An Introduction to Bioinformatics

WEILONG HAO

McMaster University
DEPARTMENT OF BIOLOGY
HAMILTON, ONTARIO
CANADA

Bio-Informatics

◇To use computers for the management and analysis of biological information

◇Internet + Programs + Data

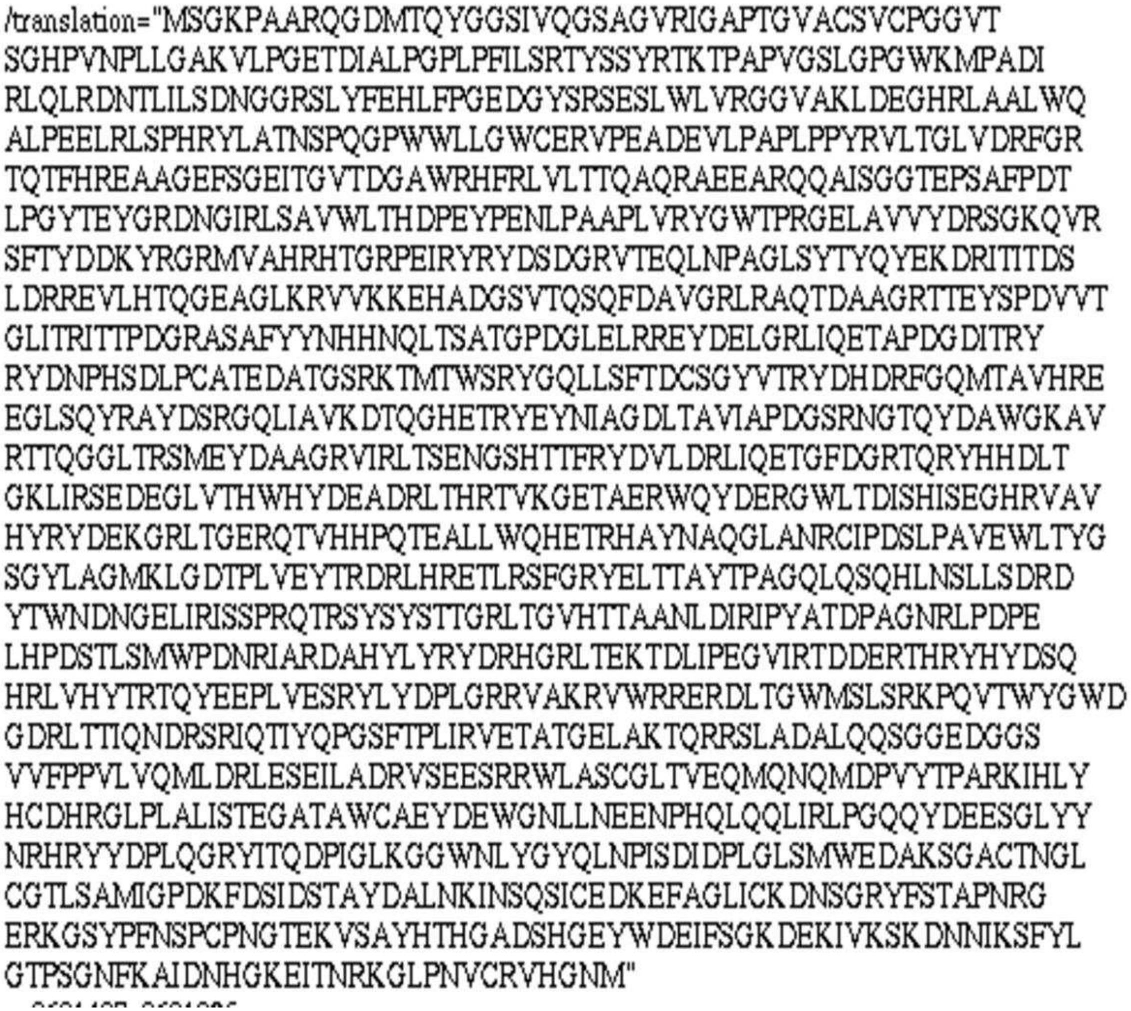
DNA and Protein

A T C G

DNA and Protein

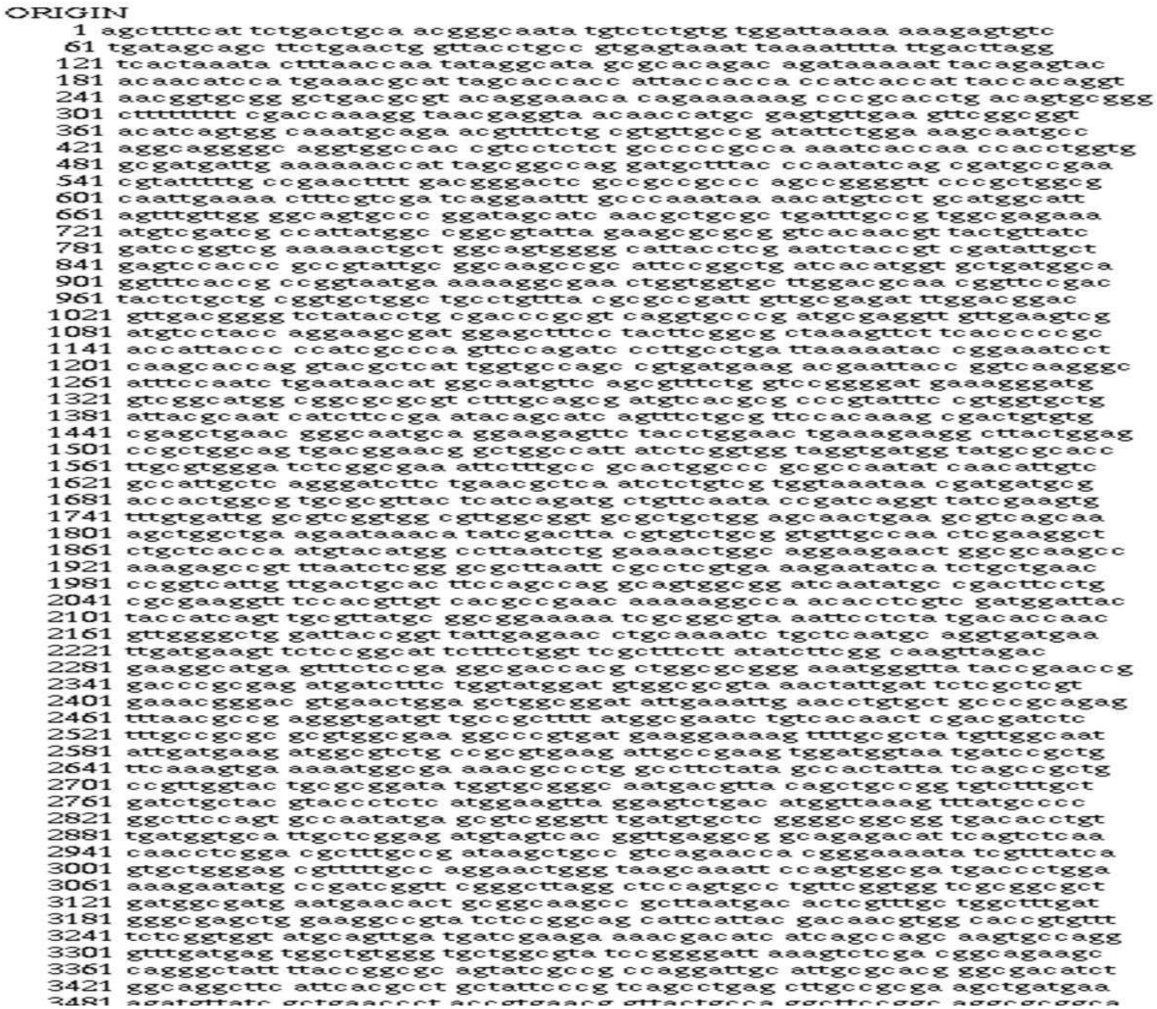
Ala	A	Leu	L
Arg	R	Lys	K
Asn	N	Met	M
Asp	D	Phe	F
Cys	C	Pro	P
Glu	Q	Ser	S
Gln	Q	Thr	T
Gly	G	Trp	W
His	H	Tyr	Y
Ile	I	Val	V

A Protein Sequence

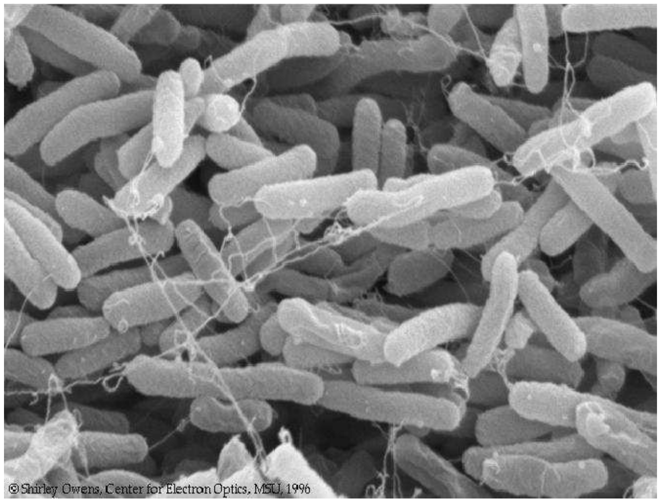


```
/translation="MSGKPAARQGDMTQYGGSIVQGSAGVRIGAPTGVACSVCPGGVT
SGHPVNPLLGAKVLPGETDIALPGPLPFILSRTYSSYRTKTPAPVGSLGPGWKMPADI
RLQLRDNTLILSDNGGRSLYFEHLFPGEDGYSRSESLWLVRGGVAKLDEGHRLAALWQ
ALPEELRLSPHRYLATNSPQGPWWLLGWCERVPEADEVLPAPLPPYRVLTGLVDRFGR
TQTFHREAAGEFSGEITGVTDGAWRHFRLVLTTQAQRAEEARQQAISGGTEPSAFPDT
LPGYTEYGRDNGIRLSAVWLTHDPEYPENLPAAPLVRYGWTPRGELAVVYDRSGKQVR
SFTYDDKYRGRMVAHRHTGRPEIRYRYDSDGRVTEQLNPAGLSYTYQYEKDRITITDS
LDRREVLHTQGEAGLKRVVKKEHADGSVTQSQFDAVGRLRAQTDAAGRTTEYSPDVVT
