

---

# **Master Biomedizin 2018**

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

# Phylogeny

---

12

a. All of the sequences in “*file1.fasta*” (<https://cbdm.uni-mainz.de/mb18/>) are homologs. How many groups of orthologs would you say there are in this file? Use Trex (<http://www.trex.uqam.ca/>).

Two groups of orthologs: Protein A & protein B.

b. What could you say about the history of this protein family?

*E.coli* has only one protein, and then it duplicated to form A and B. It is possible that *X.laevis\_B* duplicated later to form B and C.

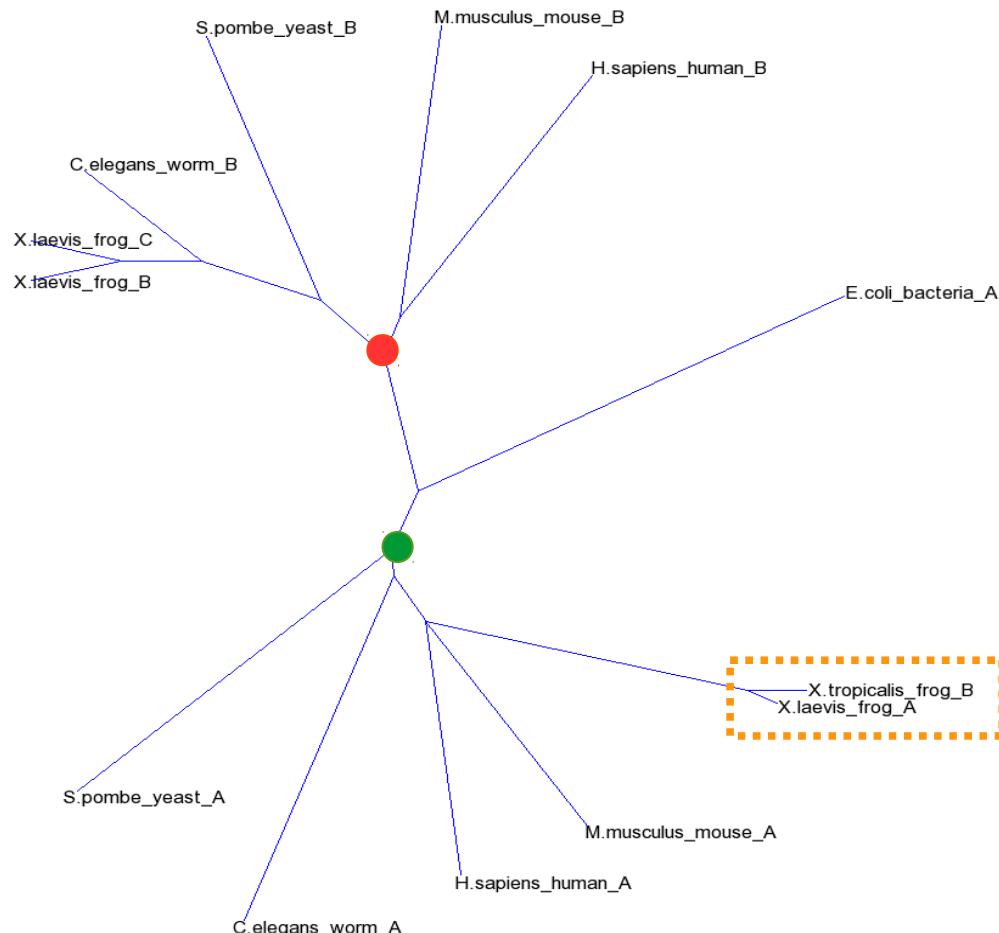
c. Would you say there is any wrongly annotated sequence?

*X.tropicalis\_B* is wrongly annotated. It should be *X.tropicalis\_A*, because they are in the same branch. The actual *X.tropicalis\_B* either is not in the dataset or was lost during evolution.

