Exercise 5 – Analysing and aligning newly discovered proteins

Objectives:

- to apply what has been learned today

Anonymous Test Proteins:

below, we provide 20 randomly chosen proteins. All have been derived from DNA on the teeth of ancient skeletons found in a german monastery (same as for the previous exercises). None of the proteins have been analyzed in detail before … Please select arbitrarily one of the proteins below, and analyze it like we did in exercises #1 through #3.

Optionally you can also study some protein sequence related to SARS-CoV-2, which are listed at the bottom below.
Questions:

- what protein family does your protein belong to?
- which domain(s), if any, does the protein contain?
- from which organism is it, likely?
- what function might it have?
- is it complete?
- how can it be best aligned to other members of its family?

```
>NODE_4178_length_1047_cov_6.240688_S6
SATSAAAMKLIAPSWPSRYVSPPRRRQGAAMAEWWSAQLVDRDGRIRGELPDIRGGSLEW
NISSAVRTGGSVEFAEPPSAGIDWVTTRIRILHHDGAEVRPMGVYRASWPNRKLRDGHTS
STLKLEAPTSRLRSQLGYWTQYEAGIVVTDRVAMTLRQLGESQLALTPSPQTLRTPLTWD
PDKTWGTLYSELLDAIGYGGIWCDANGWWRAAPYVAPMERPLAATYGGDPADYRCRTTYG
DEADWTDVPNRVLLYTRATSEAPALTSEVWITDPANPWHPDRVGPHTRCEAVEATSQEVL
DAKAKRLLAEGQERSRYITWTHPVDDTTLGDRVRIRRLGLDVAIEARK
```

```
>NODE_25515_length_1898_cov_8.371970_S6
RGGKMKIIIAGIGNIGAGLAGRLLNEGHDIVLVDRDIDRLEYNEETQDVMTVKGTCAAME
TLRKAGVEDADLLITATSSDEKNLLSCMTAHGMNPNIKTVARVRMQEYLETTSVFGEKFG
LSMIIDPGMYAAKDIEAILTYPGFLHRERFTKGMTDVVEAELHPESELCGKPVSAIQEIT
GSGALVCVVKRGDTAITPGRDFILREKDRIYVTAEEEDLSTLLRLFGKKKETVEKVMIVG
GGRIAKGLIPRLQKEGMEIIVIDTDKEICEELAMEFPKVNIVHGDGRKFALLERKNVREQ
DALICLTNDDESN
```

```
>NODE_60099_length_1027_cov_6.267770_S6
RVASVCLSTIVAVWMTSLIVPWASYSVKEGSRWAIESMYESRMVTKDDRDAFNWLAKQPH
AYDGIIFGNSADGYGWMYAYNKLPSLARHYDGVSAKPGAPSHVLRDSAYLIGAGNHGDPD
QRNRADLAAENLGVNFIMLSPPNFWWFQQSNLEMSAKLDKAPGLTLVYQKNSIRIYAVNA
KFKDAELTRMRASGPASNQLPVPQCPKDSADGKAAATAGETTQVEYDPDTGEQTTVTKPK
PCYHRPSKPDIPPRANDAKGKTPATPKSGGGDDKSNKYSEKTGLDLTEKEARRRRLDNGY
