STATISTICAL DATA ANALYSIS 2

Kamil Łopuszański

Phylogenetics of the hemoglobin

The goal of the project was to conduct phylogenetic analysis on hemoglobin DNA sequences coming from a set of animals.

Libraries used in the project:

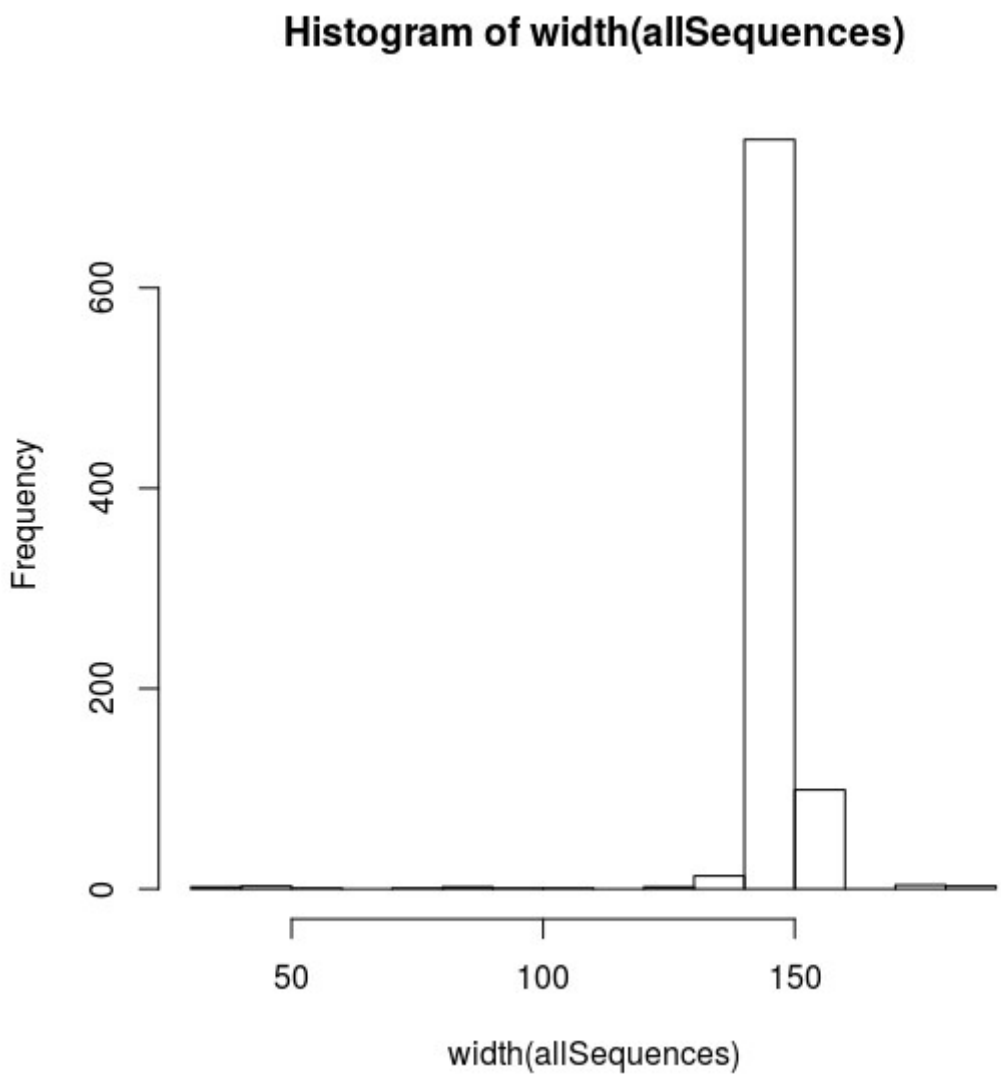
- ape
- Biostrings
- phytools
- phangorn
- msa
- httr
- XML
- stringr

After all sequences were loaded, headers were processed. Uniprot ID of each header was extracted, as substring on positions from 1 to 8. The ID was used to request an XML document using Uniprot API. Status code of the response from the server was always 200 (OK). After the XML document was ready, it was parsed. To get the scientific name, child of the child of the root tag was extracted. The scientific name was split to 2 words and a new string was created (function **toShortName(name)**). Apart from the scientific name, taxons of the organism was extracted. The author checked the class of an organism in the following way. If *Homo* was part of the taxons set, the organism belonged to Homo sapiens species. Otherwise, the following conditions (and corresponding classes) were crafted:

