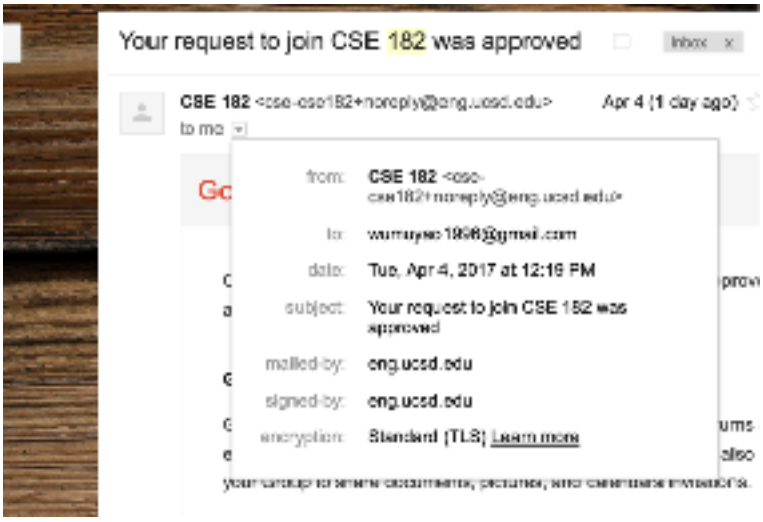
Muyao Wu
A11735974
CSE 182
April 5, 2017

Hw1

Terminal outputs and source code in the end.



1. Done!
2. Python file running on mac. Filename is HelloBio.py
3. done
4. done
5. done
6. done 18202068 is the result.

7. I spent around 5 hours on this assignment.

I'm pretty new to Python so had to look up some basic stuff:

http://stackoverflow.com/questions/4967580/how-to-get-the-size-of-a-string-in-python

http://stackoverflow.com/questions/675442/comment-out-a-python-code-block

http://stackoverflow.com/questions/4495176/nth-word-in-a-text

http://stackoverflow.com/questions/6181763/converting-a-string-to-a-list-of-words

and other basic python stuff

Learned how to download python3 online after realizing my laptop runs off python 2.6/7

I also asked Dominik Stec from our class a basic Python question: what the common did after the print statement.

Output dump with terminal commands Commands in bold:

Muyaos-MacBook-Pro:HW1 Muyao$ python3 hellobio.py

Hello Bioinformatics

Muyaos-MacBook-Pro:HW1 Muyao$ python3 cat.py

>gi|6978799|ref|NP_036683.1| early growth response 1; nerve growth factor-induced gene [Rattus norvegicus]

508

>gi|45768856|gb|AAH67618.1| Serum/glucocorticoid regulated kinase [Danio rerio]

433

>gi|45768786|gb|AAH68134.1| Unknown (protein for MGC:95907) [Mus musculus]

423

>gi|27923854|sp|P59241|STK6_RAT Serine/threonine kinase 6 (Aurora-A) (ratAurA)

397

>gi|45768720|gb|AAH67812.1| Cyclin L1 [Homo sapiens]

526

>gi|45768758|gb|AAH68160.1| Cdk7 protein [Mus musculus]

346

>gi|45219906|gb|AAH66834.1| Mastl protein [Mus musculus]

671

>gi|18202599|sp|Q63796|M3KC_RAT Mitogen-activated protein kinase kinase kinase 12 (MAPK-upstream kinase) (MUK)

888

>gi|4835224|emb|CAB42902.1| protein kinase ATN1 like protein [Arabidopsis thaliana]

370

>gi|40787731|gb|AAH64804.1| SLK protein [Homo sapiens]

617

>gi|18202068|sp|O55173|PDPK_RAT 3-phosphoinositide dependent protein kinase-1 (Protein kinase B kinase) (PkB kinase)

559

>gi|34191428|gb|AAH36504.2| C9orf96 protein [Homo sapiens]

700

>gi|29747774|gb|AAH50806.1| Gene model 711, (NCBI) [Mus musculus]

587

>gi|28856169|gb|AAH48033.1| Serine/threonine kinase 3 (STE20 homolog, yeast) [Danio rerio]

492

>gi|20071571|gb|AAH26466.1| Unknown (protein for IMAGE:4485517) [Mus musculus]

202

>gi|45709347|gb|AAH67695.1| Unknown (protein for MGC:85918) [Danio rerio]

Muyaos-MacBook-Pro:HW1 Muyao$ python3 filter.py

>gi|6978799|ref|NP_036683.1| early growth response 1; nerve

growth factor-induced gene [Rattus norvegicus]

MDNYPKLEEMMLLSNGAPQFLGAAGTPEGSGGNNSSSSSSSSSGGGGGGGSNSGSSAFNP

QGEPSEQPYEHLTTESFSDIALNNEKALVETSYPSQTTRLPPITYTGRFSLEPAPNSGNT

LWPEPLFSLVSGLVSMTNPPTSSSSAPSPAASSSSSASQSPPLSCAVPSNDSSPIYSAAP

TFPTPNTDIFPEPQSQAFPGSAGTALQYPPPAYPATKGGFQVPMIPDYLFPQQQGDLSLG

TPDQKPFQGLENRTQQPSLTPLSTIKAFATQSGSQDLKALNNTYQSQLIKPSRMRKYPNR

PSKTPPHERPYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRNFSRSDHLTTHI

RTHTGEKPFACDICGRKFARSDERKRHTKIHLRQKDKKADKSVVASSAASSLSSYPSPVA

TSYPSPATTSFPSPVPTSYSSPGSSTYPSPAHSGFPSPSVATTYASVPPAFPAQVSTFQS

AGVSNSFSTSTGLSDMTATFSPRTIEIC

>gi|45768786|gb|AAH68134.1| Unknown (protein for MGC:95907)

[Mus musculus]

MSTRNCQGTDSVIKHLDTIPEDKKVRVQRTQSTFDPFEKPANQVKRVHSENNACINFKSS

SAGKESPKVRRHSSPSSPTSPKFGKADSYEKLEKLGEGSYATVYKGKSKVNGKLVALKVI

RLQEEEGTPFTAIREASLLKGLKHANIVLLHDIIHTKETLTLVFEYVHTDLCQYMDKHPG

GLHPDNVKLFLFQLLRGLSYIHQRYILHRDLKPQNLLISDTGELKLADFGLARAKSVPSH

TYSNEVVTLWYRPPDVLLGSTEYSTCLDMWGVGCIFVEMIQGVAAFPGMKDIQDQLERIF

LVLGTPNEDTWPGVHSLPHFKPERFTVYSSKSLRQAWNKLSYVNHAEDLASKLLQCSPKN

RLSAQAALSHEYFSDLPPRLWELTDMSSIFTVPNVRLQPEAGESMRAFGKNNSYGKSLSN

SKH

>gi|27923854|sp|P59241|STK6_RAT Serine/threonine kinase 6 (A

urora-A) (ratAurA)

