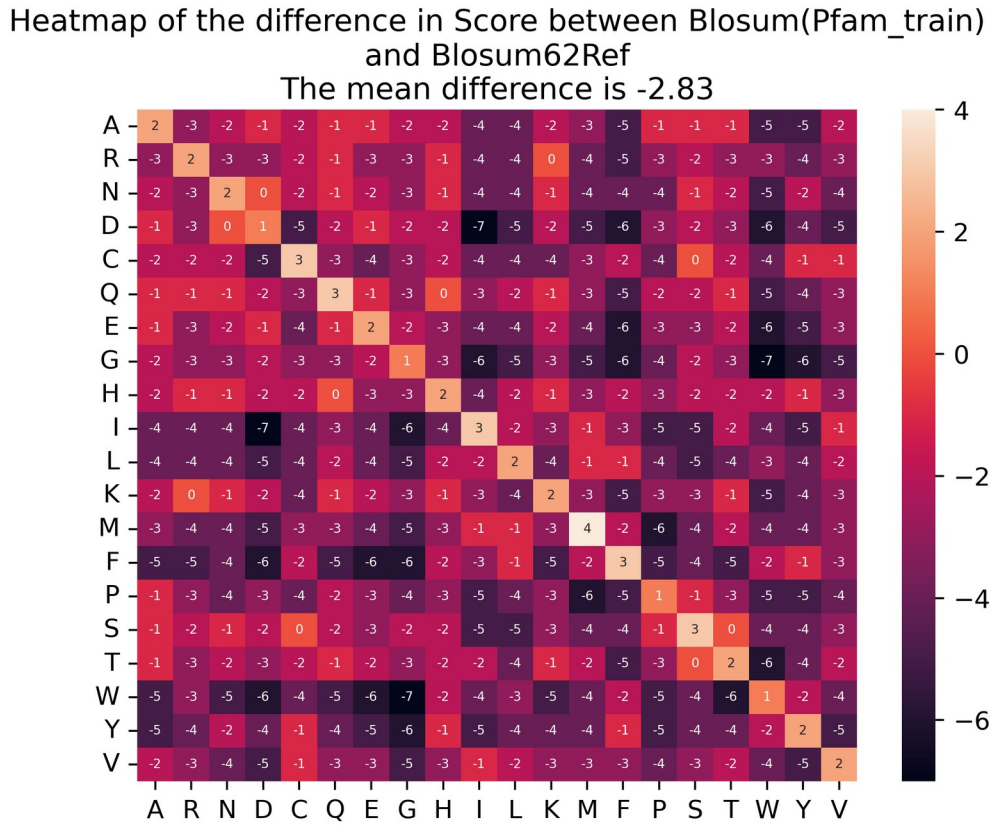
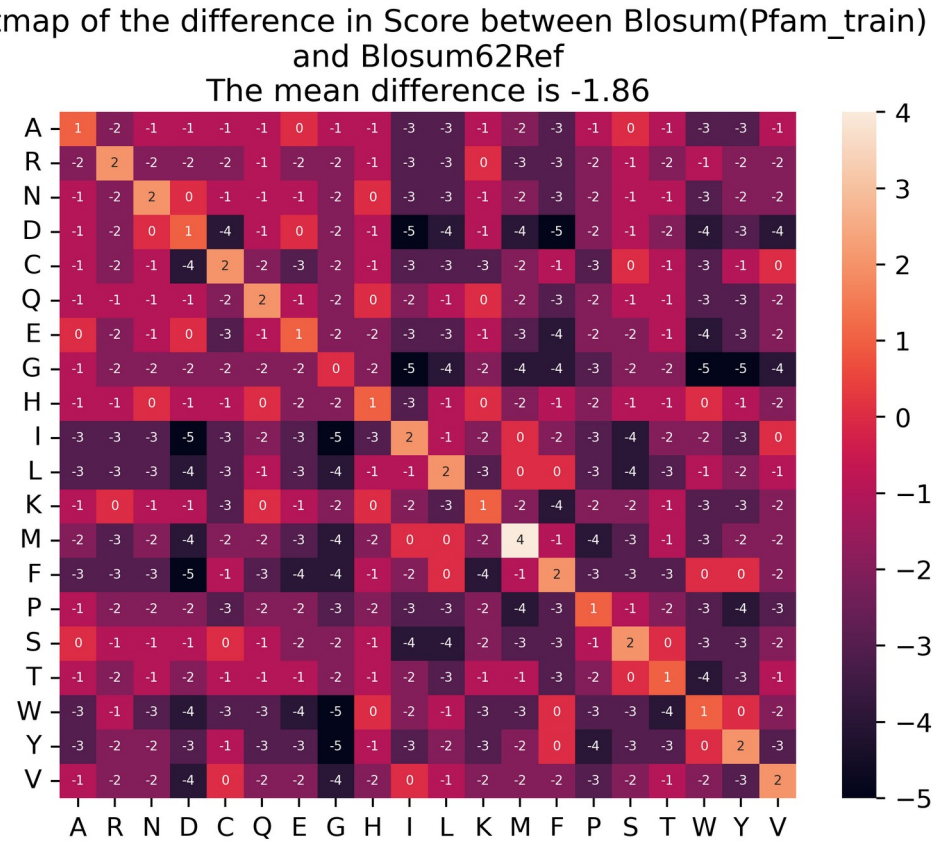
Recalcul de Blosum avec pid et clustering corrigés

