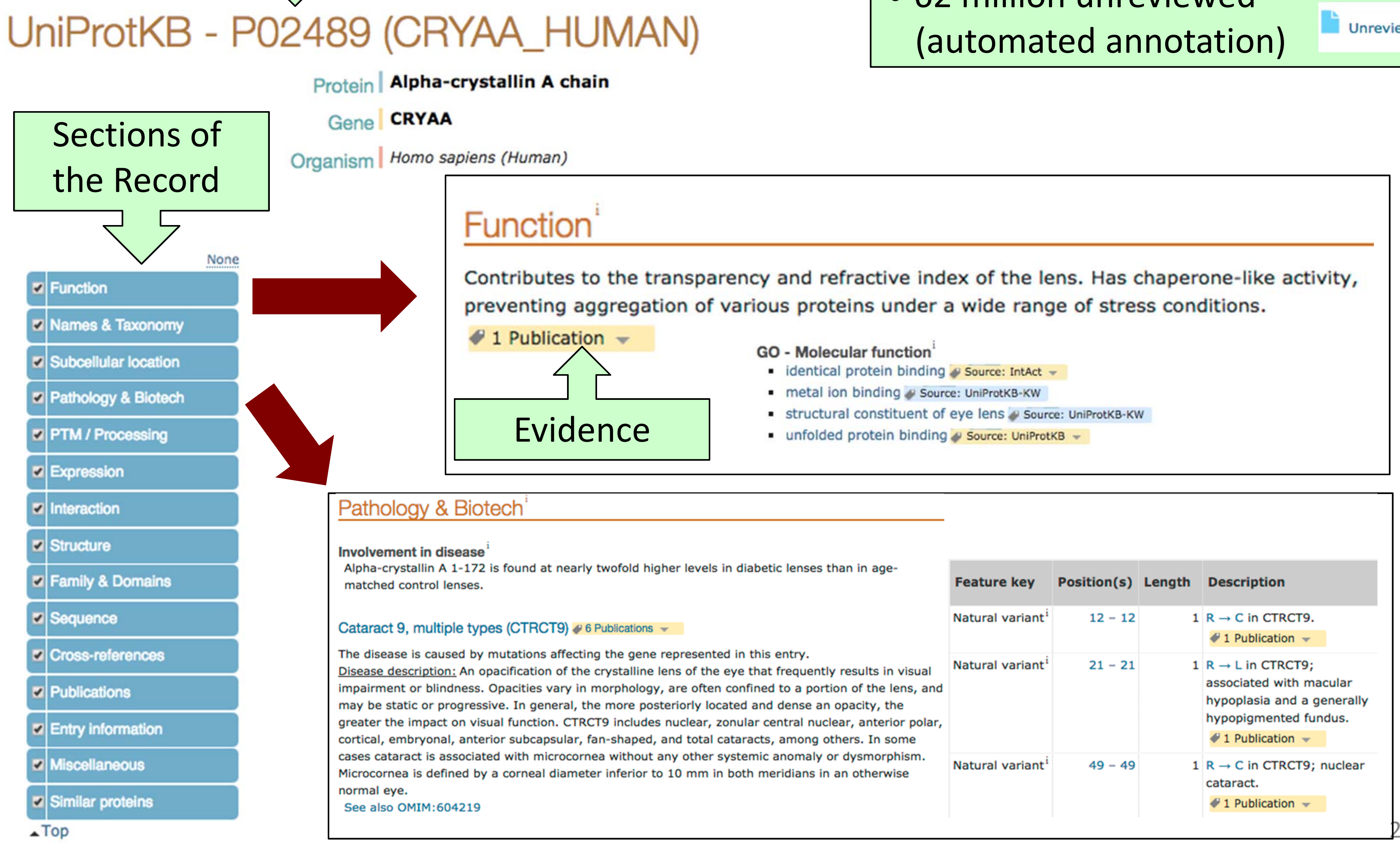
- amino acid sequence
- protein name or description
- taxonomic data
- citation information

Courtesy of the Swiss-Prot group (SIB Swiss Institute of Bioinformatics)

What’s in UniProtKB?

Unique Identifier

Records for ~62.5 million proteins
- ~500K manually reviewed (Reviewed)
- 62 million unreviewed (automated annotation) (Unreviewed)



UniProtKB - P02489 (CRYAA_HUMAN)

Protein	Alpha-crystallin A chain
Gene	CRYAA
Organism	*Homo sapiens (Human)*

Sections of the Record

Function

Contributes to the transparency and refractive index of the lens. Has chaperone-like activity, preventing aggregation of various proteins under a wide range of stress conditions.

1 Publication

Evidence

GO - Molecular function
- identical protein binding Source: IntAct
- metal ion binding Source: UniProtKB-KW
- structural constituent of eye lens Source: UniProtKB-KW
- unfolded protein binding Source: UniProtKB

Pathology & Biotech

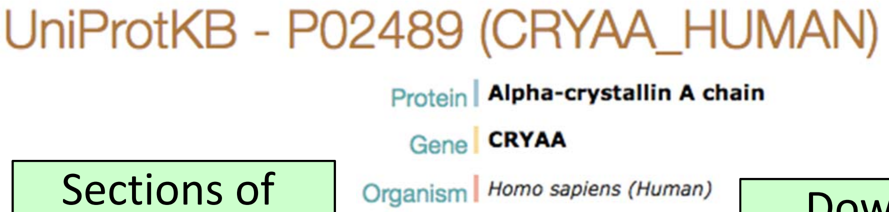
Involvement in disease
Alpha-crystallin A 1-172 is found at nearly twofold higher levels in diabetic lenses than in age-matched control lenses.

Cataract 9, multiple types (CTRCT9) 6 Publications

The disease is caused by mutations affecting the gene represented in this entry.
Disease description: An opacification of the crystalline lens of the eye that frequently results in visual impairment or blindness. Opacities vary in morphology, are often confined to a portion of the lens, and may be static or progressive. In general, the more posteriorly located and dense an opacity, the greater the impact on visual function. CTRCT9 includes nuclear, zonular central nuclear, anterior polar, cortical, embryonal, anterior subcapsular, fan-shaped, and total cataracts, among others. In some cases cataract is associated with microcornea without any other systemic anomaly or dysmorphism. Microcornea is defined by a corneal diameter inferior to 10 mm in both meridians in an otherwise normal eye.
See also OMIM:604219

Feature key	Position(s)	Length	Description
Natural variant	12 – 12	1	R → C in CTRCT9. 1 Publication
Natural variant	21 – 21	1	R → L in CTRCT9; associated with macular hypoplasia and a generally hypopigmented fundus. 1 Publication
Natural variant	49 – 49	1	R → C in CTRCT9; nuclear cataract. 1 Publication

What’s in UniProtKB?



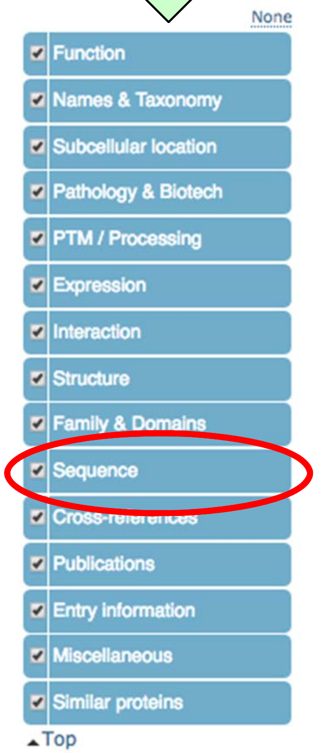
UniProtKB - P02489 (CRYAA_HUMAN)

Protein | Alpha-crystallin A chain

Gene | CRYAA

Organism | *Homo sapiens (Human)*

Sections of the Record

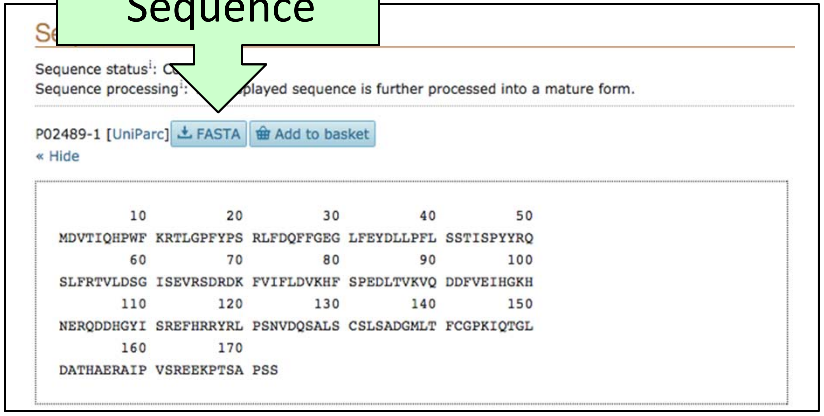


None

- ☑ Function
- ☑ Names & Taxonomy
- ☑ Subcellular location
- ☑ Pathology & Biotech
- ☑ PTM / Processing
- ☑ Expression
- ☑ Interaction
- ☑ Structure
- ☑ Family & Domains
- ☑ Sequence
- ☑ Cross-references
- ☑ Publications
- ☑ Entry information
- ☑ Miscellaneous
- ☑ Similar proteins

▲Top

Download Sequence



Se

Sequence status: C

Sequence processing: played sequence is further processed into a mature form.

P02489-1 [UniParc]

« Hide

```
        10         20         30         40         50
MDVTIQHPWF KRTLGPFYPS RLFDQFFGEG LFEYDLLPFL SSTISPYYRQ
        60         70         80         90        100
SLFRTVLDSG ISEVRSDRDK FVIFLDVKHF SPEDLTVKVQ DDFVEIHGKH
       110        120        130        140        150
NERQDDHGYI SREFHRRYRL PSNVDQSALS CSLSADGMLT FCGPKIQTGL
       160        170
DATHAERAIP VSREEKPTSA PSS
```

FASTA Format:
Common input format for sequence analysis

```
>sp|P02489|CRYAA_HUMAN Alpha-crystallin A chain OS=Homo
sapiens GN=CRYAA PE=1 SV=2
MDVTIQHPWFKRTLGPFYPSRLFDQFFGEGLFEYDLLPFLSSTISPYYRQSLFRTVLDSG
ISEVRSDRDKFVIFLDVKHFSPEDLTVKVQDDFVEIHGKHNERQDDHGYISREFHRRYRL
PSNVDQSALSCSLSADGMLTFCGPKIQTGLDATHAERAIPVSREEKPTSAPSS
```

Header Line:
- Starts with “>”
- Contains ID and description

Sequence

Hands On #2 – UniProtKB

1. Go to UniProtKB (http://www.uniprot.org/) and search for the entry for human HEXA

2. What is the UniProtKB identifier for this protein?

3. What is the function of this protein?

4. What disease is associated with defects in this protein?

5. Give an example of a genetic variant with publication support that leads to the infantile form of the disease.

6. Download the sequence of Isoform 1 in FASTA format

Hands On #2 – UniProtKB (Answers)

1. Go to UniProtKB (http://www.uniprot.org/) and search for the entry for human HEXA

2. What is the UniProtKB identifier for this protein? P06865

3. What is the function of this protein? Degrades GM2-gangliosides. A ganglioside is a type of glycolipid (sugar + lipid).

4. What disease is associated with defects in this protein? GM2-gangliosidosis 1 (aka Tay-Sachs Disease)

5. Give an example of a genetic variant with publication support that leads to the infantile form of the disease. Several examples in the variant table in the Pathology & Biotech section of the entry.

