Master Biomedizin 2019

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

Pablo Mier
munoz@uni-mainz.de

12

a. All of the sequences in "*file1.fasta*" (https://cbdm.uni-mainz.de/mb19/) are homologs. How many groups of orthologs would you say there are in this file? Use Trex (http://www.trex.uqam.ca/).
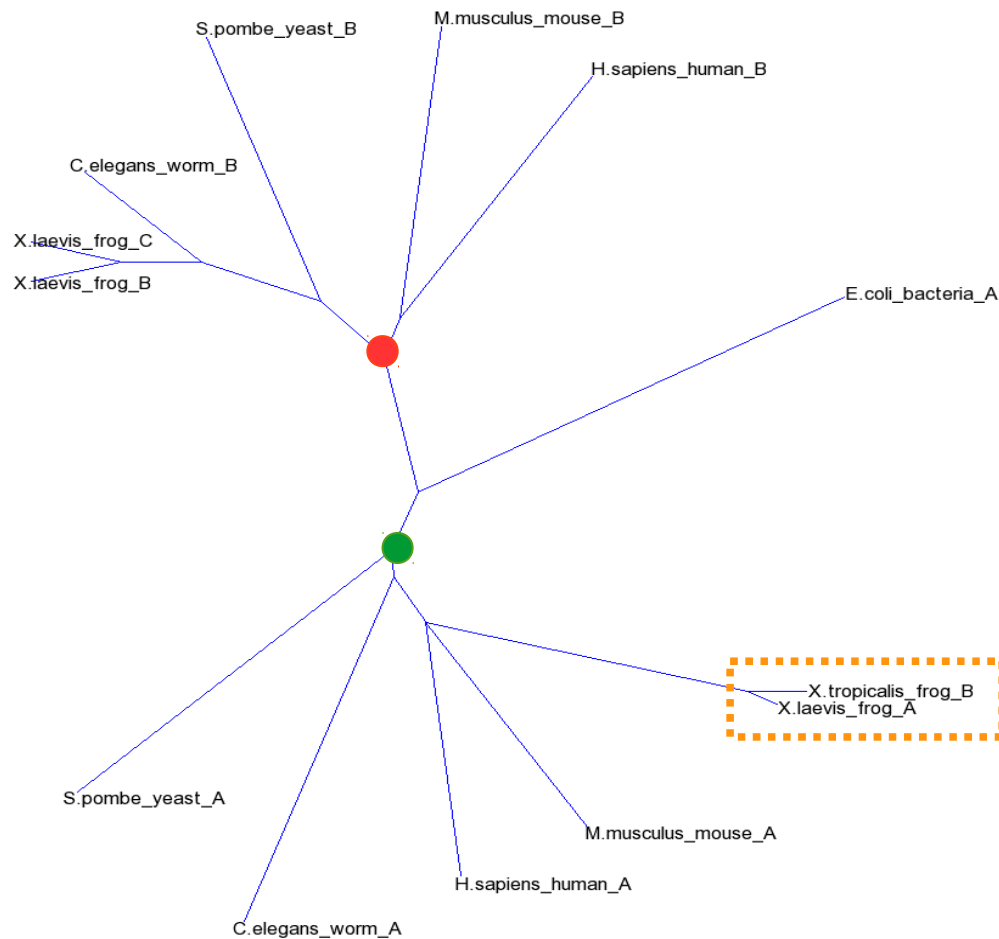Two groups of orthologs: Protein A & protein B.

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 is either not in the dataset or was lost during evolution.

