

---

# Master Biomedizin 2019

- 1) UCSC & UniProt
- 2) Homology
- 3) MSA
- 4) Phylogeny

---

## 2) Homology

## 3

- a. Using the human protein “P21741”, find its orthologous proteins in frog (*Xenopus laevis*) and get their UniProt AC. P48530, P48531
- b. Check the identity between the orthologs (human – frog proteins).  
P21741-P48530 = 61.1%, P21741-P48531 = 60.4%
- c. Check the identity between the paralogs (frog – frog proteins). P48530-P48531 = 97.9%

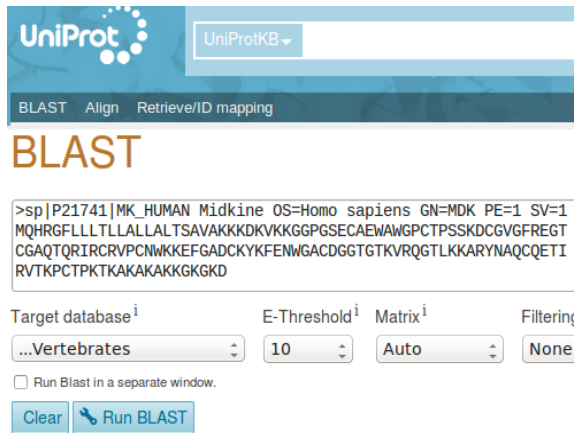


Human  
(*Homo sapiens*)



Frog  
(*Xenopus laevis*)

3



UniProtKB

BLAST Align Retrieve/ID mapping

**BLAST**

>sp|P21741|MK\_HUMAN Midkine OS=Homo sapiens GN=MDK PE=1 SV=1  
MQHRGFLLLTLLALLTSAVAKKDKVKKGGPGSECAEWAGPCTPSSKDCGVGFREGT  
CGAQTQIRICRVPCNMWKEFGADCKYKFENWGACDGGTGTQVRQGLKKARYNAQCQETI  
RVTKPCTPKTKAKAKKGGKGD

Target database<sup>i</sup> E-Threshold<sup>i</sup> Matrix<sup>i</sup> Filtering





...Vertebrates 10 Auto None

☐ Run Blast in a separate window.



Clear Run BLAST

Two options to filter results:

- View by → Taxonomy view → “*Xenopus laevis*”
- Filter by → Reviewed

P48530	 <b>Midkine-A</b> (Xenopus laevis)		61.1%
P48531	 <b>Midkine-B</b> (Xenopus laevis)		60.4%

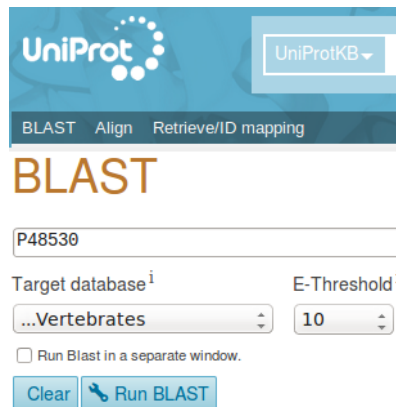
  

<input checked="" type="checkbox"/>	MKA_XENLA - Midkine-A - Xenopus laevis ... - View alignment	E-value: 3.1e-65 Score: 510 Ident.: 61.1%	
<input checked="" type="checkbox"/>	MKB_XENLA - Midkine-B - Xenopus laevis ... - View alignment	E-value: 4.4e-65 Score: 509 Ident.: 60.4%	

a. Query: P21741.  
Ortholog1: P48530.  
Ortholog2: P48531.

b. P21741-P48530 = 61.1%  
P21741-P48531 = 60.4%

c. P48530-P48531 = 97.9%  
Note: may also be done with “alignments”.



UniProtKB

BLAST Align Retrieve/ID mapping

**BLAST**







P48530

Target database<sup>i</sup> E-Threshold<sup>i</sup>

...Vertebrates 10

☐ Run Blast in a separate window.