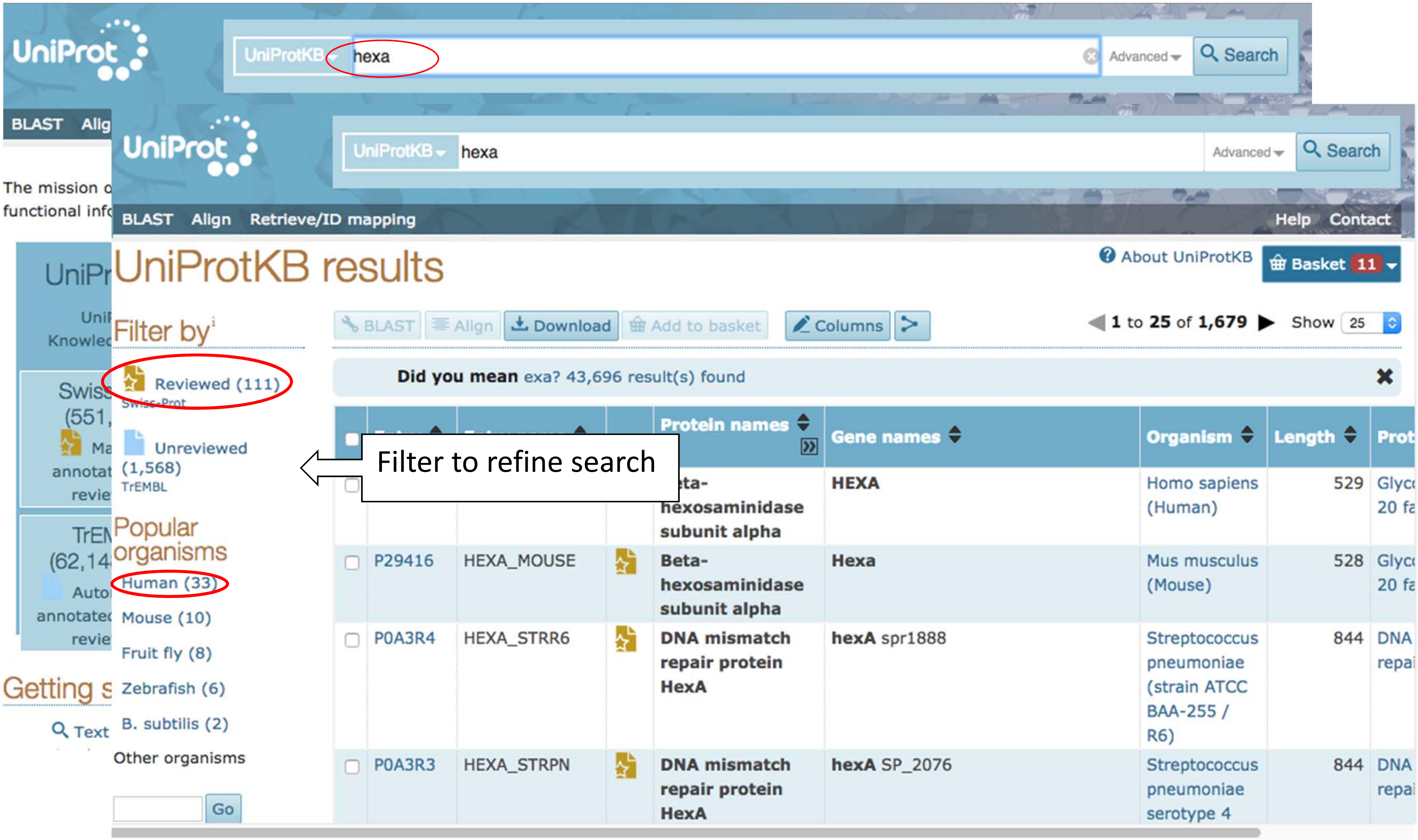
Hands On #2 – UniProtKB (Answers)

6. Download the sequence of Isoform 1 in FASTA format. Click FASTA button in above Isoform 1 sequence in Sequence section.

```
>sp|P06865|HEXA_HUMAN Beta-hexosaminidase subunit alpha
OS=Homo sapiens GN=HEXA PE=1 SV=2
MTSSRLWFSLLLAAAFAGRATALWPWPQNFQTSDQRYVLYPNNFQFQYDVSSAAQPGCSV
LDEAFQRYRDLLFGSGSWPRPYLTGKRHTLEKNVLVVSVVTPGCNQLPTLESVENYTLTI
NDDQCLLLSETVWGALRGLETFSQLVWKSAEGTFFINKTEIEDFPRFPHRGLLLDTSRHY
LPLSSILDTLDVMAYNKLNVFHWHLVDDPSFPYESFTFPELMRKGSYNPVTHIYTAQDVK
EVIEYARLRGIRVLAEFDTPGHTLSWGPGIPGLLTPCYSGSEPSGTFGPVNPSLNNTYEF
MSTFFLEVSSVFPDFYLHLGGDEVDFTCWKSNPEIQDFMRKKGFGEDFKQLESFYIQTLL
DIVSSYGKGYVVWQEVFDNKVKIQPDTIIQVWREDIPVNYMKELELVTKAGFRALLSAPW
YLNRISYGPDWKDFYIVEPLAFEGTPEQKALVIGGEACMWGEYVDNTNLVPRLWPRAGAV
AERLWSNKLTSDLTFAYERLSHFRCELLRRGVQAQPLNVGFCEQEFEQT
```

Hands On #2 – UniProtKB (Answers)

Searching for a protein in UniProtKB



Search: hexa

UniProtKB results

1 to 25 of 1,679

Filter by

Reviewed (111)
Swiss-Prot

Unreviewed (1,568)
TrEMBL

Popular organisms

Human (33)

Mouse (10)

Fruit fly (8)

Zebrafish (6)

B. subtilis (2)

Other organisms

Did you mean exa? 43,696 result(s) found

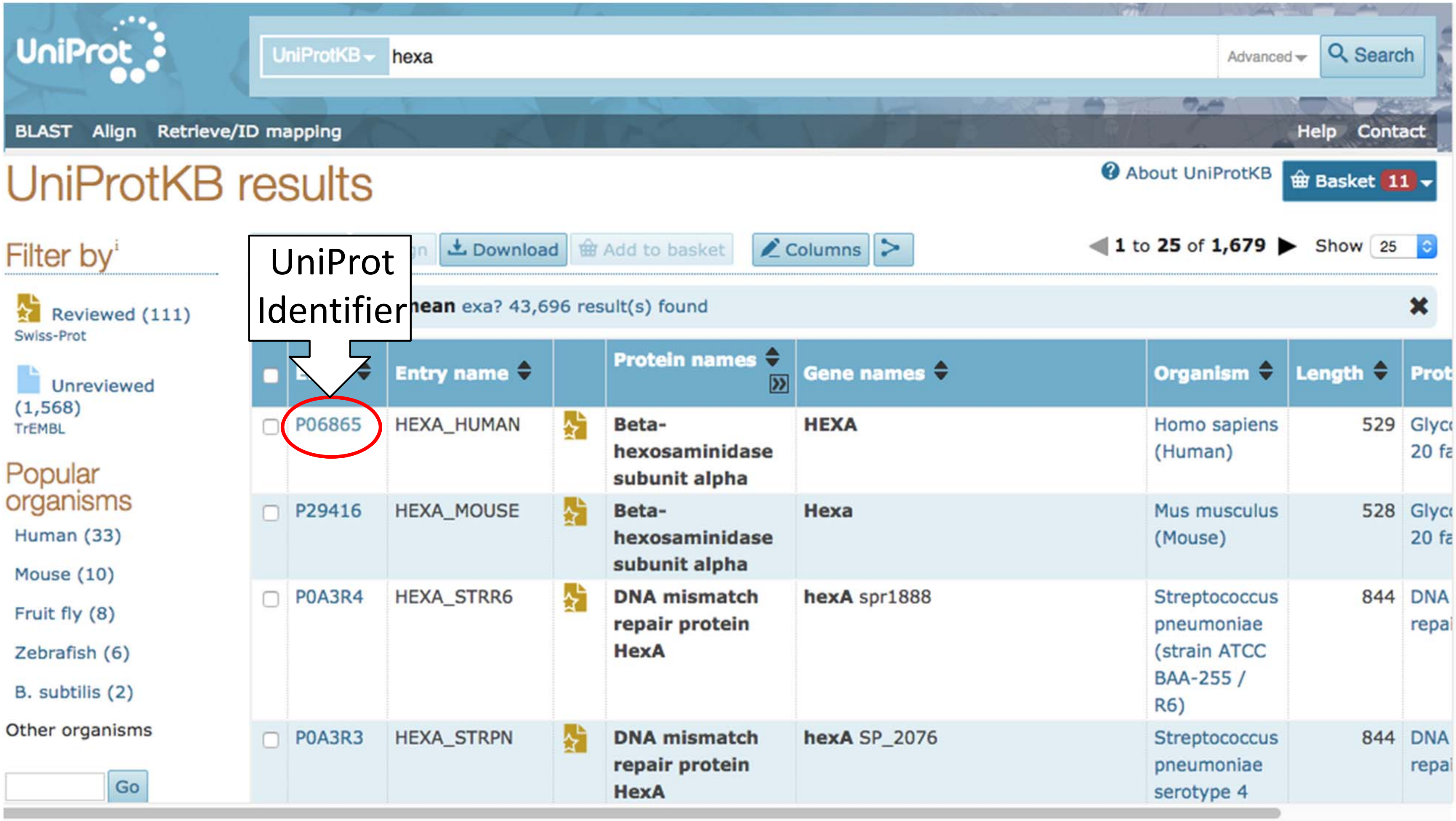
Filter to refine search

	Entry	Entry name		Protein names	Gene names	Organism	Length	Prot
□				eta-hexosaminidase subunit alpha	HEXA	Homo sapiens (Human)	529	Glyc 20 fa
□	P29416	HEXA_MOUSE	★	Beta-hexosaminidase subunit alpha	Hexa	Mus musculus (Mouse)	528	Glyc 20 fa
□	P0A3R4	HEXA_STRR6	★	DNA mismatch repair protein HexA	hexA spr1888	Streptococcus pneumoniae (strain ATCC BAA-255 / R6)	844	DNA repai
□	P0A3R3	HEXA_STRPN	★	DNA mismatch repair protein HexA	hexA SP_2076	Streptococcus pneumoniae serotype 4	844	DNA repai

UniProtKB (http://www.uniprot.org/)

Hands On #2 – UniProtKB (Answers)

Search Results



UniProtKB results

1 to 25 of 1,679

Filter by

Reviewed (111)
Swiss-Prot

Unreviewed (1,568)
TrEMBL

Popular organisms

Human (33)

Mouse (10)

Fruit fly (8)

Zebrafish (6)

B. subtilis (2)

Other organisms

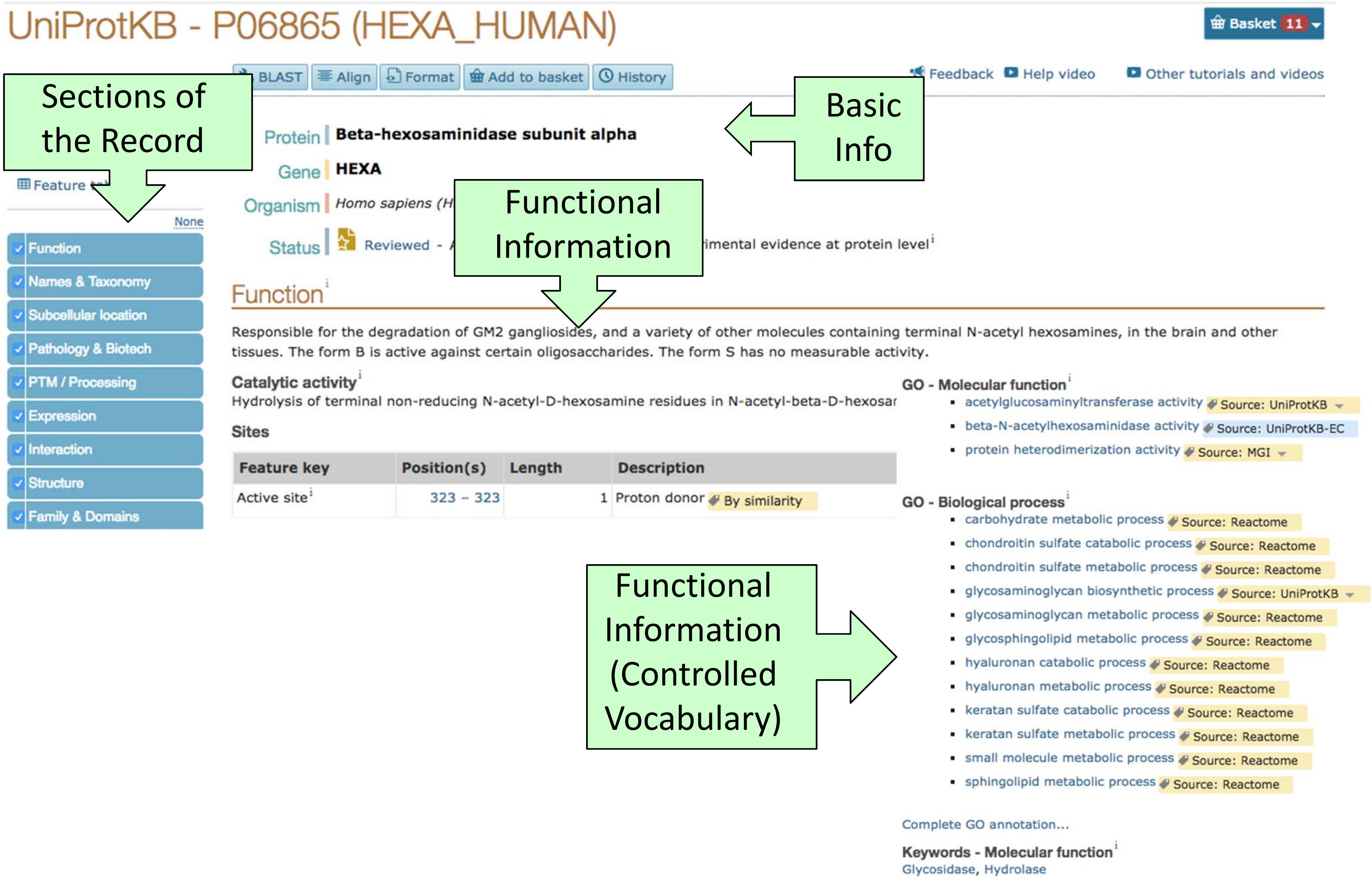
UniProt Identifier

nean exa? 43,696 result(s) found

□		Entry name		Protein names	Gene names	Organism	Length	Prot
□	P06865	HEXA_HUMAN		Beta-hexosaminidase subunit alpha	HEXA	Homo sapiens (Human)	529	Glyc 20 fa
□	P29416	HEXA_MOUSE		Beta-hexosaminidase subunit alpha	Hexa	Mus musculus (Mouse)	528	Glyc 20 fa
□	P0A3R4	HEXA_STRR6		DNA mismatch repair protein HexA	hexA spr1888	Streptococcus pneumoniae (strain ATCC BAA-255 / R6)	844	DNA repai
□	P0A3R3	HEXA_STRPN		DNA mismatch repair protein HexA	hexA SP_2076	Streptococcus pneumoniae serotype 4	844	DNA repai

Hands On #2 – UniProtKB (Answers)

UniProtKB Entry Page I



Sections of the Record

Basic Info

Functional Information

Functional Information (Controlled Vocabulary)

UniProtKB - P06865 (HEXA_HUMAN)

Protein | **Beta-hexosaminidase subunit alpha**

Gene | **HEXA**

Organism | *Homo sapiens (H*

Status | Reviewed - / imental evidence at protein level

Function

Responsible for the degradation of GM2 gangliosides, and a variety of other molecules containing terminal N-acetyl hexosamines, in the brain and other tissues. The form B is active against certain oligosaccharides. The form S has no measurable activity.

