protein seqfasta dictionary function

October 11, 2020

[35]:
```
x = """>gi|27544883|ref|NP_775396.1| putative ribosomal protein S3␣
 ↪(mitochondrion) [Lecanicillium muscarium]
MEKNIKLKNILLSKVPQKITLNNKIADIRKTKYLPSFSKEWKDTIYSYNKNIMKNIPSHHLNINKIIQSYFNLFFSTNKN
NQNKRFISFGKYITMKRRRNLLRKIYVSNPVFKFSYFSAFITLFSLSREKVFYKKNILKRIKKYKVIKCNVIYFYIKIIK
WIYKIFLAKQYYKNKNISYFLLKNKNKFIQYKLKYLSKFLLLKNLYLKRVWSKIIKNFIKRHLIFLRKYELLYSLNQLKF
NKLTLLNKLSLLLNKILGKKIEYNIINLKSIIFNSDLFTQAITLKFQKRKSFNYKKNILSILGRVNFPYFKDVVALAYNE
STNYILNKYKDGKILSYINNHQNLDNFINKIHNATNKNIHKEIFNSIQYKNIEGIRIETNGRLTKRYRADRAVHYRKWKG
GLQKTSLNSTLFRGNVNPNISYSIVNNTRRVGSFAIRGWISGK
>gi|27544884|ref|NP_775397.1| NADH dehydrogenase subunit 2 (mitochondrion)␣
 ↪[Lecanicillium muscarium]
MIIISILSLLLSNAVNLRRDVSILYNRIAILILLYCIVHDYTSLTVVTKGIGLHGGLLLINNLTQIFHIFVFIVTIFILT
LTSFYPRKVWVSEYSSIKDLEPVDPLFNKFIYYNTKIINKMGEHLKIIEYPLILLFIVRPGAIFLMSSNDLITIFLSIEL
QSYGLYILSTVYRNSELSTTGGLIYFLLGGLSSCFILLGTALLYAKSGRTSLESLYIITSISDIHSSTNDLWYSPNYISL
SLVIFTVGFLFKISAAPFHFWSPDVYDAIPTIVTTFVALIAKISILILLLQLVYYTNADITMNNWTFILLISSLFTLVIG
TVVGLTQFRIKRLFAYSTISQIGFMLLALSISSIESTQAPIFYLTQYIIINLNAFIILLAITYSLYFYTNYNKEHKDLLD
KTNSPIQLITQLKGYFFINPVLALNLTITIFSFSQIPPLLVCFCKQIVLSASLNQNLYFITLIAILTNFIERIYYLTIIK
KIFFNKSDYKINTLLCNFKLKINIYNNNNSINSVEYNYKNIFLSIPISFIISILTLTILFFLFINKKWLIMITIFVQLIF
NS
>gi|27544885|ref|NP_775398.1| NADH dehydrogenase subunit 3 (mitochondrion)␣
 ↪[Lecanicillium muscarium]
MSGTTFLFIFVCVIAILFLALNFILAPHNPYQEKYSIFECGFHSFLGQNRTQFGFKFFIFSLVYLLLDLDLEILVIYPYG
LSSYENGIYGLIIVLLFIGIITAGFVFELGKGALKIDSRQSYNYFNVQSTKNFINTFFENK"""
```

[36]:
```
def protein_dict(seq):
    pro_dict = {}
    seq_split = seq.split('>')
    seq_split.pop(0)
    for i in range(len(seq_split)):
            for pro in seq_split:
                pro_dict[seq_split[i].split('\n')[0]] = ''.join(seq_split[i].
 ↪split('\n')[1 : len(pro) + 1])
    return(pro_dict)
```