Clear Run BLAST

Entry	Protein names	Match hit	Identity
P48530	 Midkine-A (Xenopus laevis)		100.0%
Q6P8F3	 Midkine (Xenopus tropicalis)		98.6%
P48531	 Midkine-B (Xenopus laevis)		97.9%



## 4

a. Based on the sequence of the “ATP synthase subunit a” protein from the extinct mammoth (*Mammuthus primigenius*) [Q38PR7], was the mammoth closer to the asian elephant (*Elephas maximus*) or to the african elephant (*Loxodonta africana*)? Use only SwissProt proteins.

*M. primigenius* (Q38PR7) – *E. maximus* (Q2I3G9) = 95.5%

*M. primigenius* (Q38PR7) – *L. africana* (Q9TA24) = 93.2%

b. Is there evidence enough to conclude if they are / are not closer? No.

c. Could you check with the “cytochrome b” protein too? [P92658] Use only SwissProt proteins.

*M. primigenius* (P92658) – *E. maximus* (O47885) = 96.3%

*M. primigenius* (P92658) – *L. africana* (P24958) = 97.9%



Woolly mammoth  
(*Mammuthus primigenius*)



Asian elephant  
(*Elephas maximus*)



African elephant  
(*Loxodonta africana*)

4

UniProtKB

BLAST Align Retrieve/ID mapping

BLAST

>sp|Q38PR7|ATP6\_MAMPR ATP synthase subunit a OS=Mammuthus primigenius GN=MT-ATP6 PE=3 SV=1  
MNEELSAFFDVPVGTMMIAIAPPAILLPTPNRLITNRWITIQQWLVKIMKQLLSIHNTK  
GLSWSLMLITLTLFIGLTNLLGLLPYSFAPTAQLTVNLSMAIPLWTGTVLGFRYKTKIS  
LAHLLPQGTPTFLIPMIIIIETISLLIRPVTAVRLTANITAGHLLIHLTGTAALTLISI  
HSMITITVTFITVVVLTILELAVALIQAYVFALLISLYHESA

Target database<sup>i</sup> UniProtKB/Swiss-Prot<sup>i</sup> E-Threshold<sup>i</sup> 10<sup>i</sup> Matrix<sup>i</sup> Auto<sup>i</sup> Filtering<sup>i</sup> None<sup>i</sup> Gapped<sup>i</sup> yes<sup>i</sup>

Clear Run BLAST

UniProtKB

BLAST Align Retrieve/ID mapping

BLAST

>sp|P92658|CYB\_MAMPR Cytochrome b OS=Mammuthus primigenius GN=MT-CYB PE=3 SV=3  
MTHIRKSHPLKILNKSFIDLPSTNISTWNNFGSLLGACLTITQILTGLFLAMHYTPDTM  
TAFSSMSHICRDVNYGWIIRQLHNSGASIFFLCLYTHIGRNIYYGSYSETWNTGIMLL  
LITMATAFMGYVLPWGQMSFWGATVITNLSAIPYIGTDLVEWIGGFSVDKATLNRFFA  
LHFILPFTMIALAGVHLTFLEHETGSNNPLGLTSDSDKIPFPHYTIKDFLGLLILILFLL  
LLALLSPDMLGDPDNYMPADPLNPLHIKPEWYFLFAYAILRSVPNKLGGVLALLSILI  
LGIMPLLHTSKHRSMLRPLSQVLFWTLATDMLLMTWIGSQPVVEPYIIGQMASILYFS  
IILAFPLIAGMIENYLIK

Target database<sup>i</sup> UniProtKB/Swiss-Prot<sup>i</sup> E-Threshold<sup>i</sup> 10<sup>i</sup> Matrix<sup>i</sup> Auto<sup>i</sup> Filtering<sup>i</sup> None<sup>i</sup> Gapped<sup>i</sup> yes<sup>i</sup>

Clear Run BLAST

Entry	Protein names	Match hit	Identity
Q38PR7	ATP synthase subunit a (Mammuthus primigenius)	<div><div></div></div>	100.0%
Q2I3G9	ATP synthase subunit a (Elephas maximus)	<div><div></div></div>	95.5%
Q9TA24	ATP synthase subunit a (Loxodonta africana)	<div><div></div></div>	93.2%

- a. *M. primigenius* (Q38PR7) – *E. maximus* (Q2I3G9) = 95.5%  
*M. primigenius* (Q38PR7) – *L. africana* (Q9TA24) = 93.2%

b. Just this sequence similarity is not evidence enough for claiming the mammoth is closer to the asian elephant than to the african elephant,  
 BUT

the last genome sequencing work on the woolly mammoth (PMID: 19020620), in 2008, provides evidence enough to determine that it is really closer to the asian elephant; corroborating the similarity shown in exercise 4a.