# Phylogeny

\*Images from: Trex

12



a. Two groups of orthologs:

- Proteins “A”
- Proteins “B”

b. E.coli has only one protein, and then it duplicated to form A and B. It is possible that X.laevis\_B duplicated later to form B and C.

c. X.tropicalis\_B is wrongly annotated. It should be X.tropicalis\_A, because they are in the same branch. The actual X.tropicalis\_B either is not in the dataset or was lost during evolution.

# Phylogeny

---

13

a. Using “*file2.fasta*” (<https://cbdm.uni-mainz.de/mb18/>) and Trex (<http://www.trex.uqam.ca/>), can you approximate to which taxonomic division belongs “proteinX”? **Primates**.

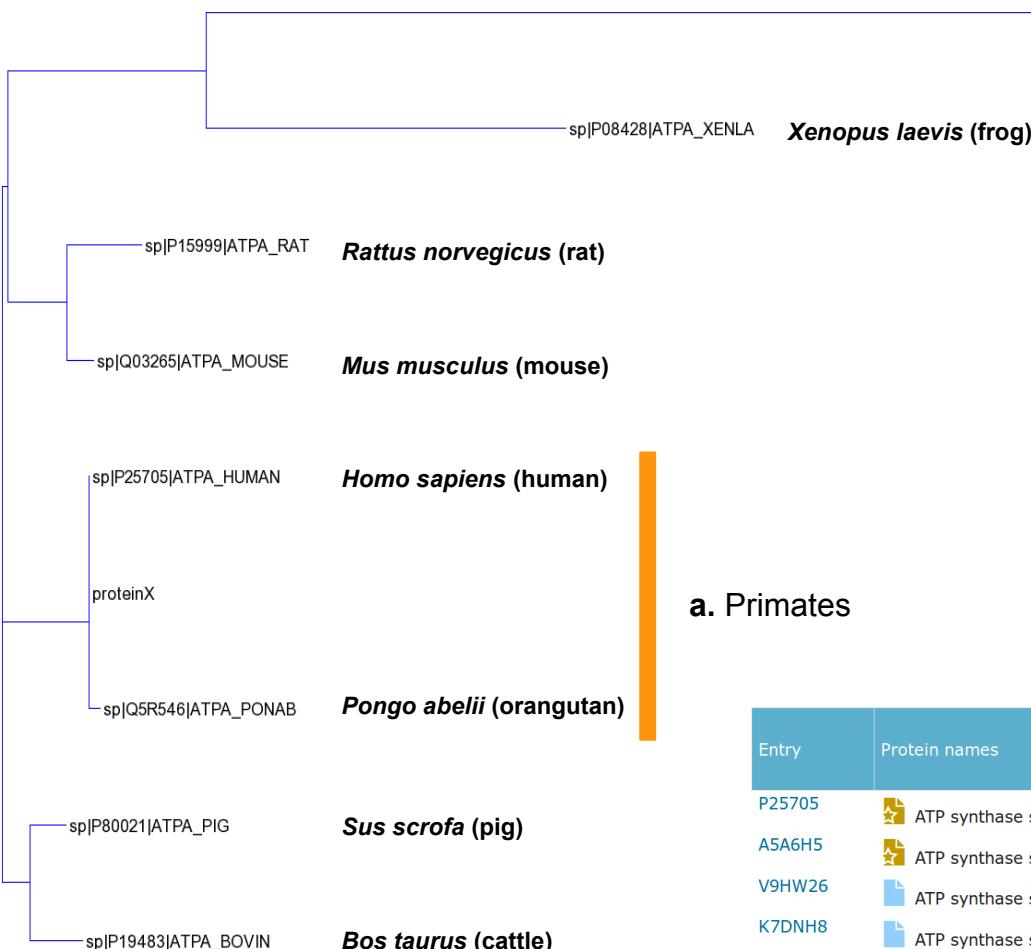
b. From which organism could it be? After guessing, check it.

*Homo sapiens* (human) or *Pan troglodytes* (chimpanzee); they are 100% identical.

