Applications du Shell (1)

1 Utilisation des utilitaires Unix

Il existe 22 acides aminés principaux. Ils sont représentés par les lettres : `A C D E F G H I K L M N O P Q R S T U V W Y`, *cf.,* http://fr.wikipedia.org/wiki/Acide_aminé

Le fichier q9660.fasta contient une suite d'acides aminés de la protéine humaine `Q9660:26S` *proteasome non-ATPase*. Dans ce fichier chacun des 22 acides aminés est représenté par son code alphabétique (une lettre). Les premières lignes de ce fichier sont :

```
$ cat q9660.fasta
         mitSAAgiis LLDEDEPQLK EFALHKLNAV VNDFWAEISE SVDKIEVLYE DEGFRSRQFA
         ALVASKVFYH LGAFEESLNY ALGAGDLFNV NDNSEYVETI IAKCIDHYTK QCVENADLPE
         GEKKPIDQRL EGIVNKMFQR CLDDHKYKQA IGIALETRRL DVFEKTILES NDVPGMLAYS
         LKLCMSLMQN KQFRNKVLRV LVKIYMNLEK PDFINVCQCL IFLDDPQAVS DILEKLVKED
         NLLMAyQICF DLYESASQQF LSSVIQNLRT VGTPIASVPG STNTGTVPGS EKDSDSMETE
         EKTSSAFVGK TPEASPEPKD QTLKMIKILS   GEMAIELHLQ  FLIRNNNTDL MILKNTKDAV
         RNSVCHTATV IANSFMHCGT TSDQFLRDNL EWLARATNWA KFTATASLGV IHKGHEKEAL
         QLMATYLPkD TSPGSAYQEG GGLYALGLIH ANHGGDIIDY LLNQLKNASN DIVRHGGSLG
         LGLAAMGTAR QDVYDLLKTN LYQDDAVTGE AAGLALGLVM LGSKNAQAIE DMVGYAQETQ
         HEKILRGLAV GIALVMYGRM EEADALIESL CRDKDPILRR SGMYTVAMAY CGSGNNKAIR
         RLLHVAVSDV nddVrrAAve SLGFILFRTP EqCPSVVSLL SESYNPHVRY GAAMALGICC
         AGTGNKEAIN LLEPMTNDPV NYVRQGALIA    SAlimIQQTE ITCPKVNQFR QLYSKVINDK
         HDDVMAKFGA ILAQGILDAG GHNVTISLQS RTGHTHMPSV VGVLVFTQFW FWFPLSHFLS
         LAYTPTCVIG LNKDLKMPKV QYKSNCKPST FAYPAPLEVP KEKEKEKVST AVLSITAKAK
         KKEKEKEKKE EEKMEVDEAE KKEEKEKKKE PEPNFQLLDN PARVMPAQLK VLTMPETCRY
         QPFKPLSIGG IIILKDTSED IEELVEPVAA HGPKIEEEEQ EPEPPEPFEY     IDD
[...]
```