GLITRITTPDGRASAFYYNHHNQLTSATGPDGLELRREYDELGRLIQETAPDGDITRY
RYDNPHSDLPCATEDATGSRKTMTWSRYGQLLSFTDCSGYVTRYDHDRFGQMTAVHRE
EGLSQYRAYDSRGQLIAVKDTQGHETRYEYNIAGDLTAVIAPDGSRNGTQYDAWGKAV
RTTQGGLTRSMEYDAAGRVIRLTSENGSHTTFRYDVLDRLIQETGFDGRTQRYHHDLT
GKLIRSEDEGLVTHWHYDEADRLTHRTVKGETAERWQYDERGWLTDISHISEGHRVAV
HYRYDEKGRLTGERQTVHHPQTEALLWQHETRHAYNAQGLANRCIPDSLPAVEWLTYG
SGYLAGMKLGDTPLVEYTRDRLHRETLRSFGRYELTTAYTPAGQLQSQHLNSLLSDRD
YTWNDNGELIRISSPRQTRSYSYSTTGRLTGVHTTAANLDIRIPYATDPAGNRLPDPE
LHPDSTLSMWPDNRIARDAHYLYRYDRHGRLTEKTDLIPEGVIRTDDERTHRYHYDSQ
HRLVHYTRTQYEEPLVESRYLYDPLGRRVAKRVWRRERDLTGWMSLSRKPQVTWYGWD
GDRLTTIQNDRSRIQTIYQPGSFTPLIRVETATGELAKTQRRSLADALQQSGGEDGGS
VVFPPVLVQMLDRLESEILADRVSEESRRWLASCGLTVEQMQNQMDPVYTPARKIHLY
HCDHRGLPLALISTEGATAWCAEYDEWGNLLNEENPHQLQQLIRLPGQQYDEESGLYY
NRHRYYDPLQGRYITQDPIGLKGGWNLYGYQLNPISDIDPLGLSMWEDAKSGACTNGL
CGTLSAMIGPDKFDSIDSTAYDALNKINSQSICEDKEFAGLICKDNSGRYFSTAPNRG
ERKGSYPFNSPCPNGTEKVSAYHTHGADSHGEYWDEIFSGKDEKIVKSKDNNIKSFYL
GTPSGNFKAIDNHGKEITNRKGLPNVCRVHGNM"
```

