Sorbonne Universités

3I005 - Statistique et Informatique

Projet 2 - Analyse statistique d'une famille de proteines

Par
Marc HU & Sabrina CHAOUCHE

Année universitaire : 2017-2018

Table des matières

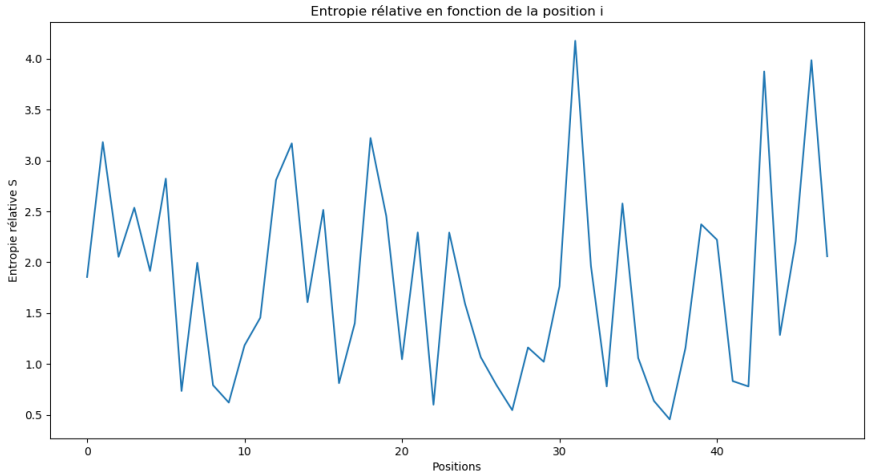
I Réponses aux questions

I.1 Données

Les données des fichiers Dtrain, test-seq, test-seq2 et distances sont lues et conservées dans des matrices pour leurs utlisations.

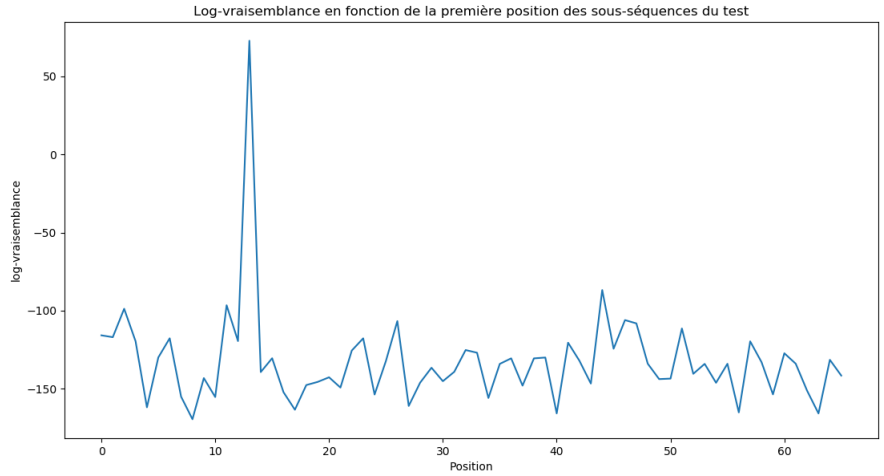
I.2 Modélisation par PSWM

L'entropie rélative S :



La valeur de l'entropie pour une position donnée "i" varie entre 0 et log2(21) : plus la valeur est maximale, plus la position est conservée. D'après le tracé de l'entropie, on remarque que la position la plus conservée est la position 31 et l'acide la plus conservée à cette position est l'acide W.