# Phylogeny

\*Images from: Trex, UniProt

13



a. Primates

sp|P35381 ***Drosophila melanogaster*** (Fruit fly)



## BLAST

### How to use this tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity evolutionary relationships between sequences as well as help identify members

```
>proteinX
MLSVRVAAVVRALPPRAGLVSRNALGSSFIARNFHASNTHLQKTGTAEAMSILEERIL
GADTSVDEETGRVLSIGDGIARVHGLRNQAEEMVFSSGLKGMSLNLEPDNVGVVFG
NDKLKEGDIVKRTGAIIDPVGEELLGRVVDALGNAIDGKPIGSKTRRVLGKAPGII
PRISVREPMQTGIKAVDSLVPIGRGQRELIGDRTGKTSIAIDTIINQKRFDNGSDEKK
KLYCIVVAIGQKRSTVQALVKRLTDADAMKYTIIVSATASDAAPLQYLAPYSGCSMGEYF
RDNGKHALLIYYDDLSKQAVAYRQMSLLRRPGREAYPGDFYLHSRRLERAAKMNDAFG
GGSLTALPVIEQTQAGDVSAIPTNVISITDGQTFLETLYKGIRPAINVGLSVRGSA
```

Target database <sup>i</sup>	E-Threshold <sup>i</sup>	Matrix <sup>i</sup>	Filterin
...Mammals	10	Auto	None
<input type="checkbox"/> Run Blast in a separate window.			
<input type="button" value="Clear"/>		<input type="button" value="Run BLAST"/>	

Entry	Protein names	Match hit	Identity
P25705	ATP synthase subunit alpha, mitochondrial (Homo sapiens)	100 200 300 400 500	100.0%
A5A6H5	ATP synthase subunit alpha, mitochondrial (Pan troglodytes)	100 200 300 400 500	100.0%
V9HW26	ATP synthase subunit alpha (Homo sapiens)	100 200 300 400 500	100.0%
K7DNH8	ATP synthase subunit alpha (Pan troglodytes)	100 200 300 400 500	100.0%

b. *Homo sapiens* (human) or *Pan troglodytes* (chimpanzee); they are 100% identical.

# Phylogeny

---

14

Human hemoglobin consists of four protein subunits: two from the alpha globin gene cluster (located on chromosome 16) and two more from the beta globin gene cluster (located on chromosome 11). But there are at least nine different globin genes in these clusters, which are: zeta, mu, alpha, theta1, epsilon, gamma1, gamma2, delta and beta. Use the proteins in “*file3.fasta*” (<https://cbdm.uni-mainz.de/mb18/>).

a. Sort them either in cluster alpha or cluster beta.

Alpha: zeta, mu, alpha, theta1. Beta: epsilon, gamma1, gamma2, delta and beta.

b. Why do you think they are clustered in either cluster alpha or cluster beta?

Paralogous expansion from one ancestral alpha and one ancestral beta.

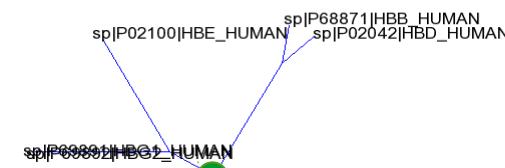
	Your list...ITXRE	Entry	Entry name		Protein names
<input type="checkbox"/>	HBAZ_HUMAN	P02008	HBAZ_HUMAN		Hemoglobin subunit zeta
<input type="checkbox"/>	HBM_HUMAN	Q6B0K9	HBM_HUMAN		Hemoglobin subunit mu
<input type="checkbox"/>	HBA_HUMAN	P69905	HBA_HUMAN		Hemoglobin subunit alpha
<input type="checkbox"/>	HBAT_HUMAN	P09105	HBAT_HUMAN		Hemoglobin subunit theta-1
<input type="checkbox"/>	HBE_HUMAN	P02100	HBE_HUMAN		Hemoglobin subunit epsilon
<input type="checkbox"/>	HBG1_HUMAN	P69891	HBG1_HUMAN		Hemoglobin subunit gamma-1
<input type="checkbox"/>	HBG2_HUMAN	P69892	HBG2_HUMAN		Hemoglobin subunit gamma-2
<input type="checkbox"/>	HBD_HUMAN	P02042	HBD_HUMAN		Hemoglobin subunit delta
<input type="checkbox"/>	HBB_HUMAN	P68871	HBB_HUMAN		Hemoglobin subunit beta