- c. Different results! (read “b” again...)  
*M. primigenius* (P92658) – *E. maximus* (O47885) = 96.3%  
*M. primigenius* (P92658) – *L. africana* (P24958) = 97.9%

Entry	Protein names	Match hit	Identity
P92658	Cytochrome b (Mammuthus primigenius)	<div><div></div></div>	100.0%
P24958	Cytochrome b (Loxodonta africana)	<div><div></div></div>	97.9%
O47885	Cytochrome b (Elephas maximus)	<div><div></div></div>	96.3%

## 5

a. Based solely on the sequence of the “Cytochrome b” protein (Q8SG72) from the extinct dodo (*Raphus cucullatus*), was the dodo closer to the Nicobar pigeon “*Caloenas nicobarica*” or to the chicken (*Gallus gallus*)? Use NCBI Blast.

*R. cucullatus* – *C. nicobarica* = 98.50%

*R. cucullatus* – *G. gallus* = 92.51%

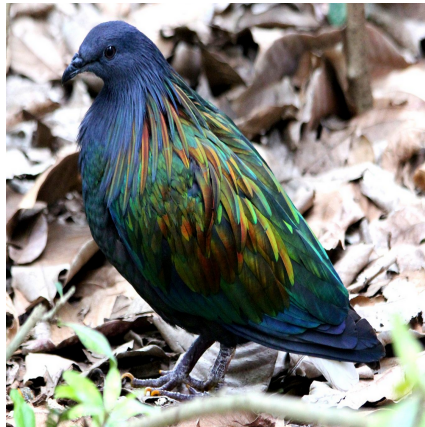
b. There are more than 300 species of pigeons. Do the results differ if you consider the street pigeon (*Columba livia*)?

*R. cucullatus* – *C. livia* = 95.88%

*R. cucullatus* – *G. gallus* = 92.51%



Dodo  
(*Raphus cucullatus*)



Nicobar pigeon  
(*Caloenas nicobarica*)



Chicken (rooster)  
(*Gallus gallus*)



Pigeon  
(*Columba livia*)

5

- a. It seems that the dodo was closer to the pigeon than to the chicken.

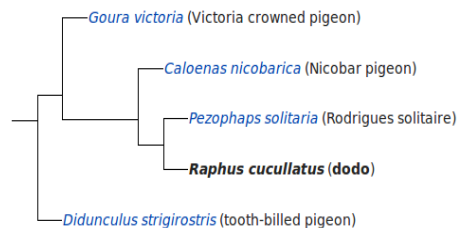
*R. cucullatus* – *C. nicobarica* = 98.50%

*R. cucullatus* – *G. gallus* = 92.51%

- b. Same results for different pigeons.

*R. cucullatus* – *C. livia* = 95.88%

*R. cucullatus* – *G. gallus* = 92.51%



	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input checked="" type="checkbox"/>	<a href="#">cytochrome b [Caloenas nicobarica]</a>	536	536	100%	0.0	98.50%	<a href="#">ATZ68926.1</a>
<input type="checkbox"/>	<a href="#">cytochrome b [Caloenas nicobarica]</a>	535	535	100%	0.0	98.50%	<a href="#">AAM19503.1</a>
<input type="checkbox"/>	<a href="#">cytochrome b [Caloenas nicobarica]</a>	534	534	100%	0.0	98.13%	<a href="#">YP_009318465.1</a>
<input checked="" type="checkbox"/>	<a href="#">cytochrome b [Columba livia]</a>	526	526	100%	0.0	95.88%	<a href="#">YP_003540719.1</a>
<input type="checkbox"/>	<a href="#">cytochrome b [Columba livia]</a>	522	522	100%	0.0	95.13%	<a href="#">AJK30555.1</a>
<input type="checkbox"/>	<a href="#">cytochrome b [Columba livia]</a>	521	521	100%	0.0	95.13%	<a href="#">AKB93366.1</a>
<input checked="" type="checkbox"/>	<a href="#">cytochrome b [Gallus gallus]</a>	509	509	100%	0.0	92.51%	<a href="#">ADB06697.1</a>