La log-vraisemblance :

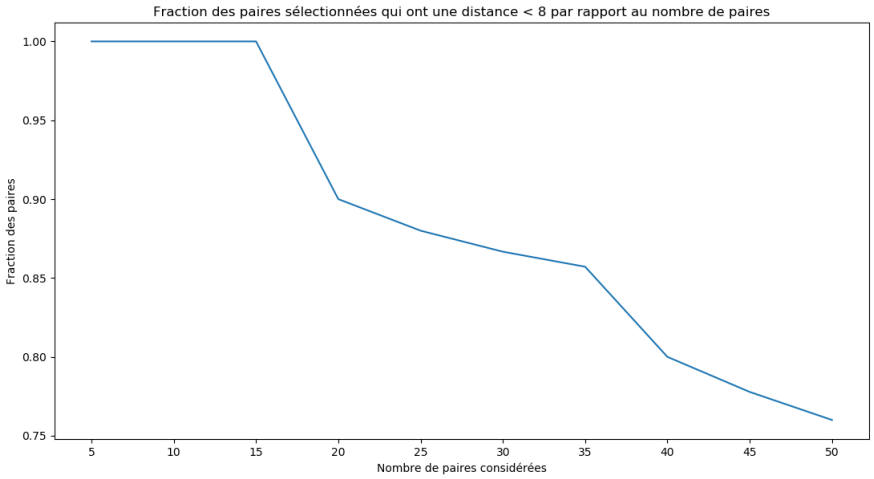


La log-vraisemblance indique l'appartenance d'une séquence à la famille donnée par Dtrain si sa valeur est positive. D'après le tracé de la log-vraisemblance des sous-séquences du fichier de test, on a une seule valeur positive donc on en déduit qu'il y a une seule séquence qui appartient à la famille donnée par Dtrain.

La sous-séquence du fichier de test appartenant à la famille donnée par Dtrain est :
KALYDFAGQSAGELSLGKDEIILVTQKENNGWWLASRLDKSASGWAPS.

I.3 Co-évolution des résidues en contact :

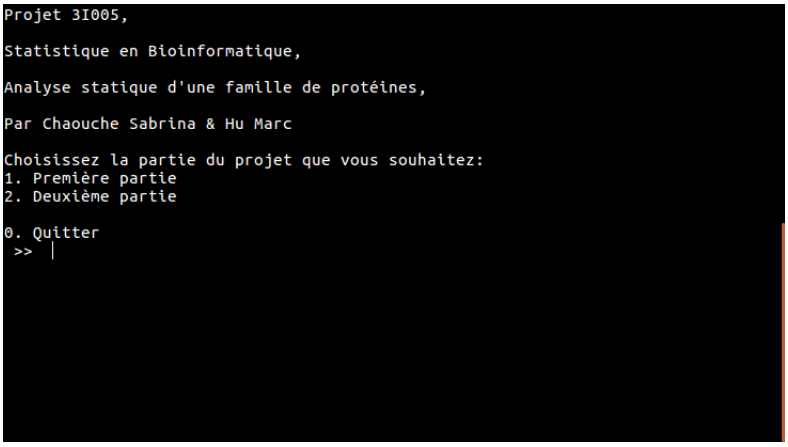
La fraction des 50 paires sélectionnées



Le tracé de la fraction des 50 paires sélectionnées montre que plus les paires sont correlées plus la probabilité qu'elles soient en contact augmente : plus l'information mutuelle Mij augmente plus la probabilité que les positions correspondantes soient statistiquement dépendantes. Pour les 10 plus grandes valeurs Mij, l'information mutuelle est très élevée donc leurs distances respectives sont < 8 et dans ce cas, on a une $fraction = 1$ ce qui signifie que les paires sont en contact. Dans le cas où on considère les 50, la fraction diminue car l'information mutuelle diminue.

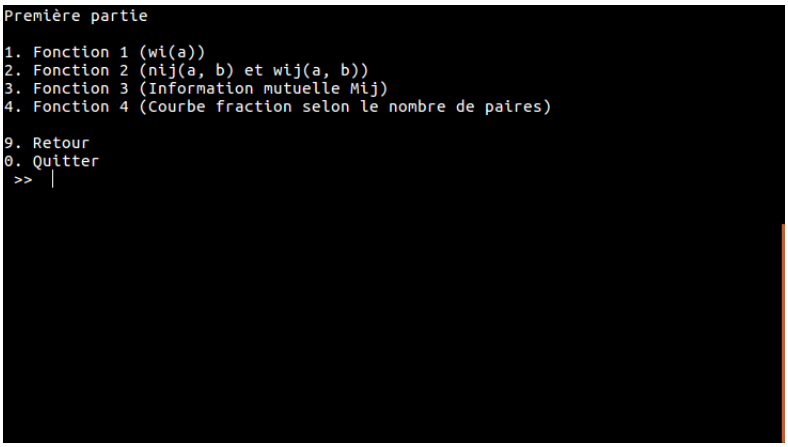
II Guide d'utilisation

Pour faciliter le test des fonctions, on a utilisé un menu depuis la ligne de commande. Pour lancer les menu, il suffit de taper la commande **"python3 projet2.py"** et le menu suivant s'affiche : pour utiliser les fonctions de la partie "Modélisation par PSWM" il suffit d'entrer **1** et pour les fonctions de la partie "Co-évolution des résidues en contact", entrer **2**. Pour quitter le programme, entrez **0**.

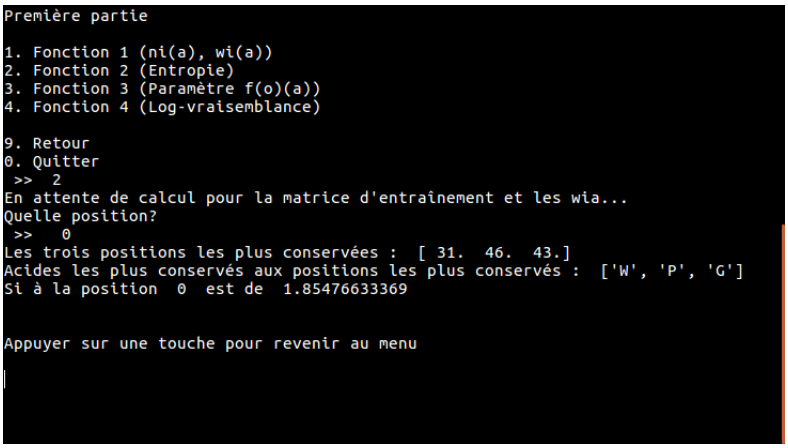


La première partie :

Pour utiliser une des fonctions de la partie "Modélisation par PSWM", il suffit de choisir un argument entre 1 et 4 correspondant au numéro de la fonction -tel est mentionné dans l'énoncé- et de suivre les instructions pour effectuer le calcul correspondant.

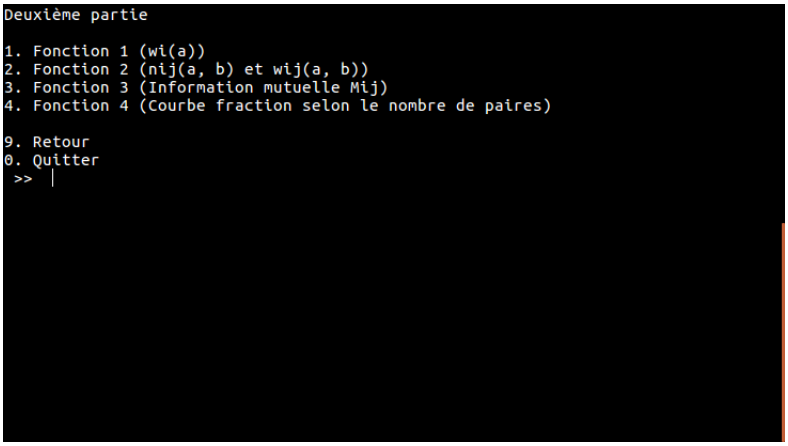


Exemple d'utilisation de la fonction 2 de la partie "Modélisation par PSWM"



La seconde partie :

Pour utiliser une des fonctions de la partie "Co-évolution des résidues en contact", il suffit de choisir un argument entre 1 et 4 correspondant au numéro de la fonction -tel est mentionné dans l'énoncé- et de suivre les instructions pour effectuer le calcul correspondant.



Exemple d'utilisation de la fonction 4 de la partie "Co-évolution des résidues en contact"