Ce fichier n'est malheureusement pas très exploitable tel quel par les biologistes, car il contient parfois de caractères en majuscule d'autres en minuscule, parfois un ou plusieurs espaces entre deux acides aminés, parfois des tabulations et il contient un retour à la ligne à la fin de chaque ligne. Les biologistes souhaiteraient pouvoir travailler sur un fichier normalisé appelé q9660.norm qui ne contienne ni espace, ni tabulation ni retour à la ligne et où tous les acides aminés sont notés en majuscule. Voici un exemple du début du fichier précédent normalisé :

```
$ cat q9660.norm
MITSAAGIISLLDEDEPQLKEFALHKLNAVVNDFWAEISESVDKIEVLYEDEGFRSRQFAALVASKVFYHLGAFEESLNYALGAGDLFNVNDNS
EYVETIIAKCIDHYTKQCVENADLPEGEKKPIDQRLEGIVNKMFQRCLDDHKYKQAIGIALETRRLDVFEKTILESNDVPGMLAYSLKLCMSLM
QNKQFRNKVLRVLVKIYMNLEKPDFINVCQCLIFLDDPQAVSDILEKLVKEDNLLMAYQICFDLYESASQQFLSSVIQNLRTVGTPIASVPGST
NTGTVPGSEKDSDSMETEEKTSSAFVGKTPEASPEPKDQTLKMIKILSGEMAIELHLQFLIRNNNTDLMILKNTKDAVRNSVCHTATVIANSFM
HCGTTSDQFLRDNLEWLARATNWAKFTATASLGVIHKGHEKEALQLMATYLPKDTSPGSAYQEGGGLYALGLIHANHGGDIIDYLLNQLKNASN
DIVRHGGSLGLGLAAMGTARQDVYDLLKTNLYQDDAVTGEAAGLALGLVMLGSKNAQAIEDMVGYAQETQHEKILRGLAVGIALVMYGRMEEAD
ALIESLCRDKDPILRRSGMYTVAMAYCGSGNNKAIRRLLHVAVSDVNDDVRRAAVESLGFILFRTPEQCPSVVSLLSESYNPHVRYGAAMALGI
CCAGTGNKEAINLLEPMTNDPVNYVRQGALIASALIMIQQTEITCPKVNQFRQLYSKVINDKHDDVMAKFGAILAQGILDAGGHNVTISLQSRT
GHTHMPSVVGVLVFTQFWFWFPLSHFLSLAYTPTCVIGLNKDLKMPKVQYKSNCKPSTFAYPAPLEVPKEKEKEKVSTAVLSITAKAKKKEKEK
EKKEEEKMEVDEAEKKEEKEKKKEPEPNFQLLDNPARVMPAQLKVLTMPETCRYQPFKPLSIGGIIILKDTSEDIEELVEPVAAHGPKIEEEEQ
EPEPPEPFEYIDD [...]
```

Notez bien que dans ce fichier l'ensemble des acides aminés sont sur une seule et même ligne, il n'y a aucun retour à la ligne dans ce fichier q9660.norm.

1.1 Exercice Écrire la suite de commandes en ligne permettant d'obtenir le fichier normalisé q9660.norm à partir du fichier initial q9660.fasta.

1.2 Exercice Faire la même chose dans un script-shell `normalise.sh` qui prendra en paramètres le nom du fichier à traiter et le nom du fichier normalisé.

Les biologistes souhaitent faire différentes statistiques à partir de ce fichier normalisé.

1.3 Exercice Écrire la commande permettant de calculer le nombre d'acide aminé présent dans ce fichier.

1.4 Exercice Écrire une suite de commande permettant de calculer le nombre d'acide aminé de type `A` (Alanine) présent dans le fichier q9660.norm.

1.5 Exercice Écrire une suite de commande permettant de calculer la proportion, i.e. le pourcentage de présence d'Alanine dans le fichier q9660.norm. Pour ce faire, vous pourrez tirer avantage de l'utilisation d'une calculatrice à précision arbitraire disponible en commande Unix comme par exemple bc .

1.6 Exercice Écrire un script-shell statbio.sh affichant sur chaque ligne de la sortie standard la lettre correspondant à chaque acide aminé suivi de son nombre d'apparition dans la séquence du fichier q9660.norm.

1.7 Exercice Écrire la ligne de commande appelant votre *script-shell* statbio.sh, afin que le résultat soit stocké dans un fichier q9660.stat.

1.8 Exercice Écrire la commande permettant de trier les données du fichier q9660.stat par ordre croissant sur le nombre d'apparition des acides aminés.

1.9 Exercice Écrire un script-shell test-sous-sequence.sh qui prend en paramètre le nom du fichier normalisé à traiter et une sous-séquence d'acides aminés (*i.e.*, une suite consécutive d'acides aminés). Ce script renvoie la chaîne de caractère `OK` si la sous-séquence est présente dans le fichier et renvoie `KO` sinon. Pour exemple, le test de présence de la sous-séquence `ADC` dans le fichier q9660.norm doit renvoyer `KO`, alors que le test de présence de la sous-séquence `WAK` doit renvoyer `OK`.

1.10 Exercice En utilisant le contenu du fichier q9660.stat, écrire une commande permettant d'afficher la liste des acides aminés non présents dans la séquence (donc sans afficher leur nombre d'apparition qui bien entendu est égal à 0).