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 is either not in the dataset or was lost during evolution.
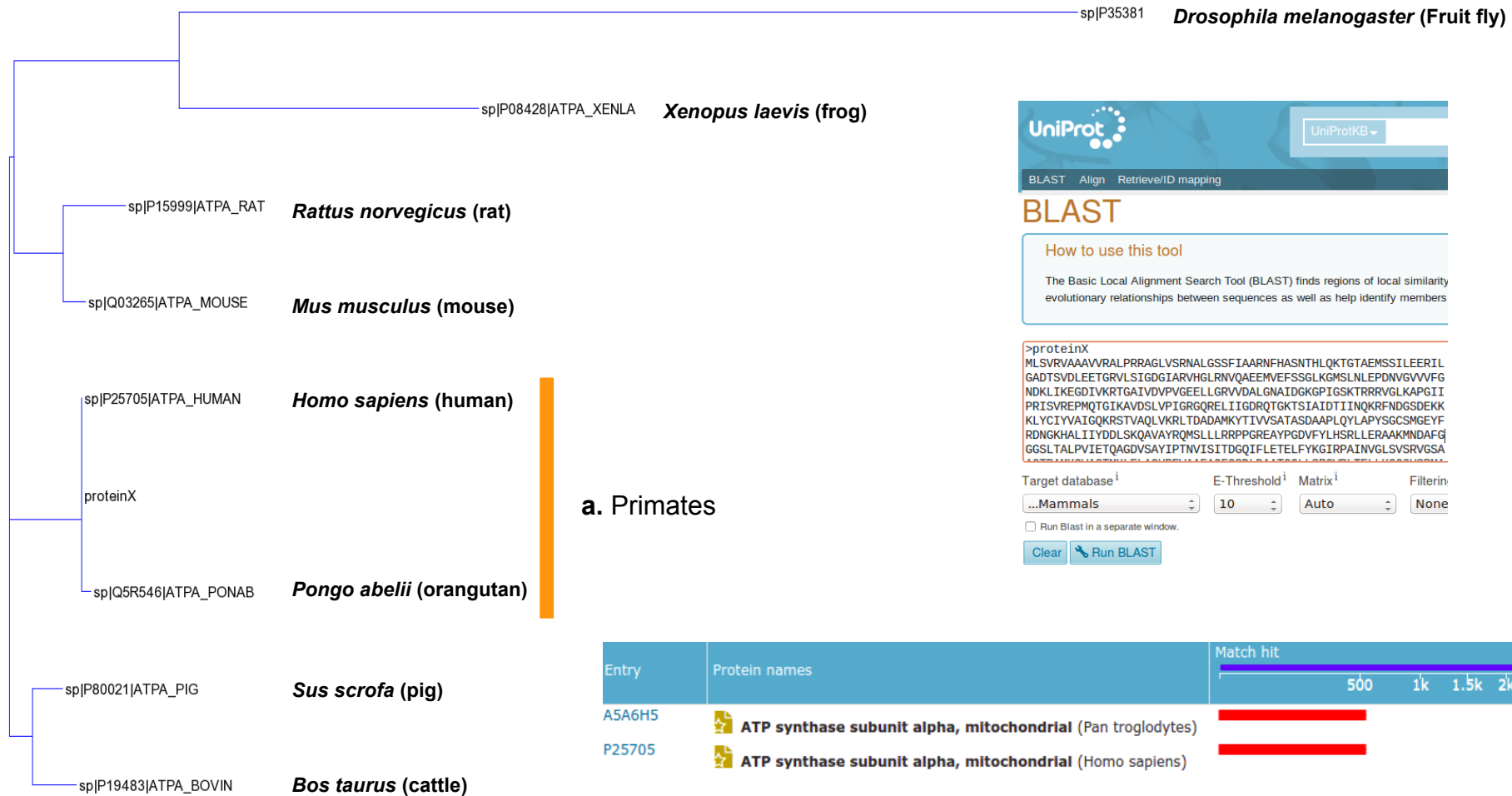
13

a. Using "*file2.fasta*" (https://cbdm.uni-mainz.de/mb19/) 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

13

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

sp|P08428|ATPA_XENLA   *Xenopus laevis* **(frog)**

sp|P15999|ATPA_RAT   *Rattus norvegicus* **(rat)**

sp|Q03265|ATPA_MOUSE   *Mus musculus* **(mouse)**

sp|P25705|ATPA_HUMAN   *Homo sapiens* **(human)**

proteinX

sp|Q5R546|ATPA_PONAB   *Pongo abelii* **(orangutan)**

a. Primates

sp|P80021|ATPA_PIG   *Sus scrofa* **(pig)**

sp|P19483|ATPA_BOVIN   *Bos taurus* **(cattle)**

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/mb19/).