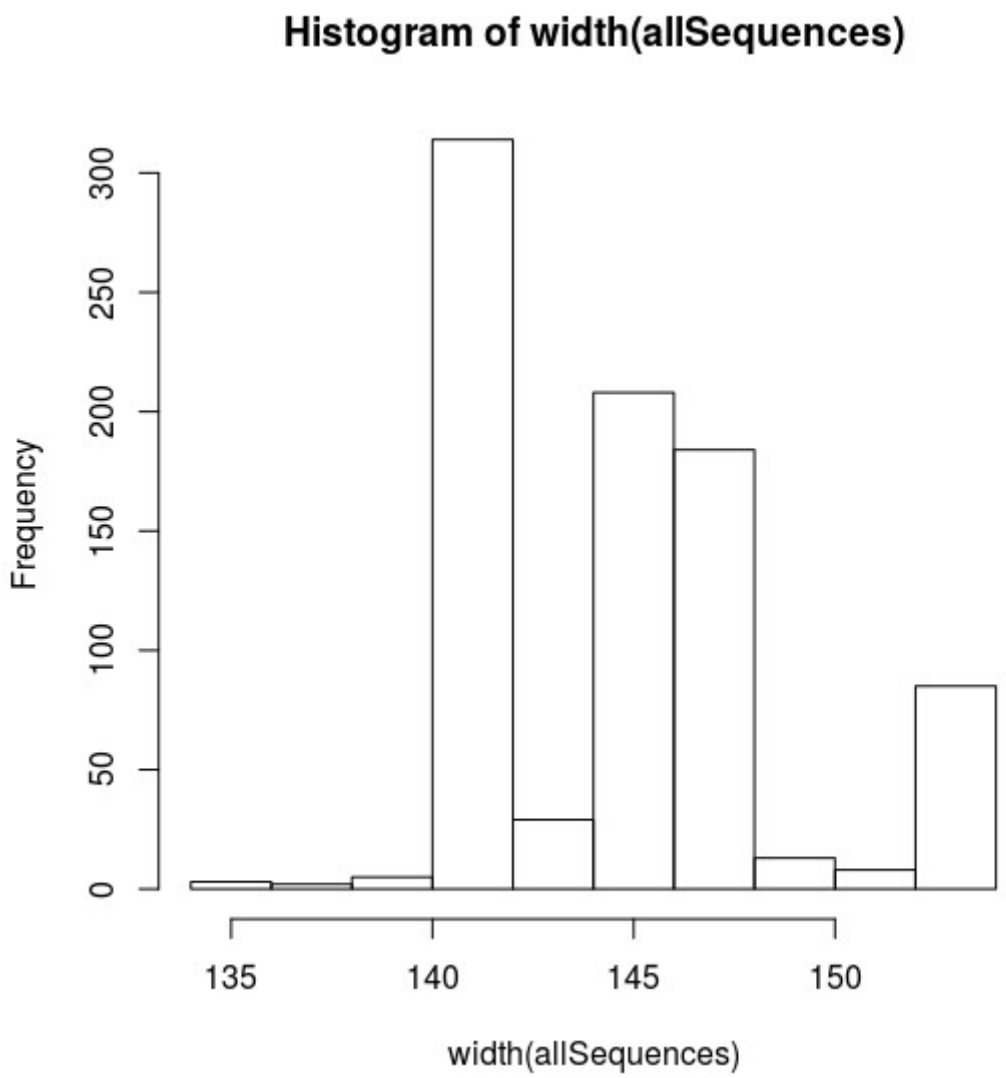
- taxons included *Mammalia* (mammals)
- taxons included *Aves* (birds)
- taxons included *Amphibia* (amphibians)
- taxons included one of *Lepidosauria, Archelosauria* (reptiles)
- taxons included one of *Actinopterygii, Elasmobranchii, Coelacanthiformes, Hyperoartia, Hyperotreti, Dipnoi* (fish)

A new file **cleaned-hemoglobin.fasta** was created, which was the file created from the original **hemoglobin.fasta**, on cleaning the data (replacing existing headers with new names).

The consequent analysis was done on the cleaned data. Mean and standard deviation of the length of sequences was calculated: $\mu = 144.52, \sigma = 10.81$. Histogram:



To remove outliers, all the sequences whose length was either shorter than 134 (about $\mu - std$) or longer than 154 (about $\mu + std$). Histogram after this operation:



And the mean was practically unchanged (the new mean equaled `145.22`), as well as the number of sequences (originally 880, then 851).