DNA Sequence



```
ORIGIN
        1 agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc
       61 tgatagcagc ttctgaactg gttacctgcc gtgagtaaat taaaatttta ttgacttagg
      121 tcactaaata ctttaaccaa tataggcata gcgcacagac agataaaaat tacagagtac
      181 acaacatcca tgaaacgcat tagcaccacc attaccacca ccatcaccat taccacaggt
      241 aacggtgcgg gctgacgcgt acaggaaaca cagaaaaaag cccgcacctg acagtgcggg
      301 cttttttttt cgaccaaagg taacgaggta acaaccatgc gagtgttgaa gttcggcggt
      361 acatcagtgg caaatgcaga acgttttctg cgtgttgccg atattctgga aagcaatgcc
      421 aggcaggggc aggtggccac cgtcctctct gcccccgcca aaatcaccaa ccacctggtg
      481 gcgatgattg aaaaaaccat tagcggccag gatgctttac ccaatatcag cgatgccgaa
      541 cgtatttttg ccgaactttt gacgggactc gccgccgccc agccggggtt cccgctggcg
      601 caattgaaaa ctttcgtcga tcaggaattt gcccaaataa aacatgtcct gcatggcatt
      661 agtttgttgg ggcagtgccc ggatagcatc aacgctgcgc tgatttgccg tggcgagaaa
      721 atgtcgatcg ccattatggc cggcgtatta gaagcgcgcg gtcacaacgt tactgttatc
      781 gatccggtcg aaaaactgct ggcagtgggg cattacctcg aatctaccgt cgatattgct
      841 gagtccaccc gccgtattgc ggcaagccgc attccggctg atcacatggt gctgatggca
      901 ggtttcaccg ccggtaatga aaaaggcgaa ctggtggtgc ttggacgcaa cggttccgac
      961 tactctgctg cggtgctggc tgcctgttta cgcgccgatt gttgcgagat ttggacggac
     1021 gttgacgggg tctatacctg cgacccgcgt caggtgcccg atgcgaggtt gttgaagtcg
     1081 atgtcctacc aggaagcgat ggagctttcc tacttcggcg ctaaagttct tcaccccgc
     1141 accattaccc ccatcgccca gttccagatc ccttgcctga ttaaaaatac cggaaatcct
     1201 caagcaccag gtacgctcat tggtgccagc cgtgatgaag acgaattacc ggtcaagggc
     1261 atttccaatc tgaataacat ggcaatgttc agcgtttctg gtccggggat gaaagggatg
     1321 gtcggcatgg cggcgcgcgt ctttgcagcg atgtcacgcg cccgtatttc cgtggtgctg
     1381 attacgcaat catcttccga atacagcatc agtttctgcg ttccacaaag cgactgtgtg
     1441 cgagctgaac gggcaatgca ggaagagttc tacctggaac tgaaagaagg cttactggag
     1501 ccgctggcag tgacggaacg gctggccatt atctcggtgg taggtgatgg tatgcgcacc
     1561 ttgcgtggga tctcggcgaa attctttgcc gcactggccc gcgccaatat caacattgtc
     1621 gccattgctc agggatcttc tgaacgctca atctctgtcg tggtaaataa cgatgatgcg
     1681 accactggcg tgcgcgttac tcatcagatg ctgttcaata ccgatcaggt tatcgaagtg
     1741 tttgtgattg gcgtcggtgg cgttggcggt gcgctgctgg agcaactgaa gcgtcagcaa
     1801 agctggctga agaataaaca tatcgactta cgtgtctgcg gtgttgccaa ctcgaaggct
     1861 ctgctcacca atgtacatgg ccttaatctg gaaaactggc aggaagaact ggcgcaagcc
     1921 aaagagccgt ttaatctcgg gcgcttaatt cgcctcgtga aagaatatca tctgctgaac
     1981 ccggtcattg ttgactgcac ttccagccag gcagtggcgg atcaatatgc cgacttcctg
     2041 cgcgaaggtt tccacgttgt cacgccgaac aaaaaggcca acacctcgtc gatggattac
     2101 taccatcagt tgcgttatgc ggcggaaaaa tcgcggcgta aattcctcta tgacaccaac
     2161 gttggggctg gattaccggt tattgagaac ctgcaaaatc tgctcaatgc aggtgatgaa
     2221 ttgatgaagt tctccggcat tctttctggt tcgctttctt atatcttcgg caagttagac
     2281 gaaggcatga gtttctccga ggcgaccacg ctggcgcggg aaatgggtta taccgaaccg
     2341 gacccgcgag atgatctttc tggtatggat gtggcgcgta aactattgat tctcgctcgt
     2401 gaaacgggac gtgaactgga gctggcggat attgaaattg aacctgtgct gcccgcagag
     2461 tttaacgccg agggtgatgt tgccgctttt atggcgaatc tgtcacaact cgacgatctc
     2521 tttgccgcgc gcgtggcgaa ggcccgtgat gaaggaaaag ttttgcgcta tgttggcaat
     2581 attgatgaag atggcgtctg ccgcgtgaag attgccgaag tggatggtaa tgatccgctg
     2641 ttcaaagtga aaaatggcga aaacgccctg gccttctata gccactatta tcagccgctg
     2701 ccgttggtac tgcgcggata tggtgcgggc aatgacgtta cagctgccgg tgtctttgct
     2761 gatctgctac gtaccctctc atggaagtta ggagtctgac atggttaaag tttatgcccc
     2821 ggcttccagt gccaatatga gcgtcgggtt tgatgtgctc ggggcggcgg tgacacctgt
     2881 tgatggtgca ttgctcggag atgtagtcac ggttgaggcg gcagagacat tcagtctcaa
     2941 caacctcgga cgctttgccg ataagctgcc gtcagaacca cgggaaaata tcgtttatca
     3001 gtgctgggag cgtttttgcc aggaactggg taagcaaatt ccagtggcga tgaccctgga
     3061 aaagaatatg ccgatcggtt cgggcttagg ctccagtgcc tgttcggtgg tcgcggcgct
     3121 gatggcgatg aatgaacact gcggcaagcc gcttaatgac actcgtttgc tggctttgat
     3181 gggcgagctg gaaggccgta tctccggcag cattcattac gacaacgtgg caccgtgttt
     3241 tctcggtggt atgcagttga tgatcgaaga aaacgacatc atcagccagc aagtgccagg
     3301 gtttgatgag tggctgtggg tgctggcgta tccggggatt aaagtctcga cggcagaagc
     3361 cagggctatt ttaccggcgc agtatcgccg ccaggattgc attgcgcacg ggcgacatct
     3421 ggcaggcttc attcacgcct gctattcccg tcagcctgag cttgccgcga agctgatgaa
     3481 agatgttatc gctgaaccct accgtgaacg gttactgcca ggcttccggc aggcgcggca
```