Catalytic activity

Hydrolysis of terminal non-reducing N-acetyl-D-hexosamine residues in N-acetyl-beta-D-hexosar

Sites

Feature key	Position(s)	Length	Description
Active site	323 – 323	1	Proton donor By similarity

GO - Molecular function

- acetylglucosaminyltransferase activity Source: UniProtKB
- beta-N-acetylhexosaminidase activity Source: UniProtKB-EC
- protein heterodimerization activity Source: MGI

GO - Biological process

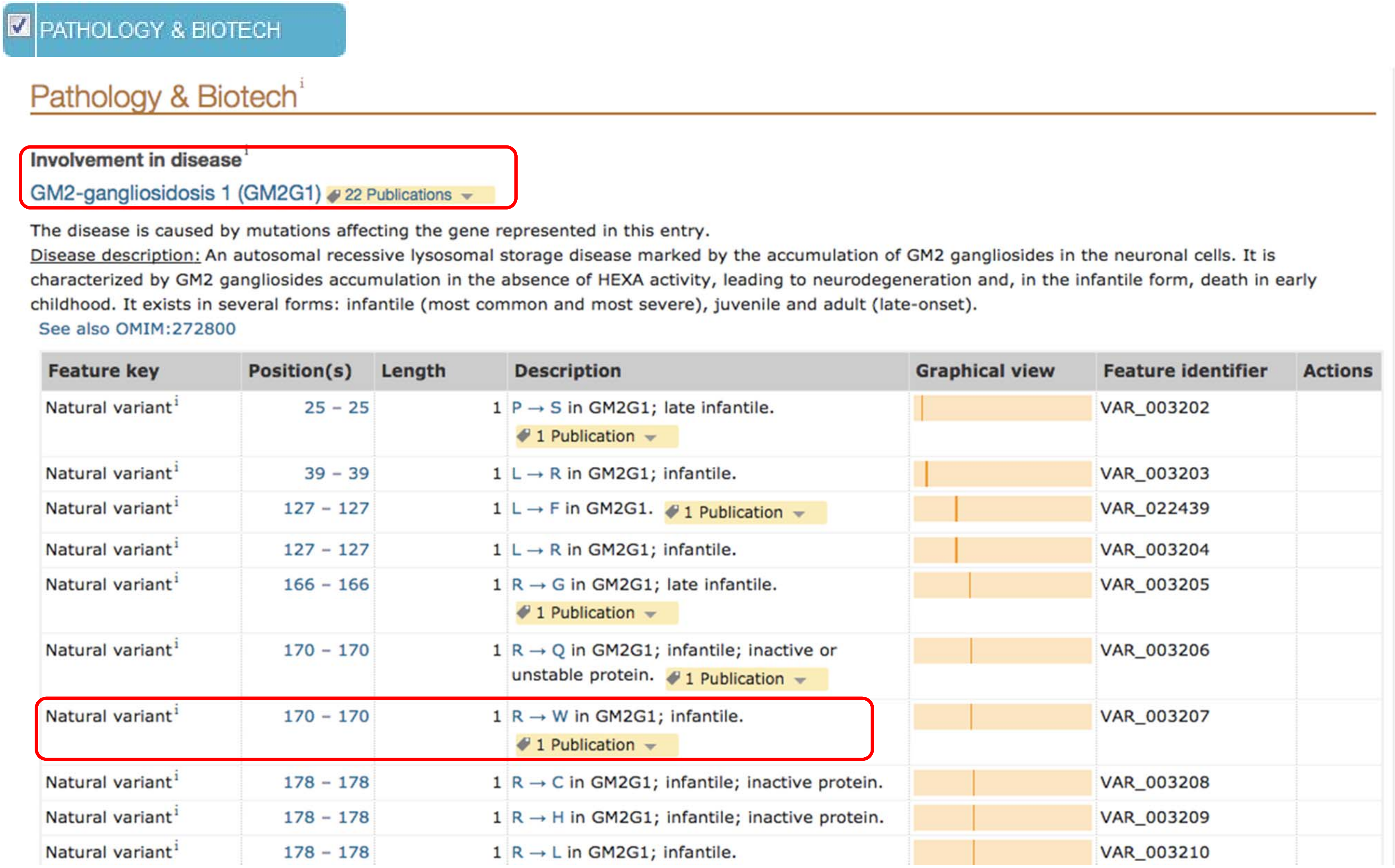
- carbohydrate metabolic process Source: Reactome
- chondroitin sulfate catabolic process Source: Reactome
- chondroitin sulfate metabolic process Source: Reactome
- glycosaminoglycan biosynthetic process Source: UniProtKB
- glycosaminoglycan metabolic process Source: Reactome
- glycosphingolipid metabolic process Source: Reactome
- hyaluronan catabolic process Source: Reactome
- hyaluronan metabolic process Source: Reactome
- keratan sulfate catabolic process Source: Reactome
- keratan sulfate metabolic process Source: Reactome
- small molecule metabolic process Source: Reactome
- sphingolipid metabolic process Source: Reactome

Complete GO annotation...

Keywords - Molecular function

Glycosidase, Hydrolase

Hands On #2 – UniProtKB (Answers)
Disease and Variant Information



☑ PATHOLOGY & BIOTECH

Pathology & Biotech

Involvement in disease

GM2-gangliosidosis 1 (GM2G1) 22 Publications

The disease is caused by mutations affecting the gene represented in this entry.

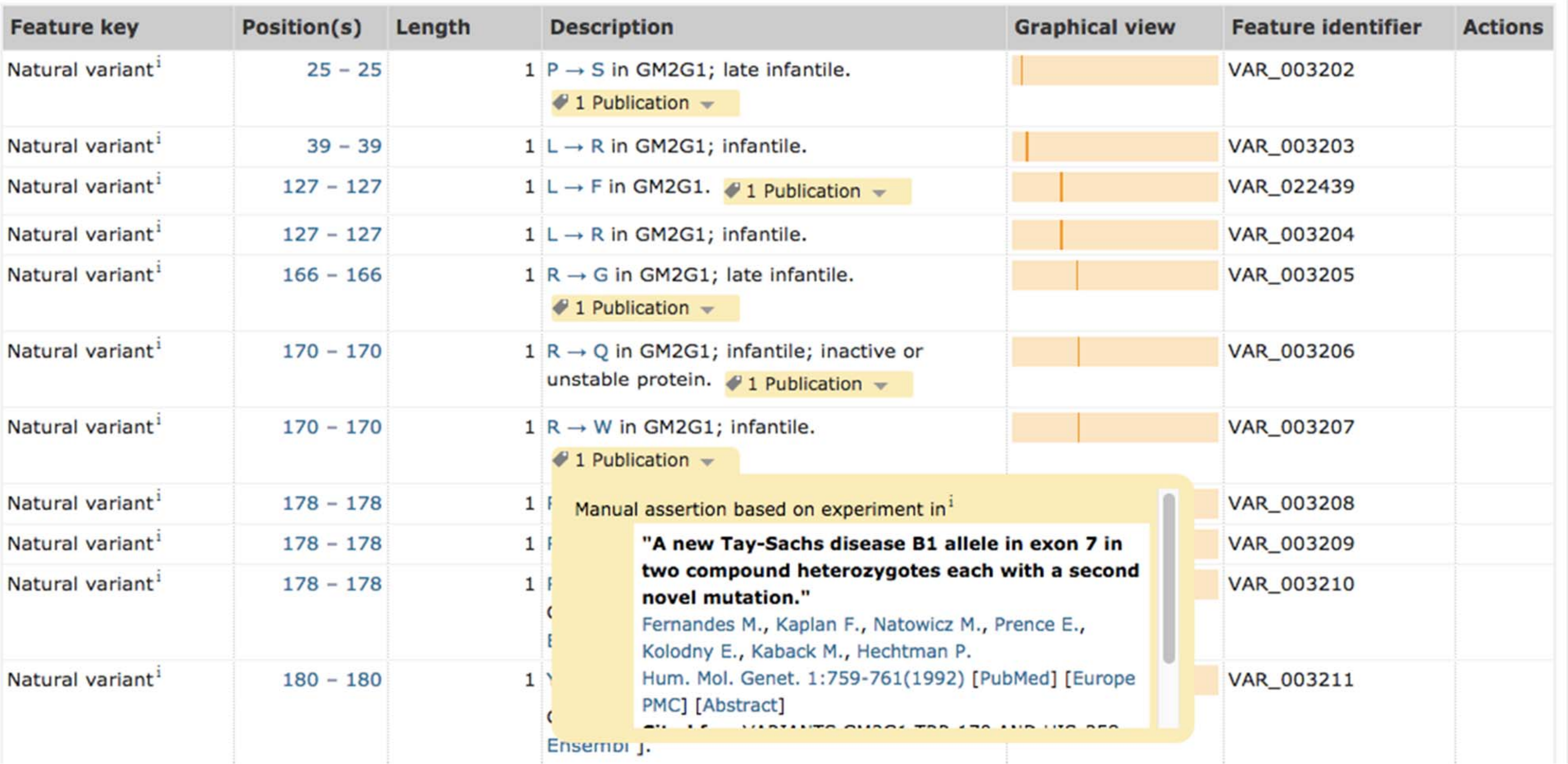
Disease description: An autosomal recessive lysosomal storage disease marked by the accumulation of GM2 gangliosides in the neuronal cells. It is characterized by GM2 gangliosides accumulation in the absence of HEXA activity, leading to neurodegeneration and, in the infantile form, death in early childhood. It exists in several forms: infantile (most common and most severe), juvenile and adult (late-onset).

See also OMIM:272800

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Natural variant	25 – 25	1	P → S in GM2G1; late infantile. 1 Publication		VAR_003202	
Natural variant	39 – 39	1	L → R in GM2G1; infantile.		VAR_003203	
Natural variant	127 – 127	1	L → F in GM2G1. 1 Publication		VAR_022439	
Natural variant	127 – 127	1	L → R in GM2G1; infantile.		VAR_003204	
Natural variant	166 – 166	1	R → G in GM2G1; late infantile. 1 Publication		VAR_003205	
Natural variant	170 – 170	1	R → Q in GM2G1; infantile; inactive or unstable protein. 1 Publication		VAR_003206	
Natural variant	170 – 170	1	R → W in GM2G1; infantile. 1 Publication		VAR_003207	
Natural variant	178 – 178	1	R → C in GM2G1; infantile; inactive protein.		VAR_003208	
Natural variant	178 – 178	1	R → H in GM2G1; infantile; inactive protein.		VAR_003209	
Natural variant	178 – 178	1	R → L in GM2G1; infantile.		VAR_003210	