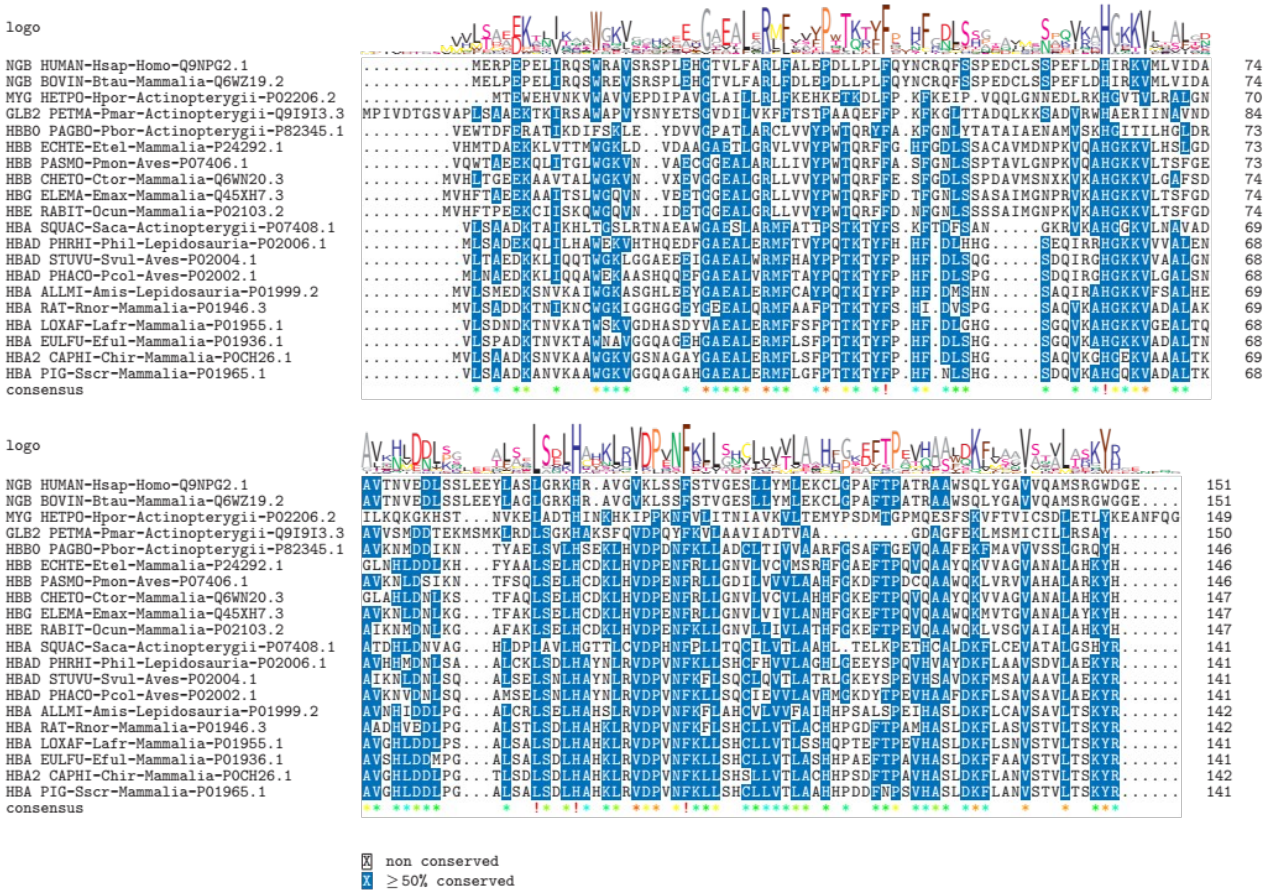
To take a sample consisting of 1 homo sapiens, 4 mammals, 3 birds, 2 reptiles and 3 fish, sets corresponding to each class were created. Then, each sequence was added to the right set. The core (set of essential organism) was created as a union of samples taken on the aforementioned sets. The picked organism were removed frmo the sets. Remaining organisms (not taken into the core) were gathered, let's call it **R**. In order to create a sample big enough, the sample was taken from **R**. The final sample was created as a union of the core and the sample from **R**. The sequences were written out to files `'subsetA.fasta'`, `'subsetB.fasta'`, `'subsetC.fasta'`.

For each subset, the alignment was performed. Used algorithms:

- ClustalW
- ClustalOmega
- MUSCLE

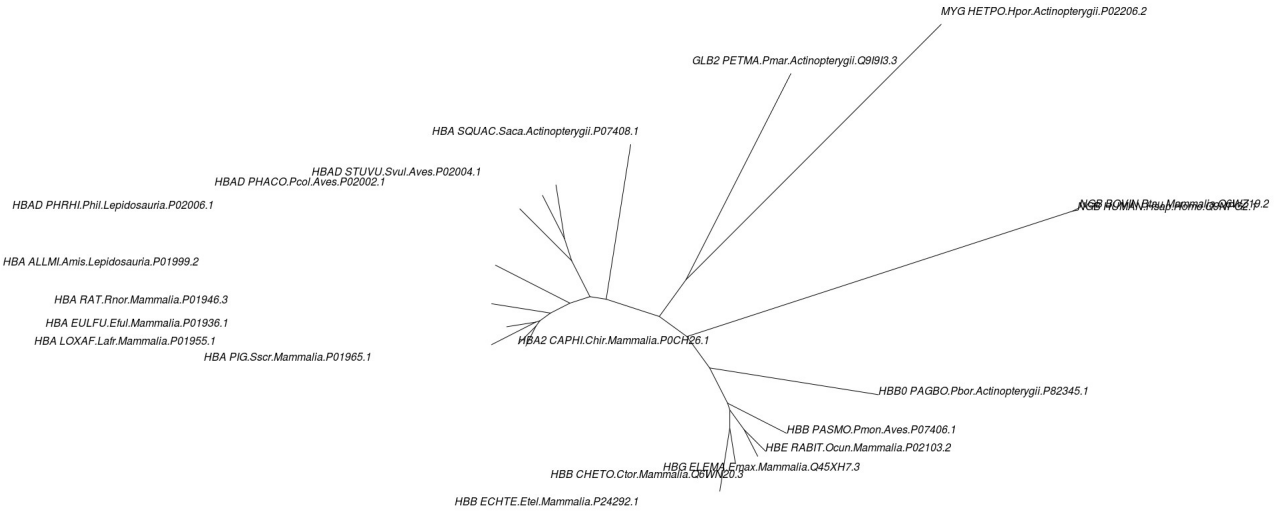
The aligned sequences were written out to the folder `out`, to files `'subsetA clustalW.fasta'`, `'subsetA clustalOmega.fasta'`, `'subsetA MUSCLE.fasta'`, `'subsetB clustalW.fasta'`, `'subsetB clustalOmega.fasta'`, `'subsetB MUSCLE.fasta'`, `'subsetC clustalW.fasta'`, `'subsetC clustalOmega.fasta'`, `'subsetC MUSCLE.fasta'`.