## 6

**a.** The UniProt entry “P04585” contains the Gag-Pol polyprotein from the virus HV1H2. Do you think it would resemble any protein in the proteome of the Zebra finch (*Taeniopygia guttata*)? Check it using NCBI Blast.

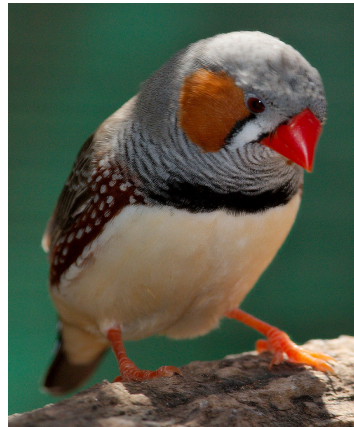
XP\_012432209.1. It has 26.99% identity with an endogenous retrovirus in *T. guttata*'s genome.

**b.** Discuss the results. What is the query coverage telling us?

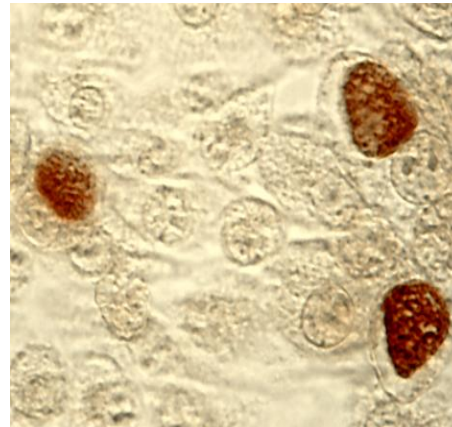
The query coverage is 50%, meaning that the viral “pol” protein (C-terminal) is integrated, while the “gag” protein (N-terminal) is not.

**c.** The Gag-Pol polyprotein is composed of many proteins. Using only protein entries from the bacteria “*Chlamydia trachomatis*”, can you identify some of the individual proteins of the Gag-Pol polyprotein?

P24, Reverse transcriptase, Ribonuclease H



Zebra finch  
(*Taeniopygia guttata*)



*Chlamydia trachomatis*

# Homology

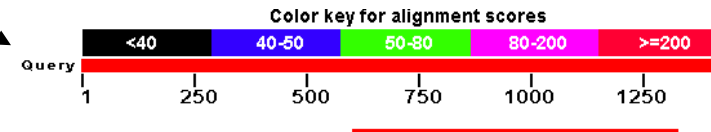
\*Images from: NCBI

6

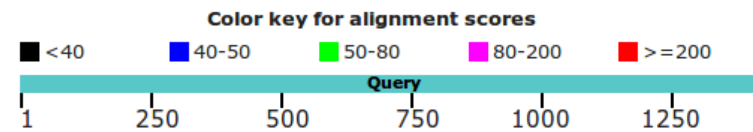
Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> PREDICTED: endogenous retrovirus group K member 18 Pol protein-like [Taeniopygia guttata]	240	240	50%	1e-67	26.99%	XP_012432209.1

a. XP\_012432209.1. It has 26.99% identity with an endogenous retrovirus in *T. guttata*'s genome.

b. The query coverage is 50%, meaning that the viral “pol” protein (C-terminal) is integrated, while the “gag” protein (N-terminal) is not.