MDRCKENCVSRPVKSTVPFGPKRVLVTEQIPSQHPGSASSGQAQRVLCPSNSQRVPPQAQ

KPVAGQKPVLKQLPAASGPRPASRLSNPQKSEQPQPAASGNNSEKEQTSIQKTEDSKKRQ

WTLEDFDIGRPLGKGKFGNVYLAREKQSKFILALKVLFKVQLEKAGVEHQLRREVEIQSH

LRHPNILRLYGYFHDATRVYLILEYAPLGTVYRELQKLSKFDEQRTATYITELANALSYC

HSKRVIHRDIKPENLLLGSNGELKIADFGWSVHAPSSRRTTLCGTLDYQPPEMIEGRMHD

EKVDLWSLGVLCYEFLVGMPPFEAHTYQETYRRISRVEFTFPDFVTEGARDLISRLLKHN

SSQRLTLAEVLEHPWIKANSSKPPTGHNSKEATSKSS

>gi|45768758|gb|AAH68160.1| Cdk7 protein [Mus musculus]

MAVDVKSRAKRYEKLDFLGEGQFATVYKARDKNTNQIVAIKKIKLGHRSEAKDGINRTAL

REIKLLQELSHPNIIGLLDAFGHKSNISLVFDFMETDLEVIIKDNSLVLTPSHIKAYMLM

TLQGLEYLHQHWILHRDLKPNNLLLDENGVLKLADFGLAKSFGSPNRAYTHQVVTRWYRA

PELLFGARMYGVGVDMWAVGCILAELLLRVPFLPGDSDLDQLTRIFETLGTPTEEQWPDM

CSLPDYVTFKSFPGVPLQHIFIAAGDDLLELIQGLFLFNPCTRTTASQALKTKYFSNRPG

PTPGCQLPRPNCPVEALKEPANPTVATKRKRAEALEQGILPKKLIF

>gi|45219906|gb|AAH66834.1| Mastl protein [Mus musculus]

SMSKPKQDYSRTPGQVLSLISSLGFFTPVGEKDQDSANMFSAPKSAAQLSRGFICPMSVD

QKEPTSYSSKLLKSCFETLSSNPEIPVKCLTSNLLQCRKRLGTSSTSSQSHTFVSSVESE

CHSNPKWERDCQSTESSGCAMSWNAVEMLYAKSTSAIKTKTELELALSPIHDSSAIPAAG

SNQVTLPRKCFREISWEARDPDNENMTIDKGQSGFCQSSQRSVNSSATSEEHLGKRNYKR

NFHLVDSSPCQEIMQSKKNCTEYEANKERQGCRANQSTGLTTEVQNLKLSGCESQQLDYA

NKENIVTYLTDRQTPEKLHIPTIAKNLMSELDEDRELSSKKDCLSSNSVCSDEDRALKTT

CVDSDSSFPGVSMMESSLEIQALEPDKSIRDYSFEEPNTEDLFVLPKCQENSLPQDDCHA

CIQDSSQVSAHPSKAPKALTSKINVVAFRSFNSHINASTNSEPSKISITSLDAMDISYDY

SGSYPMAVSPTEKGRHYTSHQTPNQVKLGTSYRTPKSVRRGAAPVDDGRILGTPDYLAPE

LLLGTAHGPAVDWWALGVCLFEFLTGIPPFNDETPQQVFQNILKRDIPWPEGEEKLSDNA

QSAMDMLLTIDDSKRAGMRELKQHPLFSEVDWENLQHQTMPFVPQPDDETDTSYFEARNN

AQHLTISGFSL

>gi|18202599|sp|Q63796|M3KC_RAT Mitogen-activated protein ki

nase kinase kinase 12 (MAPK-upstream kinase) (MUK)

MACLHETRTPSPSFGGFVSTLSEASMRKLDPDTSDCTPEKDLTPTQCVLRDVVPLGGQGG

GGPSPSPGGEPPPEPFANSVLQLHEQDTGGPGGATGSPESRASRVRADEVRLQCQSGSGF

LEGLFGCLRPVWTMIGKAYSTEHKQQQEDLWEVPFEEILDLQWVGSGAQGAVFLGRFHGE

EVAVKKVRDLKETDIKHLRKLKHPNIITFKGVCTQAPCYCILMEFCAQGQLYEVLRAGRP

VTPSLLVDWSMGIAGGMNYLHLHKIIHRDLKSPNMLITYDDVVKISDFGTSKELSDKSTK

MSFAGTVAWMAPEVIRNEPVSEKVDIWSFGVVLWELLTGEIPYKDVDSSAIIWGVGSNSL

HLPVPSSCPDGFKILLRQCWNRKPRNRPSFRQILLHLDIASADVLSTPQETYFKSQAEWR

EEVKLHFEKIKSEGTCLHRLEEELVMRRREELRHALDIREHYERKLERANNLYMELNALM

LQLELKERELLRREQALERRCPGLLKSHTSRSLLHGNTMEKLIKKRNVPQKLSPHSKRPD

ILKTESLLPKLDAALSGVGLPGCPKAPPSPGRSRRGKTRHRKASAKGSCGDLPGLRAALP

PHEPGGLGSPGGLGVGPTAWDASPPALRGLHHDLLLRKMSSSSPDLLSAALGARGRGATG

GARDPGSPPPPQGDTPPSEGSAPGSTSPDSPGGAKGEPPPPVGPGEGVGLLGTGREGTTG

RGGSRAGYQHLTPAALLYRAAVTRSQKRGISSEEEEGEVDSEVELPPSQRWPQGPNMRQS

LSTFSSENPSDVEEGTASEPSPSGTPEVGSTNTDERPDERSDDMCSQGSEIPLDLPTSEV

VPERETSSLPMQHQDDQGPNPEDSDCDSTELDNSNSIDALPPPASLPP

>gi|18202068|sp|O55173|PDPK_RAT 3-phosphoinositide dependent

protein kinase-1 (Protein kinase B kinase) (PkB kinase)

MARTTSQLYDAVPIQSSVVLCSCPSPSMVRSQTEPSSSPGIPSGVSRQGSTMDGTTAEAR

PSTNPLQQHPAQLPPQPRKKRPEDFKFGKILGEGSFSTVVLARELATSREYAIKILEKRH

IIKENKVPYVTRERDVMSRLDHPFFVKLYFTFQDDEKLYFGLSYAKNGELLKYIRKIGSF

DETCTRFYTAEIVSALEYLHGKGIIHRDLKPENILLNEDMHIQITDFGTAKVLSPDSKQA

RANSFVGTAQYVSPELLTEKSACKSSDLWALGCIIYQLVAGLPPFRAGNEYLIFQKIIKL

EYDFPEKFFPKARDLVEKLLVLDATKRLGCEEMEGYGPLKAHPFFESITWENLHQQTPPK

LTAYLPAMSEDDEDCYGNYDNLLSQFGCMQVSSSSSSHSLCAVDASLPQRSGSNIEQYIH

DLDTNSFELDLQFSEDEKRLLLEKQAGGNPWHQFVENNLILKMGPVDKRKGLFARRRQLL

LTEGPHLYYVDPVNKVLKGEIPWSQELRPEAKNFKTFFVHTPNRTYYLMDPSGNAHKWCR

KIQEVWRQQYQSSPDAAVQ

>gi|29747774|gb|AAH50806.1| Gene model 711, (NCBI) [Mus musc

ulus]