Hands On #2 – UniProtKB (Answers)
Disease and Variant Information

☑ PATHOLOGY & BIOTECH

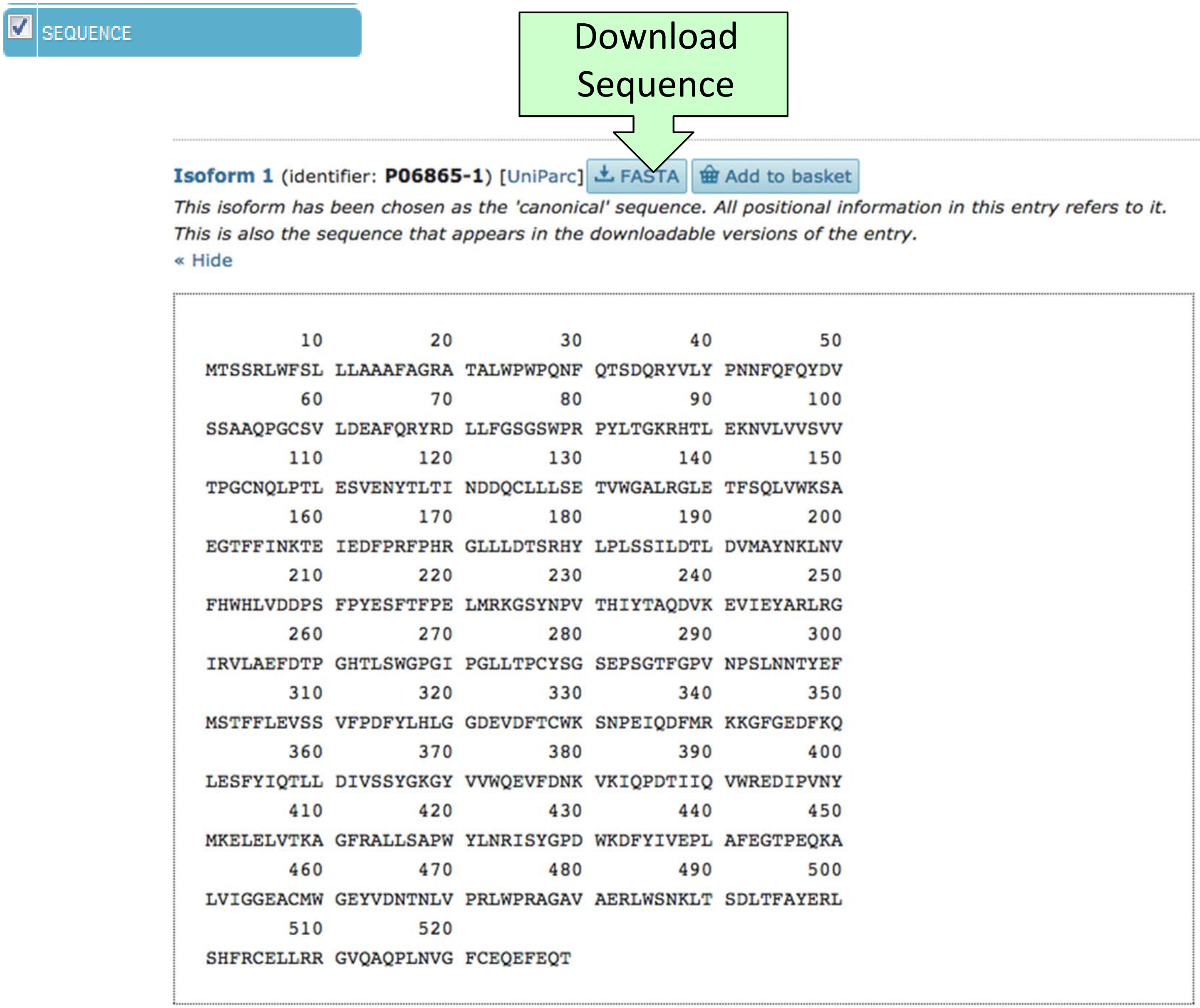


Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Natural variant	25 – 25	1	P → S in GM2G1; late infantile. 1 Publication		VAR_003202	
Natural variant	39 – 39	1	L → R in GM2G1; infantile.		VAR_003203	
Natural variant	127 – 127	1	L → F in GM2G1. 1 Publication		VAR_022439	
Natural variant	127 – 127	1	L → R in GM2G1; infantile.		VAR_003204	
Natural variant	166 – 166	1	R → G in GM2G1; late infantile. 1 Publication		VAR_003205	
Natural variant	170 – 170	1	R → Q in GM2G1; infantile; inactive or unstable protein. 1 Publication		VAR_003206	
Natural variant	170 – 170	1	R → W in GM2G1; infantile. 1 Publication		VAR_003207	
Natural variant	178 – 178	1	R		VAR_003208	
Natural variant	178 – 178	1	R		VAR_003209	
Natural variant	178 – 178	1	R (E		VAR_003210	
Natural variant	180 – 180	1	Y (Ensembl].		VAR_003211	

Manual assertion based on experiment in

"A new Tay-Sachs disease B1 allele in exon 7 in two compound heterozygotes each with a second novel mutation."
Fernandes M., Kaplan F., Natowicz M., Prence E., Kolodny E., Kaback M., Hechtman P.
Hum. Mol. Genet. 1:759-761(1992) [PubMed] [Europe PMC] [Abstract]

Hands On #2 – UniProtKB (Answers) Sequence Information



☑ SEQUENCE

Download Sequence

Isoform 1 (identifier: **P06865-1**) [UniParc] FASTA Add to basket

This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

« Hide

```
        10         20         30         40         50
MTSSRLWFSL LLAAAFAGRA TALWPWPQNF QTSDQRYVLY PNNFQFQYDV
        60         70         80         90        100
SSAAQPGCSV LDEAFQRYRD LLFGSGSWPR PYLTGKRHTL EKNVLVVSVV
       110        120        130        140        150
TPGCNQLPTL ESVENYTLTI NDDQCLLLSE TVWGALRGLE TFSQLVWKSA
       160        170        180        190        200
EGTFFINKTE IEDFPRFPHR GLLLDTSRHY LPLSSILDTL DVMAYNKLNV
       210        220        230        240        250
FHWHLVDDPS FPYESFTFPE LMRKGSYNPV THIYTAQDVK EVIEYARLRG
       260        270        280        290        300
IRVLAEFDTP GHTLSWGPGI PGLLTPCYSG SEPSGTFGPV NPSLNNTYEF
       310        320        330        340        350
MSTFFLEVSS VFPDFYLHLG GDEVDFTCWK SNPEIQDFMR KKGFGEDFKQ
       360        370        380        390        400
LESFYIQTLL DIVSSYGKGY VVWQEVFDNK VKIQPDTIIQ VWREDIPVNY
       410        420        430        440        450
MKELELVTKA GFRALLSAPW YLNRISYGPD WKDFYIVEPL AFEGTPEQKA
       460        470        480        490        500
LVIGGEACMW GEYVDNTNLV PRLWPRAGAV AERLWSNKLT SDLTFAYERL
       510        520
SHFRCELLRR GVQAQPLNVG FCEQEFEQT
```

Question

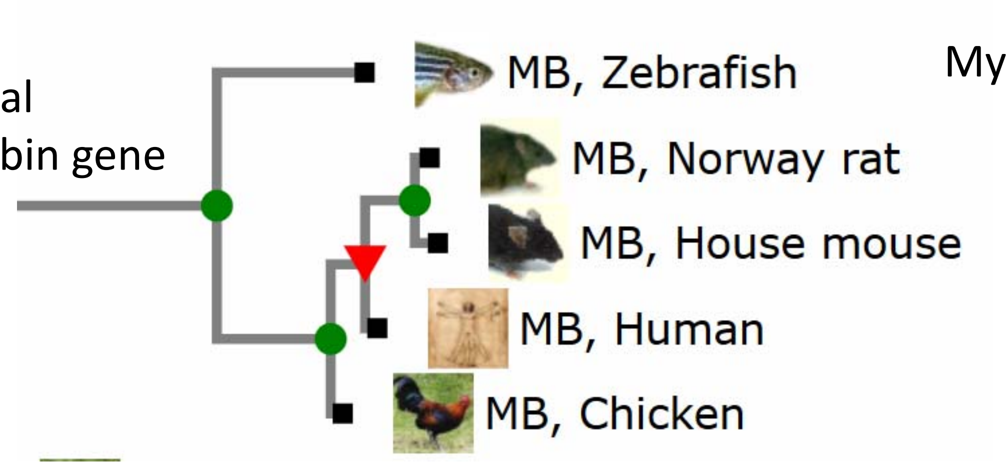
How did searching UniProtKB compare to searching other resources (e.g., Google, PubMed) for finding information about a protein and associated diseases?

IV. Protein Sequence Similarity Search

Important Concepts

Homologous sequences share a common ancestor

Ancestral Myoglobin gene



From: http://www.treefam.org/family/TF332967

Myoglobin gene

Known function in these organisms:
Serves as a reserve supply of oxygen and facilitates the movement of oxygen within muscles.

At the molecular level, homology is similarity between sequences that is due to their shared ancestry

We use sequence similarity that is statistically significant as evidence of homology

Searching for similar sequences

Query Protein Sequence

MRASLLSAGALIALLAALCPASRA
LEEKKVCQGTSNKLTQLGTFEDH
FLSLNKMFNNCEVGNLEITYVQR
NYDLSFLKTIQEVAGYVLIALNTL

?

Sequence Database (Target)

>protein 1
MLSPDDIEQWFTEDPGP
QIIRGNMYYENSYALA

>protein 2
MRPSGTAGAALLALLAALCPASRAL
EEKKVCQGTSNKLTQLGTFEDHFLS
LQRMFNNCEVGNLEITYVQRNYDLS
FLKTIQEVAGYVLIALNTVERIPLENL

>protein 3
MTEQMTLRGTLKGHNGWVTQIA
TTPQFPDMILSASRDKTIIMWKLT
RDETNYGIPQRA

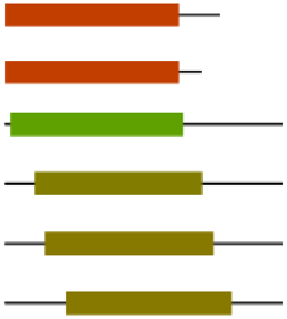
....

>protein n
MEVQLGLGRVYPRPPSKTYR
GAFQNLFQSVREVIQNPGPR
HPEAASAAPPGASLLLLQQQ
QQQQQQQQQQQQQQQQQ
QQQETSPRQQ

BLAST

List of entries similar to the query

Protein 2
Protein 5
Protein 180
:
:
:



% identity
Score
E-value

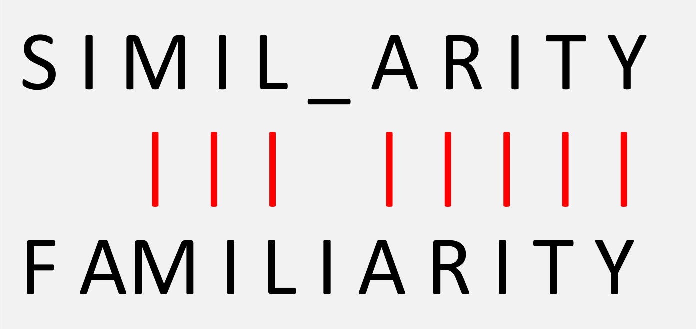
Basic **L**ocal **A**lignment **S**earch **T**ool (BLAST)

- Use of a set of algorithms to **compare a query sequence to all the sequences** in a specific database and find high scoring pairs of alignments
- Based on **Pair-Wise Alignments**
- The **score** of each comparison reflects the degree of similarity between the two sequences (the higher the greater the degree of similarity)
- The **Expectation value or E-value** tells you how many alignments with a given score are expected by chance (the closer to 0 the better)

Sequence comparison

Sequences are compared directly, position by position.

Score Key
match: +1
no match: -1
gap: –1



```
S I M I L _ A R I T Y
    | | |   | | | | |
F A M I L I A R I T Y
```

Matches=	8
No matches=	2
Gaps=	1

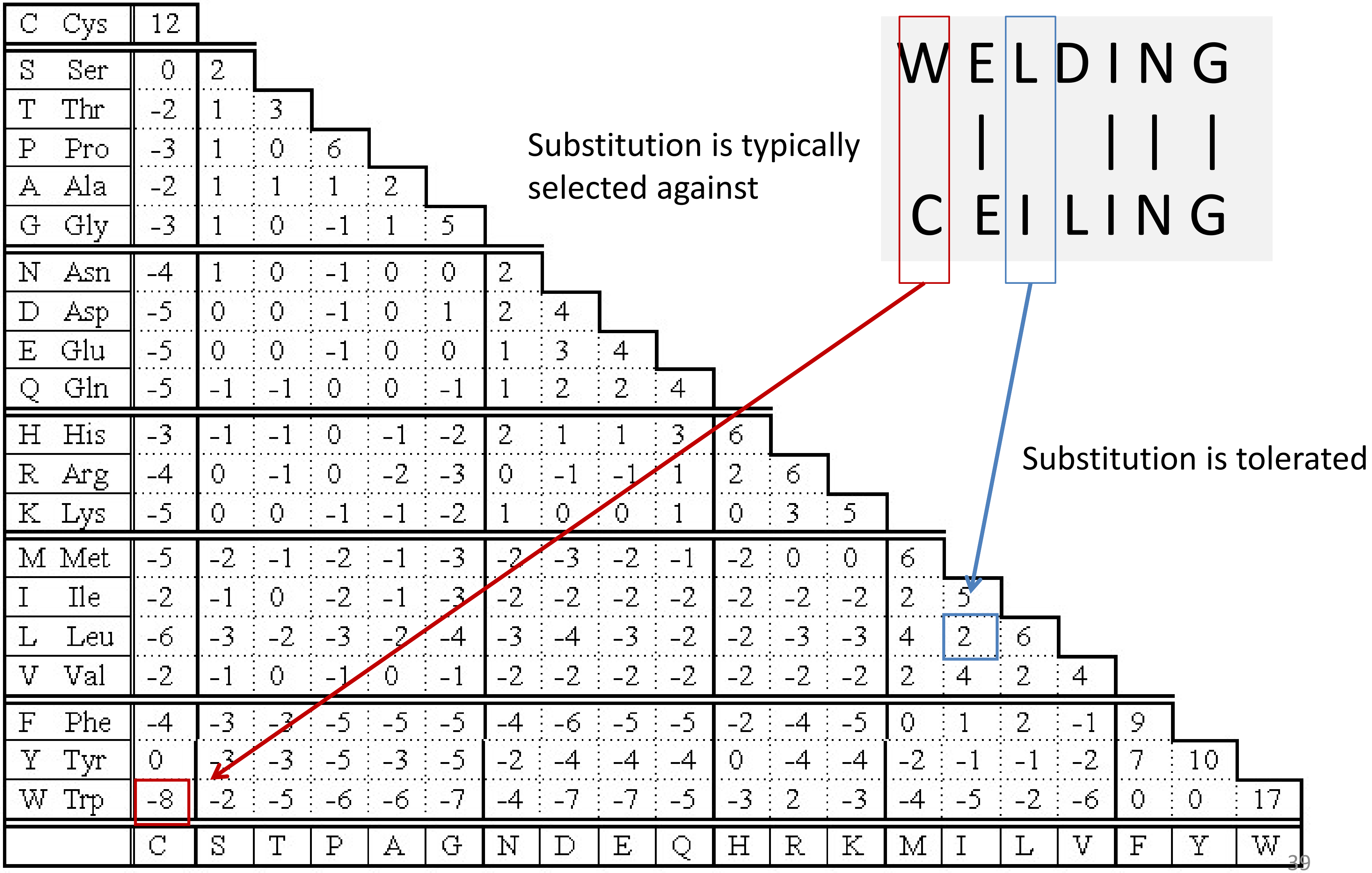
Score= 8*(+1)+2*(-1)+1*(-1)=5

- In reality, **some amino acid substitutions are more likely** than others **to be tolerated** during natural selection/evolution
- The **frequency of occurrence of the 20 amino acids** within proteins **varies** a lot