[37]:
```
protein_dict(x)
```

[37]:
```
{'gi|27544883|ref|NP_775396.1| putative ribosomal protein S3 (mitochondrion)
[Lecanicillium muscarium]': 'MEKNIKLKNILLSKVPQKITLNNKIADIRKTKYLPSFSKEWKDTIYSYNKN
```

```
IMKNIPSHHLNINKIIQSYFNLFFSTNKNNQNKRFISFGKYITMKRRRNLLRKIYVSNPVFKFSYFSAFITLFSLSREKV
FYKKNILKRIKKYKVIKCNVIYFYIKIIKWIYKIFLAKQYYKNKNISYFLLKNKNKFIQYKLKYLSKFLLLKNLYLKRVW
SKIIKNFIKRHLIFLRKYELLYSLNQLKFNKLTLLNKLSLLLNKILGKKIEYNIINLKSIIFNSDLFTQAITLKFQKRKS
FNYKKNILSILGRVNFPYFKDVVALAYNESTNYILNKYKDGKILSYINNHQNLDNFINKIHNATNKNIHKEIFNSIQYKN
IEGIRIETNGRLTKRYRADRAVHYRKWKGGLQKTSLNSTLFRGNVNPNISYSIVNNTRRVGSFAIRGWISGK',
 'gi|27544884|ref|NP_775397.1| NADH dehydrogenase subunit 2 (mitochondrion)
[Lecanicillium muscarium]': 'MIIISILSLLLSNAVNLRRDVSILYNRIAILILLYCIVHDYTSLTVVTKGI
GLHGGLLLINNLTQIFHIFVFIVTIFILTLTSFYPRKVWVSEYSSIKDLEPVDPLFNKFIYYNTKIINKMGEHLKIIEYP
LILLFIVRPGAIFLMSSNDLITIFLSIELQSYGLYILSTVYRNSELSTTGGLIYFLLGGLSSCFILLGTALLYAKSGRTS
LESLYIITSISDIHSSTNDLWYSPNYISLSLVIFTVGFLFKISAAPFHFWSPDVYDAIPTIVTTFVALIAKISILILLLQ
LVYYTNADITMNNWTFILLISSLFTLVIGTVVGLTQFRIKRLFAYSTISQIGFMLLALSISSIESTQAPIFYLTQYIIIN
LNAFIILLAITYSLYFYTNYNKEHKDLLDKTNSPIQLITQLKGYFFINPVLALNLTITIFSFSQIPPLLVCFCKQIVLSA
SLNQNLYFITLIAILTNFIERIYYLTIIKKIFFNKSDYKINTLLCNFKLKINIYNNNNSINSVEYNYKNIFLSIPISFII
SILTLTILFFLFINKKWLIMITIFVQLIFNS',
 'gi|27544885|ref|NP_775398.1| NADH dehydrogenase subunit 3 (mitochondrion)
[Lecanicillium muscarium]': 'MSGTTFLFIFVCVIAILFLALNFILAPHNPYQEKYSIFECGFHSFLGQNRT
QFGFKFFIFSLVYLLLDLDLEILVIYPYGLSSYENGIYGLIIVLLFIGIITAGFVFELGKGALKIDSRQSYNYFNVQSTK
NFINTFFENK'}
```