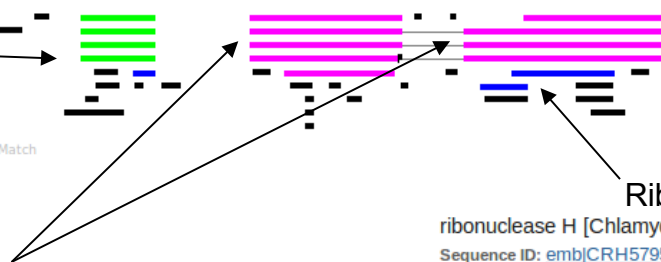
c. Database: protein entries from “*Chlamydia trachomatis*”.



P24

gag gene protein p24 (core nucleocapsid protein) [*Chlamydia trachomatis*]  
Sequence ID: [emb|CRH58881.1](#) Length: 382 Number of Matches: 1

Score	Expect	Method	Identities	Positives	Gaps
74.3 bits(181)	2e-13	Compositional matrix adjust.	52/152(34%)	73/152(48%)	20/152(13%)



Ribonuclease H

ribonuclease H [*Chlamydia trachomatis*]  
Sequence ID: [emb|CRH57958.1](#) Length: 831 Number of Matches: 1

Score	Expect	Method	Identities	Positives	Gaps
40.8 bits(94)	0.007	Compositional matrix adjust.	57/222(26%)	92/222(41%)	24/222(10%)

Reverse transcriptase

Reverse transcriptase (RNA-dependent DNA polymerase) [*Chlamydia trachomatis*]  
Sequence ID: [emb|CRH45814.1](#) Length: 867 Number of Matches: 2

Score	Expect	Method	Identities	Positives	Gaps
164 bits(414)	6e-41	Compositional matrix adjust.	105/295(36%)	157/295(53%)	8/295(2%)

Score	Expect	Method	Identities	Positives	Gaps
97.1 bits(240)	4e-20	Compositional matrix adjust.	105/431(24%)	178/431(41%)	57/431(13%)

## 7

Using the protein “P38398”, perform a “tblastn” search in NCBI against human entries.

a. What would this search be used for? To look for the gene encoding the query protein.

b. Is there any difference between the second and the third result?

Second result: NM\_007294.3  
7224 bp, transcript variant 1, mRNA

Third result: U14680.1  
5711bp, complete CDS

## 7

a. Query: protein. Database: nucleotide. To look for the gene encoding the query protein.

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	<a href="#">Synthetic construct DNA, clone: pF1KB5593, Homo sapiens BRCA1 gene for breast cancer type 1 susceptibility protein, c</a>	3578	3578	100%	0.0	93.72%	<a href="#">AB385129.1</a>
<input checked="" type="checkbox"/>	<a href="#">Homo sapiens BRCA1 DNA repair associated (BRCA1), transcript variant 1, mRNA</a>	3577	3577	100%	0.0	93.72%	<a href="#">NM_007294.3</a>
<input checked="" type="checkbox"/>	<a href="#">Homo sapiens breast and ovarian cancer susceptibility (BRCA1) mRNA, complete cds</a>	3577	3577	100%	0.0	93.72%	<a href="#">U14680.1</a>

Homo sapiens BRCA1 DNA repair associated (BRCA1), transcript variant 1, mRNA

Sequence ID: [NM\\_007294.3](#) Length: 7224 Number of Matches: 1

Range 1: 233 to 5821 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
3577 bits(9275)	0.0	Compositional matrix adjust.	1863/1863(100%)	1863/1863(100%)	0/1863(0%)	+2
Query 1	MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQ	60				
Sbjct 233	MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQ	412				



b. Second result: NM\_007294.3  
7224 bp, transcript variant 1, mRNA

Query: 1 \_\_\_\_\_ 1863  
Subject: 233 \_\_\_\_\_ 5821

Homo sapiens breast and ovarian cancer susceptibility (BRCA1) mRNA, complete cds

Sequence ID: [U14680.1](#) Length: 5711 Number of Matches: 1

Range 1: 120 to 5708 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
3577 bits(9275)	0.0	Compositional matrix adjust.	1863/1863(100%)	1863/1863(100%)	0/1863(0%)	+3
Query 1	MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQ	60				
Sbjct 120	MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQ	299				



Third result: U14680.1  
5711bp, complete CDS

Query: 1 \_\_\_\_\_ 1863  
Subject: 120 \_\_\_\_\_ 5708