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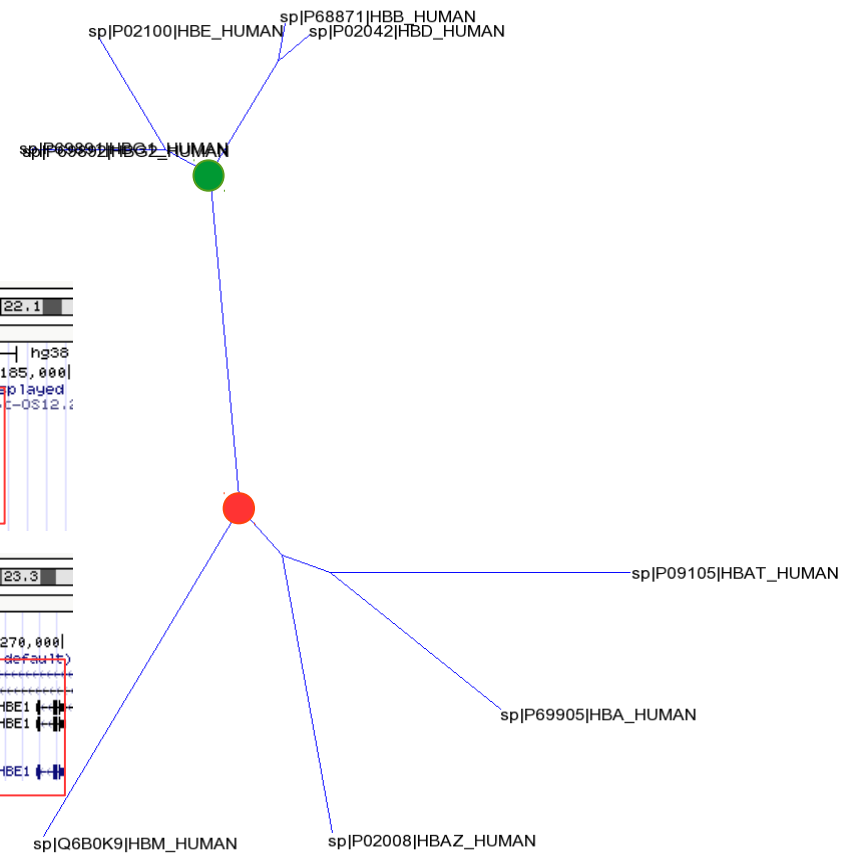Paralogous expansion from one ancestral alpha and one ancestral beta.

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

Phylogeny

14

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

a. Using Trex:

🔴 Cluster alpha

🟢 Cluster beta

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

Pablo Mier
munoz@uni-mainz.de

7

JG|U
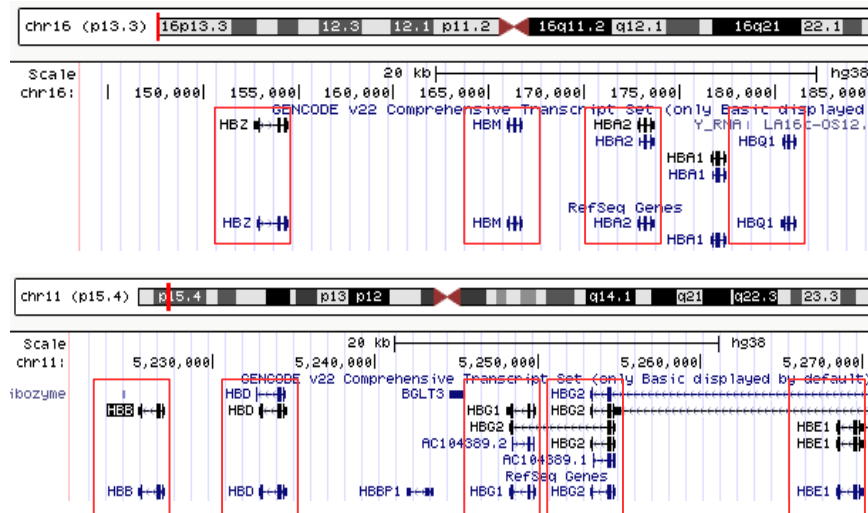
15

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? Build the tree with Trex and 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