[38]:
```
fastaSeqs = """>CCDS2.2|Hs109|chr1
MSKGILQVHPPICDCPGCRISSPVNRGRLADKRTVALPAARNLKKERTPSFSASDGDSDG
SGPTCGRRPGLKQEDGPHIRIMKRRVHTHWDVNISFREASCSQDGNLPTLISSVHRSRHL
VMPEHQSRCEFQRGSLEIGLRPAGDLLGKRLGRSPRISSDCFSEKRARSESPQEALLLPR
ELGPSMAPEDHYRRLVSALSEASTFEDPQRLYHLGLPSHGEDPPWHDPPHHLPSHDLLRV
RQEVAAAALRGPSGLEAHLPSSTAGQRRKQGLAQHREGAAPAAAPSFSERELPQPPPLLS
PQNAPHVALGPHLRPPFLGVPSALCQTPGYGFLPPAQAEMFAWQQELLRKQNLARLELPA
DLLRQKELESARPQLLAPETALRPNDGAEELQRRGALLVLNHGAAPLLALPPQGPPGSGP
PTPSRDSARRAPRKGGPGPASARPSESKEMTGARLWAQDGSEDEPPKDSDGEDPETAAVG
CRGPTPGQAPAGGAGAEGKGLFPGSTLPLGFPYAVSPYFHTGAVGGLSMDGEEAPAPEDV
TKWTVDDVCSFVGGLSGCGEYTRVFREQGIDGETLPLLTEEHLLTNMGLKLGPALKIRAQ
VARRLGRVFYVASFPVALPLQPPTLRAPERELGTGEQPLSPTTATSPYGGGHALAGQTSP
KQENGTLALLPGAPDPSQPLC
>CCDS3.1|Hs109|chr1
MAAAGSRKRRLAELTVDEFLASGFDSESESESENSPQAETREAREAARSPDKPGGSPSAS
RRKGRASEHKDQLSRLKDRDPEFYKFLQENDQSLLNFSDSDSSEEEEGPFHSLPDVLEEA
SEEEDGAEEGEDGDRVPRGLKGKKNSVPVTVAMVERWKQAAKQRLTPKLFHEVVQAFRAA
VATTRGDQESAEANKFQVTDSAAFNALVTFCIRDLIGCLQKLLFGKVAKDSSRMLQPSSS
PLWGKLRVDIKAYLGSAIQLVSCLSETTVLAAVLRHISVLVPCFLTFPKQCRMLLKRMVI
VWSTGEESLRVLAFLVLSRVCRHKKDTFLGPVLKQMYITYVRNCKFTSPGALPFISFMQW
TLTELLALEPGVAYQHAFLYIRQLAIHLRNAMTTRKKETYQSVYNWQYVHCLFLWCRVLS
TAGPSEALQPLVYPLAQVIIGCIKLIPTARFYPLRMHCIRALTLLSGSSGAFIPVLPFIL
EMFQQVDFNRKPGRMSSKPINFSVILKLSNVNLQEKAYRDGLVEQLYDLTLEYLHSQAHC
IGFPELVLPVVLQLKSFLRECKVANYCRQVQQLLGKVQENSAYICSRRQRVSFGVSEQQA
VEAWEKLTREEGTPLTLYYSHWRKLRDREIQLEISGKERLEDLNFPEIKRRKMADRKDED
RKQFKDLFDLNSSEEDDTEGFSERGILRPLSTRHGVEDDEEDEEEGEEDSSNSEDGDPDA
EAGLAPGELQQLAQGPEDELEDLQLSEDD
>CCDS4.1|Hs109|chr1
```

```
MGNSHCVPQAPRRLRASFSRKPSLKGNREDSARMSAGLPGPEAARSGDAAANKLFHYIPG
TDILDLENQRENLEQPFLSVFKKGRRRVPVRNLGKVVHYAKVQLRFQHSQDVSDCYLELF
PAHLYFQAHGSEGLTFQGLLPLTELSVCPLEGSREHAFQITGPLPAPLLVLCPSRAELDR
WLYHLEKQTALLGGPRRCHSAPPQRRLTRLRTASGHEPGGSAVCASRVKLQHLPAQEQWD
RLLVLYPTSLAIFSEELDGLCFKGELPLRAVHINLEEKEKQIRSFLIEGPLINTIRVVCA
SYEDYGHWLLCLRAVTHREGAPPLPGAESFPGSQVMGSGRGSLSSGGQTSWDSGCLAPPS
TRTSHSLPESSVPSTVGCSSQHTPDQANSDRASIGRRRTELRRSGSSRSPGSKARAEGRG
PVTPLHLDLTQLHRLSLESSPDAPDHTSETSHSPLYADPYTPPATSHRRVTDVRGLEEFL
SAMQSARGPTPSSPLPSVPVSVPASDPRSCSSGPAGPYLLSKKGALQSRAAQRHRGSAKD
GGPQPPDAPQLVSSAREGSPEPWLPLTDGRSPRRSRDPGYDHLWDETLSSSHQKCPQLGG
PEASGGLVQWI
>CCDS5.1|Hs109|chr1
MAADTPGKPSASPMAGAPASASRTPDKPRSAAEHRKSSKPVMEKRRRARINESLAQLKTL
ILDALRKESSRHSKLEKADILEMTVRHLRSLRRVQVTAALSADPAVLGKYRAGFHECLAE
VNRFLAGCEGVPADVRSRLLGHLAACLRQLGPSRRPASLSPAAPAEAPAPEVYAGRPLLP
SLGGPFPLLAPPLLPGLTRALPAAPRAGPQGPGGPWRPWLR
>CCDS6.1|Hs109|chr1
MGWDLTVKMLAGNEFQVSLSSSMSVSELKAQITQKIGVHAFQQRLAVHPSGVALQDRVPL
ASQGLGPGSTVLLVVDKCDEPLSILVRNNKGRSSTYEVRLTQTVAHLKQQVSGLEGVQDD
LFWLTFEGKPLEDQLPLGEYGLKPLSTVFMNLRLRGGGTEPGGRS
>CCDS7.2|Hs109|chr1
MALRHLALLAGLLVGVASKSMENTAQLPECCVDVVGVNASCPGASLCGPGCYRRWNADGS
ASCVRCGNGTLPAYNGSECRSFAGPGAPFPMNRSSGTPGRPHPGAPRVAASLFLGTFFIS
SGLILSVAGFFYLKRSSKLPRACYRRNKAPALQPGEAAAMIPPPQSSVRKPRYVRRERPL
DRATDPAAFPGEARISNV
>CCDS8.1|Hs109|chr1
MGSSQEEGLRCQPSQPDHDADGHCGPDLEGAERASATPGPPGLLNSHRPADSDDTNAAGP
SAALLEGLLLGGGKPSPHSTRPGPFFYIGGSNGATIISSYCKSKGWQRIHDSRRDDYTLK
WCEVKSRDSYGSFREGEQLLYQLPNNKLLTTKIGLLSTLRGRARAMSKASKVPGGVQARL
EKDAAAPALEDLPWTSPGYLRPQRVLRMEEFFPETYRLDLKHEREAFFTLFDETQIWICK
PTASNQGKGIFLLRNQEEVAALQAKTRSMEDDPIHHKTPFRGPQARVVQRYIQNPLLVDG
RKFDVRSYLLIACTTPYMIFFGHGYARLTLSLYDPHSSDLGGHLTNQFMQKKSPLYMLLK
EHTVWSMEHLNRYISDTFWKARGLAKDWVFTTLKVRPLCPPVWE"""
```