MDYYSQGTFQNIMENKRKLKAVVDTEWMHTMLSQVLDAIEYLHKLNIVHRNLKPSNIVLV

NSGYCKLQDMSSQALMTHEAKWNVRAEEDPCQKSWMAPEALKFSFSTKSDIWSLGCIILD

MATCSFLNDTEAMQLRKAIRHHPGSLKPILKTMEEKQIPGTDVYYLLLPFMLHINPSDRL

AIKDVMQVTFMSNSFKSSSVALNMQRQKVPIFITDVLLEGNMANILDVMQNFSSRPEVQL

RAINKLLTMPEDQLGLPWPTELLEEVISIIKQHGRILDILLSTCSLLLRVLGQALAKDPE

AEIPRSSLIISFLMDTLRSHPNSERLVNVVYNVLAIISSQGQISEELEEEGLFQLAQENL

EHFQEDRDICLSILSLLWSLLVDVVTVDKEPLEQLSGMVTWVLATHPEDVEIAEAGCAVL

WLLSLLGCIKESQFEQVVVLLLRSIQLCPGRVLLVNNAFRGLASLAKVSELVAFRIVVLE

EGSSGLHLIQDIYKLYKDDPEVVENLCMLLAHLTSYKEILPEMESGGIKDLVQVIRGRFT

SSLELISYADEILQVLEANAQPGLQEDQLEPPAGQEAPLQGEPLFRP

>gi|20071571|gb|AAH26466.1| Unknown (protein for IMAGE:44855

17) [Mus musculus]

PTRPTRLIVSNFSQAKQKSHLVDPQILRDQSRLAPEIITATQYKKCDEFQTGILIYEMLH

LPNPFDENPELKEKEYTRTDLPRIPLRSPYSWGLQQLASCLLNPNPSERILISDAKGILQ

CLLWGPREDLFQIFTTSATLAQKNALLQNWLDIKRTLLMIKFAEKSLDREGGISLEDWLC

AQYLAFATTDSLSYIVKILQYR

Muyaos-MacBook-Pro:HW1 Muyao$ cat data.seq

MDNYPKLEEMMLLSNGAPQFLGAAGTPEGSGGNNSSSSSSSSSGGGGGGGSNSGSSAFNPQGEPSEQPYEHLTTESFSDIALN
NEKALVETSYPSQTTRLPPITYTGRFSLEPAPNSGNTLWPEPLFSLVSGLVSMTNPPTSSSSAPSPAASSSSSASQSPPLSCAVPSNDSSPIYS
AAPTFPTPNTDIFPEPQSQAFPGSAGTALQYPPPAYPATKGGFQVPMIPDYLFPQQQGDLSLGTPDQKPFQGLENRTQQPSLTPLSTIK
AFATQSGSQDLKALNNTYQSQLIKPSRMRKYPNRPSKTPPHERPYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRNFSRS
DHLTTHIRTHTGEKPFACDICGRKFARSDERKRHTKIHLRQKDKKADKSVVASSAASSLSSYPSPVATSYPSPATTSFPSPVPTSYSSPG
SSTYPSPAHSGFPSPSVATTYASVPPAFPAQVSTFQSAGVSNSFSTSTGLSDMTATFSPRTIEIC@MTIQTETSVSAPDLTYSKTRGLVANL
SAFMKQRKMGLNDFIQKLSANSYACKHPEVQSILNLTPPQDVELMNSNPSPPPSPSQQINLGPSSNPTAKPSDFDFLKVIGKGSFGKVL
LARHRSDEKFYAVKVLQKKAILKKKEEKHIMSERNVLLKNVKHPFLVGLHYSFQTTDKLYFVLDYINGGELFYHLQRERCFLEPRA
RFYAAEIASALGYLHSLNIVYRDLKPENILLDSQGHIILTDFGLCKENIEPNGTTSTFCGTPEYLAPEVLHKQPYDRTVDWWCLGAVLY
EMLYGLPPFYSRNTAEMYDNILNKPLQLKPNISNAARHLLEGLLQKDRTKRLGFTDDFTEIKNHMFFSPINWDDLNAKKLTPPFNPN
VTGPNDLRHFDPEFTDEPVPNSIGCSPDSALVTSSITEATEAFLGFSYAPAMDSYL@MSTRNCQGTDSVIKHLDTIPEDKKVRVQRTQ
STFDPFEKPANQVKRVHSENNACINFKSSSAGKESPKVRRHSSPSSPTSPKFGKADSYEKLEKLGEGSYATVYKGKSKVNGKLVALKV
IRLQEEEGTPFTAIREASLLKGLKHANIVLLHDIIHTKETLTLVFEYVHTDLCQYMDKHPGGLHPDNVKLFLFQLLRGLSYIHQRYIL
HRDLKPQNLLISDTGELKLADFGLARAKSVPSHTYSNEVVTLWYRPPDVLLGSTEYSTCLDMWGVGCIFVEMIQGVAAFPGMKDIQ
DQLERIFLVLGTPNEDTWPGVHSLPHFKPERFTVYSSKSLRQAWNKLSYVNHAEDLASKLLQCSPKNRLSAQAALSHEYFSDLPPRL
WELTDMSSIFTVPNVRLQPEAGESMRAFGKNNSYGKSLSNSKH@MDRCKENCVSRPVKSTVPFGPKRVLVTEQIPSQHPGSASSGQ
AQRVLCPSNSQRVPPQAQKPVAGQKPVLKQLPAASGPRPASRLSNPQKSEQPQPAASGNNSEKEQTSIQKTEDSKKRQWTLEDFDIG
RPLGKGKFGNVYLAREKQSKFILALKVLFKVQLEKAGVEHQLRREVEIQSHLRHPNILRLYGYFHDATRVYLILEYAPLGTVYRELQ
KLSKFDEQRTATYITELANALSYCHSKRVIHRDIKPENLLLGSNGELKIADFGWSVHAPSSRRTTLCGTLDYQPPEMIEGRMHDEKV
DLWSLGVLCYEFLVGMPPFEAHTYQETYRRISRVEFTFPDFVTEGARDLISRLLKHNSSQRLTLAEVLEHPWIKANSSKPPTGHNSKE
ATSKSS@MASGPHSTATAAAAASSAAPSAGGSSSGTTTTTTTTTTGGILIGDRLYSEVSLTIDHSLIPEERLSPTPSMQDGLDLPSETDLR
ILGCELIQAAGILLRLPQVAMATGQVLFHRFFYSKSFVKHSFEIVAMACINLASKIEEAPRRIRDLINVFHHLRQLRGKRTPSPLILDQNY
INTKNQVIKAERRVLKELGFCVHVKHPHKIIVMYLQVLECERNQTLVQTAWNYMNDSLRTNVFVRFQPETIACACIYLAARALQIPL
PTRPHWFLLFGTTEEEIQEICIETLRLYTRKKPNYELLEKEVEKRKVALQEAKLKAKGLNPDGTPALSTLGGFSPASKPSSPREVKAEE
KSPISINVKTVKKEPEDRQQASKSPYNGVRKDSKRSRNSRSASRSRSRTRSRSRSHTPRRHYNNRRSRSGTYSSRSRSRSRSHSESPR
RHHNHGSPHLKAKHTRDDLKSSNRHGHKRKKSRSRSQSKSRDHSDAAKKHRHERGHHRDRRERSRSFERSHKSKHHGGSRSG
HGRHRR@MAVDVKSRAKRYEKLDFLGEGQFATVYKARDKNTNQIVAIKKIKLGHRSEAKDGINRTALREIKLLQELSHPNIIGLLD
AFGHKSNISLVFDFMETDLEVIIKDNSLVLTPSHIKAYMLMTLQGLEYLHQHWILHRDLKPNNLLLDENGVLKLADFGLAKSFGSPN
RAYTHQVVTRWYRAPELLFGARMYGVGVDMWAVGCILAELLLRVPFLPGDSDLDQLTRIFETLGTPTEEQWPDMCSLPDYVTFKSF
PGVPLQHIFIAAGDDLLELIQGLFLFNPCTRTTASQALKTKYFSNRPGPTPGCQLPRPNCPVEALKEPANPTVATKRKRAEALEQGILP
KKLIF@SMSKPKQDYSRTPGQVLSLISSLGFFTPVGEKDQDSANMFSAPKSAAQLSRGFICPMSVDQKEPTSYSSKLLKSCFETLSSNP
EIPVKCLTSNLLQCRKRLGTSSTSSQSHTFVSSVESECHSNPKWERDCQSTESSGCAMSWNAVEMLYAKSTSAIKTKTELELALSPIH
DSSAIPAAGSNQVTLPRKCFREISWEARDPDNENMTIDKGQSGFCQSSQRSVNSSATSEEHLGKRNYKRNFHLVDSSPCQEIMQSKK
NCTEYEANKERQGCRANQSTGLTTEVQNLKLSGCESQQLDYANKENIVTYLTDRQTPEKLHIPTIAKNLMSELDEDRELSSKKDCL
SSNSVCSDEDRALKTTCVDSDSSFPGVSMMESSLEIQALEPDKSIRDYSFEEPNTEDLFVLPKCQENSLPQDDCHACIQDSSQVSAHPS