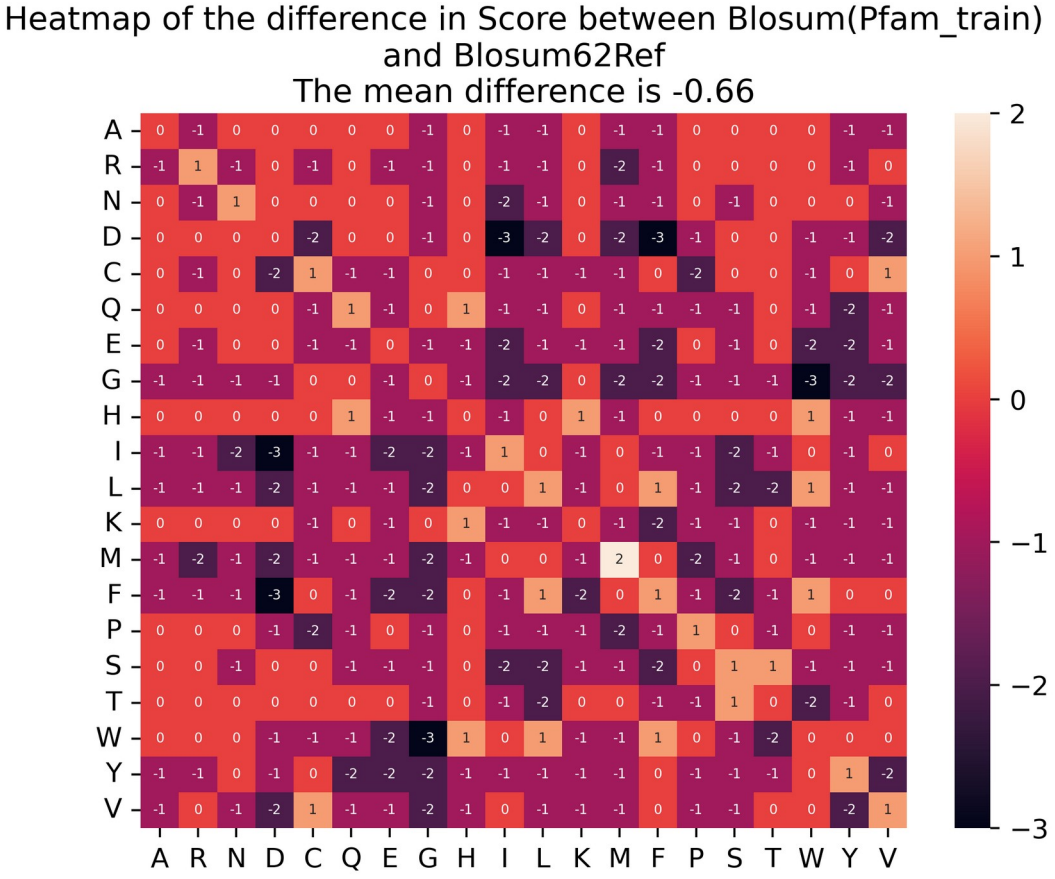
Variante blosum62 (5min)



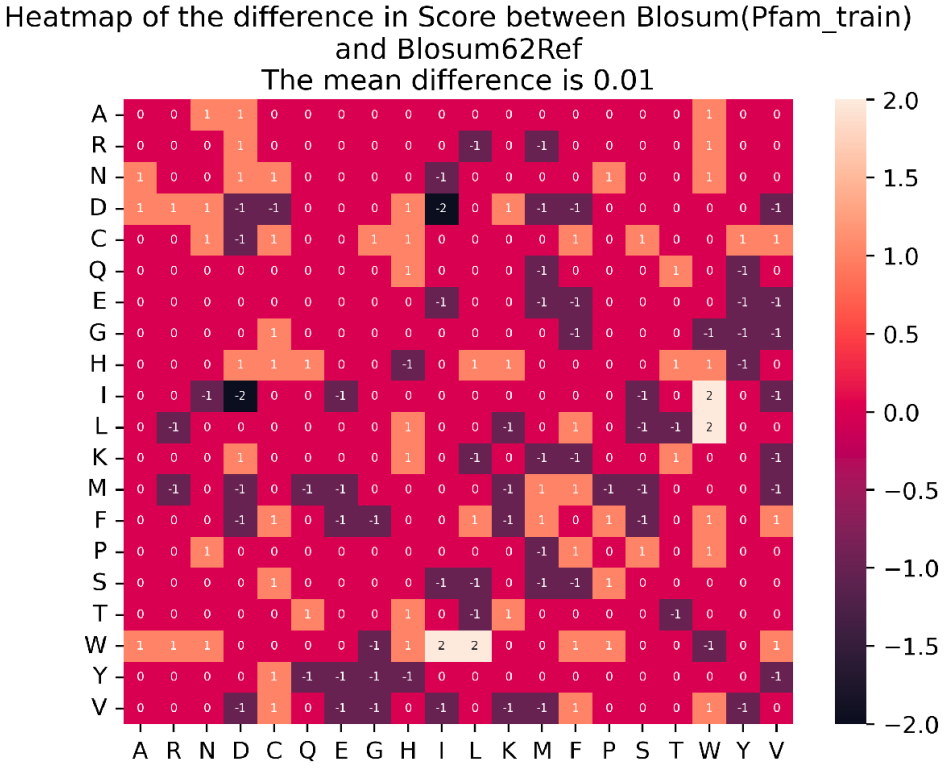
Variante blosum50 (20min)