The output of alignment for subset C, using ClustalOmega:

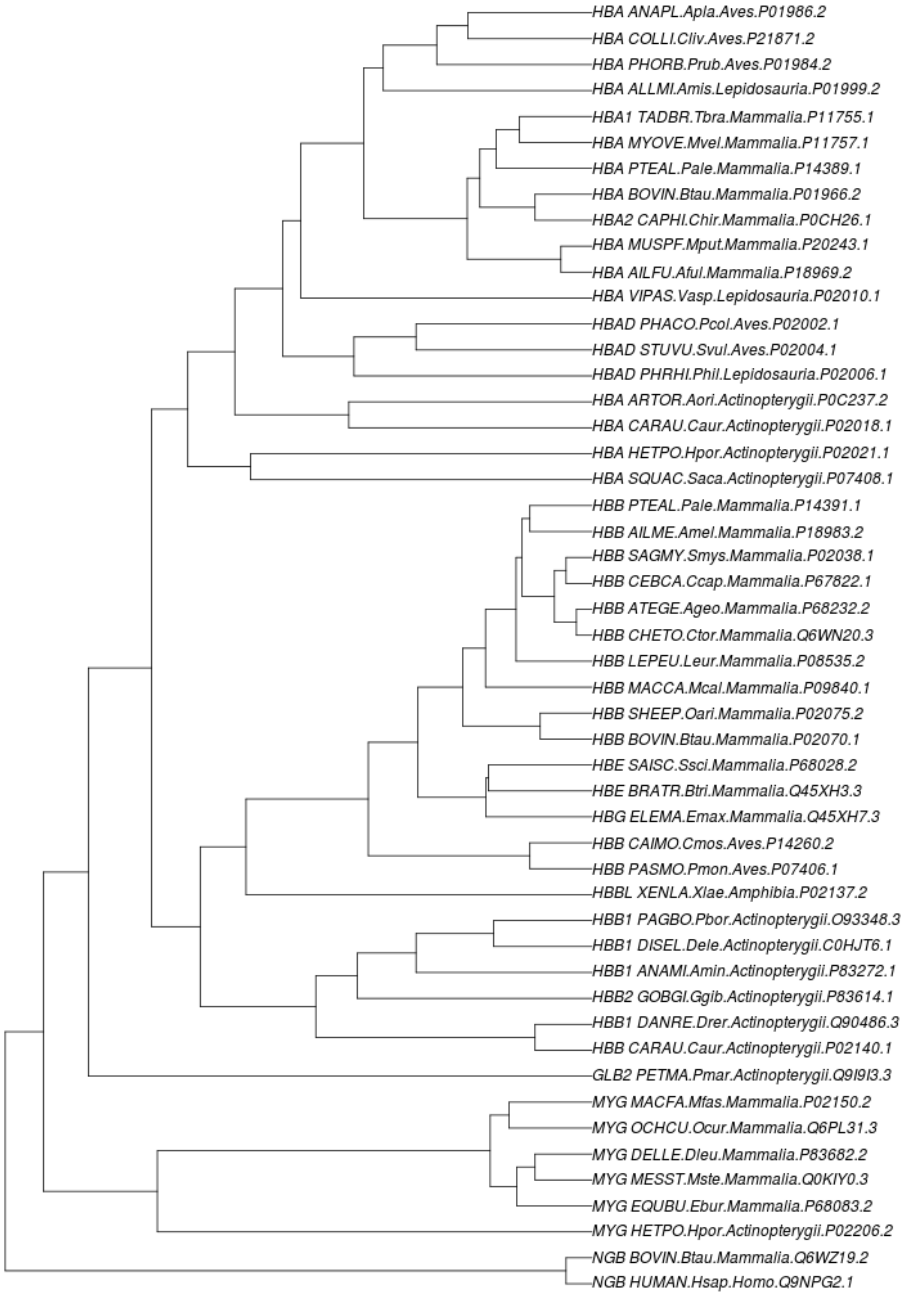


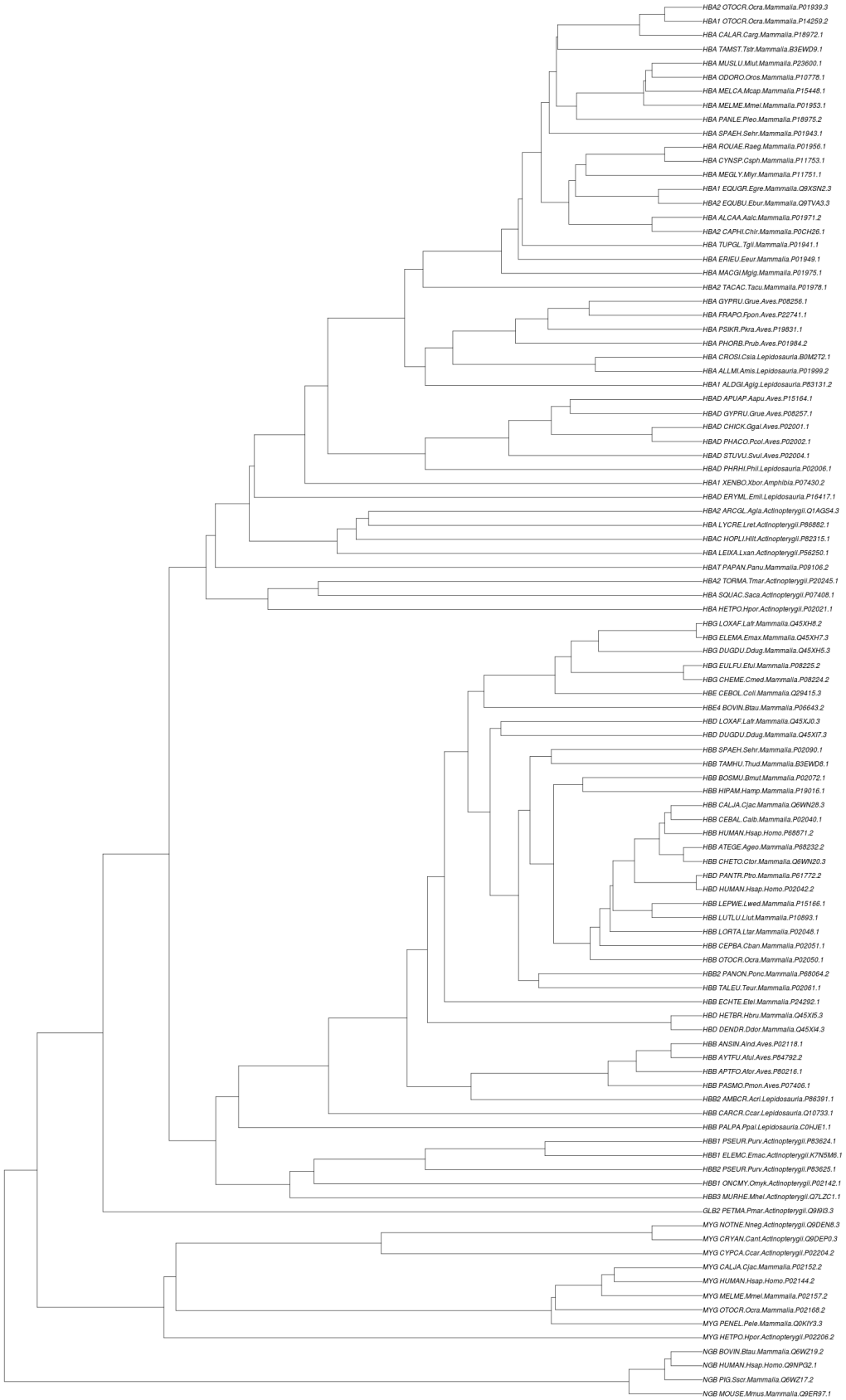
The aligned sequences were loaded and transformed to objects using functions `read.aa` and `phyDat`. Afterwards, distance matrix were calculated, using Hamming and ML metrics. The result was written out to the folder `out/distance`, to files `'distance A clustalW