# Phylogeny

\*Images from: UniProt, Trex, UCSC

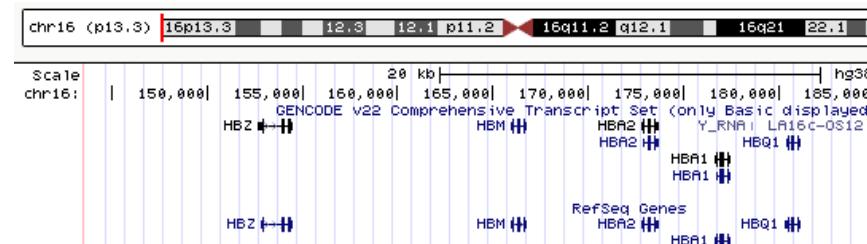
14

	Your list:...ITXRE	Entry	Entry name		Protein names
□	HBAZ_HUMAN	P02008	HBAZ_HUMAN	star	Hemoglobin subunit zeta
□	HBM_HUMAN	Q6B0K9	HBM_HUMAN	star	Hemoglobin subunit mu
□	HBA_HUMAN	P69905	HBA_HUMAN	star	Hemoglobin subunit alpha
□	HBAT_HUMAN	P09105	HBAT_HUMAN	star	Hemoglobin subunit theta-1
□	HBE_HUMAN	P02100	HBE_HUMAN	star	Hemoglobin subunit epsilon
□	HBG1_HUMAN	P69891	HBG1_HUMAN	star	Hemoglobin subunit gamma-1
□	HBG2_HUMAN	P69892	HBG2_HUMAN	star	Hemoglobin subunit gamma-2
□	HBD_HUMAN	P02042	HBD_HUMAN	star	Hemoglobin subunit delta
□	HBB_HUMAN	P68871	HBB_HUMAN	star	Hemoglobin subunit beta

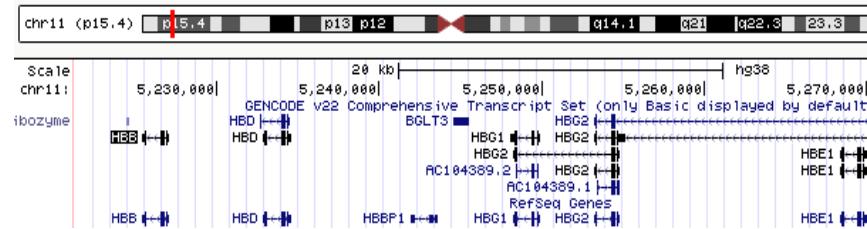


a. Using Trex:

● Cluster alpha



● Cluster beta

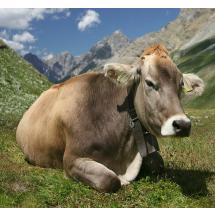


b. Paralogous expansion from one ancestral alpha and one ancestral beta.

15

Consider the following sequences: Q9T4B2 (CYB\_DELDE, dolphin), P00156 (CYB\_HUMAN, human), P00157 (CYB\_BOVIN, cattle), Q9MIX8 (CYB\_DANRE, zebra fish), P18946 (CYB\_CHICK, chicken), Q36461 (CYB\_ORNAN, platypus) and P00160 (CYB\_XENLA, frog).

Do you think that a phylogenetic tree built using these sequences would be coherent with evolution? Check with the taxonomy provided by UniProt (<http://www.uniprot.org/taxonomy/>).