Variante blosum30 (1,5h)



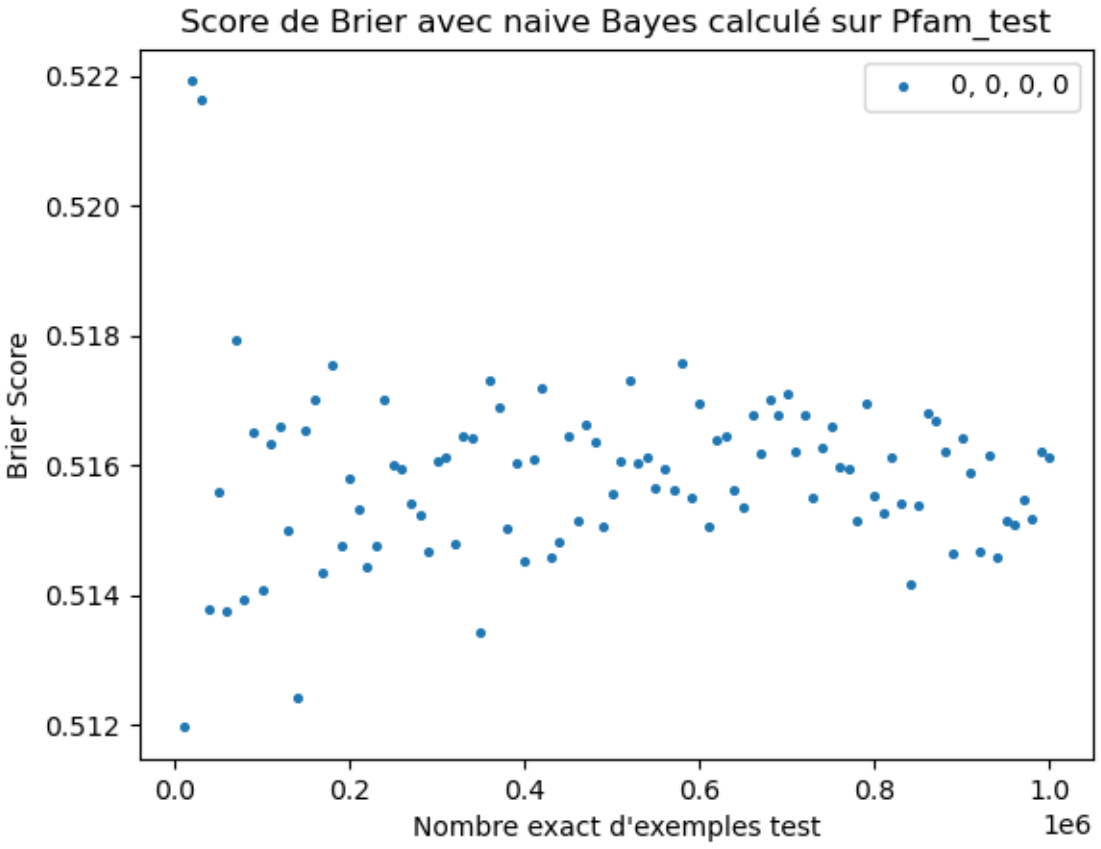
Variante blosum0 (3,7h)



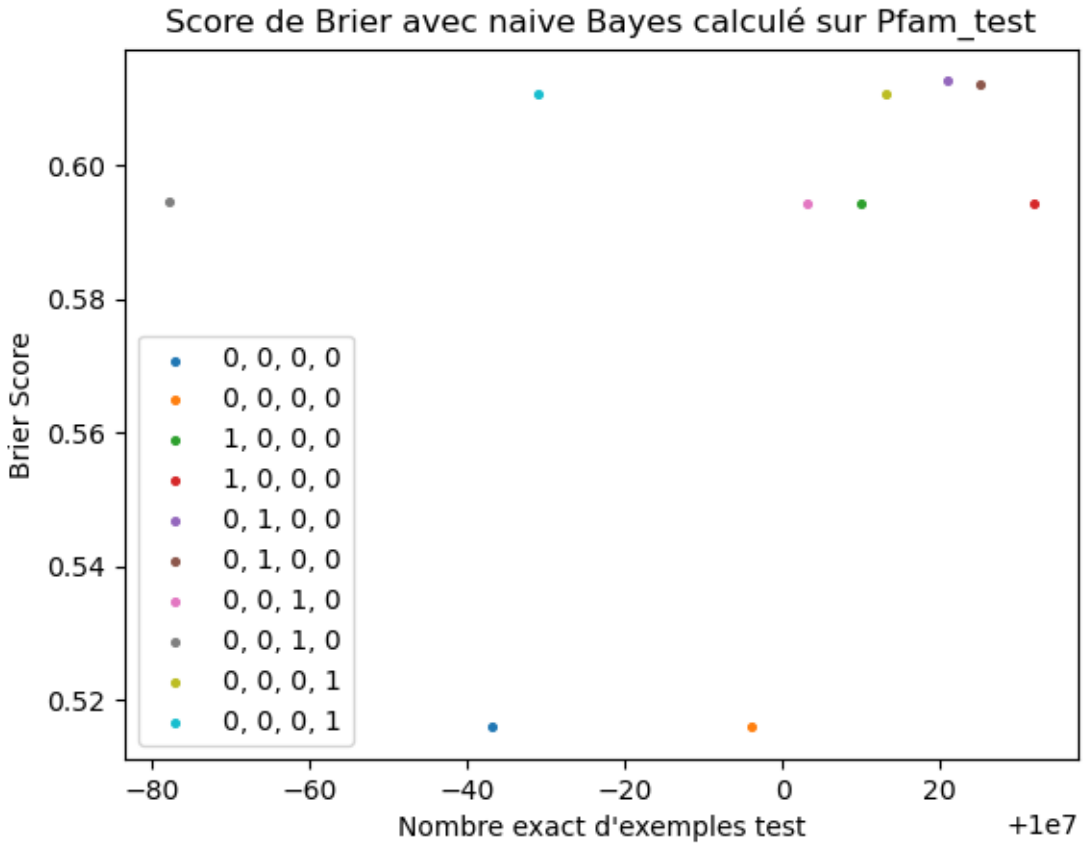
Recalcul de Scores de Brier avec Bayes Naif

(avec pid = 62, temps de calcul 13min/cube)

Stabilisation du Score (test de 10m à 1M d'exemples)



Sur 10M d’exemples demandés



Tests préliminaires supplémentaires

seeds identité

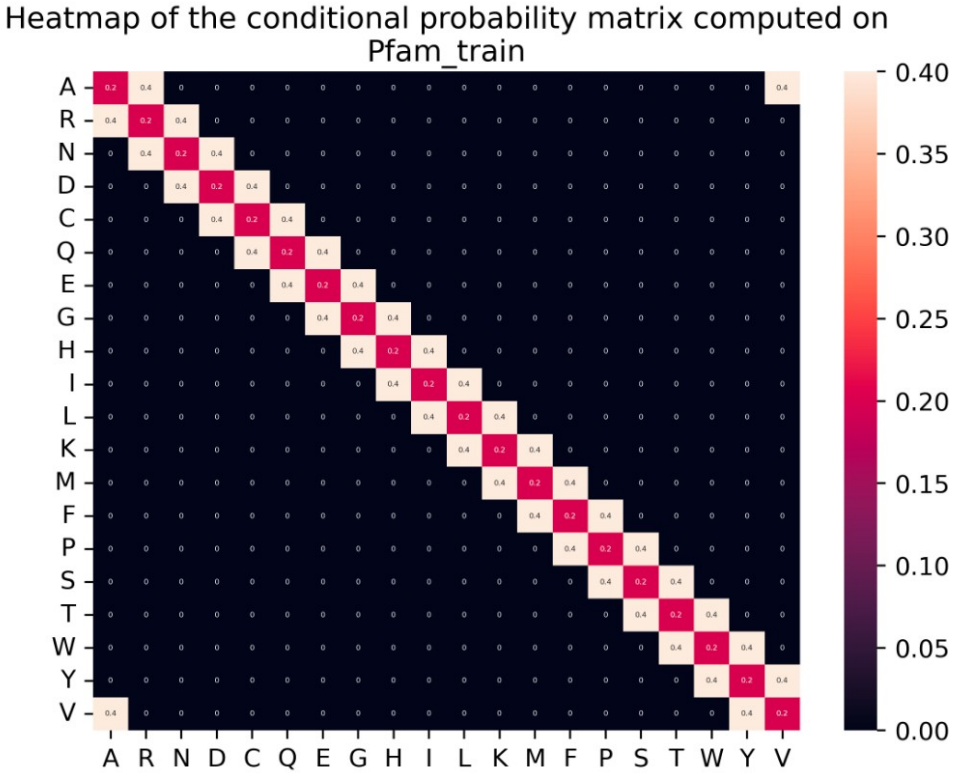
```
>seq1
ARNDCQEGHILKMFPSTWYV
>seq2
ARNDCQEGHILKMFPSTWYV
```

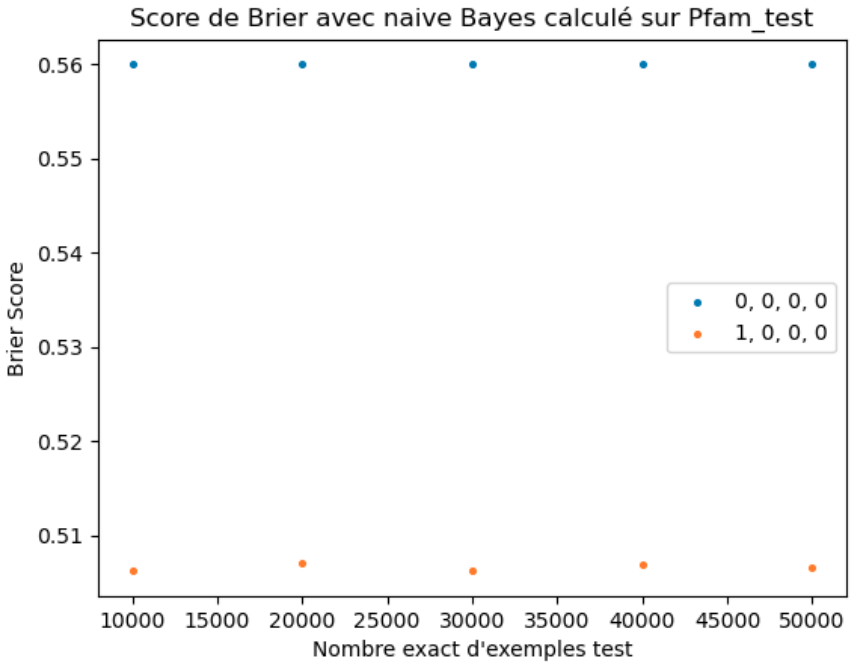
seeds décalés

```
>seq1
ARNDCQEGHILKMFPSTWYV
>seq2
RNDCQEGHILKMFPSTWYVA
```

- 10 seeds identité/décalés (5 train, 5 test)
- Pas de clustering 99 %
- pid_inf = 0
- Imposer ordre seq1, seq2

1 identité, 9 décalés
(id dans le train)

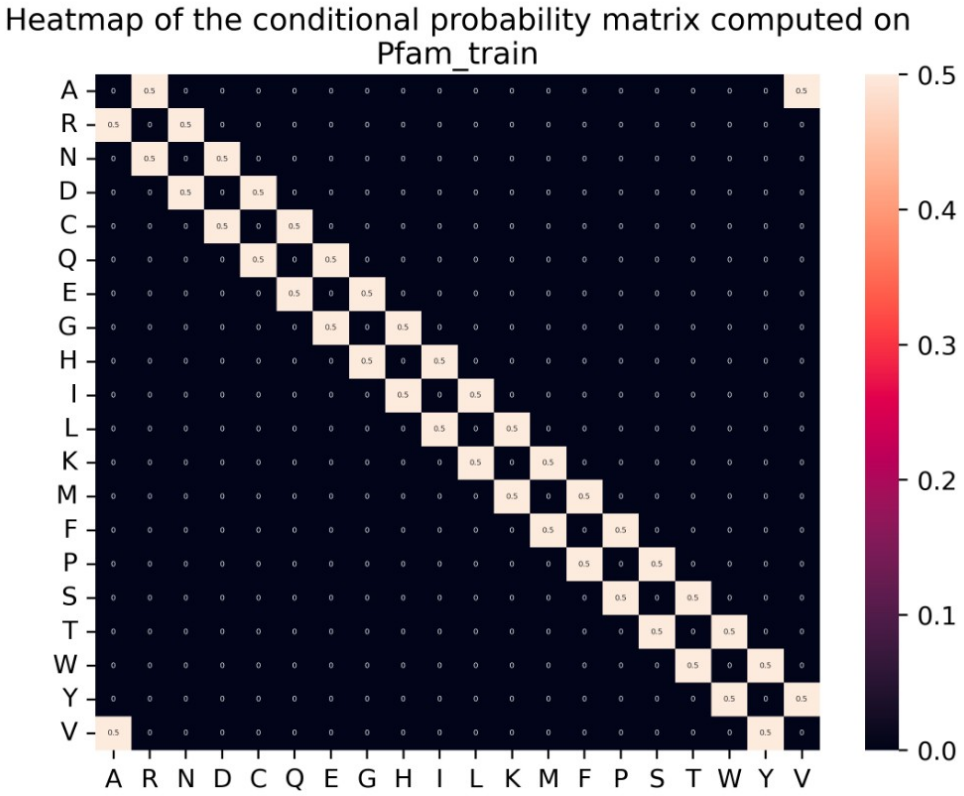


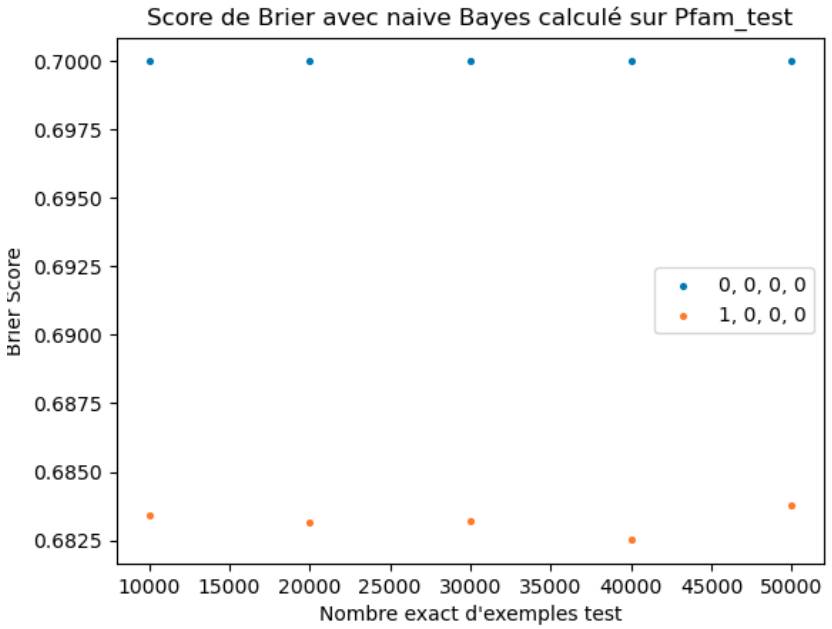


0-0- … -0-**0,4-0,2-0,4**-0-0 ...0

1 identité, 9 décalés
(id pas dans le train)

CAS 2

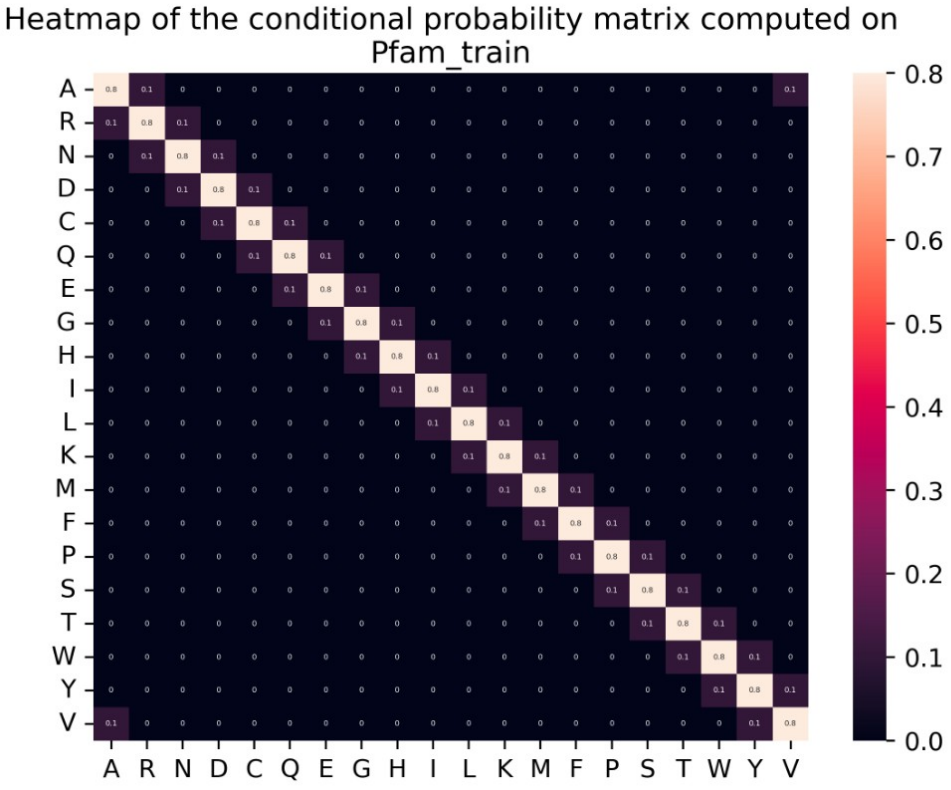




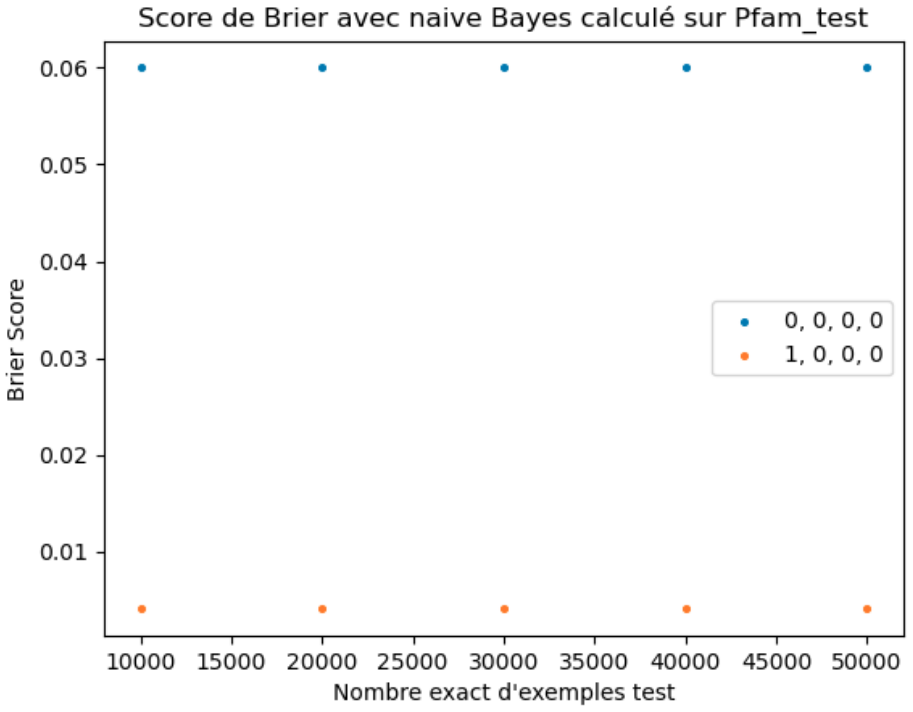
0-0- … -0-**0,5-0-0,5**-0-0 ...0

9 seeds identités, 1 seed décalé (cas décalé dans train)

CAS 3



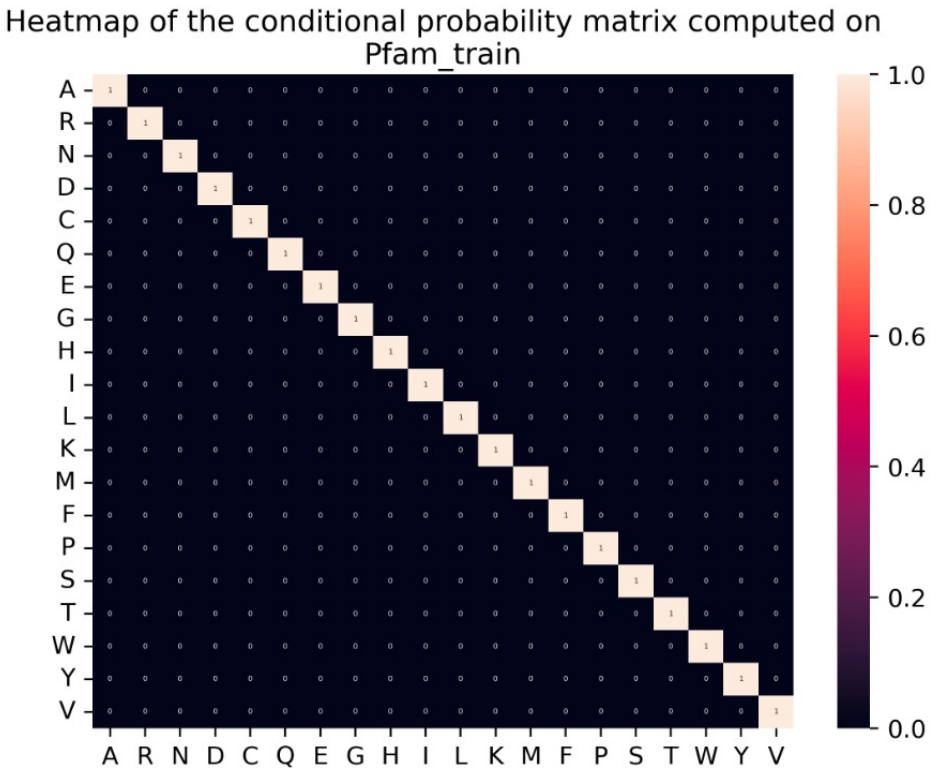
0-0- … -0-**0,1-0,8-0,1**-0-0 ...0



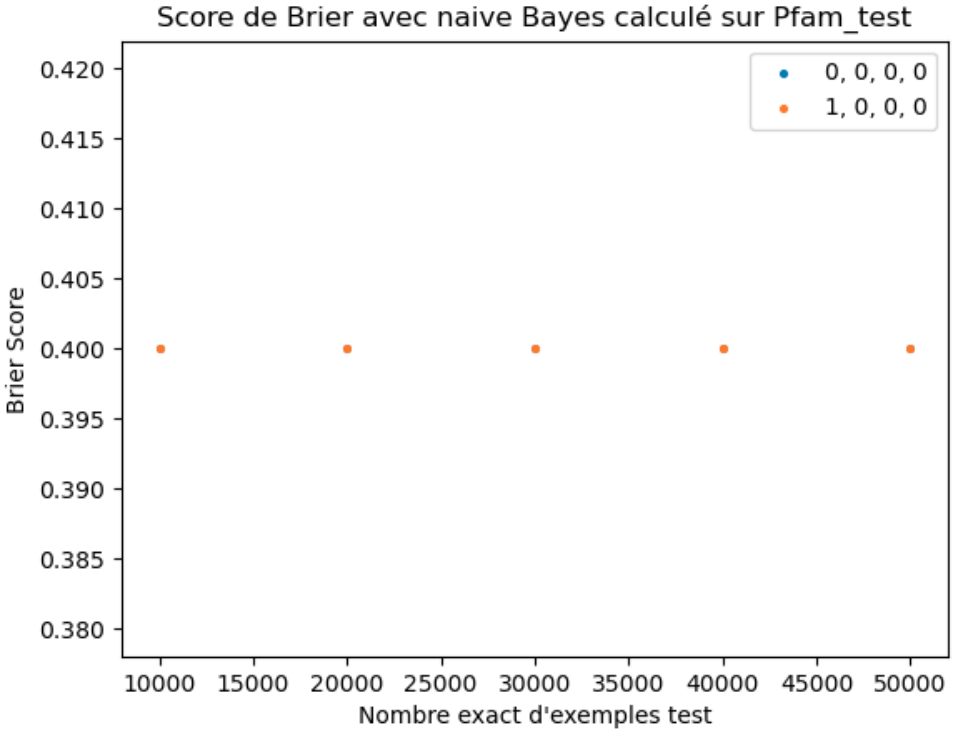
(0,0,0,0) : [0.05999999999999378, 0.05999999999996961, 0.05999999999996155, 0.059999999999957525, 0.05999999999995511]
(1,0,0,0) : [0.004068480393967478, 0.004067730587873158, 0.004066591993432518, 0.0040691885441681355, 0.004067813899663722]

9 seeds identités, 1 seed décalé (cas décalé pas dans train)

CAS 4



0-0- … 0-0-**1**-0-0 ...0

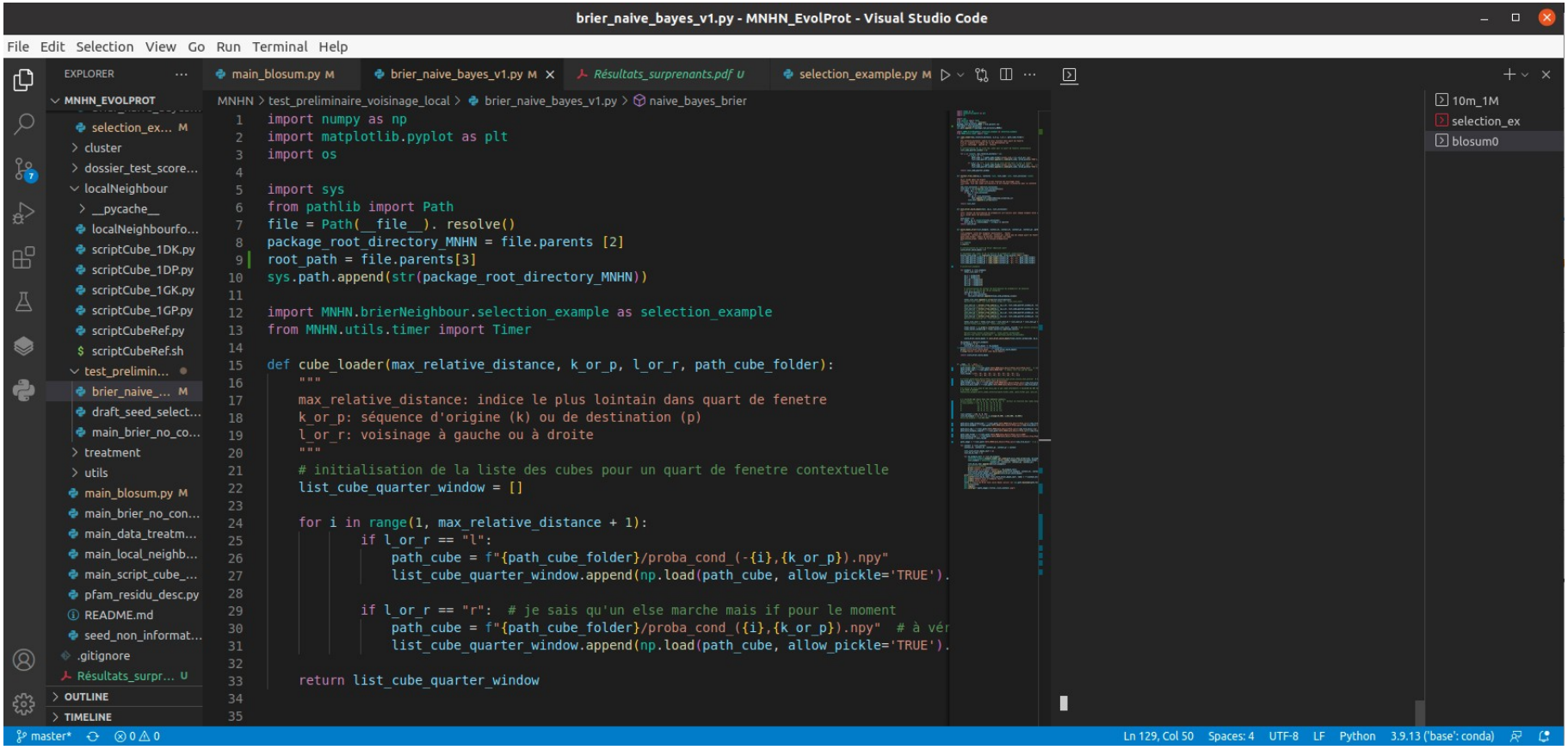


(0,0,0,0) : [0.4, 0.4, 0.4, 0.4, 0.4]

(1,0,0,0) : [0.4, 0.4, 0.4, 0.4, 0.4]

ANNEXES

Prob affichage/ veille



Temps clustering 99 % Pfam et Pfam en sortie de traitement

```
99.99490627546862, PF17407.5
---> time redundant: 2.03332 s
100.0, PF02713.17
---> time redundant: 0.02593 s
---> Compute and save non-redundant files: 59154.88105 s
nbre seed:                19 632.00
nbre seq:                 1 235 590.00
nbre position:            4 448 999.00
total character:          346 322 402.00
total character included: 192 394 396.00
mean len seq:             155.71
mean nbre seq:            62.94
---> Split data total in data A and data B: 2.01514 s
(base) pauline@abiboom:~/MNHN_EvolProt$
```