© Shirley Owens, Center for Electron Optics, MSU, 1996

E. coli **K12 genome is 4,639,675 bp**

Human is about 3,000,000,000 bp

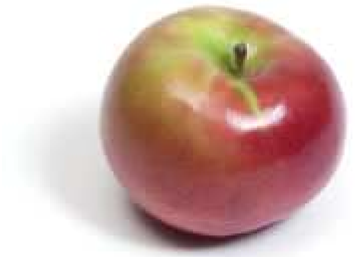
How to analyze the sequences

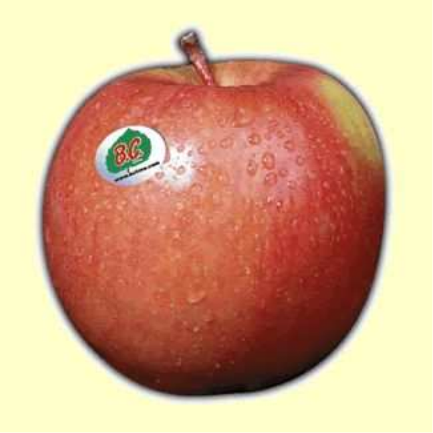
>Sample1 MAKAAIGIDLGTTYSCVGVFQHGKVEIIANDQGNRTTPS

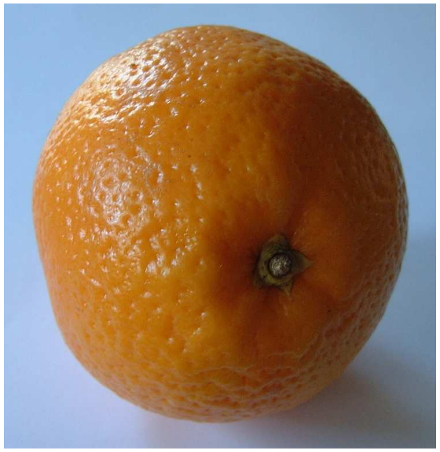
>Sample2 TMAKAAISIDLGTTYSCVGVFQHGKVEIIANDQGNRTTPS

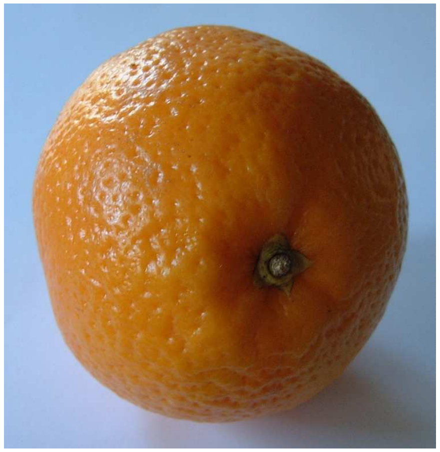
>Sample3 MAKAAIGIDLGTFTYSCVGVFQHGKVEIIANDQGNRTTPS

Which two are more closely related?



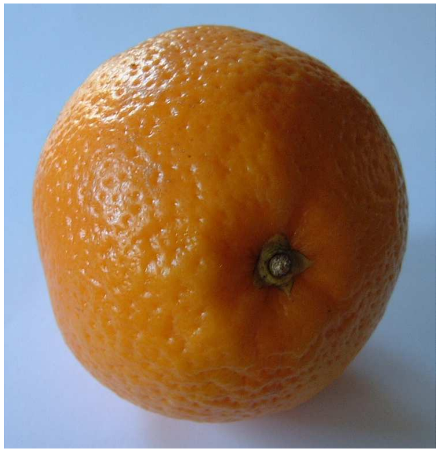






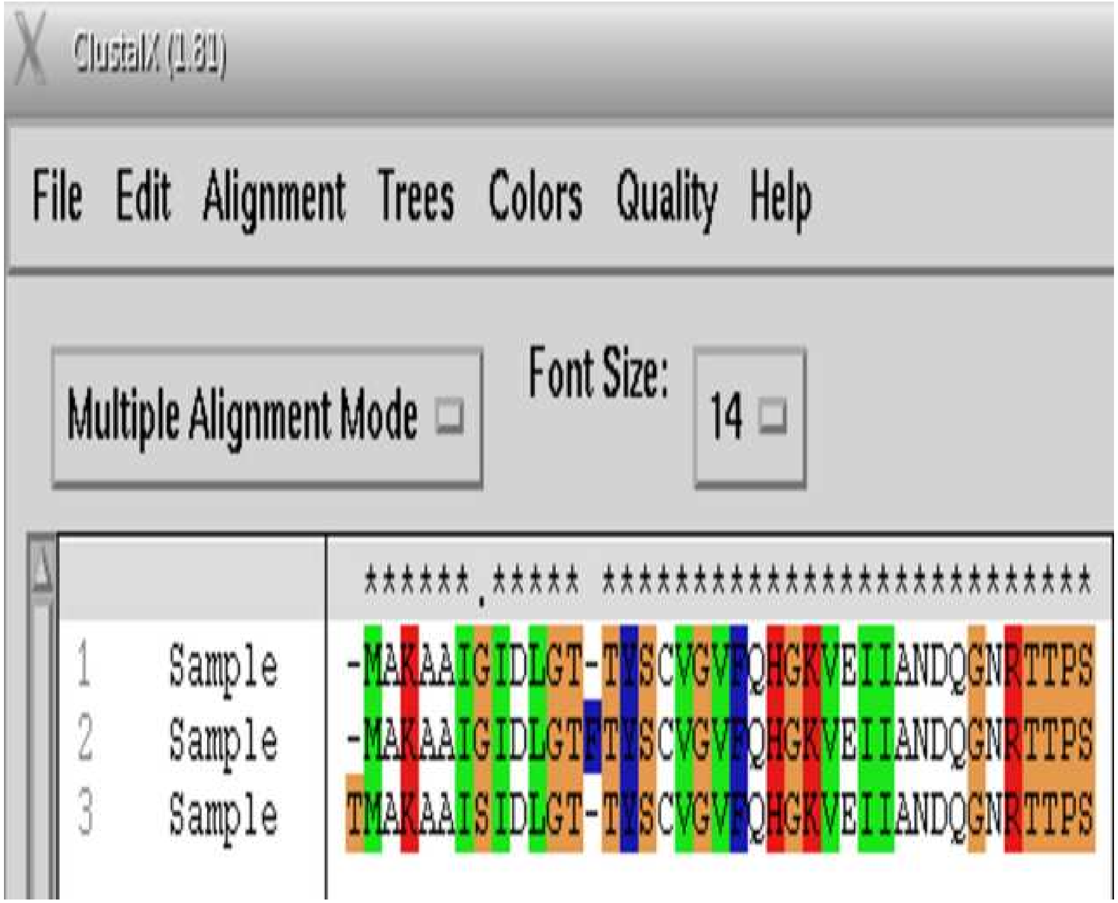
?

B.C.



?

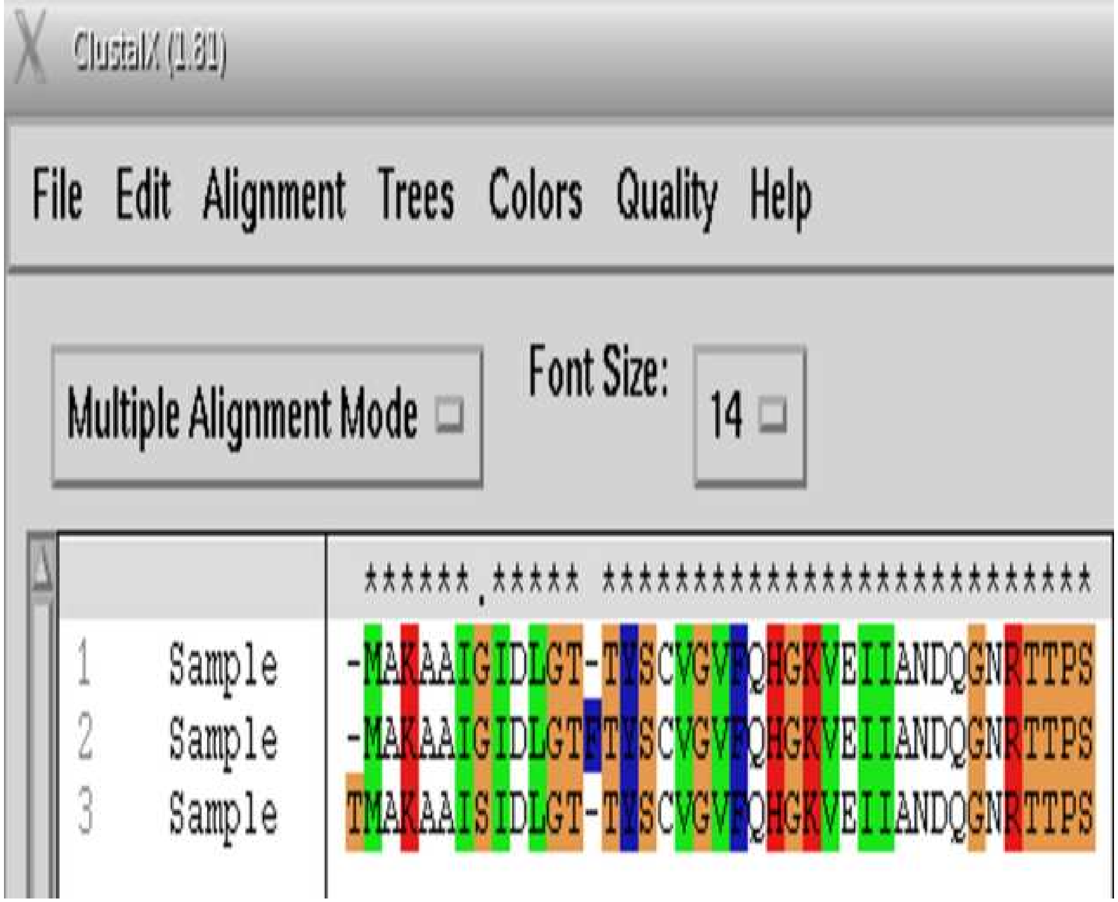
Alignment



Multiple Alignment Mode □ Font Size: 14 □

```
                    ******.***** ***************************
1    Sample   -MAKAAIGIDLGT-TYSCVGVFQHGKVEIIANDQGNRTTPS
2    Sample   -MAKAAIGIDLGTFTYSCVGVFQHGKVEIIANDQGNRTTPS
3    Sample   TMAKAAISIDLGT-TYSCVGVFQHGKVEIIANDQGNRTTPS
```

Alignment



Multiple Alignment Mode

Font Size: 14

```
                  ******.***** ***************************
1     Sample     -MAKAAIGIDLGT-TYSCVGVFQHGKVEIIANDQGNRTTPS
2     Sample     -MAKAAIGIDLGTFTYSCVGVFQHGKVEIIANDQGNRTTPS
3     Sample     TMAKAAISIDLGT-TYSCVGVFQHGKVEIIANDQGNRTTPS
```

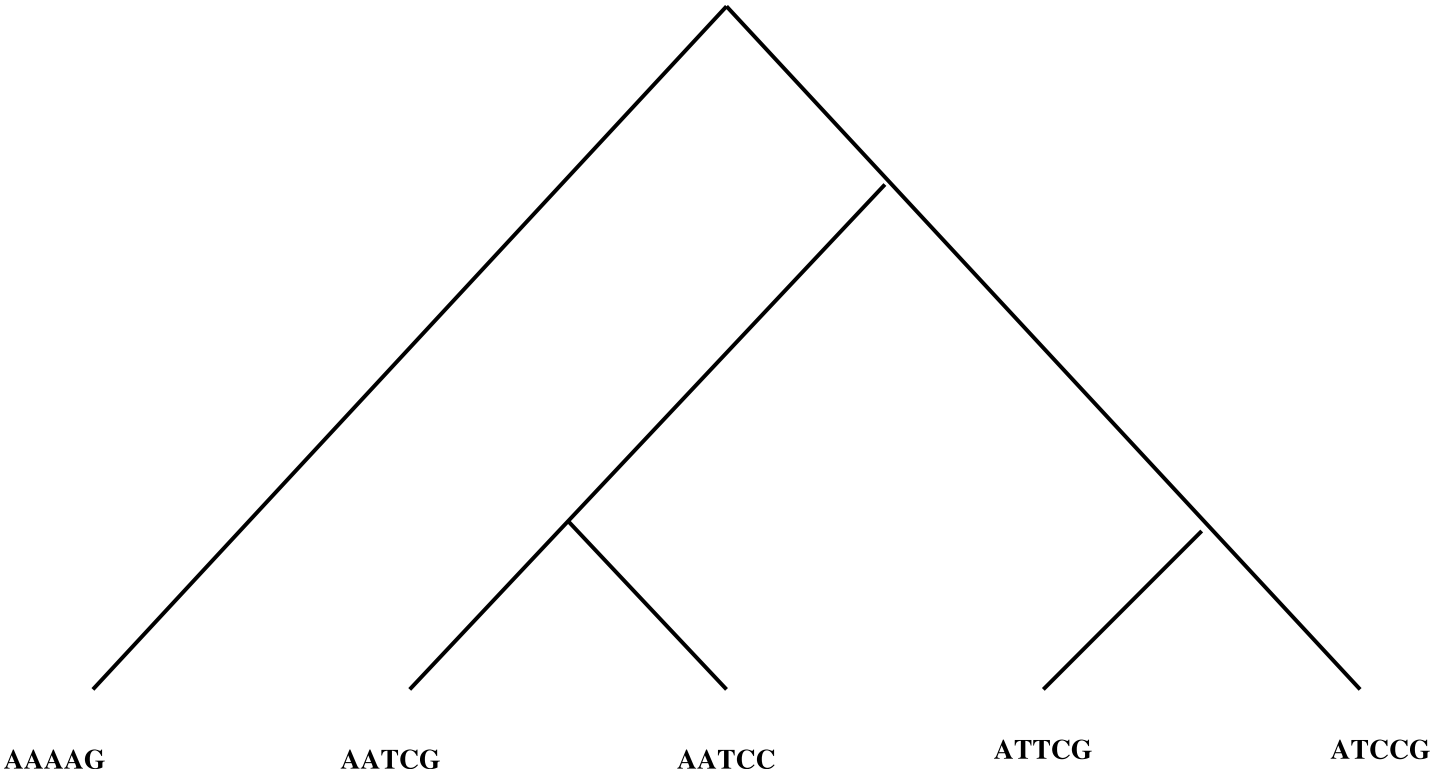
Difference ⟶ Distance

Distance → Tree

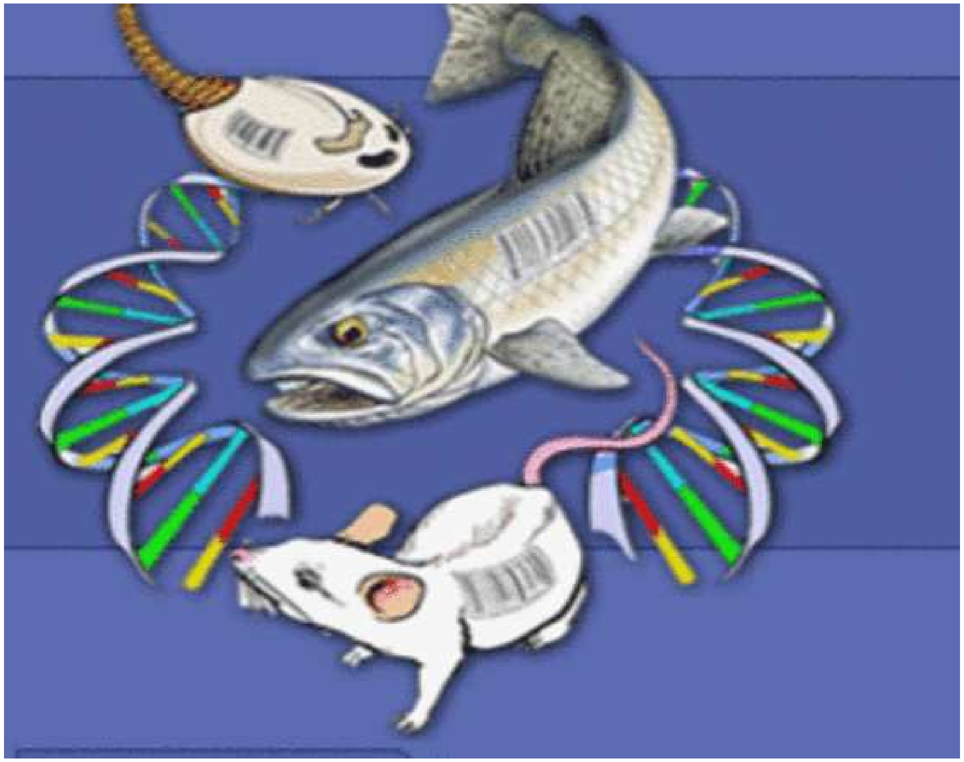
Parsimony Tree

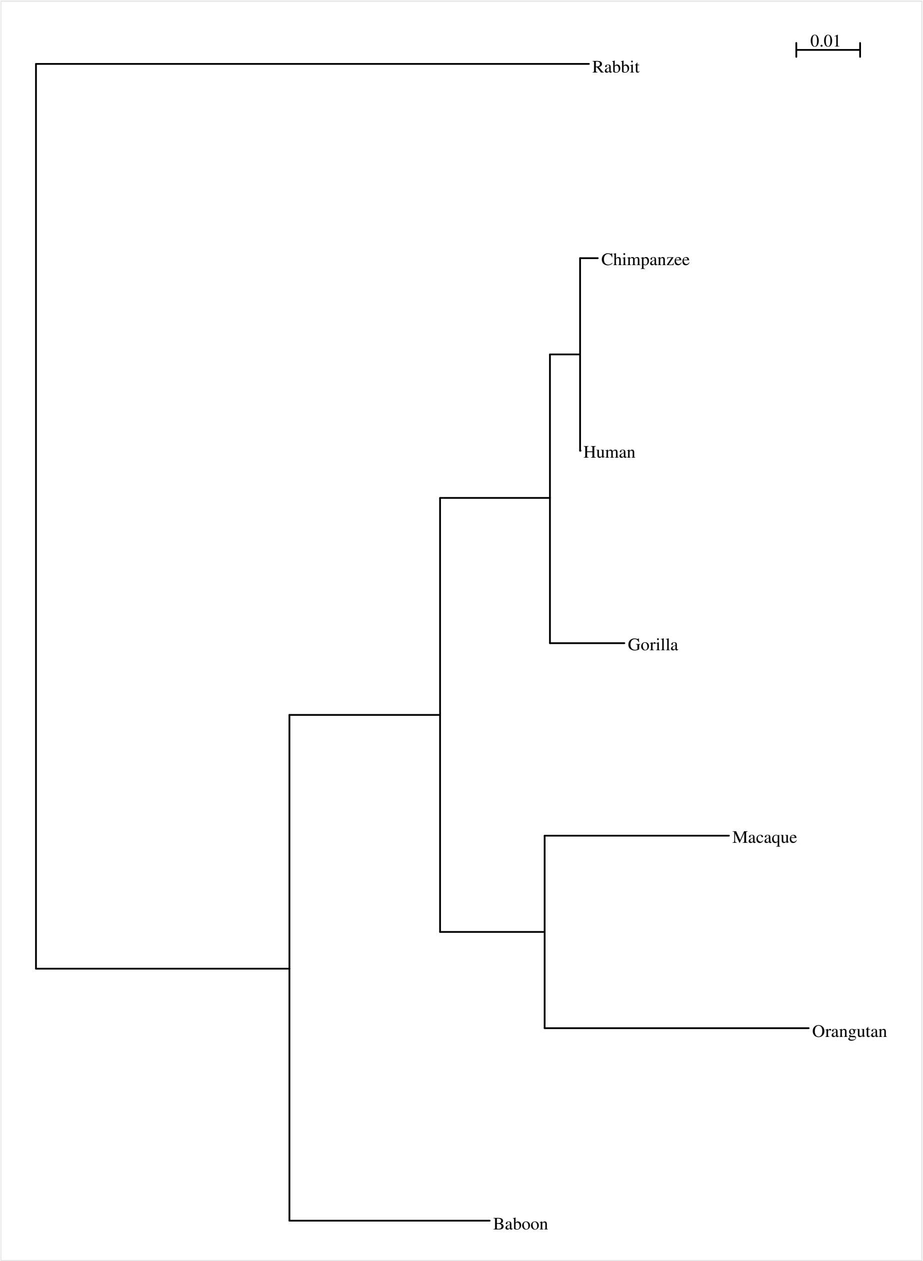
AAAAG AATCG AATCC ATTCC ATCCC

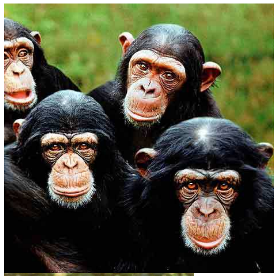
Parsimony Tree

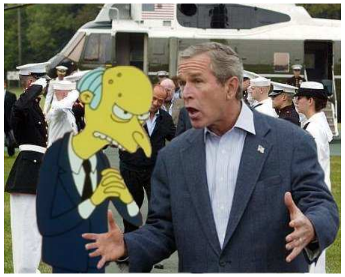


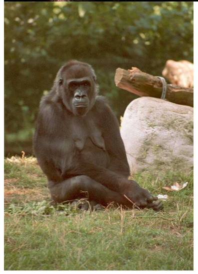
One application



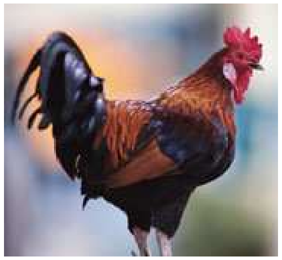

0.01
Rabbit
Chimpanzee
Human
Gorilla
Macaque
Orangutan
Baboon

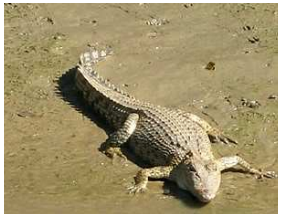


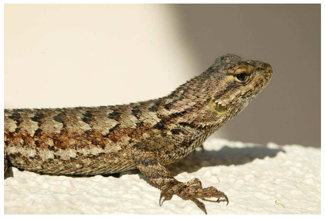


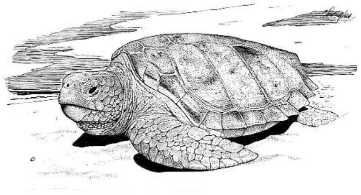


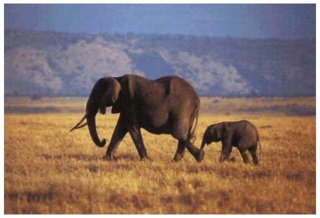
Relationship?

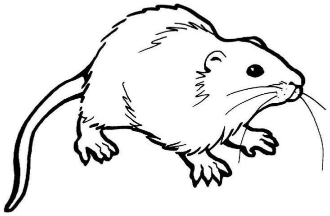


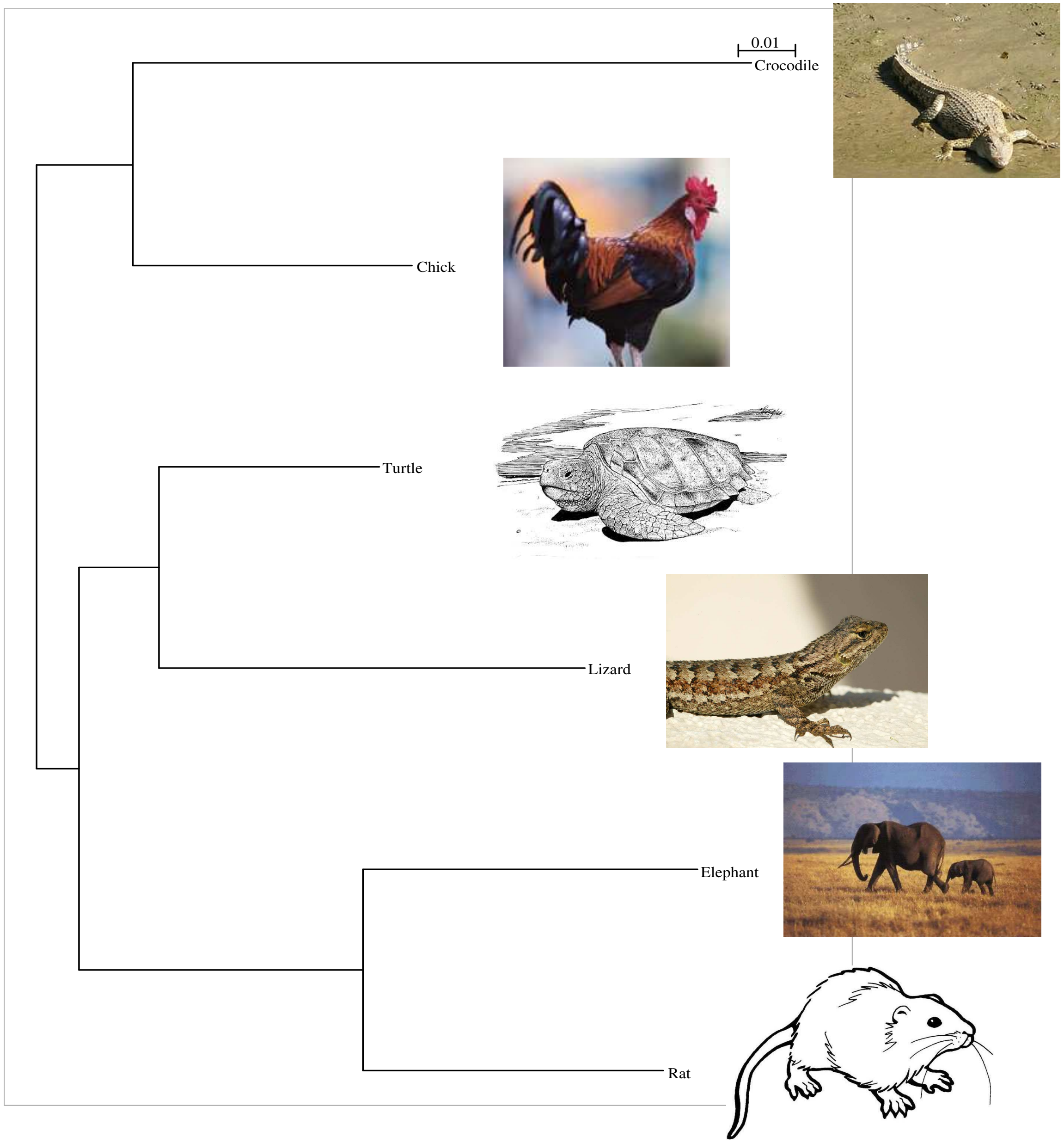











0.01
Crocodile
Chick
Turtle
Lizard
Elephant
Rat

Something to think about

Why should we do the alignment before we compare genes?

How do we get the relationship of genes?

More applications...