hamming.txt'`, `'distance A clustalW ml.txt'`, `'distance A clustalOmega hamming.txt'`, `'distance A clustalOmega ml.txt'`, `'distance A MUSCLE hamming.txt'`, `'distance A MUSCLE ml.txt'` etc.

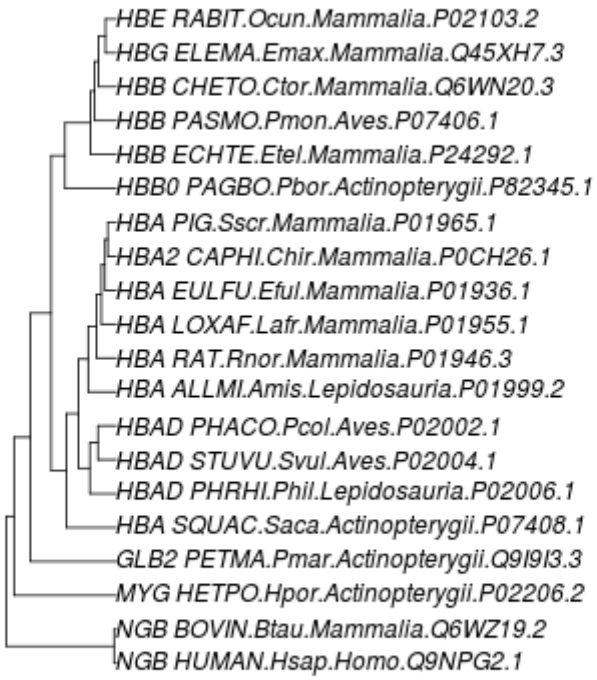
Trees were constructed, based on Neighbor joining algorithm, using all the before calculated distance matrices. Unrooted plot of tree, distance ML, alignment MUSCLE:



The parsimony scores of all trees were computed and compared. ClustalW and ClustalOmega performed better (depending on the set). UPGMA algorithm performed even better (to reduce complexity, only best matrices were considered for UPGMA).
The figures show trees produced using UPGMA:




HBA2 OTOCR.Ocra.Mammalia.P01939.3
HBA1 OTOCR.Ocra.Mammalia.P14259.2
HBA CALAR.Carg.Mammalia.P18972.1
HBA TAMST.Tstr.Mammalia.B3EWD9.1
HBA MUSLU.Mlut.Mammalia.P23600.1
HBA ODORO.Oros.Mammalia.P10778.1
HBA MELCA.Mcap.Mammalia.P15448.1
HBA MELME.Mmel.Mammalia.P01953.1
HBA PANLE.Pleo.Mammalia.P18975.2
HBA SPAEH.Sehr.Mammalia.P01943.1
HBA ROUAE.Raeg.Mammalia.P01956.1
HBA CYNSP.Csph.Mammalia.P11753.1
HBA MEGLY.Mlyr.Mammalia.P11751.1
HBA1 EQUGR.Egre.Mammalia.Q9XSN2.3
HBA2 EQUBU.Ebur.Mammalia.Q9TVA3.3
HBA ALCAA.Aalc.Mammalia.P01971.2
HBA2 CAPHI.Chir.Mammalia.P0CH26.1
HBA TUPGL.Tgli.Mammalia.P01941.1
HBA ERIEU.Eeur.Mammalia.P01949.1
HBA MACGI.Mgig.Mammalia.P01975.1
HBA2 TACAC.Tacu.Mammalia.P01978.1
HBA GYPRU.Grue.Aves.P08256.1
HBA FRAPO.Fpon.Aves.P22741.1
HBA PSIKR.Pkra.Aves.P19831.1
HBA PHORB.Prub.Aves.P01984.2
HBA CROSI.Csla.Lepidosauria.B0M2T2.1
HBA ALLMI.Amis.Lepidosauria.P01999.2
HBA1 ALDGI.Agig.Lepidosauria.P83131.2
HBAD APUAP.Aapu.Aves.P15164.1
HBAD GYPRU.Grue.Aves.P08257.1
HBAD CHICK.Ggal.Aves.P02001.1
HBAD PHACO.Pcol.Aves.P02002.1
HBAD STUVU.Svul.Aves.P02004.1
HBAD PHRHI.Phil.Lepidosauria.P02006.1
HBA1 XENBO.Xbor.Amphibia.P07430.2
HBAD ERYML.Emil.Lepidosauria.P16417.1
HBA2 ARCGL.Agla.Actinopterygii.Q1AGS4.3
HBA LYCRE.Lret.Actinopterygii.P86882.1
HBAC HOPLI.Hlit.Actinopterygii.P82315.1
HBA LEIXA.Lxan.Actinopterygii.P56250.1
HBAT PAPAN.Panu.Mammalia.P09106.2
HBA2 TORMA.Tmar.Actinopterygii.P20245.1
HBA SQUAC.Saca.Actinopterygii.P07408.1
HBA HETPO.Hpor.Actinopterygii.P02021.1
HBG LOXAF.Lafr.Mammalia.Q45XH8.2
HBG ELEMA.Emax.Mammalia.Q45XH7.3
HBG DUGDU.Ddug.Mammalia.Q45XH5.3
HBG EULFU.Etul.Mammalia.P08225.2
HBG CHEME.Cmed.Mammalia.P08224.2
HBE CEBOL.Coli.Mammalia.Q29415.3
HBE4 BOVIN.Btau.Mammalia.P06643.2
HBD LOXAF.Lafr.Mammalia.Q45XJ0.3
HBD DUGDU.Ddug.Mammalia.Q45XI7.3
HBB SPAEH.Sehr.Mammalia.P02090.1
HBB TAMHU.Thud.Mammalia.B3EWD6.1
HBB BOSMU.Bmut.Mammalia.P02072.1
HBB HIPAM.Hamp.Mammalia.P19016.1
HBB CALJA.Cjac.Mammalia.Q6WN26.3
HBB CEBAL.Calb.Mammalia.P02040.1
HBB HUMAN.Hsap.Homo.P68871.2
HBB ATEGE.Ageo.Mammalia.P68232.2
HBB CHETO.Ctor.Mammalia.Q6WN20.3
HBD PANTR.Ptro.Mammalia.P61772.2
HBD HUMAN.Hsap.Homo.P02042.2
HBB LEPWE.Lwed.Mammalia.P15166.1
HBB LUTLU.Llut.Mammalia.P10893.1
HBB LORTA.Ltar.Mammalia.P02046.1
HBB CEPBA.Cban.Mammalia.P02051.1
HBB OTOCR.Ocra.Mammalia.P02050.1
HBB2 PANON.Ponc.Mammalia.P68064.2
HBB TALEU.Teur.Mammalia.P02061.1
HBB ECHTE.Etel.Mammalia.P24292.1
HBD HETBR.Hbru.Mammalia.Q45XI5.3
HBD DENDR.Ddor.Mammalia.Q45XI4.3
HBB ANSIN.Alnd.Aves.P02118.1
HBB AYTFU.Aful.Aves.P84792.2
HBB APTFO.Ator.Aves.P80216.1
HBB PASMO.Pmon.Aves.P07406.1
HBB2 AMBCR.Acri.Lepidosauria.P86391.1
HBB CARCR.Ccar.Lepidosauria.Q10733.1
HBB PALPA.Ppal.Lepidosauria.C0HJE1.1
HBB1 PSEUR.Purv.Actinopterygii.P83624.1
HBB1 ELEMC.Emac.Actinopterygii.K7N5M6.1
HBB2 PSEUR.Purv.Actinopterygii.P83625.1
HBB1 ONCMY.Omyk.Actinopterygii.P02142.1
HBB3 MURHE.Mhel.Actinopterygii.Q7LZC1.1
GLB2 PETMA.Pmar.Actinopterygii.Q9I9I3.3
MYG NOTNE.Nneg.Actinopterygii.Q9DEN8.3
MYG CRYAN.Cant.Actinopterygii.Q9DEP0.3
MYG CYPCA.Ccar.Actinopterygii.P02204.2
MYG CALJA.Cjac.Mammalia.P02152.2
MYG HUMAN.Hsap.Homo.P02144.2
MYG MELME.Mmel.Mammalia.P02157.2
MYG OTOCR.Ocra.Mammalia.P02168.2
MYG PENEL.Pele.Mammalia.Q0KIY3.3
MYG HETPO.Hpor.Actinopterygii.P02206.2
NGB BOVIN.Btau.Mammalia.Q6WZ19.2
NGB HUMAN.Hsap.Homo.Q9NPG2.1
NGB PIG.Sscr.Mammalia.Q6WZ17.2
NGB MOUSE.Mmus.Mammalia.Q9ER97.1



UPGMA performed better than NJ (bigger parsimony score). Best trees were saved to the folders `out/nexus` and `out/newick`, in nexus and newick format respectively.

Then, maximum parsimony algorithm was performed.

```
tree.max.pars.A <- optim.parsimony(tree.upgma.A, aligned.A.seqs.clustalOmega)
tree.max.pars.B <- optim.parsimony(tree.upgma.B, aligned.B.seqs.clustalW)
tree.max.pars.C <- optim.parsimony(tree.upgma.C, aligned.C.seqs.clustalOmega)
```

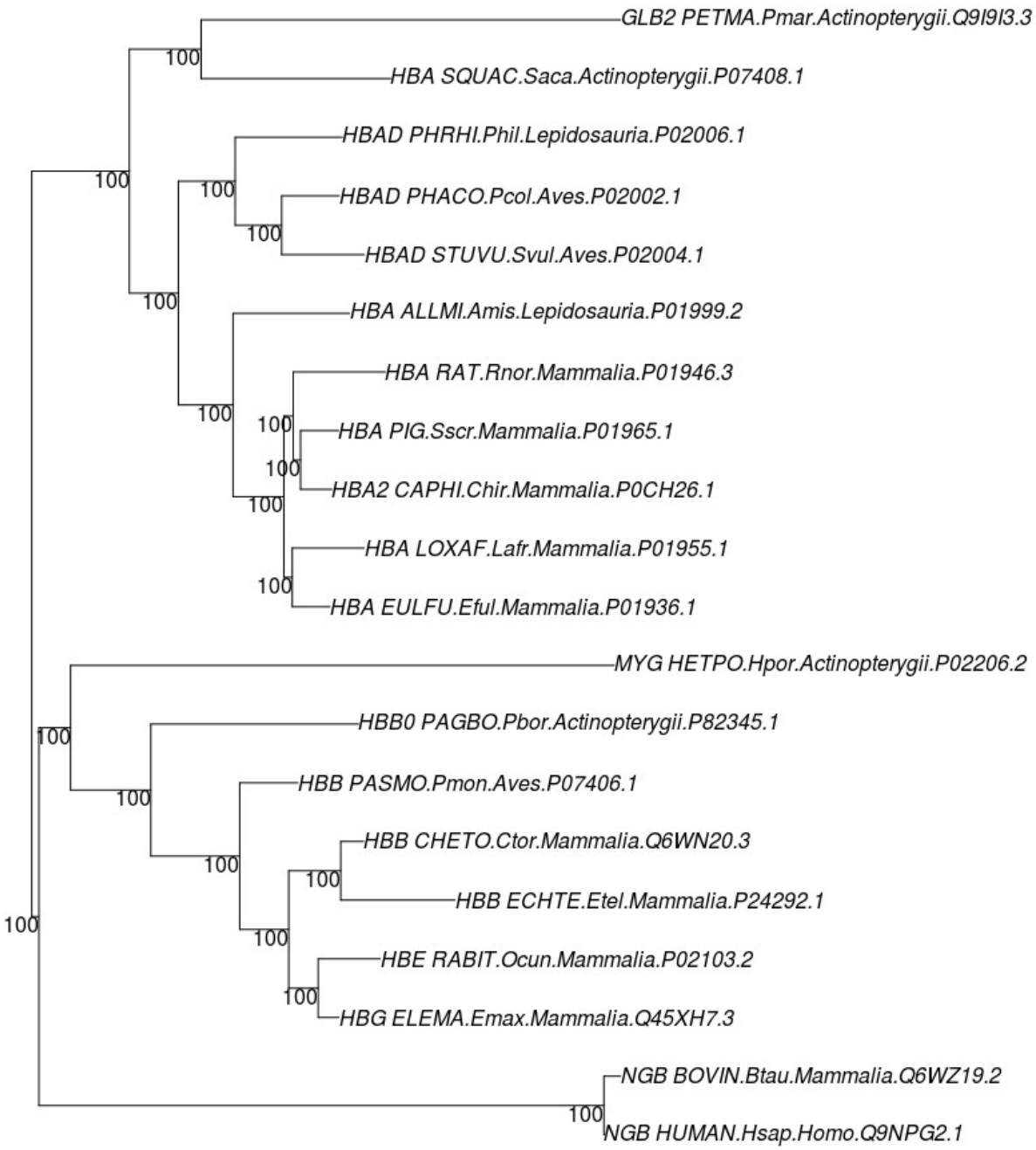
Using maximum likelihood algorithm, one can decrease the loglikelihood:

```
fit.A <- pml(tree.nj.hamming.A.clustalOmega, data = aligned.A.seqs.clustalOmega)
fit.B <- pml(tree.nj.hamming.B.clustalW, data = aligned.B.seqs.clustalW)
fit.C <- pml(tree.nj.ml.C.clustalOmega, data = aligned.C.seqs.clustalOmega)

fit.A.JC <- optim.pml(fit.A, TRUE)
fit.B.JC <- optim.pml(fit.B, TRUE)
fit.C.JC <- optim.pml(fit.C, TRUE)
```