  Leucine, Isoleucine, Alanine are frequently found in proteins

  Tryptophan and Cysteine occur with less frequency

Scoring Matrix Example



C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

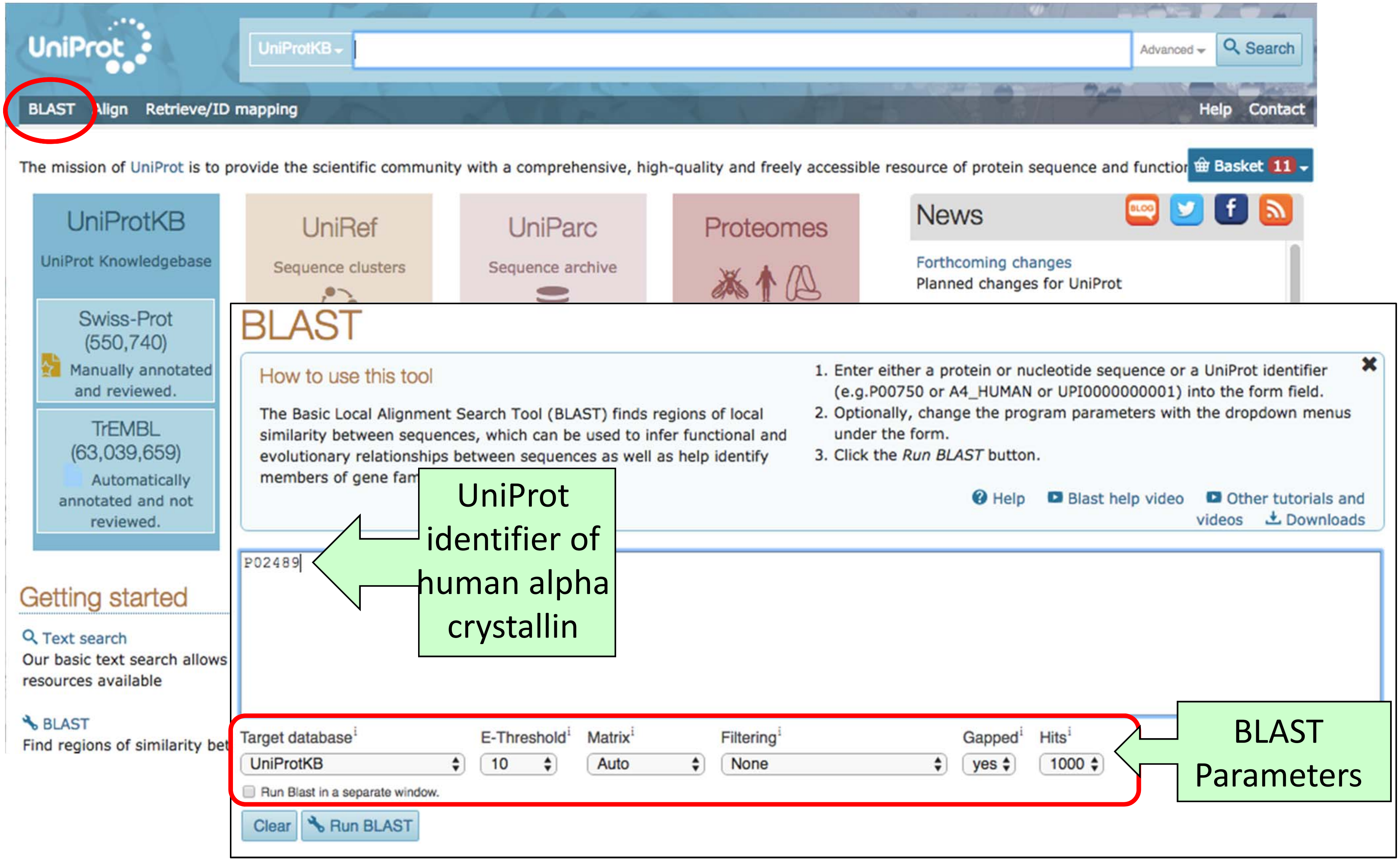
```
W E L D I N G
  |     | | |
C E I L I N G
```

Substitution is typically selected against

Substitution is tolerated

Running Protein BLAST from UniProt

(http://www.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 functio

Our basic text search allows
resources available

Find regions of similarity bet

BLAST

How to use this tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences, which can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene fam

1. Enter either a protein or nucleotide sequence or a UniProt identifier (e.g.P00750 or A4_HUMAN or UPI0000000001) into the form field.
2. Optionally, change the program parameters with the dropdown menus under the form.
3. Click the *Run BLAST* button.

```
P02489
```

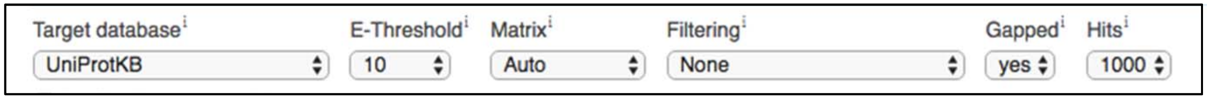
Target database	E-Threshold	Matrix	Filtering	Gapped	Hits
UniProtKB	10	Auto	None	yes	1000

□ Run Blast in a separate window.

UniProt identifier of human alpha crystallin

BLAST Parameters

Protein BLAST Parameters

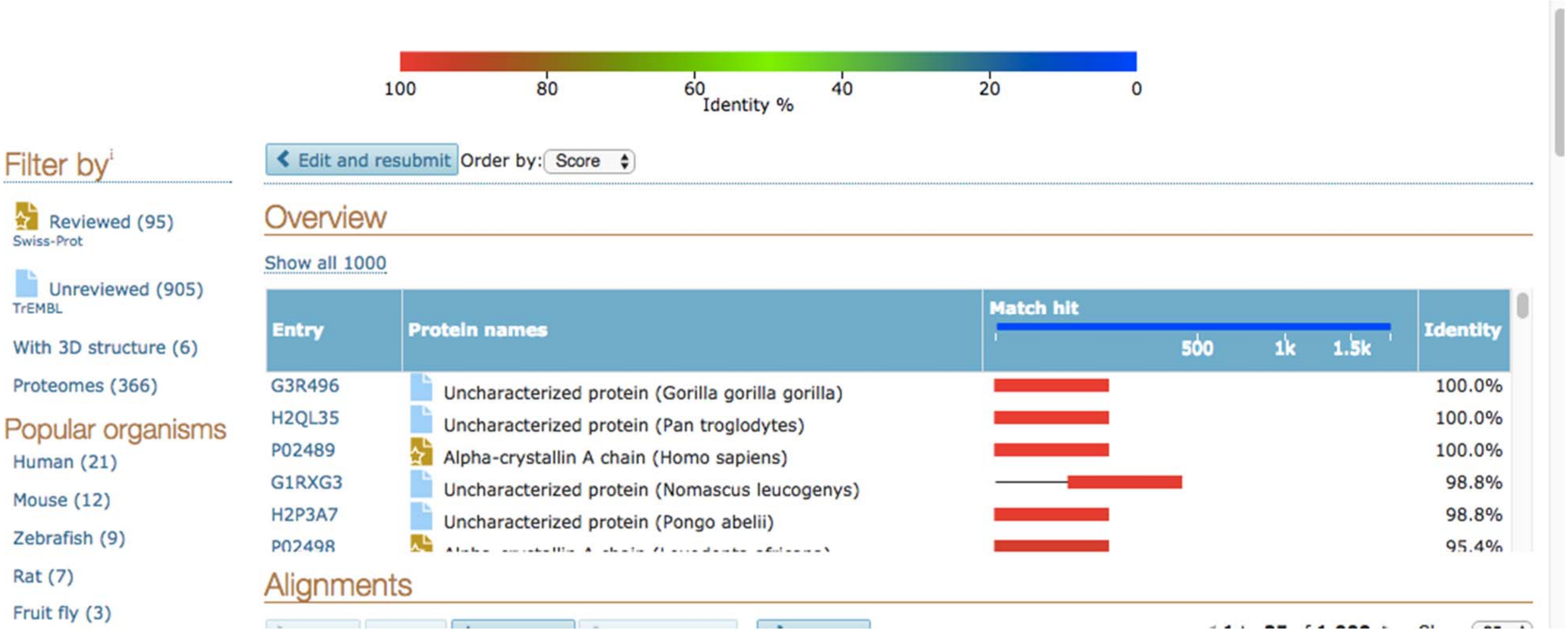


Target database: UniProtKB
E-Threshold: 10
Matrix: Auto
Filtering: None
Gapped: yes
Hits: 1000

Adjusting BLAST parameters can affect the outcome of your analysis.

- Target database: database against which the search is performed
- E-Threshold: maximum E-value that will be displayed
- Matrix: scoring matrix that will be used (common scoring matrices are call PAM or BLOSUM followed by a number (e.g., BLOSUM62)
- Filtering: allows you to filter out regions of low-complexity sequence that tend to give meaningless matches
- Gaps: sets whether gaps in alignments are allowed
- Hits: sets the number of returned hits

BLAST Results I



Filter by

Reviewed (95) Swiss-Prot

Unreviewed (905) TrEMBL

With 3D structure (6)

Proteomes (366)

Popular organisms

Human (21)

Mouse (12)

Zebrafish (9)

Rat (7)

Fruit fly (3)

Overview

Show all 1000

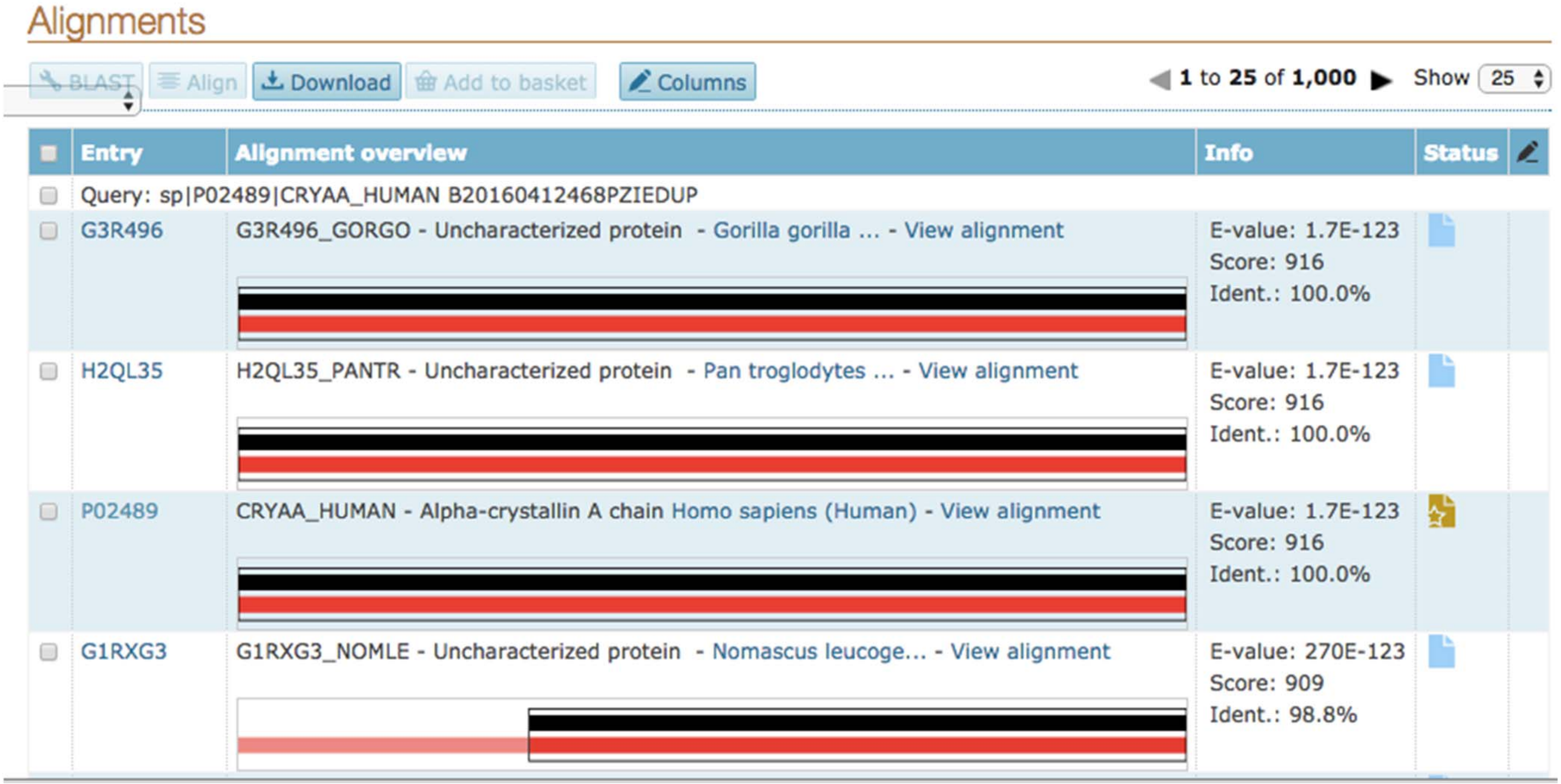
Entry	Protein names	Match hit	Identity
G3R496	Uncharacterized protein (Gorilla gorilla gorilla)	🟥	100.0%
H2QL35	Uncharacterized protein (Pan troglodytes)	🟥	100.0%
P02489	Alpha-crystallin A chain (Homo sapiens)	🟥	100.0%
G1RXG3	Uncharacterized protein (Nomascus leucogenys)	🟥	98.8%
H2P3A7	Uncharacterized protein (Pongo abelii)	🟥	98.8%
P02498	Alpha-crystallin A chain (Loxodonta africana)	🟥	95.4%

Alignments

- Overview box shows UniProt identifier, name, matched region, and percent identity ranked in order of score

- Filters can be used to restrict results to reviewed entries only or to particular organisms of interest

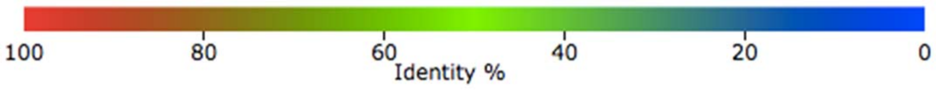
BLAST Results II



☐	Entry	Alignment overview	Info	Status
☐	Query: sp|P02489|CRYAA_HUMAN B20160412468PZIEDUP			
☐	G3R496	G3R496_GORGO - Uncharacterized protein - Gorilla gorilla ... - View alignment	E-value: 1.7E-123 Score: 916 Ident.: 100.0%	
☐	H2QL35	H2QL35_PANTR - Uncharacterized protein - Pan troglodytes ... - View alignment	E-value: 1.7E-123 Score: 916 Ident.: 100.0%	
☐	P02489	CRYAA_HUMAN - Alpha-crystallin A chain Homo sapiens (Human) - View alignment	E-value: 1.7E-123 Score: 916 Ident.: 100.0%	
☐	G1RXG3	G1RXG3_NOMLE - Uncharacterized protein - Nomascus leucoge... - View alignment	E-value: 270E-123 Score: 909 Ident.: 98.8%	

- Alignments panel shows matching regions as well as E-value, score, and percent identity for each result.

- Color indicates percent identity.



Hands On #3 – Protein BLAST

Go to http://www.uniprot.org and run protein BLAST on the following sequence. The sequence is an uncharacterized protein from *Latimeria chalumnae* (West Indian ocean coelacanth).

```
>Mystery Sequence
ALEYKCNINMTAETADCFNSSQITASEQEALVKPKQLLLKLLKCAGAQKDIFTMKEVIYY
LGQYIMAKQLYDKNQQHIVHCSNDLLGELFGVQSFSVKEPRRLYAMISKNLLPVNQEDPI
GIHVSMKETRCHRGSETGVKDNTQEVAGEKPAAPVTASCSTTSCRRTFSETEDAVSDDPL
SERRRKRHKSDSISLTFDDSLSWCVISGLRRDRSSSESTESPSNPDSDVVSVSENSKDSW
FDQDSDSDHFSVEFEVESVYSENYSDNEEAQDVTDEDDEFYQVTIYEAEDSDDSFTEDTE
ISVADYWTCTECEEVNPPLPRHCNRCWALRKDWLPENTKSSSCKSLDLKEPDREEGIDVP
DCKKTKEDPSCDSNVDVNEEDMTVQSSESQETNISQPSTSSSFIGGSQEESRETEREESS
ESTLPLTCLEPCVICQSRPKNGCIVHGRTGHLMACYTCAKKLKRRNKPCPVCRQPIQMVV
LTYFS
```

Latimeria chalumnae

Hands On #3 – Protein BLAST

Look at the BLAST results

1. What is the top result?

Filter by reviewed entries, and look over the top 10 results.

1. Do they have significant e-values?
2. Approximately what % identity do they have to the query protein?
3. Is the similarity over the full length of the protein?
4. What are the names of the top hits?

Do not close the window with the BLAST results!! You will need it again.

Hands On #3 – Protein BLAST (Answers)

Look at the BLAST results

1. What is the top result?

   H3APM8 Uncharacterized Protein from *Latimeria chalumnae*. This is the UniProt record for the sequence you input into BLAST. It shows 100% identity over the full length (as you would expect).

Filter by reviewed entries, and look over the top 10 results.

1. Do they have significant e-values? Yes—0 to 11E-165.
2. Approximately what % identity do they have to the query protein? 53%-62%
3. Is the similarity over the full length of the protein?

   For the first eight results, yes. In the last two cases, Q00987-8 and P56950-2 the match does not include the N-terminal region of the query protein.
4. What are the names of the top hits?

   E3 ubiquitin-protein ligase Mdm2

V. Multiple Sequence Alignment (MSA)

Multiple Sequence Alignment

- So far, we have talked about BLAST, which aligns pairs of sequences and comes up with a relatedness score based on how similar the amino acids are at each position.

Pairwise alignment:

Protein 1	a	b	a	c	d
Protein 2	a	b	e	c	d

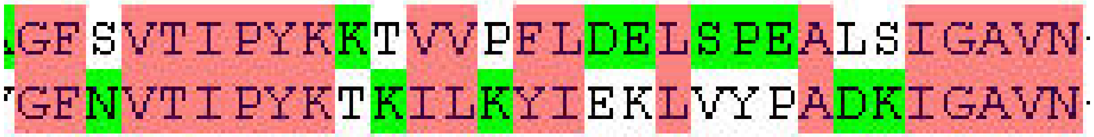
- Multiple sequence alignment (MSA) extends the same idea and provides more information

Multiple sequence alignment:

Protein 1	a	b	a	c	d
Protein 2	a	b	e	c	d
Protein 3	c	b	a	c	f
⋮					

Multiple Sequence Alignment

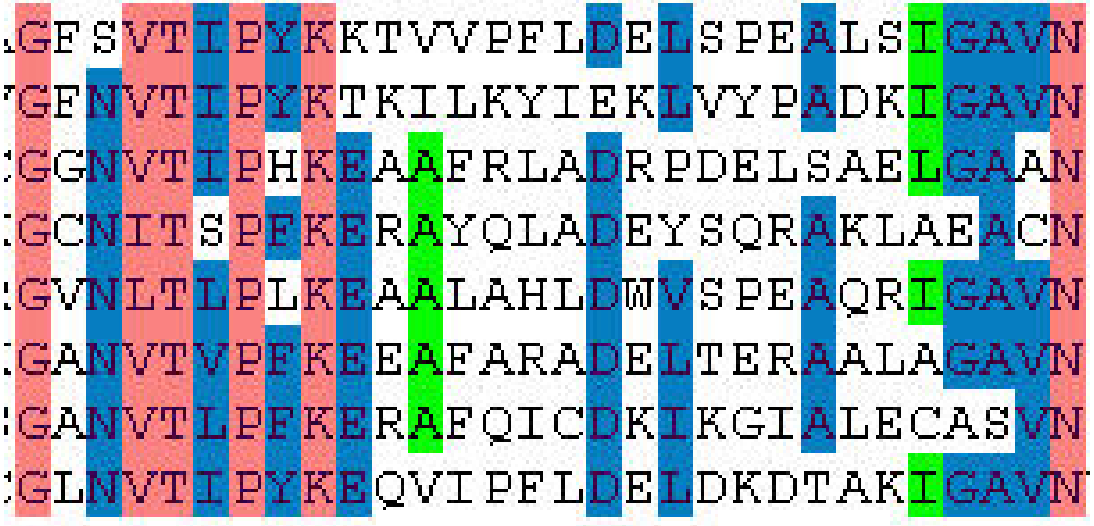
“Two homologous
sequences whisper…



```
GFSVTIPYKKTVVPFLDELSPEALSIGAVN
GFNVTIPYKTKILKYIEKLVYPADKIGAVN
```

A multiple sequence
alignment shouts”

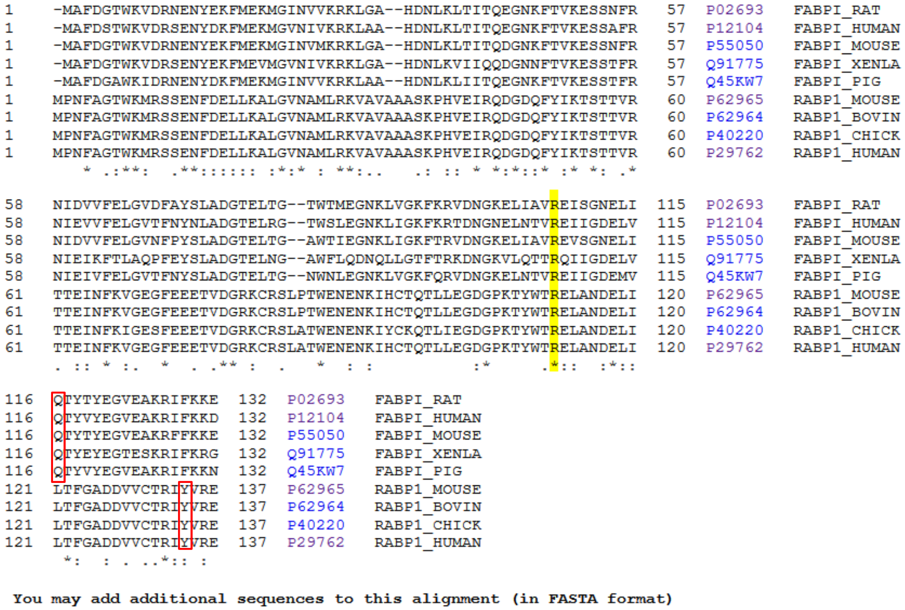
Prof. Arthur M Lesk



```
GFSVTIPYKKTVVPFLDELSPEALSIGAVN
GFNVTIPYKTKILKYIEKLVYPADKIGAVN
GGNVTIPHKEAAFRLADRPDELSAELGAAN
GCNITSPFKERAYQLADEYSQRAKLAEACN
GVNLTLPLKEAALAHLDWVSPEAQRIGAVN
GANVTVPFKEEAFARADELTERAALAGAVN
GANVTLPFKERAFQICDKIKGIALECASVN
GLNVTIPYKEQVIPFLDELDKDTAKIGAVN
```

Multiple Sequence Alignment

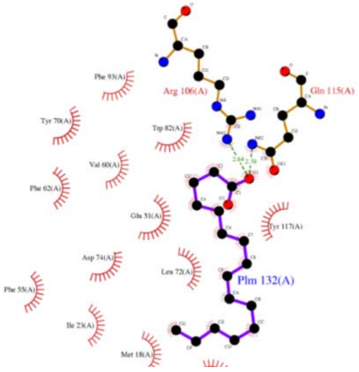
MSA can reveal patterns of conservation in sequences that allow us to determine which residues are under selective constraint (may be important for protein function)



```
1     -MAFDGTWKVDRNENYEKFMEKMGINVVKRKLGA--HDNLKLTITQEGNKFTVKESSNFR   57  P02693   FABPI_RAT
1     -MAFDSTWKVDRSENYDKFMEKMGVNIVKRKLAA--HDNLKLTITQEGNKFTVKESSAFR   57  P12104   FABPI_HUMAN
1     -MAFDGTWKVDRNENYEKFMEKMGINVMKRKLGA--HDNLKLTITQDGNKFTVKESSNFR   57  P55050   FABPI_MOUSE
1     -MAFDGTWKVDRSENYEKFMEVMGVNIVKRKLGA--HDNLKVIIQQDGNNFTVKESSTFR   57  Q91775   FABPI_XENLA
1     -MAFDGAWKIDRNENYDKFMEKMGINVVKRKLAA--HDNLKLIITQEGNKFTVKESSTFR   57  Q45KW7   FABPI_PIG
1     MPNFAGTWKMRSSENFDELLKALGVNAMLRKVAVAAASKPHVEIRQDGDQFYIKTSTTVR   60  P62965   RABP1_MOUSE
1     MPNFAGTWKMRSSENFDELLKALGVNAMLRKVAVAAASKPHVEIRQDGDQFYIKTSTTVR   60  P62964   RABP1_BOVIN
1     MPNFAGTWKMRSSENFDELLKALGVNAMLRKVAVAAASKPHVEIRQDGDQFYIKTSTTVR   60  P40220   RABP1_CHICK
1     MPNFAGTWKMRSSENFDELLKALGVNAMLRKVAVAAASKPHVEIRQDGDQFYIKTSTTVR   60  P29762   RABP1_HUMAN
         * .:**:  .**::::::: :*:* : **:..   .: :: * *:*::* :* *: .*

58    NIDVVFELGVDFAYSLADGTELTG--TWTMEGNKLVGKFKRVDNGKELIAVREISGNELI  115  P02693   FABPI_RAT
58    NIEVVFELGVTFNYNLADGTELRG--TWSLEGNKLIGKFKRTDNGNELNTVREIIGDELV  115  P12104   FABPI_HUMAN
58    NIDVVFELGVNFPYSLADGTELTG--AWTIEGNKLIGKFTRVDNGKELIAVREVSGNELI  115  P55050   FABPI_MOUSE
58    NIEIKFTLAQPFEYSLADGTELNG--AWFLQDNQLLGTFTRKDNGKVLQTTRQIIGDELV  115  Q91775   FABPI_XENLA
58    NIEIVFELGVTFNYSLADGTELTG--NWNLEGNKLVGKFQRVDNGKELNTVREIIGDEMV  115  Q45KW7   FABPI_PIG
61    TTEINFKVGEGFEEETVDGRKCRSLPTWENENKIHCTQTLLEGDGPKTYWTRELANDELI  120  P62965   RABP1_MOUSE
61    TTEINFKVGEGFEEETVDGRKCRSLPTWENENKIHCTQTLLEGDGPKTYWTRELANDELI  120  P62964   RABP1_BOVIN
61    TTEINFKIGESFEEETVDGRKCRSLATWENENKIYCKQTLIEGDGPKTYWTRELANDELI  120  P40220   RABP1_CHICK
61    TTEINFKVGEGFEEETVDGRKCRSLATWENENKIHCTQTLLEGDGPKTYWTRELANDELI  120  P29762   RABP1_HUMAN
      . :: * :.  *  . .** :  .   * : :          :*     .*:: :*::

116   QTYTYEGVEAKRIFKKE  132  P02693   FABPI_RAT
116   QTYVYEGVEAKRIFKKD  132  P12104   FABPI_HUMAN
116   QTYTYEGVEAKRFFKKE  132  P55050   FABPI_MOUSE
116   QTYEYEGTESKRIFKRG  132  Q91775   FABPI_XENLA
116   QTYVYEGVEAKRIFKKN  132  Q45KW7   FABPI_PIG
121   LTFGADDVVCTRIYVRE  137  P62965   RABP1_MOUSE
121   LTFGADDVVCTRIYVRE  137  P62964   RABP1_BOVIN
121   LTFGADDVVCTRIYVRE  137  P40220   RABP1_CHICK
121   LTFGADDVVCTRIYVRE  137  P29762   RABP1_HUMAN
       *:  : . ..*:: :

You may add additional sequences to this alignment (in FASTA format)
```

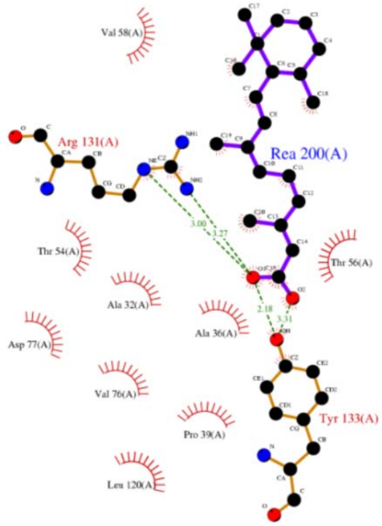
Gunasekaran et. al, 2004, Proteins: Structure, Function, and Bioinformatics, 54, 2:179-194.

FABP



Arg and Gln conserved in all FABPs

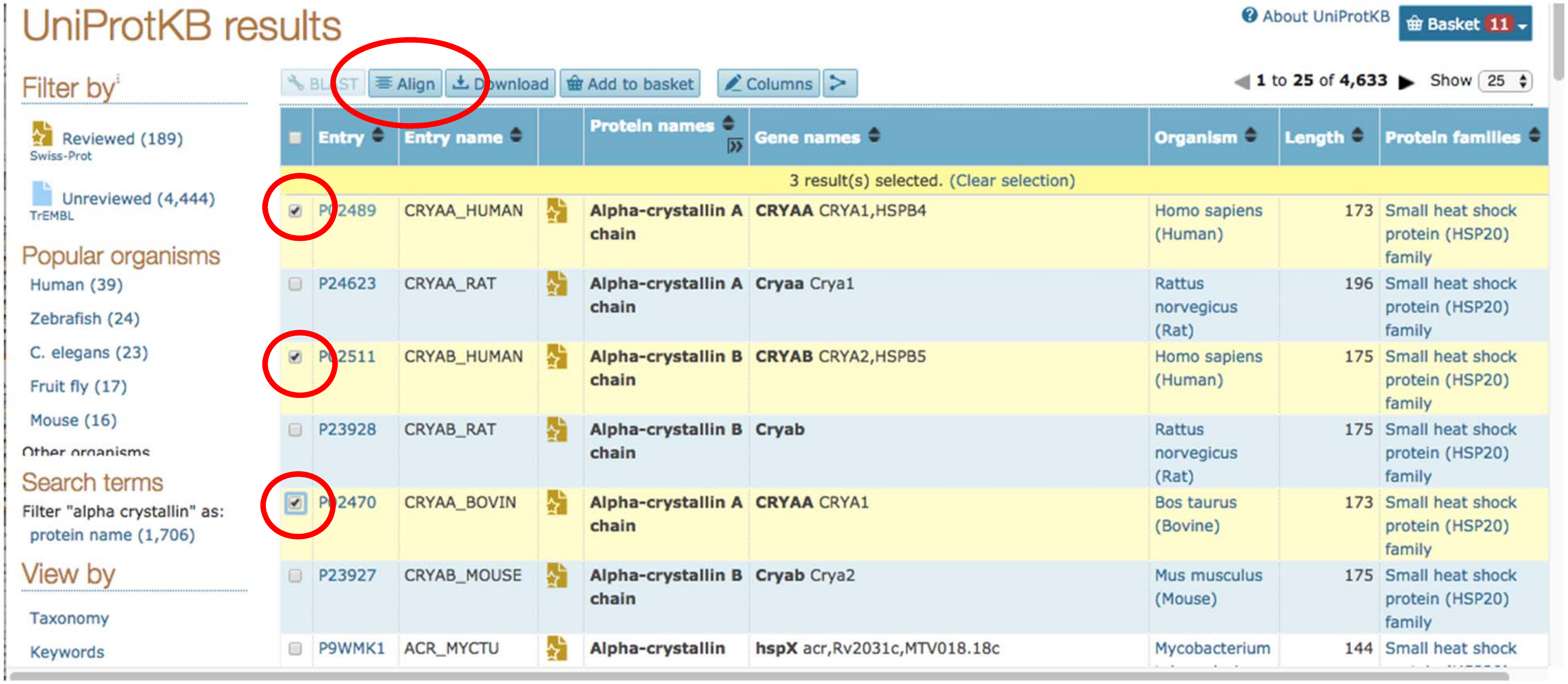
CRABP



Arg and Tyr conserved in all CRABPs

Performing MSA with UniProt

Select several UniProt search results to align…



UniProtKB results

? About UniProtKB | Basket 11

Filter by

Reviewed (189)
Swiss-Prot

Unreviewed (4,444)
TrEMBL

Popular organisms

Human (39)

Zebrafish (24)

C. elegans (23)

Fruit fly (17)

Mouse (16)

Other organisms

Search terms

Filter "alpha crystallin" as:
protein name (1,706)

View by

Taxonomy

Keywords

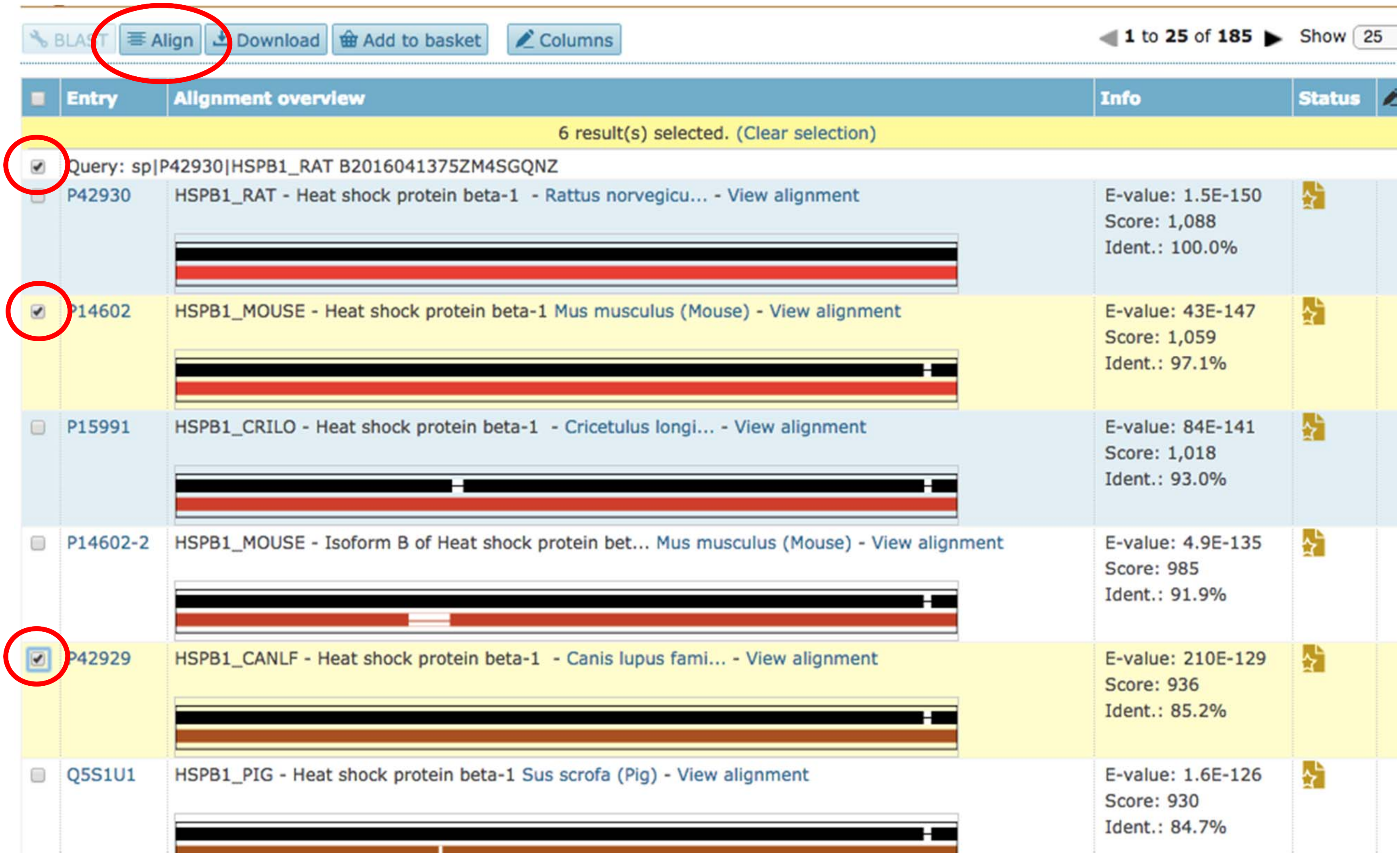
◀ 1 to 25 of 4,633 ▶ Show 25

3 result(s) selected. (Clear selection)

□	Entry	Entry name		Protein names	Gene names	Organism	Length	Protein families
☑	P02489	CRYAA_HUMAN		**Alpha-crystallin A chain**	**CRYAA** CRYA1,HSPB4	Homo sapiens (Human)	173	Small heat shock protein (HSP20) family
□	P24623	CRYAA_RAT		**Alpha-crystallin A chain**	**Cryaa** Crya1	Rattus norvegicus (Rat)	196	Small heat shock protein (HSP20) family
☑	P02511	CRYAB_HUMAN		**Alpha-crystallin B chain**	**CRYAB** CRYA2,HSPB5	Homo sapiens (Human)	175	Small heat shock protein (HSP20) family
□	P23928	CRYAB_RAT		**Alpha-crystallin B chain**	**Cryab**	Rattus norvegicus (Rat)	175	Small heat shock protein (HSP20) family
☑	P02470	CRYAA_BOVIN		**Alpha-crystallin A chain**	**CRYAA** CRYA1	Bos taurus (Bovine)	173	Small heat shock protein (HSP20) family
□	P23927	CRYAB_MOUSE		**Alpha-crystallin B chain**	**Cryab** Crya2	Mus musculus (Mouse)	175	Small heat shock protein (HSP20) family
□	P9WMK1	ACR_MYCTU		**Alpha-crystallin**	**hspX** acr,Rv2031c,MTV018.18c	Mycobacterium	144	Small heat shock

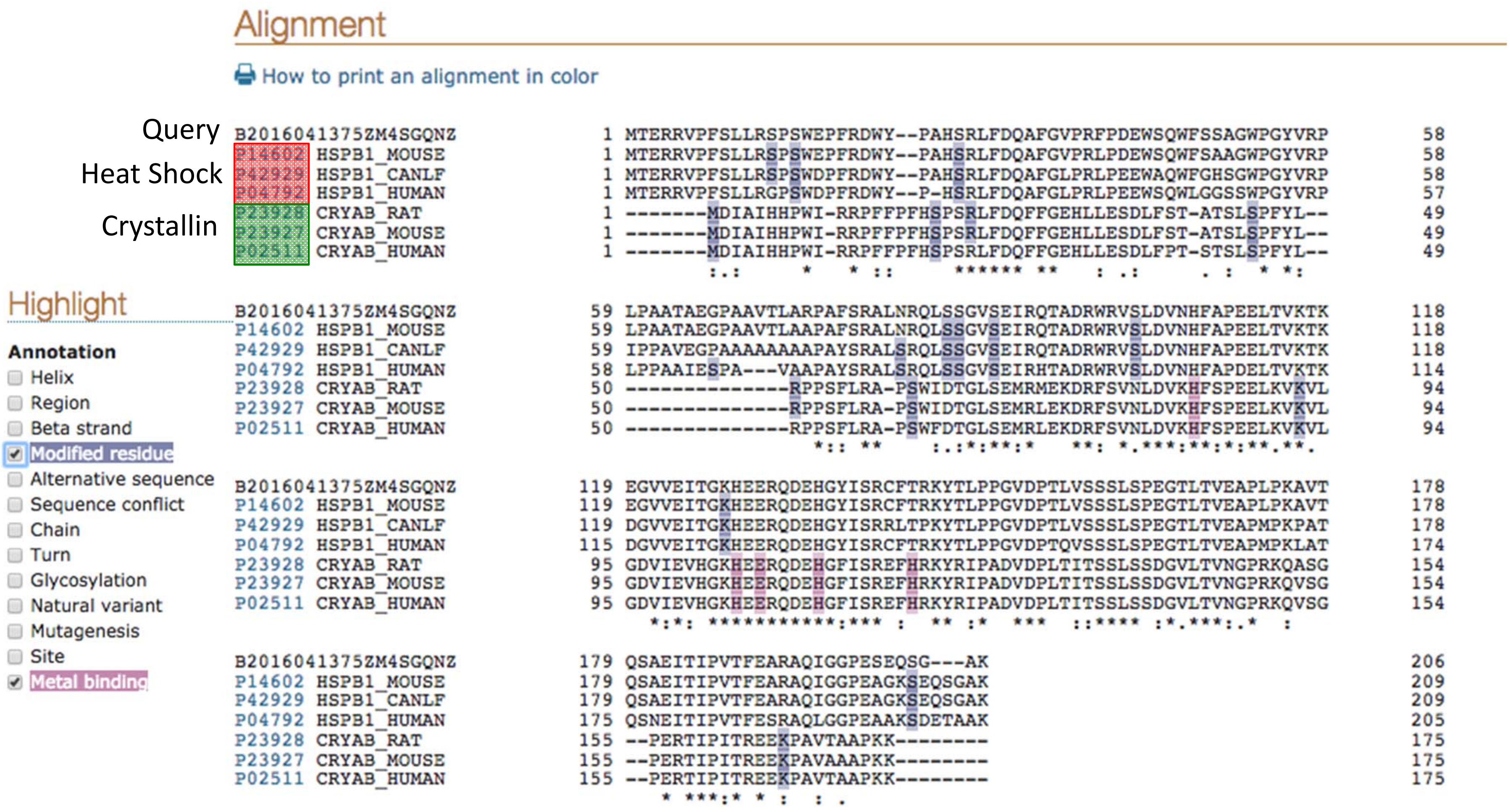
Performing MSA with UniProt

…or select several BLAST results to align



☐	Entry	Alignment overview	Info	Status
		6 result(s) selected. (Clear selection)		
☑	Query: sp|P42930|HSPB1_RAT B2016041375ZM4SGQNZ			
☐	P42930	HSPB1_RAT - Heat shock protein beta-1 - Rattus norvegicu... - View alignment	E-value: 1.5E-150 Score: 1,088 Ident.: 100.0%	
☑	P14602	HSPB1_MOUSE - Heat shock protein beta-1 Mus musculus (Mouse) - View alignment	E-value: 43E-147 Score: 1,059 Ident.: 97.1%	
☐	P15991	HSPB1_CRILO - Heat shock protein beta-1 - Cricetulus longi... - View alignment	E-value: 84E-141 Score: 1,018 Ident.: 93.0%	
☐	P14602-2	HSPB1_MOUSE - Isoform B of Heat shock protein bet... Mus musculus (Mouse) - View alignment	E-value: 4.9E-135 Score: 985 Ident.: 91.9%	
☑	P42929	HSPB1_CANLF - Heat shock protein beta-1 - Canis lupus fami... - View alignment	E-value: 210E-129 Score: 936 Ident.: 85.2%	
☐	Q5S1U1	HSPB1_PIG - Heat shock protein beta-1 Sus scrofa (Pig) - View alignment	E-value: 1.6E-126 Score: 930 Ident.: 84.7%	

UniProt Alignment Results



Query

Heat Shock

Crystallin

Alignment

How to print an alignment in color

```
B2016041375ZM4SGQNZ         1 MTERRVPFSLLRSPSWEPFRDWY--PAHSRLFDQAFGVPRFPDEWSQWFSSAGWPGYVRP     58
P14602 HSPB1_MOUSE          1 MTERRVPFSLLRSPSWEPFRDWY--PAHSRLFDQAFGVPRLPDEWSQWFSAAGWPGYVRP     58
P42929 HSPB1_CANLF          1 MTERRVPFSLLRSPSWDPFRDWY--PAHSRLFDQAFGLPRLPEEWAQWFGHSGWPGYVRP     58
P04792 HSPB1_HUMAN          1 MTERRVPFSLLRGPSWDPFRDWY--P-HSRLFDQAFGLPRLPEEWSQWLGGSSWPGYVRP     57
P23928 CRYAB_RAT            1 -------MDIAIHHPWI-RRPFFPFHSPSRLFDQFFGEHLLESDLFST-ATSLSPFYL--     49
P23927 CRYAB_MOUSE          1 -------MDIAIHHPWI-RRPFFPFHSPSRLFDQFFGEHLLESDLFST-ATSLSPFYL--     49
P02511 CRYAB_HUMAN          1 -------MDIAIHHPWI-RRPFFPFHSPSRLFDQFFGEHLLESDLFPT-STSLSPFYL--     49
                                     :.:     *   * ::     ****** **   : .:     . :  * *:

B2016041375ZM4SGQNZ        59 LPAATAEGPAAVTLARPAFSRALNRQLSSGVSEIRQTADRWRVSLDVNHFAPEELTVKTK    118
P14602 HSPB1_MOUSE         59 LPAATAEGPAAVTLAAPAFSRALNRQLSSGVSEIRQTADRWRVSLDVNHFAPEELTVKTK    118
P42929 HSPB1_CANLF         59 IPPAVEGPAAAAAAAAPAYSRALSRQLSSGVSEIRQTADRWRVSLDVNHFAPEELTVKTK    118
P04792 HSPB1_HUMAN         58 LPPAAIESPA---VAAPAYSRALSRQLSSGVSEIRHTADRWRVSLDVNHFAPDELTVKTK    114
P23928 CRYAB_RAT           50 --------------RPPSFLRA-PSWIDTGLSEMRMEKDRFSVNLDVKHFSPEELKVKVL     94
P23927 CRYAB_MOUSE         50 --------------RPPSFLRA-PSWIDTGLSEMRLEKDRFSVNLDVKHFSPEELKVKVL     94
P02511 CRYAB_HUMAN         50 --------------RPPSFLRA-PSWFDTGLSEMRLEKDRFSVNLDVKHFSPEELKVKVL     94
                                              *:: **    :.:*:**:*   **: *.***:**:*:**.**.

B2016041375ZM4SGQNZ       119 EGVVEITGKHEERQDEHGYISRCFTRKYTLPPGVDPTLVSSSLSPEGTLTVEAPLPKAVT    178
P14602 HSPB1_MOUSE        119 EGVVEITGKHEERQDEHGYISRCFTRKYTLPPGVDPTLVSSSLSPEGTLTVEAPLPKAVT    178
P42929 HSPB1_CANLF        119 DGVVEITGKHEERQDEHGYISRRLTPKYTLPPGVDPTLVSSSLSPEGTLTVEAPMPKPAT    178
P04792 HSPB1_HUMAN        115 DGVVEITGKHEERQDEHGYISRCFTRKYTLPPGVDPTQVSSSLSPEGTLTVEAPMPKLAT    174
P23928 CRYAB_RAT           95 GDVIEVHGKHEERQDEHGFISREFHRKYRIPADVDPLTITSSLSSDGVLTVNGPRKQASG    154
P23927 CRYAB_MOUSE         95 GDVIEVHGKHEERQDEHGFISREFHRKYRIPADVDPLTITSSLSSDGVLTVNGPRKQVSG    154
P02511 CRYAB_HUMAN         95 GDVIEVHGKHEERQDEHGFISREFHRKYRIPADVDPLTITSSLSSDGVLTVNGPRKQVSG    154
                                *:*: ***********:*** :  ** :*  ***  ::**** :*.***:.*  :

B2016041375ZM4SGQNZ       179 QSAEITIPVTFEARAQIGGPESEQSG---AK                                206
P14602 HSPB1_MOUSE        179 QSAEITIPVTFEARAQIGGPEAGKSEQSGAK                                209
P42929 HSPB1_CANLF        179 QSAEITIPVTFEARAQIGGPEAGKSEQSGAK                                209
P04792 HSPB1_HUMAN        175 QSNEITIPVTFESRAQLGGPEAAKSDETAAK                                205
P23928 CRYAB_RAT          155 --PERTIPITREEKPAVTAAPKK--------                                175
P23927 CRYAB_MOUSE        155 --PERTIPITREEKPAVAAAPKK--------                                175
P02511 CRYAB_HUMAN        155 --PERTIPITREEKPAVTAAPKK--------                                175
                                 * ***:* * :  : .
```

Highlight

Annotation

- □ Helix
- □ Region
- □ Beta strand
- ☑ Modified residue
- □ Alternative sequence
- □ Sequence conflict
- □ Chain
- □ Turn
- □ Glycosylation
- □ Natural variant
- □ Mutagenesis
- □ Site
- ☑ Metal binding

- Modified residues of heat shock group are also found in query
- 3 of 4 metal binding residues of crystallin group are conserved in heat shock group
  -> Maybe heat shock group also binds metal at these sites?

Hands On #4 - MSA

- Go back to the results page for your BLAST of the *Latimeria chalumnae* uncharacterized protein.
- If you have not already done so, filter results for reviewed entries
- Select the query sequence and the top five BLAST results and perform an alignment.
- Experiment with highlighting the alignment according to different annotations or amino acid properties. Observe whether your mystery sequence is conserved in the highlighted regions.

Hands On #4 – MSA Part II

Highlight the alignment according to the mutagenesis annotation. (This means that the UniProt entry has information about mutagenesis experiments for these residues)

1. Find the highlighted residue at position 374 of the human sequence Q00987. Is this residue conserved in the mystery protein?
2. In a separate tab, go to the UniProtKB record for Q00987. What was the consequence of mutagenesis at position 374?
3. In the alignment, find the highlighted residues at positions 452, 455, and 457. Are these conserved in the mystery protein?
4. What are the consequences of mutagenesis of positions 452, 455, and 457?
5. Based on these results, do you think it is possible that the mystery protein has ubiquitin ligase activity like the human protein Q00987?

Hands On #4 – MSA Part II (Answers)

Highlight the alignment according to the mutagenesis annotation. (This means that the UniProt entry has information about mutagenesis experiments for these residues)

1. Find the highlighted residue at position 374 of the human sequence Q00987. Is this residue conserved in the mystery protein? No
2. In a separate tab, go to the UniProtKB record for Q00987. What was the consequence of mutagenesis at position 374? No loss of ubiquitin ligase activity.
3. In the alignment, find the highlighted residues at positions 452, 455, and 457. Are these conserved in the mystery protein? Yes
4. What are the consequences of mutagenesis of positions 452, 455, and 457? Loss or significant decrease in ubiquitin ligase activity.

Hands On #4 – MSA Part II (Answers)

5. Based on these results, do you think it is possible that the mystery protein has ubiquitin ligase activity like the human protein Q00987?

   At least some of the residues that are important for ubiquitin ligase activity (452, 455, and 457) are conserved in the mystery protein. The one residue that we checked that was not conserved (374) seems to be less important for activity. These results are consistent with the possibility that the mystery protein has ubiquitin ligase activity, but we would need to check other critical residues and ultimately do experiments on the mystery protein to see if it really does have the activity.

Take Home Messages

- Pubmed (http://www.ncbi.nlm.nih.gov/pubmed) is an excellent resource for searching high-quality scientific literature. Using advanced querying techniques can help to target your searches to articles you are most interested in.

- UniProtKB (http://www.uniprot.org/) is a centralized resource for protein sequence and function information.

- Protein BLAST and multiple sequence alignments (MSA) can help in assigning functions to uncharacterized proteins and in determining evolutionary relationships among proteins.

More Resources

- UniProtKB tutorials on YouTube:
  https://www.youtube.com/user/uniprotvideos

- A good introductory paper on BLAST:
  Using BLAST to Teach "E-value-tionary" Concepts
  Cheryl A. Kerfeld and Kathleen M. Scott
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3032543/

- Our contact information:
  Cecilia Arighi: arighi@dbi.udel.edu
  Hongzhan Huang: huang@dbi.udel.edu
  Karen Ross: ker25@georgetown.edu
  C.R. Vinayaka: CR.Vinayaka@georgetown.edu
  Cathy Wu: wuc@dbi.udel.edu