[39]:
```
protein_dict(fastaSeqs)
```

[39]:
```
{'CCDS2.2|Hs109|chr1': 'MSKGILQVHPPICDCPGCRISSPVNRGRLADKRTVALPAARNLKKERTPSFSASDG
DSDGSGPTCGRRPGLKQEDGPHIRIMKRRVHTHWDVNISFREASCSQDGNLPTLISSVHRSRHLVMPEHQSRCEFQRGSL
EIGLRPAGDLLGKRLGRSPRISSDCFSEKRARSESPQEALLLPRELGPSMAPEDHYRRLVSALSEASTFEDPQRLYHLGL
PSHGEDPPWHDPPHHLPSHDLLRVRQEVAAAALRGPSGLEAHLPSSTAGQRRKQGLAQHREGAAPAAAPSFSERELPQPP
PLLSPQNAPHVALGPHLRPPFLGVPSALCQTPGYGFLPPAQAEMFAWQQELLRKQNLARLELPADLLRQKELESARPQLL
APETALRPNDGAEELQRRGALLVLNHGAAPLLALPPQGPPGSGPPTPSRDSARRAPRKGGPGPASARPSESKEMTGARLW
AQDGSEDEPPKDSDGEDPETAAVGCRGPTPGQAPAGGAGAEGKGLFPGSTLPLGFPYAVSPYFHTGAVGGLSMDGEEAPA
PEDVTKWTVDDVCSFVGGLSGCGEYTRVFREQGIDGETLPLLTEEHLLTNMGLKLGPALKIRAQVARRLGRVFYVASFPV
ALPLQPPTLRAPERELGTGEQPLSPTTATSPYGGGHALAGQTSPKQENGTLALLPGAPDPSQPLC',
 'CCDS3.1|Hs109|chr1': 'MAAAGSRKRRLAELTVDEFLASGFDSESESESENSPQAETREAREAARSPDKPGGS
PSASRRKGRASEHKDQLSRLKDRDPEFYKFLQENDQSLLNFSDSDSSEEEEGPFHSLPDVLEEASEEEDGAEEGEDGDRV
```

```
PRGLKGKKNSVPVTVAMVERWKQAAKQRLTPKLFHEVVQAFRAAVATTRGDQESAEANKFQVTDSAAFNALVTFCIRDLI
GCLQKLLFGKVAKDSSRMLQPSSSPLWGKLRVDIKAYLGSAIQLVSCLSETTVLAAVLRHISVLVPCFLTFPKQCRMLLK
RMVIVWSTGEESLRVLAFLVLSRVCRHKKDTFLGPVLKQMYITYVRNCKFTSPGALPFISFMQWTLTELLALEPGVAYQH
AFLYIRQLAIHLRNAMTTRKKETYQSVYNWQYVHCLFLWCRVLSTAGPSEALQPLVYPLAQVIIGCIKLIPTARFYPLRM
HCIRALTLLSGSSGAFIPVLPFILEMFQQVDFNRKPGRMSSKPINFSVILKLSNVNLQEKAYRDGLVEQLYDLTLEYLHS
QAHCIGFPELVLPVVLQLKSFLRECKVANYCRQVQQLLGKVQENSAYICSRRQRVSFGVSEQQAVEAWEKLTREEGTPLT
LYYSHWRKLRDREIQLEISGKERLEDLNFPEIKRRKMADRKDEDRKQFKDLFDLNSSEEDDTEGFSERGILRPLSTRHGV
EDDEEDEEEGEEDSSNSEDGDPDAEAGLAPGELQQLAQGPEDELEDLQLSEDD',
 'CCDS4.1|Hs109|chr1': 'MGNSHCVPQAPRRLRASFSRKPSLKGNREDSARMSAGLPGPEAARSGDAAANKLFH