VHNEKATLRF
```

```
>NODE_77700_length_886_cov_21.930023_S6
WANMCPRKCKVMSMKRNQSAVHQLITYGMVIAAYILCQILVENGSMTRSLKGQLIPIAVY
IVMAVSLNLTVGISGELSLGHAGFMSVGAFSGIVVSQWMGTVYPNVHVYVRLVFAIVTGG
IAAGIAGVLIGIPVLRLRGDYLAIVTLAFGEIIRNIMNLLYVSVDQGRLRMAFNDGALPG
EQVIAGPKGAVGIEKIATFTMGFILVMITLFVVLNLINSRSGRAIMAIRDSRIAAESVGI
NVTKYKMMAFVISSVLAGMAGALFGLNYSTVSAGKFKFDMSILVLVFVVLGGIGNIRGSV

>NODE_87482_length_1095_cov_5.276712_S6
NPLIARTRGQQRDAHSVHARYGDKYLPFSDLENSMRDMEGLLNKVADLAVKAGSIMLSDS
DVEVGNKGTKENYVTSTDLKVQRFLREGLATLLPGAVFRGEEDDLPREDEGTRGEYVWIV
DPIDGTANYARGFGESAVSIALAKDDEPVLGVVRNPYARETYCAIKGRGAFLNGTPIHVS
GRSKENAMICLSWSAYDKSRSADCFRISQDLYAVCEDIRRTGSAAYELCLLARGSVDMHF
EIRLAPWDYAAGGLIIEEAGGRTGSLEGRLDMRRQCLVMAANSEKNFAFLKGVVSENLSL
RRRLAPVHV

>NODE_107984_length_1345_cov_80.271378_S6
ERKAYSMGKRTIIPFGPQHPVLPEPVHLDLVIEDETVVEAIPSIGFIHRGLEKLVEKKEY
PEMVYVIERICGICSFGHGWGYCAAVEGAMNVEIPERAMYLRTILHELGRMHSHLLWLGL
LADGFGFESLFQHCWRIRETVLDLFEQTTGGRVIFSICKVGGLNKDIDNETLNKIVKTLR
GIEKEIREYTSVFINDTSVKNRLTGVGVLSREDAEALCTVGPMARASGLRQDMRLAGEGK
YLELGFEPVLEEAGDCMARCKVRIGELLQAIDIIEKAVAQIPDGDIAVAVKGNVDGEFIN
RLEQPRGEAFYYCKGQGTKFLERIRVRTPTNMNIPAMVKILQGCDLADVPMIVLTIDPCI
SCTER

>NODE_123020_length_4291_cov_7.623631_S6
AVFEERWGDRPFMRSYRIPSIPVRPIWICVSRQNRAVLCLKTYIQMEQAILGAKREPVCQ
AASHALGPSAEDSCLTARPDPMRVDYDTDVRAFAQRLLGGNVFEPVTFAGITLPLISFIL
FGAALAFLLIVQVARTMISNKLQNLFASKLYDEFLDTVDEPLTRFFIPAYNRTYLRLNAF
MAKGSVEKAMEAFDQLLAMRSTRAQRDDLLFKAFQFYMQQEDFKGAKAVLDEMQSYGRHE
KRVEECVQAYEIFGNNSYAYIDEMEAAFDEAPYALKVSYALMLAAQYTSKKDGEAAEKWQ
DTARELLENPPKKGPAETR

>NODE_182329_length_1939_cov_4.566271_S6
APDDPPRHRRREQREKLFRRTTCLHPWGRVLLRGDHGAAQRRSTRAYRTASQTARREKVR
DHRAAARSGRARPAARSGARRGATGGYRELGGKRAVRCRAVRRPRIRVLGRPRGRLRAHH
GRNRPSERALLPSRSRHLRSRVGRTPHRTRNALLYRAYRTGELHDCRPGSNGARVRGKHR
ATAQRGICRMTALRSIALAFTLFSRVPMPHVEWNPENMRYTMLAFPLVGCVIGTAVATWC
ALCATLGLNGAAFGAGTVLVPLFVTGGIHMDGFADVVDAQSSHAAPERKREILADPHIGA
FAAIGIGGYLLAWAALAS

>NODE_212586_length_1033_cov_30.919651_S6