Human  
(*Homo sapiens*)

Dolphin  
(*Delphinus delphis*)

Cattle  
(*Bos taurus*)

Chicken (rooster)  
(*Gallus gallus*)

Zebra fish  
(*Danio rerio*)

Platypus  
(*Ornithorhynchus anatinus*)

Frog  
(*Xenopus laevis*)

# Phylogeny

\*Images from: UniProt, Trex

15

**UniProt**

BLAST Align Retrieve/ID mapping

## Retrieve/ID mapping

How to use this tool

Enter or upload a list of identifiers to do one of the following:  
 Retrieve the corresponding UniProt entries to download  
 Convert identifiers which are of a different type to UniProt

### 1. Provide your identifiers

Q9T4B2  
P00156  
P00157  
Q9MIX8  
P18946  
Q36461  
P00160

BLAST Align Download Add to basket Columns >

Your list... B30

Q9T4B2  
P00156  
P00157  
Q9MIX8  
P18946  
Q36461  
P00160

Format: FASTA (canonical)   
 Compressed  Uncompressed  
 Preview first 10

	Q9MIX8	CYB_DANRE
Q9T4B2		
P00156		
P00157		
Q36461		
P00160		

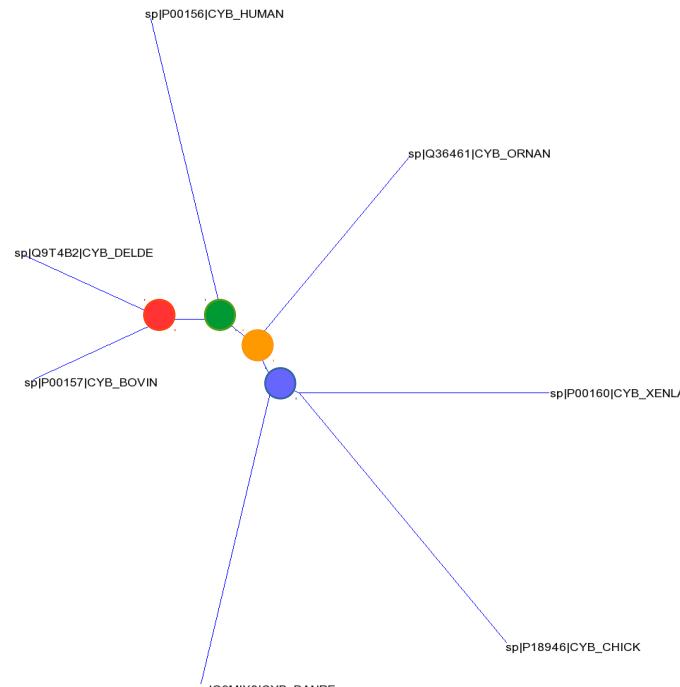
ClustalW is a widely used sequence alignment tool.

Paste your sequence into the window:  
 (7 input formats are accepted: FASTA, NBRF/PIR, EMBL/Swiss-Prot, GDE, Clustal, GCG/MSF, RSF)

```
>sp|Q9T4B2|CYB_DELDE Cytochrome b OS=Delphinus delphis GN=MT-CYB PE=3 SV=1
MINIRKTHLMIKLNDAFIQLPTPSNISSWWNEGSSLGLCLIMQILTGLFLAMHYTPDTS
TAFFSVAHICRDVNYGWFIRYLHANGASMFICLYAHAYGRGLYYGSYMFQETNNIGVLL
LTWMATAEVGVVLPGOMSEFWATVITULLSAIPYIGITLVENINGEFSVOKATLTREFA
FHILPFIATAALAAVHLLFHETGSMPGSPNMMDIPEHPYJIKDILGALLLILLL
ALTIFTPDLLGDPDNYTPANPLSTPAHKPEWYELFAYAIIERSIPNKLGGVLALLSIL
LIFIPMLQTSKORSMMERPESSOLLFWTLIAADLLLTWIGGOPVEHPYIIVGOLASILYFL
LILVLMPTAGLIENKLKW
>sp|P00156|CYB_HUMAN Cytochrome b OS=Homo sapiens GN=MT-CYB PE=1 SV=2
MTCPMRKTNPMLKLINHSEFIQLPTPSNISSWWNEGSSLGACLIQITGLFLAMHYSPDAS
TAFFSVAHICRDVNYGWFIRYLHANGASMFICLYAHAYGRGLYYGSYMFQETNNIGVLL
LTWMATAEVGVVLPGOMSEFWATVITULLSAIPYIGITLVENINGEFSVOKATLTREFA
FHILPFIATAALAAVHLLFHETGSMPGSPNMMDIPEHPYJIKDILGALLLILLL
Human → Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Laurasiatheria › Cetartiodactyla › Cetacea › Odontoceti › Delphinidae › Delphinus
Cattle → Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Laurasiatheria › Cetartiodactyla › Ruminantia › Pecora › Bovidae › Bovinae › Bos
Human → Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo
Platypus → Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Monotremata › Ornithorhynchidae › Ornithorhynchus
Frog → Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Amphibia › Batrachia › Anura › Pipoidea › Pipidae › Xenopodinae › Xenopus › Xenopus
Chicken → Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Archelosauria › Archosauria › Dinosauria › Saurischia › Theropoda › Coelurosauria › Aves › ... > Gallus
Zebrafish → Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Actinopterygii › Neopterygii › Teleostei › Ostariophysi › Cypriniformes › Cyprinidae › Danio
```