YIPGTDILDLENQRENLEQPFLSVFKKGRRRVPVRNLGKVVHYAKVQLRFQHSQDVSDCYLELFPAHLYFQAHGSEGLTF
QGLLPLTELSVCPLEGSREHAFQITGPLPAPLLVLCPSRAELDRWLYHLEKQTALLGGPRRCHSAPPQRRLTRLRTASGH
EPGGSAVCASRVKLQHLPAQEQWDRLLVLYPTSLAIFSEELDGLCFKGELPLRAVHINLEEKEKQIRSFLIEGPLINTIR
VVCASYEDYGHWLLCLRAVTHREGAPPLPGAESFPGSQVMGSGRGSLSSGGQTSWDSGCLAPPSTRTSHSLPESSVPSTV
GCSSQHTPDQANSDRASIGRRRTELRRSGSSRSPGSKARAEGRGPVTPLHLDLTQLHRLSLESSPDAPDHTSETSHSPLY
ADPYTPPATSHRRVTDVRGLEEFLSAMQSARGPTPSSPLPSVPVSVPASDPRSCSSGPAGPYLLSKKGALQSRAAQRHRG
SAKDGGPQPPDAPQLVSSAREGSPEPWLPLTDGRSPRRSRDPGYDHLWDETLSSSHQKCPQLGGPEASGGLVQWI',
 'CCDS5.1|Hs109|chr1': 'MAADTPGKPSASPMAGAPASASRTPDKPRSAAEHRKSSKPVMEKRRRARINESLAQ
LKTLILDALRKESSRHSKLEKADILEMTVRHLRSLRRVQVTAALSADPAVLGKYRAGFHECLAEVNRFLAGCEGVPADVR
SRLLGHLAACLRQLGPSRRPASLSPAAPAEAPAPEVYAGRPLLPSLGGPFPLLAPPLLPGLTRALPAAPRAGPQGPGGPW
RPWLR',
 'CCDS6.1|Hs109|chr1': 'MGWDLTVKMLAGNEFQVSLSSSMSVSELKAQITQKIGVHAFQQRLAVHPSGVALQD
RVPLASQGLGPGSTVLLVVDKCDEPLSILVRNNKGRSSTYEVRLTQTVAHLKQQVSGLEGVQDDLFWLTFEGKPLEDQLP
LGEYGLKPLSTVFMNLRLRGGGTEPGGRS',
 'CCDS7.2|Hs109|chr1': 'MALRHLALLAGLLVGVASKSMENTAQLPECCVDVVGVNASCPGASLCGPGCYRRWN
ADGSASCVRCGNGTLPAYNGSECRSFAGPGAPFPMNRSSGTPGRPHPGAPRVAASLFLGTFFISSGLILSVAGFFYLKRS
SKLPRACYRRNKAPALQPGEAAAMIPPPQSSVRKPRYVRRERPLDRATDPAAFPGEARISNV',
 'CCDS8.1|Hs109|chr1': 'MGSSQEEGLRCQPSQPDHDADGHCGPDLEGAERASATPGPPGLLNSHRPADSDDTN
AAGPSAALLEGLLLGGGKPSPHSTRPGPFFYIGGSNGATIISSYCKSKGWQRIHDSRRDDYTLKWCEVKSRDSYGSFREG
EQLLYQLPNNKLLTTKIGLLSTLRGRARAMSKASKVPGGVQARLEKDAAAPALEDLPWTSPGYLRPQRVLRMEEFFPETY
RLDLKHEREAFFTLFDETQIWICKPTASNQGKGIFLLRNQEEVAALQAKTRSMEDDPIHHKTPFRGPQARVVQRYIQNPL
LVDGRKFDVRSYLLIACTTPYMIFFGHGYARLTLSLYDPHSSDLGGHLTNQFMQKKSPLYMLLKEHTVWSMEHLNRYISD
TFWKARGLAKDWVFTTLKVRPLCPPVWE'}
```

[ ]: