MASTER BIBS M1 2024
Séance 7

Le ribosome est l'une des protéines les plus abondantes du cytosol. Sa surface est chargée négativement. Un article récent a montré que le coefficient de diffusion des protéines chargées positivement en surface était 100 fois inférieur à celui des autres protéines [1]. En effet, ces dernières se trouvent piégées dans des interactions non-spécifiques avec les ribosomes chargés négativement. On peut donc émettre l'hypothèse qu'une pression de sélection s'exerce sur la charge des protéines les plus abondantes afin d'éviter la formation d'agrégats avec les ribosomes. On se propose donc d'étudier la corrélation entre la fréquence de chacun des 20 acides aminés et l'abondance des protéines de *S. cerevisiae*.

Vous disposez :

- du fichier « 4932-WHOLE_ORGANISM-integrated.txt » contenant les abondances des protéines de *S. cerevisiae* issues d'une expérience de Spectrométrie de Masse. L'identifiant « gene name » de chaque protéine se trouve en colonne 2 après le pattern « 4932. » et son abondance en colonne 3.

- du fichier fasta « CDS.multifasta » contenant les séquences protéiques des CDS (Coding Sequences) de *S. cerevisiae*. Attention, les identifiants des CDS ne correspondent pas aux identifiants « gene name » du fichier d'abondance. Dans ce fichier, nous avons défini notre propre nomenclature correspondant aux coordonnées génomiques de chaque CDS, par ex : chrI_+_36509-37147 pour une CDS localisée sur le chromosome I, brin sens (ou +) commençant au nucléotide 36509 et terminant en 37147.

- du fichier « CDS_gene_correspondance.gff » qui permet d'établir la correspondance entre les identifiants « gene name » et les identifiants indiqués dans le fichier « CDS.multifasta ». Chaque CDS est renseignée sur 1 ligne, où la colonne 1, indique le chromosome sur laquelle se trouve la CDS, les colonnes (4, 5) indiquent les coordonnées chromosomiques (positions dans le chromosome), le sens du brin est indiqué en colonne 7 et permet de retrouver l'identifiant du fichier fasta, la dernière colonne contient l'identifiant « gene name » (identifiant précédé du pattern « Name= ») tel qu'il est indiqué dans le fichier d'abondances *4932-WHOLE_ORGANISM-integrated.txt*.

Par exemple, la première CDS du fichier d'identifiant "YAL069W" (notez que le "_CDS" ne fait pas partie de l'identifiant) dans le fichier d'abondance aura pour identifiant dans le fichier fasta "chrI_+_335-649" car elle est localisée sur le chromosome I, brin "+" et de coordonnées 335-649.

Avant toute chose :

Les fichiers lus par votre programme doivent être donnés en arguments et non renseignés en "dur" dans votre programme.

exemple d'utilisation :

```
script_exam2022.py -ab 4932-WHOLE_ORGANISM-integrated.txt -fasta
CDS.multifasta -gff CDS_gene_correspondance.gff
```

Votre programme vérifiera en premier lieu que les fichiers renseignés par l'utilisateur en argument existent bien et renverra un message d'erreur si ce n'est pas le cas en quittant proprement.

1/ Créez une fonction (Parse_gff()) qui lira le fichier « CDS_gene_correspondance.gff ». Cette fonction devra créer un dictionnaire (d_abondseq) contenant pour chaque "gene name" (clé du dictionnaire), l'identifiant correspondant dans le fichier fasta. On rappelle que la tabulation se matérialise par "\t".

2/ Créez une fonction (GetAbondance()) qui lira le fichier d'abondance et récupèrera l'abondance de chaque CDS (lorsque l'info est présente) puis la stockera dans le dictionnaire d_abondseq au niveau de la clé de la CDS dans une nouvelle clé "abondance" (d_abondseq["YAL069W"][" abondance "] = 1.12.

3/ Créez une fonction (GetSeq()) qui lira le fichier fasta et récupèrera pour chaque CDS, la séquence protéique associée puis la stockera dans le dictionnaire d_abondseq au niveau de la clé de la CDS dans une nouvelle clé "seq" (d_abondseq["YAL069W"][" seq "] = " MIVNNTHVLTLPLYTTTTCHTHPHLYTDFTYAHGCYSIYHLKLTLLSDSTSLHGPSLTESVPNALTSL

CTALASAVYTLCHLPITPIIIHILISISHSAVPNIV".

4/ Pour chaque CDS, calculez la fréquence des 20 acides aminés et stockez pour chaque CDS, la fréquence de chaque acide aminé dans le dictionnaire d_abondance. Ainsi, **chaque CDS et finalement chaque valeur d'abondance sera associée à 20 fréquences d'acides aminés**.

5/ Visualisation et analyse :

- pour chaque acide aminé GLU (E), ASP (D), LYS (K) et ARG (R), tracez le plot (**abondance** (ord) vs **fréquence de l'acide-aminé en question** (abs)).

- utilisez la librairie scipy pour calculer les corrélations (Spearman) entre l'abondance de chaque CDS et la fréquence de chacun des 20 acides aminés cette fois. Que constatez-vous ?

6/ Sortie :

Votre script écrira dans un fichier "corr_ab_aa.txt", les coefficients et p-values des corrélations de Spearman calculées pour les 20 acides aminés vs abondances de CDS en suivant le format suivant :

aa (code une lettre) TABULATION corr.coef TABULATION pvalue
A 0.12 0.01
C 0.07 0.07
etc

Refaire le même calcul avec la fréquence de charges négatives ensemble (E+D). Idem pour les charges positives (K + R). Que concluez-vous ?!

7/ Stickiness :

On se pose alors la question de savoir si les protéines les plus abondantes seraient de façon générale, moins "sticky" pour ne pas être piégées dans des interactions non-fonctionnelles avec les protéines de la cellule ce qui serait fortement délétère à haute concentration. Pour cela, créez une fonction (CompStk()) qui calcule le score de stickiness de chaque CDS. Pour cela, il suffit de sommer les valeurs de stickiness (données dans le fichier "stickiness.txt") des acides aminés de la séquence. Chaque CDS sera alors associée à une nouvelle clé "stk" pointant vers la valeur de stickiness de la séquence. Calculez la corrélation (Spearman) abondance vs stickiness et tracez le plot correspondant.

Exemple d'utilisation de la librairie scipy :

```
>>>from scipy import stats

>>>a = [2,3,4,5,6,8,12]

>>>b = [100, 200, 300, 400,500,800,1200]

>>>stats.spearmanr(a,b)

SpearmanrResult(correlation=1.0, pvalue=0.0)
```

Références

[1] Schavemaker PE, Śmigiel WM, Poolman B. Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome. Elife 2017;6. doi:10.7554/eLife.30084.