15

Dolphin → 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 › Mammalia › 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

Pablo Mier
munoz@uni-mainz.de

JG|U
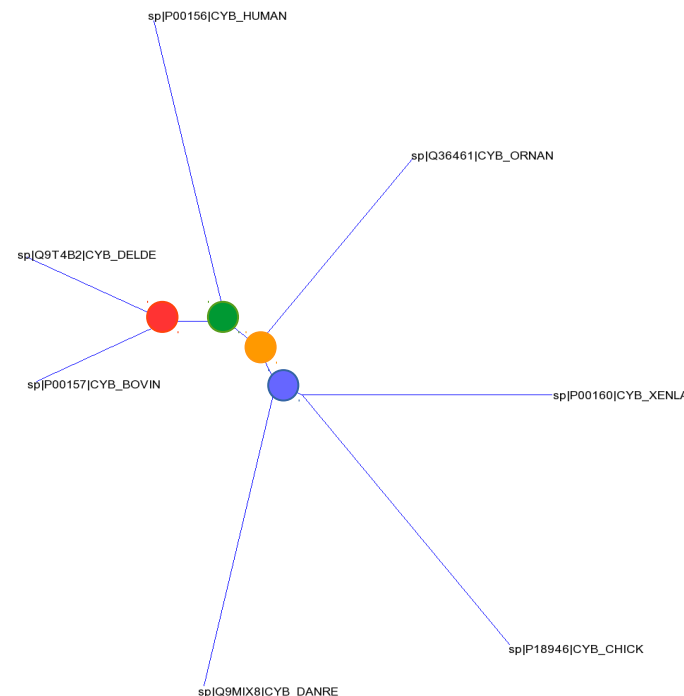
16

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"; the "H" is annotated as active site.

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/mb19/)? 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 lupus familiaris* (dog).

16

(P50225)

Feature key	Position(s)	Length	Description
Region [i]	106 – 108	3	Substrate binding

```
        10         20         30         40         50
MELIQDTSRP PLEYVKGVPL IKYFAEALGP LQSFQARPDD LLISTYPKSG
        60         70         80         90        100
TTWVSQILDM IYQGGDLEKC HRAPIFMRVP FLEFKAPGIP SGMETLKDTP
       110        120        130        140        150
APRLLKTHLP LALLPQTLLD QKVKVVYVAR NAKDVAVSYY HFYHMAKVHP
       160        170        180        190        200
EPGTWDSFLE KFMVGEVSYG SWYQHVQEWW ELSRTHPVLY LFYEDMKENP
       210        220        230        240        250
KREIQKILEF VGRSLPEETV DFVVQHTSFK EMKKNPMTNY TTVPQEFMDH
       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 ST1A1_HUMAN    1  --------MELIQDTSRPPLEYVKGVPLIKYFAEALGPLQSFQARPDDLLISTYPKSGTT   52
P50226 ST1A2_HUMAN    1  --------MELIQDTSRPPLEYVKGVPLIKYFAEALGPLQSFQARPDDLLISTYPKSGTT   52
P0DMM9 ST1A3_HUMAN    1  --------MELIQDTSRPPLEYVKGVPLIKYFAEALGPLQSFQARPDDLLINTYPKSGTT   52
P0DMN0 ST1A4_HUMAN    1  --------MELIQDTSRPPLEYVKGVPLIKYFAEALGPLQSFQARPDDLLINTYPKSGTT   52
O43704 ST1B1_HUMAN    1  --------MLSPKDILRKDLKLVHGYPMTCAFASNWEKIEQFHSRPDDIVIATYPKSGTT   52
O00338 ST1C2_HUMAN    1  ------MALT-SDLGKQIKLKEVEGTLLQPATVDNWSQIQSFEAKPDDLLICTYPKAGTT   53
Q6IMI6 ST1C3_HUMAN    1  MAKIEKNAPTMEKKPELFNIMEVDGVPTLILSKEWWEKVCNFQAKPDDLILATYPKSGTT   60
O75897 ST1C4_HUMAN    1  MALHDMEDFT-FDGTKRLSVNYVKGILQPTDTCDIWDKIWNFQAKPDDLLISTYPKAGTT   59
                                    :   *.*           :.  .*.::***::: ****:***

P50225 ST1A1_HUMAN   53  WVSQILDMIYQGGDLEKCHRAPIFMRVPFLEFKAPGIP-SGMETLKDTPAPRLLKTHLPL  111
P50226 ST1A2_HUMAN   53  WVSQILDMIYQGGDLEKCHRAPIFMRVPFLEFKVPGIP-SGMETLKNTPAPRLLKTHLPL  111
P0DMM9 ST1A3_HUMAN   53  WVSQILDMIYQGGDLEKCNRAPIYVRVPFLEVNDPGEP-SGLETLKDTPPPRLLKSHLPL  111
P0DMN0 ST1A4_HUMAN   53  WVSQILDMIYQGGDLEKCNRAPIYVRVPFLEVNDPGEP-SGLETLKDTPPPRLLKSHLPL  111
O43704 ST1B1_HUMAN   53  WVSEIIDMILNDGDIEKCKRGFITEKVPMLEMTLPGLRTSGIEQLEKNPSPRIVKTHLPT  112
O00338 ST1C2_HUMAN   54  WIQEIVDMIEQNGDVEKCQRAIIQHRHPFIEWARPPQP-SGVEKAKAMPSPRILKTHLST  112
Q6IMI6 ST1C3_HUMAN   61  WMHEILDMILNDGDVEKCKRAQTLDRHAFLELKFPHKEKPDLEFVLEMSSPQLIKTHLPS  120
O75897 ST1C4_HUMAN   60  WTQEIVELIQNEGDVEKSKRAPTHQRFPFLEMKIPSLG-SGLEQAHAMPSPRILKTHLPF  118
                         *  :*:::* : **:**..*.  .  ::* *   :*      *:::*:**
```