File  Pasted Examinar... No se ha seleccionado ning n archivo.

Align sequences Reset Clear



## 16

Consider the following sequences: P50225, P50226, P0DMM9, P0DMN0, O43704, O00338, Q6IMI6, O75897. All of them are part of a human paralog family (sulfotransferases).

- a. Look for the coordinates of the substrate binding region in the UniProt entry “P50225”. Do you expect all the paralogs to share this substrate binding region? Why? Check it.

Yes, because they share the motif “K[ST]H”.

- b. Using the previous sequences, could you determine which human paralog is the ortholog of the sequence in “*file4.fasta*” (<https://cbdm.uni-mainz.de/mb18/>)? Do not use BLAST.

The ortholog of “protein\_of\_interest” is ST1B1\_HUMAN.

- c. Which protein is the one found in “*file4.fasta*”?

ST1B1\_CANLF, from *Canis familiaris* (dog).

# MSA + Phylogeny + Homology

\*Images from: UniProt, Trex

16

(P50225)

Feature key	Position(s)	Length	Description	
Region <sup>i</sup>	106 – 108	3	Substrate binding	
	10 20 30 40 50			
MELIQDTSRP	PLEYVKGPL	IKYFAEALGP	LQSFQARPDD	LLISTYPKSG
60	70	80	90	100
TTWVSQLILDM	IYQGGDLEKC	HRAPIFMRVP	FLEFKAPGIP	SGMETLKDTP
110	120	130	140	150
APRLLKTHLP	LALLPQTLLD	QKVKVYYVAR	NAKDVAVSYY	HFYHMAKVHP
160	170	180	190	200
EPGTWDSFLE	KFMVGEVSYG	SWYQHVQEWW	ELSRTHPVLY	LFYEDMKENP
210	220	230	240	250
KREIQKILEF	VGRSLPEETV	DFVVQHTSFK	EMKKNPMTNY	TTVPQEFDMDH
260	270	280	290	
SISPFMRKGM	AGDWKTTFTV	AQNERFDADY	AEKMAGCSLS	FRSEL

Highlight

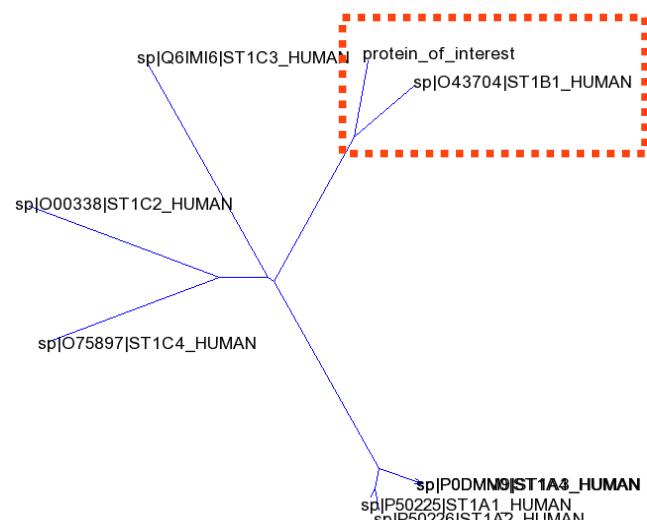
- Annotation
- Natural variant
  - Region
  - Binding site
  - Nucleotide binding
  - Beta strand
  - Active site
  - Mutagenesis
  - Modified residue
  - Alternative sequence



Align

P50225  
P02250  
P0DMM9  
P0DMN0  
043704  
000338  
Q6IMI6  
075897

P50225 ST1A1_HUMAN	1	-----MELIQDTSRPPLLEYVKGVPLIKYFAEALGPLQSQARPDDLLISTYPKSGTT	52
P50226 ST1A2_HUMAN	1	-----MELIQDTSRPPLLEYVKGVPLIKYFAEALGPLQSQARPDDLLINTYPKSGTT	52
P0DMM9 ST1A3_HUMAN	1	-----MELIQDTSRPPLLEYVKGVPLIKYFAEALGPLQSQARPDDLLINTYPKSGTT	52
P0DMN0 ST1A4_HUMAN	1	-----MELIQDTSRPPLLEYVKGVPLIKYFAEALGPLQSQARPDDLLINTYPKSGTT	52
043704 ST1B1_HUMAN	1	-----MLSPKDILRKDQLKLVHGYPMTCAFASNWKEIEQFHSPRDPDIVATYPKSGTT	52
000338 ST1C2_HUMAN	1	-----MALT-SDLGKQIKLKEVEGTLLOPATVDNWSQIQSFEAKPDDLLICTYPKAGTT	53
Q6IMI6 ST1C3_HUMAN	1	MAKIEKNAPTMEEKKPELFNIMEVDGVPTLILSKIEWEKVCNFQAKPDDLLLATYPKSGTT	60
075897 ST1C4_HUMAN	1	MALHDMDFT-FDGTKRSLVNVVKGILQPTDTCDIWDKIWNFQAKPDDLLISTYPKAGTT	59
		: *.* : *.*:*****: : *.*:*****: : *.*:*****: : *.*:*****:	
P50225 ST1A1_HUMAN	53	WVSQILDIMIYQGGDLEKCHRAPIFMRVPFLEFKAPGIP-SGMETLKDTPAPRLLKTHLPL	111
P50226 ST1A2_HUMAN	53	WVSQILDIMIYQGGDLEKCHRAPIFMRVPFLEFKVPGIP-SGMETLKNTPAPRLLKTHLPL	111
P0DMM9 ST1A3_HUMAN	53	WVSQILDIMIYQGGDLEKCNRAPIYVRVPFLEVNDPGEP-SGLETLKDTPPPRLIKSHLPL	111
P0DMN0 ST1A4_HUMAN	53	WVSQILDIMIYQGGDLEKCNRAPIYVRVPFLEVNDPGEP-SGLETLKDTPPPRLIKSHLPL	111
043704 ST1B1_HUMAN	53	WVSEIIDMILNDGDIKECKCRGFITEKVPMLEMTLPGLRTSGIEQLEKNPSPRIVKTHLPT	112
000338 ST1C2_HUMAN	54	WIOEQIVDMIEQNGDVEKCKRAOTLDRHAFFELKFPHKEKPDLEFVLEMSSPQLIKTHLST	112
Q6IMI6 ST1C3_HUMAN	61	WMHEILDMILNDGDVEKCKRAOTLDRHAFFELKFPHKEKPDLEFVLEMSSPQLIKTHLPS	120
075897 ST1C4_HUMAN	60	WTQEIVELIQNEGDVEKSKRAPTHQRFPFLEMKIPSLG-SGLEQAHAMPSPRILKTHLPS	118
		* : *.* : *.*:*****: : *.*:*****: : *.*:*****: : *.*:*****:	



a. Yes, because they share the motif "K[ST]H".

b. The ortholog of "protein\_of\_interest" is ST1B1\_HUMAN.

c. ST1B1\_CANLF, from *Canis familiaris* (dog).