ISKTDESYPDFLRPSDGALHPAVNEYRSLWISLSLKGALPGLYPIHIVVEQDGEECYRAT
LCVRVCTAPLEKQKLIHTEWLHADCLCSYYNVEAFSERHFALLENFIRAAVQDYGINMIL
TPVFTPPLDTQVGGERRTVQLVDIACDSRGYHFDFSKLARWADICKRCGVEYLEIAHLFT
QWGAQHCPKIIVTEKGRERKKFGWQSDAAGTEYRKFLEQFLPALRSALQGMGYPDEKVYY
HISDEPSEDNLEHYRRAKAQVADLLEGANVVDALSSYRFYQEGLVTEPIVSSDHIQAFLD
AGVPNLWVYYCCGQDKLVPNRFFAMPSPRNRVFGVLLYLSGVKGFLHWGYNFY

>NODE_238737_length_1166_cov_7.374785_S6
NRHQTMFKGEIVMNSLIIVSAALGLCALLFALVLAARVKSQDSGTERMTEIAAYIHQGAK
AFLMAEYRILVIFVAILFVLIGLGISWITAVCFLVGAAFSTVAGYIGMNVATAANVRTAA
AAKDKGMNAALSVAFSGGAVMGMCVVGFGLLGASLIYFVTGNSEILSGFSLGASTIALFA
RVGGGIYTKAADVGADLVGKVEAGIPEDDPRNPAVIADNVGDNVGDVAGMGADLFESYVG
SVVSAVTLGLVAYNQEGAVFPLLIAALGIGASIIGSFFVKGDEKSSPHKALKFGSYASSV
LVAVGSLALSYKFFGNLNAGMAIVFGLVVGLLIGLVTEIYTSSDYKFVKKIADQSETGAA
TTVISGIAVGMQ
```

```
>NODE_264747_length_1361_cov_29.963263_S6
GICQGGHSSRQPYHRLLWHRTGGYMIRLLLKRRELSALFFLILLFLIAGIVNPAFLTLNN
VFLSINSSVVYAVVAMGIAFVIITGEIDVSVGAIVGISATVVGSMIRDGQPWLLALLAGI
GIGMLIGLINGFGVVTLRIPSIIMTLGTSSIIRGLMYVYTDGKWVENVPFEFKQLSQQKF
LDSFTYFYLAILLFMLLVHLIMMRSKRGKYYAAVGDNAAGANLLGIPVARTKLTAFVICG
VLSALGGVIFVSRVGFVTPIAGVGYEMKVIAACVIGGISLSGGVGNILGACIGAAFMASI
SRVLVFIGLSSDLDDTITGVLLIIIVVVDALLRKRSIEHARRERLSAKTLDLGGINNEAK
TV

>NODE_301074_length_916_cov_4.279476_S6
VVVGTMARSAELPLIIQIGATFNSIFGNFLGFCIPLIIIGFVVSGIAELGDGAGKTLGLT
VLIAYASTLFAGLLAYFVDVSVFPSFLKVGSIVLEDAQNAEETMLKGLFSIDMPPLMGVM
TALLLSFIFGIGIAVTHSTSLKNGFSEVQHIIEKLVAGVLIPLLPLHVYGIFANMTYAGT
VMDIMSVFIRVFAIIILLHVAVILIQYTIAGTVVGRNPIKLIRRMLPAYFTAIGTQSSAA
TIPVTVACTKSNDVSDRIAEFVCPLCATIHLSGSTITLTSCSIALMMLNGMDVTLGGLFP
FILMLGITMVAAPG

>NODE_313178_length_2508_cov_7.222488_S6
MLNKYGADATRWYLLHVSPAWSPTKFDEGGLQELASKFFGTLRNVYNFFVLYGNLDKIDV
KKLSVPYEKRSELDRWILSKYNKLIAEVTEHMDRYDHMKTVRAITDFVNEDLSNWYIRRA
RRRFYTPGMSADKESVFATTFEVLEGVARLIAPIAPFISDEMYSKLTGEETVHIAYYPKT
NAALIDEKVEKRMDIVRSVCNLGRGIREKKGLKVRQPLSEILVDGKYKDLISDMIPLIMD
ELNVKQVVFADELGEYMNFELKPNFKVAGPALGKKINTFAGVLAKEDAEKFTEKLEKDGF
VTCKMDGEDFKIEKEFVDIGINAKQGFAVAMENNVFVIIDTNLSQELIDEGIAREVISKI