KAPKALTSKINVVAFRSFNSHINASTNSEPSKISITSLDAMDISYDYSGSYPMAVSPTEKGRHYTSHQTPNQVKLGTSYRTPKSVRRGA
APVDDGRILGTPDYLAPELLLGTAHGPAVDWWALGVCLFEFLTGIPPFNDETPQQVFQNILKRDIPWPEGEEKLSDNAQSAMDMLLTI
DDSKRAGMRELKQHPLFSEVDWENLQHQTMPFVPQPDDETDTSYFEARNNAQHLTISGFSL@MACLHETRTPSPSFGGFVSTLSEA
SMRKLDPDTSDCTPEKDLTPTQCVLRDVVPLGGQGGGGPSPSPGGEPPPEPFANSVLQLHEQDTGGPGGATGSPESRASRVRADEV
RLQCQSGSGFLEGLFGCLRPVWTMIGKAYSTEHKQQQEDLWEVPFEEILDLQWVGSGAQGAVFLGRFHGEEVAVKKVRDLKETDI
KHLRKLKHPNIITFKGVCTQAPCYCILMEFCAQGQLYEVLRAGRPVTPSLLVDWSMGIAGGMNYLHLHKIIHRDLKSPNMLITYDD
VVKISDFGTSKELSDKSTKMSFAGTVAWMAPEVIRNEPVSEKVDIWSFGVVLWELLTGEIPYKDVDSSAIIWGVGSNSLHLPVPSSCPD
GFKILLRQCWNRKPRNRPSFRQILLHLDIASADVLSTPQETYFKSQAEWREEVKLHFEKIKSEGTCLHRLEEELVMRRREELRHALD
IREHYERKLERANNLYMELNALMLQLELKERELLRREQALERRCPGLLKSHTSRSLLHGNTMEKLIKKRNVPQKLSPHSKRPDILK
TESLLPKLDAALSGVGLPGCPKAPPSPGRSRRGKTRHRKASAKGSCGDLPGLRAALPPHEPGGLGSPGGLGVGPTAWDASPPALRGL
HHDLLLRKMSSSSPDLLSAALGARGRGATGGARDPGSPPPPQGDTPPSEGSAPGSTSPDSPGGAKGEPPPPVGPGEGVGLLGTGREGT
TGRGGSRAGYQHLTPAALLYRAAVTRSQKRGISSEEEEGEVDSEVELPPSQRWPQGPNMRQSLSTFSSENPSDVEEGTASEPSPSGTPE
VGSTNTDERPDERSDDMCSQGSEIPLDLPTSEVVPERETSSLPMQHQDDQGPNPEDSDCDSTELDNSNSIDALPPPASLPP@MISRMI
FRNYPSHNESDDEPFHFSISRELLLDRNDVVVGEMIGEGAYSIVYKGLLRNQFPVAVKIMDPSTTSAVTKAHKKTFQKEVLLLSKMK
HDNIVKFVGACIEPQLIIVTELVEGGTLQRFMHSRPGPLDLKMSLSFALDISRAMEFVHSNGIIHRDLNPRNLLVTGDLKHVKLADFG
IAREETRGGMTCEAGTSKWMAPEVYSPEPLRVGEKKEYDHKADIYSFAIVLWQLVTNEEPFPDVPNSLFVPYLVSQGRRPILTKTPDV
FVPIVESCWAQDPDARPEFKEISVMLTNLLRRMSSDSSIGTTLPDGEAYEGEMEESENSPLLQEHFCKVKKPKEKKKKKKLVKMRFP
FFKKFKVWLYNYKP@MSFFNFRKIFKLGSEKKKKQYEHVKRDLNPEDFWEIIGELGDGAFGKVYKAQNKETSVLAAAKVIDTKSE
EELEDYMVEIDILASCDHPNIVKLLDAFYYENNLWILIEFCAGGAVDAVMLELERPLTESQIQVVCKQTLDALNYLHDNKIIHRDLKA
GNILFTLDGDIKLADFGVSAKNTRTIQRRDSFIGTPYWMAPEVVMCETSKDRPYDYKADVWSLGITLIEMAEIEPPHHELNPMRVLL
KIAKSEPPTLAQPSRWSSNFKDFLKKCLEKNVDARWTTSQLLQHPFVTVDSNKPIRELIAEAKAEVTEEVEDGKEEDEEEETENSLPI
PASKRASSDLSIASSEEDKLSQNACILESVSEKTERSNSEDKLNSKILNEKPTTDEPEKAVEDINEHITDAQLEAMTELHDRTAVIKENE
REKRPKLENLPDTEDQETVDINSVSEGKENNIMITLETNIEHNLKSEEEKDQEKQQMFENKLIKSEEIKDTILQTVDLVSQETGEKE
ANIQAVDSEVGLTKEDTQEKLGEDDKTQKDVISNTSDVIGTCEAADVAQKVDEDSAEDTQSNDGKEVVEVGQKLINKPMVGPEAG
GTKEVPIKEIVEMNEIEEKKKK@MARTTSQLYDAVPIQSSVVLCSCPSPSMVRSQTEPSSSPGIPSGVSRQGSTMDGTTAEARPSTNP
LQQHPAQLPPQPRKKRPEDFKFGKILGEGSFSTVVLARELATSREYAIKILEKRHIIKENKVPYVTRERDVMSRLDHPFFVKLYFTFQ
DDEKLYFGLSYAKNGELLKYIRKIGSFDETCTRFYTAEIVSALEYLHGKGIIHRDLKPENILLNEDMHIQITDFGTAKVLSPDSKQARA
NSFVGTAQYVSPELLTEKSACKSSDLWALGCIIYQLVAGLPPFRAGNEYLIFQKIIKLEYDFPEKFFPKARDLVEKLLVLDATKRLGCEE
MEGYGPLKAHPFFESITWENLHQQTPPKLTAYLPAMSEDDEDCYGNYDNLLSQFGCMQVSSSSSSHSLCAVDASLPQRSGSNIEQYIH
DLDTNSFELDLQFSEDEKRLLLEKQAGGNPWHQFVENNLILKMGPVDKRKGLFARRRQLLLTEGPHLYYVDPVNKVLKGEIPWSQ
ELRPEAKNFKTFFVHTPNRTYYLMDPSGNAHKWCRKIQEVWRQQYQSSPDAAVQ@LTHAGWGQGWTLARTRSLLIMLGPGSNRR
RPTQGERGPGSPGEPMEKYQVLYQLNPGALGVNLVVEEMETKVKHVIKQVECMDDHYASQALEELMPLLKLRHAHISVYQELFIT
WNGEISSLYLCLVMEFNELSFQEVIEDKRKAKKIIDSEWMQNVLGQVLDALEYLHHLDIIHRNLKPSNIILISSDHCKLQDLSSNVLM
TDKAKWNIRAEEDPFRKSWMAPEALNFSFSQKSDIWSLGCIILDMTSCSFMDGTEAMHLRKSLRQSPGSLKAVLKTMEEKQIPDVE
TFRNLLPLMLQIDPSDRITIKDVVHITFLRGSFKSSCVSLTLHRQMVPASITDMLLEGNVASILEVMQKFSGWPEVQLRAMKRLLKMP
ADQLGLPWPPELVEVVVTTMELHDRVLDVQLCACSLLLHLLGQALVHHPEAKAPCNQAITSTLLSALQSHPEEEPLLVMVYSLLAIT
TTQESESLSEELQNAGLLEHILEHLNSSLESRDVCASGLGLLWALLLDGIIVNKAPLEKVPDLISQVLATYPADGEMAEASCGVFWLLS
LLGCIKEQQFEQVVALLLQSIRLCQDRALLVNNAYRGLASLVKVSELAAFKVVVQEEGGSGLSLIKETYQLHRDDPEVVENVGMLLV
HLASYEEILPELVSSSMKALLQEIKERFTSSLVSDSSAFSKPGLPPGGSPQLGCTTSGGLE@MDYYSQGTFQNIMENKRKLKAVVDTE
WMHTMLSQVLDAIEYLHKLNIVHRNLKPSNIVLVNSGYCKLQDMSSQALMTHEAKWNVRAEEDPCQKSWMAPEALKFSFSTKSDI
WSLGCIILDMATCSFLNDTEAMQLRKAIRHHPGSLKPILKTMEEKQIPGTDVYYLLLPFMLHINPSDRLAIKDVMQVTFMSNSFKSSS
VALNMQRQKVPIFITDVLLEGNMANILDVMQNFSSRPEVQLRAINKLLTMPEDQLGLPWPTELLEEVISIIKQHGRILDILLSTCSLLL
RVLGQALAKDPEAEIPRSSLIISFLMDTLRSHPNSERLVNVVYNVLAIISSQGQISEELEEEGLFQLAQENLEHFQEDRDICLSILSLLWS
LLVDVVTVDKEPLEQLSGMVTWVLATHPEDVEIAEAGCAVLWLLSLLGCIKESQFEQVVVLLLRSIQLCPGRVLLVNNAFRGLASLAK
VSELVAFRIVVLEEGSSGLHLIQDIYKLYKDDPEVVENLCMLLAHLTSYKEILPEMESGGIKDLVQVIRGRFTSSLELISYADEILQVLEA
NAQPGLQEDQLEPPAGQEAPLQGEPLFRP@MEHSVPKNKLKKLSEDSLTKQPEEVFDVLEKLGEGSYGSVFKAIHKESGQVVAIKQ
VPVESDLQEIIKEISIMQQCDSPYVVKYYGSYFKNTDLWIVMEYCGAGSVSDIIRLRNKTLTEDEIATVLKSTLKGLEYLHFMRKIHR
DIKAGNILLNTEGHAKLADFGVAGQLTDTMAKRNTVIGTPFWMAPEVIQEIGYNCVADIWSLGITSIEMAEGKPPYADIHPMRAIFMI