---

a. Yes, because they share the motif "K[ST]H"; the "H" is annotated as active site.

b. The ortholog of "protein_of_interest" is ST1B1_HUMAN.

c. ST1B1_CANLF, from *Canis lupus familiaris* (dog).
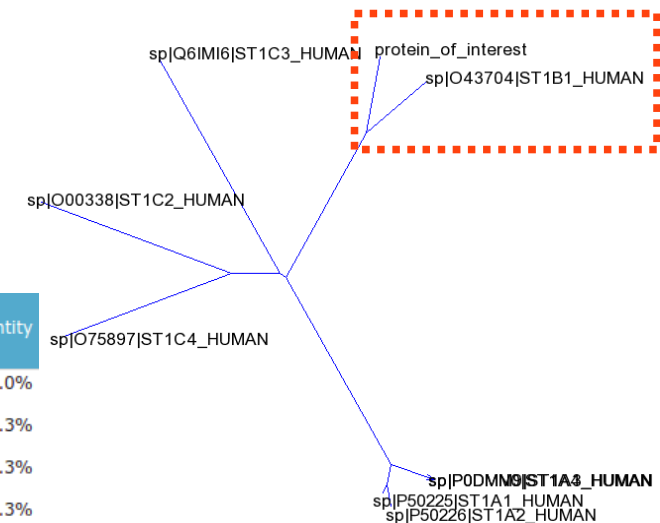
Entry	Protein names	Match hit	Identity
Q95JD5	⭐ **Sulfotransferase family cytosolic 1B member 1** (Canis lupus familiaris)		100.0%
F6V3V8	**Sulfotransferase** (Canis lupus familiaris)		98.3%
A0A2U3VE78	**Sulfotransferase** (Odobenus rosmarus divergens)		94.3%
D2I2Q5	**Sulfotransferase** (Ailuropoda melanoleuca)		94.3%

sp|Q6IMI6|ST1C3_HUMAN   protein_of_interest
sp|O43704|ST1B1_HUMAN

sp|O00338|ST1C2_HUMAN

sp|O75897|ST1C4_HUMAN

sp|P0DMN0|ST1A4_HUMAN
sp|P50225|ST1A1_HUMAN
sp|P50226|ST1A2_HUMAN