QQMRKQNDYDMMDNINVYISADAEVLGAVSKHEAYIKSETLAKTLEEAANLPEVDINGHK
TGLQVERVQN

>NODE_338494_length_1128_cov_14.833333_S6
HGRLRDEHLQRGPRLQDDPGRQPAHQRPAAPGADQPLPGPGVLRGHRRADDPARPGRVLR
GRLRLRGLPLARQGHRHEHPPARDDDPLRRHDDPAVPALREGRARQLPVGRHPADDLHAL
PHPAVPAGLALLPARDHRGGPSRRSERDRHLRAYVRAYNEVDLRGGRRRHFHERVEQLHV
AQDHPRRRQVPDDADARVQPRGRVRHRLRRPHARRPHRVAARDGGLPRPAALLRQRNHGI
SQVNTELSHLTDPTCFADNRLPAHSDHLWYATEAEVASGRSSFQVCLDGVWKLHYATNPS
QAVEGFEVPSYDVSEWDDIAVPAHLQLHGYDKPQYANIQYPWDGHEQLEPGQVPSRYNPT
ASYVRAFTLPQVLPEGERLVLRLE

>NODE_377851_length_1918_cov_6.185089_S6
LRALARLDEAHRAARTHLHPLETGRKDRIMTMLSRRAFLSTCSGLGAAALAGCAPASGTD
DDATPDGGADGPSGLTKVSFVLDYSPNVNHTGIYVAIDQGFFAKEGIEVEIVPVPADGSD
ALIGAGGADMGLTYQDYIANSLSSANPLPYTAVAAVVQHNTSGIMSRAEDGIVRPKDMEG
HSYATWGLPIEQATVKQVVEEDGGDFSKVALVPYEVDDEVMGLQAGLFDTVWVYEWWAVQ
NAKLQEYPVNYFAFADISPQFDFYTPVIAANDAFAAADPELVRAFLRACEQGYELAATSP
ERAAEILCGAVPELDPALIAAAQASISPQYTADASRWGVIDRSRWTRFYEWLNDTGLVEN
GFDPALGFTNEYLEG

>NODE_414935_length_1586_cov_4.661412_S6
GKKNDMGMTMTQKILAAHAGLPQVKAGQLIEAKLDMVLANDITGPVSIGEFYRSGFENVF
DRKKIALVMDHFVPNKDIKSAEQCKKCRTFAKRLDIENYYDVGEMGIEHALLPEKGLVAS
GEAIIGADSHTCTYGALGAFSTGVGSTDVTAAIATGKTWFKVPQAVRFVLRGALKPYVCG
KDVILHIIGMIGVDGALYKSMEFTGDGVRSLTIDDRLTIANMAIEAGAKNGIFPVDSVTE
EYMAGRVTRPYKVCEADEDAEYEKTYNIDLSSIEPTVSFPHLPENTKAISECPDIEIDQV
IIGSCTNGRMQDMKQAADILRGKHMAKGVRGIVIPATMTVYKECIRLGYINDFIDAGCIV
STPTCGPCLGGYMGILADGERCVSTTNRNFVGRMGASGSEVYLAGPAVAAASGIAGKIAD
PRKTL
```

```
>NODE_458259_length_940_cov_5.839362_S6
RLYELTNKIAKPAVSFGGKYRIIDFPLSNCANSNINIVGVLTQYESVFLNSYVTADARWG
LDASDSGIFVLPPREKAGEDLNVYRGTADAISQNIDFVDQYEPDFVLILSGDHIYKMNYE
KMLEEHKASYADASIAVIEVPMKEASRFGIMNADATGRILEFEEKPEKPKSNLASMGIYI
FNWKVLRRMLVSDQKNDLSSHDFGKDIIPKMLDENKILHAYKFSGYWKDVGTVDSFWEAN
MDLLDPHNELSMFDPTWKIYTEDSYTLPQYIGKEAKISSAFITQGCVVEGRIERSVLFTG
VRVAKGAKIVDSVLMPGVEIGE

>NODE_515146_length_1002_cov_3.901198_S6
IFMKKHLVIVESPSKSKTIEKYLGNEYRVVSSKGHICDLATRGKERLGIDVDNNFEATYS
ISKEKKEVVKELQAFVKKSKDVYLASDPDREGEAIAWHLARVLDLDIENTNRIVFHEITK