Bootstrapping gives 100% significance on all branches, unfortunately:

```
bs.C = bootstrap.pml(fit.C.JC, bs=1000, multicore = TRUE, mc.cores = 6)
plotBS(midpoint(fit.A.JC$tree), bs.A, p = 90, type = 'p')
```



At the end, `modelTest()` analysis was performed, using JTT model. Result for subset A:

Model	df	logLik	AIC	AICw	AICc	AICcw	BIC
JTT	197	-12417.97	25229.93	1.181493e-174	NaN	NaN	25845.35
JTT+I	198	-12379.87	25155.74	1.523547e-158	NaN	NaN	25774.28
JTT+G	198	-12019.27	24434.53	6.176710e-02	NaN	NaN	25053.08
JTT+G+I	199	-12015.55	24429.09	9.382329e-01	NaN	NaN	25050.76

Result for subset B:

Model	df	logLik	AIC	AICw	AICc	AICcw	BIC
JTT	97	-7140.691	14475.38	7.539854e-80	14739.44	1.240583e-77	14779.55
JTT+I	98	-7123.369	14442.74	9.240354e-73	14716.03	1.497833e-72	14750.05
JTT+G	98	-6957.993	14111.99	6.136733e-01	14385.28	9.947458e-01	14419.29
JTT+G+I	99	-6957.455	14112.91	3.863267e-01	14395.77	5.254178e-03	14423.35

Result for subset C:

Model	df	logLik	AIC	AICw	AICc	AICcw	BIC
JTT	37	-4084.124	8242.248	3.607868e-33	8264.047	8.687541e-33	8357.614
JTT+I	38	-4069.933	8215.867	1.931479e-27	8239.023	2.358814e-27	8334.351
JTT+G	38	-4008.865	8093.730	6.420081e-01	8116.886	7.840507e-01	8212.214
JTT+G+I	39	-4008.449	8094.898	3.579919e-01	8119.465	2.159493e-01	8216.500

Conclustions

In the opinion of the author, the best algorithm for creating phylogenetic tree is UPGMA, because it had the biggest parsimony score. Bootstrapping wasn't successful, because all branches were rendered in 100% of cases.

The alignments made using different algorithms differed. The difference wasn't big, though. In distance matrices, the typical difference was about 0.01 or 2%. From the diagram made by **msa** package, we can see that the sequence coding hemoglobin is conserved well, although it depends on the species. For homo sapiens, only a dozen (about 15%) amino acids were conserved. The synteny depends how far apart the species in the phylogenetic tree are.

From the phylogenetic trees, we can see that the results agree with taxonomy of the organism quite well. Mammals are clustered together, homo sapiens is on the neighbor branch with another mammal, organisms of the same class have often the LCA quite close. It seems not ideal, though: for example, Lepidosauria specimen should be closer on the phylogenetic tree, providing that the same class implies the phylogenetic affinity.