**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 is reason why we have to do this course online …

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://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