PTNPPPTFRKPEHWSDDFTDFVKKCLVKNPEQRATATQLLQHPFIVGAKPVSILRDLITEAMDMKAKRQQEQQRELEEDDENSEEE
VEVDSHTMVKSGSESAGTMRATGTMSDGAQTMIEHGSTMLESNLGTMVINSDDEEEEEDLGSMRRNPTSQQIQRPSFMDYFDKQ
DSNKAQEGFNHNQQDPCLISKTAFPDNWKVPQDGDFDFLKNLDFEELQMRLTALDPMMEREIEELRQRYTAKRQPILDAMDAKKR
RQQNF@PTRPTRLIVSNFSQAKQKSHLVDPQILRDQSRLAPEIITATQYKKCDEFQTGILIYEMLHLPNPFDENPELKEKEYTRTDLP
RIPLRSPYSWGLQQLASCLLNPNPSERILISDAKGILQCLLWGPREDLFQIFTTSATLAQKNALLQNWLDIKRTLLMIKFAEKSLDREG
GISLEDWLCAQYLAFATTDSLSYIVKILQYR@MQNKENREPRVQQTPSAGVGPLRVEMNPDTHAVSGPGRVPVKSNSKVLSIDDFDI
GRPLGKGKFGNVYLARERKLKVVIALKVLFKSQMVKEGVEHQLRREIEIQSHLRHPNILRFYNYFHDDTRVFLILEYAPRGEMYKEL
QRYGRFDDQRTATYMEEVSDALQYCHEKKVIHRDIKPENLLLGYRGELKIADFGWSVHAPSLRRRTMCGTLDYLPPEMIEGHSHDE
KVDLWSIGVLCYECLVGNPPFETASHAETYKRITKVDLQFPKLVSEGARDLISKLLRHSPSMRLPLRSVMEHPWVKANSRRVLPPVC
SSEPH

Muyaos-MacBook-Pro:HW1 Muyao$ cat data.in

6978799 0

45768856 509

45768786 943

27923854 1367

45768720 1765

45768758 2292

45219906 2639

18202599 3311

4835224 4200

40787731 4571

18202068 5189

34191428 5749

29747774 6450

28856169 7038

20071571 7531

45709347 7734

Muyaos-MacBook-Pro:HW1 Muyao$ python3 getSeq.py

Enter Sequence Here: MHIQITDFGTAKVLSPDS

18202068

Python source files names in bold:

#hellobio.py

```
print("Hello Bioinformatics")
```

#cat.py

```
data = open("datafile.txt","r")

counter = 0
for line in data:
    if line[0] == '>':
            if counter != 0:
                    print (counter)
            counter=0
            print (line),###
    else:
            counter+= len(line)-1

            #len(line)-1,
```

#filter.py

```
data = open("datafile.txt","r")


bool = 0
cs = ""
for line in data:
    if line[0] == '>':
            bool = 0
            if "mus" in line.lower() or "rat" in line.lower():
                    bool=1

                    cs += line

    elif bool == 1:
            cs += line

result = ""
counter = 0
```

```
pcount = 0
for char in cs:
    if char == '>':
                result += '\n' '\n'
                counter =0
                pcount = 0

    if char != '\n':
                result+=char
                counter+=1
    if char == ']':
                result += '\n' '\n'
                counter = 0

    if pcount == 0 and char == ')':
                pcount = 1
    elif pcount == 1 and char == ')':
                result += '\n' '\n'
                counter = 0
    if counter == 60:
                result += '\n' '\n'
                counter =0


print (result)
```

I wrote two generation programs to print out what I want for data.seq and data.in they're titled generation.py and generation2.py I then copy and pasted the outputs into their respective files.

#generation.py

```
data = open("datafile.txt","r")


bool = 0
cs = ""
for line in data:
    if line[0] == '>':
                if bool ==1 :
```

```
                cs += "@"
            bool =1

    else:
            cs += line

print ("".join(cs.split()))
```

#generation2.py

```
data = open("datafile.txt","r")


counter = 0
for line in data:
    if line[0] == '>':

            barCount = 0
            barSpot = 0
            k=0
            while k < len(line):
                if line[k]=="|" :
                    if barCount ==1:
                        barSpot = k
                    barCount+=1
                k+=1


            print (line[4:barSpot], counter)###

            counter +=1
    else:
            counter+= len(line)-1

            #len(line)-1,
```

The two python files generated the following data files:

data.seq

MDNYPKLEEMMLLSNGAPQFLGAAGTPEGSGGNNSSSSSSSSSGGGGGGGSNSGSSAFNPQGEPSEQPYEHLTTESFSDIALN
NEKALVETSYPSQTTRLPPITYTGRFSLEPAPNSGNTLWPEPLFSLVSGLVSMTNPPTSSSSAPSPAASSSSSASQSPPLSCAVPSNDSSPIYS
AAPTFPTPNTDIFPEPQSQAFPGSAGTALQYPPPAYPATKGGFQVPMIPDYLFPQQQGDLSLGTPDQKPFQGLENRTQQPSLTPLSTIK
AFATQSGSQDLKALNNTYQSQLIKPSRMRKYPNRPSKTPPHERPYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRNFSRS

DHLTTHIRTHTGEKPFACDICGRKFARSDERKRHTKIHLRQKDKKADKSVVASSAASSLSSYPSPVATSYPSPATTSFPSPVPTSYSSPG
SSTYPSPAHSGFPSPSVATTYASVPPAFPAQVSTFQSAGVSNSFSTSTGLSDMTATFSPRTIEIC@MTIQTETSVSAPDLTYSKTRGLVANL
SAFMKQRKMGLNDFIQKLSANSYACKHPEVQSILNLTPPQDVELMNSNPSPPPSPSQQINLGPSSNPTAKPSDFDFLKVIGKGSFGKVL
LARHRSDEKFYAVKVLQKKAILKKKEEKHIMSERNVLLKNVKHPFLVGLHYSFQTTDKLYFVLDYINGGELFYHLQRERCFLEPRA
RFYAAEIASALGYLHSLNIVYRDLKPENILLDSQGHIILTDFGLCKENIEPNGTTSTFCGTPEYLAPEVLHKQPYDRTVDWWCLGAVLY
EMLYGLPPFYSRNTAEMYDNILNKPLQLKPNISNAARHLLEGLLQKDRTKRLGFTDDFTEIKNHMFFSPINWDDLNAKKLTPPFNPN
VTGPNDLRHFDPEFTDEPVPNSIGCSPDSALVTSSITEATEAFLGFSYAPAMDSYL@MSTRNCQGTDSVIKHLDTIPEDKKVRVQRTQ
STFDPFEKPANQVKRVHSENNACINFKSSSAGKESPKVRRHSSPSSPTSPKFGKADSYEKLEKLGEGSYATVYKGKSKVNGKLVALKV
IRLQEEEGTPFTAIREASLLKGLKHANIVLLHDIIHTKETLTLVFEYVHTDLCQYMDKHPGGLHPDNVKLFLFQLLRGLSYIHQRYIL
HRDLKPQNLLISDTGELKLADFGLARAKSVPSHTYSNEVVTLWYRPPDVLLGSTEYSTCLDMWGVGCIFVEMIQGVAAFPGMKDIQ
DQLERIFLVLGTPNEDTWPGVHSLPHFKPERFTVYSSKSLRQAWNKLSYVNHAEDLASKLLQCSPKNRLSAQAALSHEYFSDLPPRL
WELTDMSSIFTVPNVRLQPEAGESMRAFGKNNSYGKSLSNSKH@MDRCKENCVSRPVKSTVPFGPKRVLVTEQIPSQHPGSASSGQ
AQRVLCPSNSQRVPPQAQKPVAGQKPVLKQLPAASGPRPASRLSNPQKSEQPQPAASGNNSEKEQTSIQKTEDSKKRQWTLEDFDIG
RPLGKGKFGNVYLAREKQSKFILALKVLFKVQLEKAGVEHQLRREVEIQSHLRHPNILRLYGYFHDATRVYLILEYAPLGTVYRELQ
KLSKFDEQRTATYITELANALSYCHSKRVIHRDIKPENLLLGSNGELKIADFGWSVHAPSSRRTTLCGTLDYQPPEMIEGRMHDEKV
DLWSLGVLCYEFLVGMPPFEAHTYQETYRRISRVEFTFPDFVTEGARDLISRLLKHNSSQRLTLAEVLEHPWIKANSSKPPTGHNSKE
ATSKSS@MASGPHSTATAAAAASSAAPSAGGSSSGTTTTTTTTTTGGILIGDRLYSEVSLTIDHSLIPEERLSPTPSMQDGLDLPSETDLR
ILGCELIQAAGILLRLPQVAMATGQVLFHRFFYSKSFVKHSFEIVAMACINLASKIEEAPRRIRDLINVFHHLRQLRGKRTPSPLILDQNY
INTKNQVIKAERRVLKELGFCVHVKHPHKIIVMYLQVLECERNQTLVQTAWNYMNDSLRTNVFVRFQPETIACACIYLAARALQIPL
PTRPHWFLLFGTTEEEIQEICIETLRLYTRKKPNYELLEKEVEKRKVALQEAKLKAKGLNPDGTPALSTLGGFSPASKPSSPREVKAEE
KSPISINVKTVKKEPEDRQQASKSPYNGVRKDSKRSRNSRSASRSRSRTRSRSRSHTPRRHYNNRRSRSGTYSSRSRSRSRSHSESPR
RHHNHGSPHLKAKHTRDDLKSSNRHGHKRKKSRSRSQSKSRDHSDAAKKHRHERGHHRDRRERSRSFERSHKSKHHGGSRSRSG
HGRHRR@MAVDVKSRAKRYEKLDFLGEGQFATVYKARDKNTNQIVAIKKIKLGHRSEAKDGINRTALREIKLLQELSHPNIIGLLD
AFGHKSNISLVFDFMETDLEVIIKDNSLVLTPSHIKAYMLMTLQGLEYLHQHWILHRDLKPNNLLLDENGVLKLADFGLAKSFGSPN
RAYTHQVVTRWYRAPELLFGARMYGVGVDMWAVGCILAELLLRVPFLPGDSDLDQLTRIFETLGTPTEEQWPDMCSLPDYVTFKSF
PGVPLQHIFIAAGDDLLELIQGLFLFNPCTRTTASQALKTKYFSNRPGPTPGCQLPRPNCPVEALKEPANPTVATKRKRAEALEQGILP
KKLIF@SMSKPKQDYSRTPGQVLSLISSLGFFTPVGEKDQDSANMFSAPKSAAQLSRGFICPMSVDQKEPTSYSSKLLKSCFETLSSNP
EIPVKCLTSNLLQCRKRLGTSSTSSQSHTFVSSVESECHSNPKWERDCQSTESSGCAMSWNAVEMLYAKSTSAIKTKTELELALSPIH
DSSAIPAAGSNQVTLPRKCFREISWEARDPDNENMTIDKGQSGFCQSSQRSVNSSATSEEHLGKRNYKRNFHLVDSSPCQEIMQSKK
NCTEYEANKERQGCRANQSTGLTTEVQNLKLSGCESQQLDYANKENIVTYLTDRQTPEKLHIPTIAKNLMSELDEDRELSSKKDCL
SSNSVCSDEDRALKTTCVDSDSSFPGVSMMESSLEIQALEPDKSIRDYSFEEPNTEDLFVLPKCQENSLPQDDCHACIQDSSQVSAHPS
KAPKALTSKINVVAFRSFNSHINASTNSEPSKISITSLDAMDISYDYSGSYPMAVSPTEKGRHYTSHQTPNQVKLGTSYRTPKSVRRGA
APVDDGRILGTPDYLAPELLLGTAHGPAVDWWALGVCLFEFLTGIPPFNDETPQQVFQNILKRDIPWPEGEEKLSDNAQSAMDMLLTI
DDSKRAGMRELKQHPLFSEVDWENLQHQTMPFVPQPDDETDTSYFEARNNAQHLTISGFSL@MACLHETRTPSPSFGGFVSTLSEA
SMRKLDPDTSDCTPEKDLTPTQCVLRDVVPLGGQGGGGPSPSPGGEPPPEPFANSVLQLHEQDTGGPGGATGSPESRASRVRADEV
RLQCQSGSGFLEGLFGCLRPVWTMIGKAYSTEHKQQQEDLWEVPFEEILDLQWVGSGAQGAVFLGRFHGEEVAVKKVRDLKETDI
KHLRKLKHPNIITFKGVCTQAPCYCILMEFCAQGQLYEVLRAGRPVTPSLLVDWSMGIAGGMNYLHLHKIIHRDLKSPNMLITYDD
VVKISDFGTSKELSDKSTKMSFAGTVAWMAPEVIRNEPVSEKVDIWSFGVVLWELLTGEIPYKDVDSSAIIWGVGSNSLHLPVPSSCPD
GFKILLRQCWNRKPRNRPSFRQILLHLDIASADVLSTPQETYFKSQAEWREEVKLHFEKIKSEGTCLHRLEEELVMRRREELRHALD
IREHYERKLERANNLYMELNALMLQLELKERELLRREQALERRCPGLLKSHTSRSLLHGNTMEKLIKKRNVPQKLSPHSKRPDILK
TESLLPKLDAALSGVGLPGCPKAPPSPGRSRRGKTRHRKASAKGSCGDLPGLRAALPPHEPGGLGSPGGLGVGPTAWDASPPALRGL
HHDLLLRKMSSSSPDLLSAALGARGRGATGGARDPGSPPPPQGDTPPSEGSAPGSTSPDSPGGAKGEPPPPVGPGEGVGLLGTGREGT
TGRGGSRAGYQHLTPAALLYRAAVTRSQKRGISSEEEEGEVDSEVELPPSQRWPQGPNMRQSLSTFSSENPSDVEEGTASEPSPSGTPE
VGSTNTDERPDERSDDMCSQGSEIPLDLPTSEVVPERETSSLPMQHQDDQGPNPEDSDCDSTELDNSNSIDALPPPASLPP@MISRMI
FRNYPSHNESDDEPFHFSISRELLLDRNDVVVGEMIGEGAYSIVYKGLLRNQFPVAVKIMDPSTTSAVTKAHKKTFQKEVLLLSKMK
HDNIVKFVGACIEPQLIIVTELVEGGTLQRFMHSRPGPLDLKMSLSFALDISRAMEFVHSNGIIHRDLNPRNLLVTGDLKHVKLADFG
IAREETRGGMTCEAGTSKWMAPEVYSPEPLRVGEKKEYDHKADIYSFAIVLWQLVTNEEPFPDVPNSLFVPYLVSQGRRPILTKTPDV
FVPIVESCWAQDPDARPEFKEISVMLTNLLRRMSSDSSIGTTLPDGEAYEGEMEESENSPLLQEHFCKVKKPKEKKKKKKLVKMRFP

FFKKFKVWLYNYKP@MSFFNFRKIFKLGSEKKKKQYEHVKRDLNPEDFWEIIGELGDGAFGKVYKAQNKETSVLAAAKVIDTKSE
EELEDYMVEIDILASCDHPNIVKLLDAFYYENNLWILIEFCAGGAVDAVMLELERPLTESQIQVVCKQTLDALNYLHDNKIIHRDLKA
GNILFTLDGDIKLADFGVSAKNTRTIQRRDSFIGTPYWMAPEVVMCETSKDRPYDYKADVWSLGITLIEMAEIEPPHHELNPMRVLL
KIAKSEPPTLAQPSRWSSNFKDFLKKCLEKNVDARWTTSQLLQHPFVTVDSNKPIRELIAEAKAEVTEEVEDGKEEDEEEETENSLPI
PASKRASSDLSIASSEEDKLSQNACILESVSEKTERSNSEDKLNSKILNEKPTTDEPEKAVEDINEHITDAQLEAMTELHDRTAVIKENE
REKRPKLENLPDTEDQETVDINSVSEGKENNIMITLETNIEHNLKSEEEKDQEKQQMFENKLIKSEEIKDTILQTVDLVSQETGEKE
ANIQAVDSEVGLTKEDTQEKLGEDDKTQKDVISNTSDVIGTCEAADVAQKVDEDSAEDTQSNDGKEVVEVGQKLINKPMVGPEAG
GTKEVPIKEIVEMNEIEEKKKK@MARTTSQLYDAVPIQSSVVLCSCPSPSMVRSQTEPSSSPGIPSGVSRQGSTMDGTTAEARPSTNP
LQQHPAQLPPQPRKKRPEDFKFGKILGEGSFSTVVLARELATSREYAIKILEKRHIIKENKVPYVTRERDVMSRLDHPFFVKLYFTFQ
DDEKLYFGLSYAKNGELLKYIRKIGSFDETCTRFYTAEIVSALEYLHGKGIIHRDLKPENILLNEDMHIQITDFGTAKVLSPDSKQARA
NSFVGTAQYVSPELLTEKSACKSSDLWALGCIIYQLVAGLPPFRAGNEYLIFQKIIKLEYDFPEKFFPKARDLVEKLLVLDATKRLGCEE
MEGYGPLKAHPFFESITWENLHQQTPPKLTAYLPAMSEDDEDCYGNYDNLLSQFGCMQVSSSSSSHSLCAVDASLPQRSGSNIEQYIH
DLDTNSFELDLQFSEDEKRLLLEKQAGGNPWHQFVENNLILKMGPVDKRKGLFARRRQLLLTEGPHLYYVDPVNKVLKGEIPWSQ
ELRPEAKNFKTFFVHTPNRTYYLMDPSGNAHKWCRKIQEVWRQQYQSSPDAAVQ@LTHAGWGQGWTLARTRSLLIMLGPGSNRR
RPTQGERGPGSPGEPMEKYQVLYQLNPGALGVNLVVEEMETKVKHVIKQVECMDDHYASQALEELMPLLKLRHAHISVYQELFIT
WNGEISSLYLCLVMEFNELSFQEVIEDKRKAKKIIDSEWMQNVLGQVLDALEYLHHLDIIHRNLKPSNIILISSDHCKLQDLSSNVLM
TDKAKWNIRAEEDPFRKSWMAPEALNFSFSQKSDIWSLGCIILDMTSCSFMDGTEAMHLRKSLRQSPGSLKAVLKTMEEKQIPDVE
TFRNLLPLMLQIDPSDRITIKDVVHITFLRGSFKSSCVSLTLHRQMVPASITDMLLEGNVASILEVMQKFSGWPEVQLRAMKRLLKMP
ADQLGLPWPPEIVEVVVTTMELHDRVLDVQLCACSLLLHLLGQALVHHPEAKAPCNQAITSTLLSALQSHPEEEPLLVMVYSLLAIT
TTQESESLSEELQNAGLLEHILEHLNSSLESRDVCASGLGLLWALLLDGIIVNKAPLEKVPDLISQVLATYPADGEMAEASCGVFWLLS
LLGCIKEQQFEQVVALLLQSIRLCQDRALLVNNAYRGLASLVKVSELAAFKVVVQEEGGSGLSLIKETYQLHRDDPEVVENVGMLLV
HLASYEEILPELVSSSMKALLQEIKERFTSSLVSDSSAFSKPGLPPGGSPQLGCTTSGGLE@MDYYSQGTFQNIMENKRKLKAVVDTE
WMHTMLSQVLDAIEYLHKLNIVHRNLKPSNIVLVNSGYCKLQDMSSQALMTHEAKWNVRAEEDPCQKSWMAPEALKFSFSTKSDI
WSLGCIILDMATCSFLNDTEAMQLRKAIRHHPGSLKPILKTMEEKQIPGTDVYYLLLPFMLHINPSDRLAIKDVMQVTFMSNSFKSSS
VALNMQRQKVPIFITDVLLEGNMANILDVMQNFSSRPEVQLRAINKLLTMPEDQLGLPWPTELLEEVISIIKQHGRILDILLSTCSLLL
RVLGQALAKDPEAEIPRSSLIISFLMDTLRSHPNSERLVNVVYNVLAIISSQGQISEELEEEGLFQLAQENLEHFQEDRDICLSILSLLWS
LLVDVVTVDKEPLEQLSGMVTWVLATHPEDVEIAEAGCAVLWLLSLLGCIKESQFEQVVVLLLRSIQLCPGRVLLVNNAFRGLASLAK
VSELVAFRIVVLEEGSSGLHLIQDIYKLYKDDPEVVENLCMLLAHLTSYKEILPEMESGGIKDLVQVIRGRFTSSLELISYADEILQVLEA
NAQPGLQEDQLEPPAGQEAPLQGEPLFRP@MEHSVPKNKLKKLSEDSLTKQPEEVFDVLEKLGEGSYGSVFKAIHKESGQVVAIKQ
VPVESDLQEIIKEISIMQQCDSPYVVKYYGSYFKNTDLWIVMEYCGAGSVSDIIRLRNKTLTEDEIATVLKSTLKGLEYLHFMRKIHR
DIKAGNILLNTEGHAKLADFGVAGQLTDTMAKRNTVIGTPFWMAPEVIQEIGYNCVADIWSLGITSIEMAEGKPPYADIHPMRAIFMI
PTNPPPTFRKPEHWSDDFTDFVKKCLVKNPEQRATATQLLQHPFIVGAKPVSILRDLITEAMDMKAKRQQEQQRELEEDDENSEEE
VEVDSHTMVKSGSESAGTMRATGTMSDGAQTMIEHGSTMLESNLGTMVINSDDEEEEEDLGSMRRNPTSQQIQRPSFMDYFDKQ
DSNKAQEGFNHNQQDPCLISKTAFPDNWKVPQDGDFDFLKNLDFEELQMRLTALDPMMEREIEELRQRYTAKRQPILDAMDAKKR
RQQNF@PTRPTRLIVSNFSQAKQKSHLVDPQILRDQSRLAPEIITATQYKKCDEFQTGILIYEMLHLPNPFDENPELKEKEYTRTDLP
RIPLRSPYSWGLQQLASCLLNPNPSERILISDAKGILQCLLWGPREDLFQIFTTSATLAQKNALLQNWLDIKRTLLMIKFAEKSLDREG
GISLEDWLCAQYLAFATTDSLSYIVKILQYR@MQNKENREPRVQQTPSAGVGPLRVEMNPDTHAVSGPGRVPVKSNSKVLSIDDFDI
GRPLGKGKFGNVYLARERKLKVVIALKVLFKSQMVKEGVEHQLRREIEIQSHLRHPNILRFYNYFHDDTRVFLILEYAPRGEMYKEL
QRYGRFDDQRTATYMEEVSDALQYCHEKKVIHRDIKPENLLLGYRGELKIADFGWSVHAPSLRRRTMCGTLDYLPPEMIEGHSHDE
KVDLWSIGVLCYECLVGNPPFETASHAETYKRITKVDLQFPKLVSEGARDLISKLLRHSPSMRLPLRSVMEHPWVKANSRRVLPPVC
SSEPH

data.in

6978799 0
45768856 509
45768786 943
27923854 1367

45768720 1765
45768758 2292
45219906 2639
18202599 3311
4835224 4200
40787731 4571
18202068 5189
34191428 5749
29747774 6450
28856169 7038
20071571 7531
45709347 7734

#getSeq.py

```
data = open("data.in","r")

data2 = open("data.seq","r")
seq = input("Enter Sequence Here: ")

string = "".join(data2)
count = string.find(seq, 0, len(string))



result = 0
for line in data:
    pair = line.split()
    if count > int(pair[1]):
                result = (pair[0])

print(result)
```

18202068 is the result.