PAVLEALKHPTHIDMDLVRSQETRRFLDRIIGFKLSRLLQNKIHSKSAGRVQSVALRLIV
ERENEIKAFQPQEYWTIHADVTKGKKKFEAVLSKVDGKKPKLNNEEDSHVILERCKEGDF
IVGKRTKRAKKKQARIPFTTSTLQQEASTKLNFGARRTMSIAQKLYEGIDLGGQQEGLIS
YMRTDSTRLSPMFVDDTLKYIEQTYGKEYKGTIRQKNSANAQD

>NODE_1060560_length_4372_cov_6.979186_S6
PVMERIIQDIVSAVRSAHRPPDEAWLAKLIRRYNKDVRDVARHTKKQQILAFYRKAREER
GQLWESWGIGAEEDRQILRLLKVKPRRTASGVATITVLTMPHPCSSACLYCPNDIRMPKS
YLANEPACQRAERNFFDPYLQVRARLALLESNGHITDKIELIVLGGTWSDYDPSYQIWFI
SELFRALNDGDGEAERICAERAAFYRSCGLIAEADTLAEQTRDLQRCVTAGALSYNQAIA
RLYASEAWVRARARQTATFGELEEQQRINESAHHRTVGLCVETRPDLVDDASAQLMRHLG
CTKVQMGIQSLDQDILDACGRHIRVEQIARAFSVLRLHGFKILAHMMVNLVGSTPEHDRL
DYGRLVGDPRFLPDEIKLYPCVLVESAALARLYDQGIWRPYTEDELLDVLAADVAATPAY
VRISRMIRDISSGDIVAGNKKTNLRQMVDARTEAAESAIAEIRSREIATGDVSACDVRLD
CISYTTAVSEERFLQWITDAGSIAGFLRLSLPHGRSTAMIREVHIYGRVAELGSIEAGGA
QHLGLGSALVETACKQASAAGCSAINVISSVGTRAYYRKLGFIDDGLYQRRVLGT

>NODE_1102966_length_2142_cov_5.032213_S6
WLRAVPAVSRCEYLTPLLRAVCVRCQFVTLPPLASKADRKRDASRYSRERACELPACFLG
WNKQPQLLFIYSTRDCRSRARPYFLHAGECAGRPCGSMRNRGHMAISVGIVGAAGFAGIE
LVRLVLRHSPFDLMAVTSTELSGRRLDEAYPAFAGQCDLAFSPHDADDLQSCDVVFLAVP
HTAALTFAPALIARGATVIDLSADFRLKDPAIYEEWYRVPHTEPELLARAAFGLPELFGE
ELAALAQRRSAGEGVLVACAGCYPTATSLAAAPVLRAGLSPAGLVVVDAVSGVTGAGRKA
TERTHFCFANEGVEAYGVGAHRHTPEIEQILGLEGRLIFTPHLAPYNRGLLSTVTMPVTR
GAFDQAELAEMYRSFFKDAPFVTVLPEGRQPRTVSVAGTNYAHVSACYNERAGAVVATCA
IDNIGKGAAGQAVQCANIVCGLPETCGLDAVALPI
```

SARS-CoV-2 related proteins:

```
>pdb|6YLA|A Chain A, SARS-CoV-2 RBD
ETGPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVY
ADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFER
DISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNKHHH
HHH
```

Or go to this URL:

https://tinyurl.com/2f8h5vwf

(https://www.ncbi.nlm.nih.gov/protein/?term=Severe+acute+respiratory+syndrome+coronavirus+2%5Borganism%5D+AND+protein_structure_direct%5BFilt%5D)

click any SARS-CoV-2 related protein name and then click “FASTA” button at the top of the new page and use them to repeat exercise 1-3