



Projet de Bio-informatique M1 IDIAG

ELABORE PAR :
ARFAOUI YOUSSEF
SECRAFI MOUNIR

A.U:2017/2018

Sommaire

Introduction	1
Processus général de la classification des protéines	2
Prétraitement	3
1. Extraction n-grammes.....	3
2. Construction du Tableau d'apprentissage (Booléen, Occurrence, Fréquence)	4
Traitement Datamining	7
➤ Data-Mining	7
➤ Classification	7
➤ Cross validation	8
Application des techniques de datamining (SVM, KNN, C4.5) avec Tanagra	9
➤ Tableau Booléen(2-grammes)	9
➤ Tableau Occurrence(2-grammes)	10
➤ Tableau Fréquences(2-grammes)	11
➤ Tableau Booléen(3-grammes)	12
➤ Tableau Fréquence(3-grammes)	13
➤ Tableau Occurrences(3-grammes).....	14
➤ Tableau Booléen([2-3]-grammes)	15
➤ Tableau fréquences ([2-3]-grammes)	15
➤ Tableau occurrences([2-3]-grammes)	16
➤ Tableau Booléen (4-grammes)	16
➤ Tableau Occurrences (4-grammes).....	17
➤ Tableau Fréquences(4-grammes)	18
➤ Tableau Booléen ([3-4]-grammes)	19
➤ Tableau Occurrences ([3-4]-grammes).....	20
➤ Tableau Fréquences([3-4]-grammes)	21
➤ Tableau Booléen(5-grammes)	22
➤ Tableau Occurrences (5-grammes).....	22
➤ Tableau Fréquences(5-grammes)	23
Sélection d'attributs	25
Conclusion	27

Tableau 1.Nombre de descripteurs en fonction de la longueur n de n-grammes	6
Tableau 2.Résultats détaillés pour l'algorithme KNN	24
Tableau 3.Résultats détaillés pour l'algorithme SVM	24
Tableau 4.Résultats détaillés pour l'algorithme C4.5	24

Introduction

La bio-informatique est un champ de recherche multi-disciplinaire où travaillent de concert biologistes, médecins, informaticiens, mathématiciens, physiciens et bio-informaticiens, dans le but de résoudre un problème scientifique posé par la biologie. Le spécialiste qui travaille à mi-chemin entre ces sciences et l'informatique est appelé bio-informaticien ou bionaute.

Depuis quelques années les progrès de l'informatique (et en particulier de la bio-informatique mise au service de la biodiversité ou « Biodiversity informatics » pour les anglophones) dopent la biologie évolutive en offrant aux chercheurs un accès à un nombre croissant de données sur la diversité et les variations des gènes, ainsi que des génomes, des organismes et de l'environnement en général.

Tout cela peut être relié à la phylogénie (de l'étude des populations à celle de clades entiers), via de nouveaux protocoles et réseaux dans le domaine de l'informatique de la biodiversité

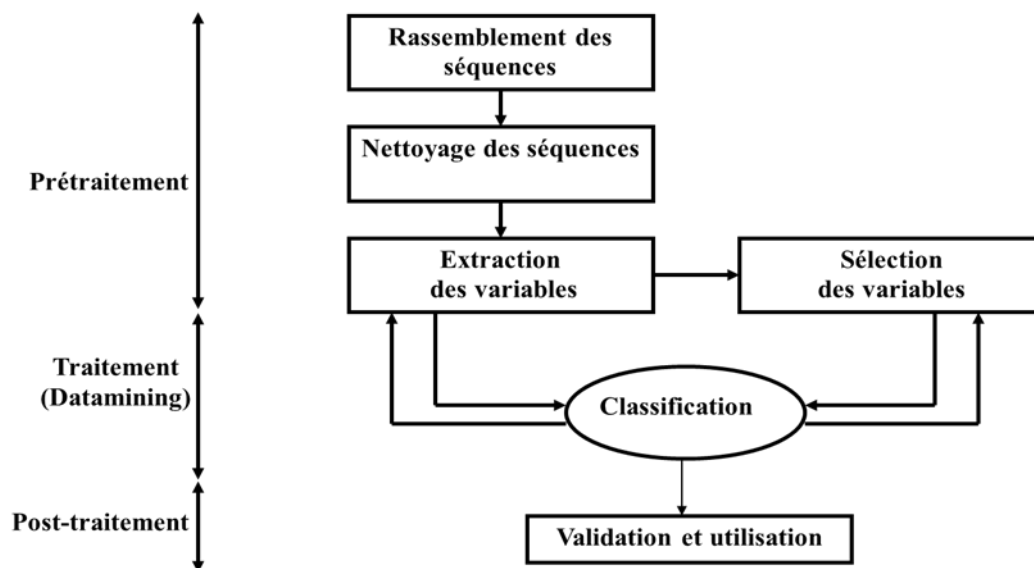
Le terme bio-informatique peut également décrire (par abus de langage) toutes les applications informatiques résultant de ces recherches Note 1. Plus généralement, la bio-informatique est l'application de la statistique et de l'informatique à la science biologique.

L'utilisation du terme bio-informatique est documentée pour la première fois en 1970 dans une publication¹ de Paulien Hogeweg et Ben Hesper (université d'Utrecht, Pays-Bas), en référence à l'étude des processus d'information dans les systèmes biotiques.

Cela va de l'analyse du génome à la modélisation de l'évolution d'une population animale dans un environnement donné, en passant par la modélisation moléculaire, l'analyse d'image, l'assemblage de génome et la reconstruction d'arbres phylogénétiques (phylogénie). Cette discipline constitue la « biologie in silico », par analogie avec in vitro ou in vivo.

Processus général de la classification des protéines

Le projet consiste à classer des séquences de protéines en se basant sur leurs structures primaires. La figure suivante explique les étapes du projet.



Processus général de la classification des protéines

Prétraitement

1. Extraction n-grammes

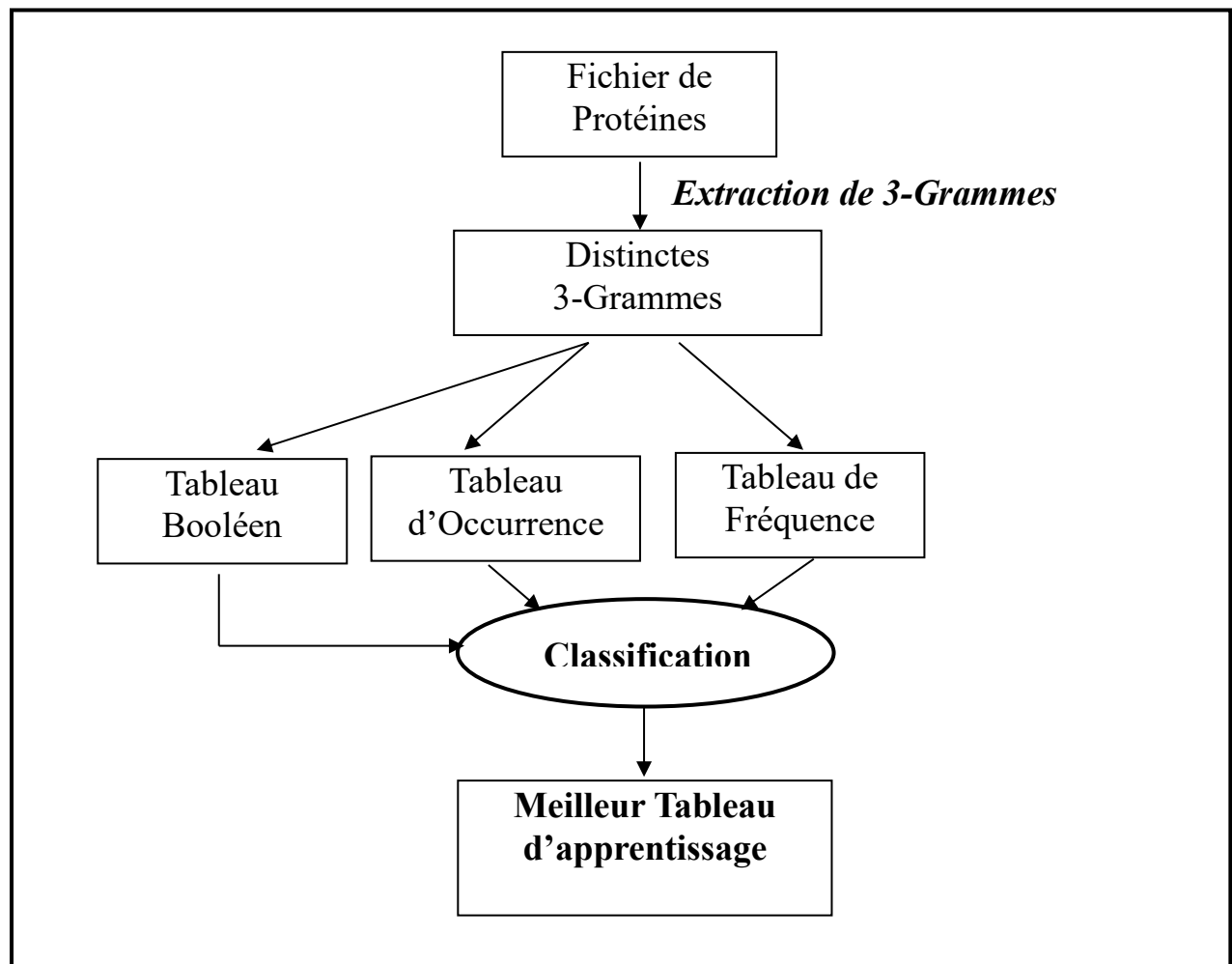
Un n-gramme est une sous-séquence de n éléments construite à partir d'une séquence donnée. L'idée semble provenir des travaux de Claude Shannon en théorie de l'information. Son idée était que, à partir d'une séquence de lettres donnée (par exemple « par exemple ») il est possible d'obtenir la fonction de vraisemblance de l'apparition de la lettre suivante. À partir d'un corpus d'apprentissage, il est facile de construire une distribution de probabilité pour la prochaine lettre avec un historique de taille n. Cette modélisation correspond en fait à un modèle de Markov d'ordre n où seules les n dernières observations sont utilisées pour la prédiction de la lettre suivante. Ainsi un bi gramme est un modèle de Markov d'ordre 2.

Exemple (K=3)

ACAAGATGCCATTGTCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC

ACA	CAA	AAG	AGA	GAT	ATG	TGC	GCC	CCA	CAT	ATT	TTG	TGT	GTC	TCC	CCC	CCG	CGG	GGC	CCT
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

2. Construction du Tableau d'apprentissage (Booléen, Occurrence, Fréquence)



- **Fichier protéine**

```

2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
10
```

Exemple N-grammes=3

- Tableau Boolean (Ligne= Ensemble de sequences, Colonne =variable(3-grammes))

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1		MLD	LDN	DNT	NTR	TRL	RLR	LRI	RIA	IAI	AIQ	IQK	QKS	KSG	SGR	GRL	RLS	LSD	SDD	DDS	DSR
2	seq1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	seq2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	seq3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
5	seq4	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	1	0	0
6	seq5	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0
7	seq6	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
8	seq7	1	0	0	0	1	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0
9	seq8	1	0	0	0	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
10	seq9	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1	0	0
11	seq10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
12	seq11	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
13	seq12	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
14	seq13	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
15	seq14	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
16	seq15	1	0	0	0	0	1	1	1	0	0	0	1	1	1	1	1	0	0	0	0
17	seq16	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
18	seq17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	seq18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	seq19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	seq20	1	0	0	0	0	1	1	1	0	0	0	1	1	1	1	1	0	0	0	0
22	seq21	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0
23	seq22	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0

- Tableau Occurrence

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1		MLD	LDN	DNT	NTR	TRL	RLR	LRI	RIA	IAI	AIQ	IQK	QKS	KSG	SGR	GRL	RLS	LSD	SDD	DDS	DSR
2	seq1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1
3	seq2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	seq3	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
5	seq4	0	0	0	0	0	1	0	2	0	0	0	0	0	0	1	2	1	0	0	0
6	seq5	0	0	0	0	0	1	0	2	0	0	0	0	0	0	1	1	0	0	0	0
7	seq6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	seq7	1	0	0	0	1	1	1	2	0	0	0	1	1	1	1	2	1	0	0	0
9	seq8	1	0	0	0	1	1	1	2	0	0	0	1	1	1	1	2	1	1	1	1
10	seq9	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	1	0	0	0
11	seq10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
12	seq11	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
13	seq12	1	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0
14	seq13	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
15	seq14	1	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0
16	seq15	1	0	0	0	0	1	1	2	0	0	0	1	1	1	1	1	0	0	0	0
17	seq16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
18	seq17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	seq18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	seq19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	seq20	1	0	0	0	0	1	1	2	0	0	0	1	1	1	1	2	0	0	0	0
22	seq21	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0
23	seq22	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0

➤ Tableau fréquence

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1		MLD	LDN	DNT	NTR	TRL	RLR	LRI	RIA	IAI	AIQ	IQK	QKS	KSG	SGR	GRL	RLS	LSD	SDD	DDS	DSR
2	seq1	0,003367	0,003367	0,003367	0,003367	0,003367	0,003367	0,003367	0,006734	0,003367	0,003367	0,003367	0,003367	0,003367	0,003367	0,003367	0,006734	0,003367	0,003367	0,003367	0,003367
3	seq2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	seq3	0	0	0	0	0	0,002618	0	0	0	0	0	0	0	0,002618	0	0	0	0	0	0
5	seq4	0	0	0	0	0	0,003436	0	0,006873	0	0	0	0	0	0	0,003436	0,006873	0,003436	0	0	0
6	seq5	0	0	0	0	0	0,003436	0	0,006873	0	0	0	0	0	0	0,003436	0,003436	0	0	0	0
7	seq6	0	0	0	0	0	0,00463	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	seq7	0,003367	0	0	0	0,003367	0,003367	0,003367	0,006734	0	0	0	0,003367	0,003367	0,003367	0,003367	0,006734	0,003367	0	0	0
9	seq8	0,003367	0	0	0	0,003367	0,003367	0,003367	0,006734	0	0	0	0,003367	0,003367	0,003367	0,003367	0,006734	0,003367	0,003367	0,003367	0,003367
10	seq9	0	0	0	0	0	0	0,002591	0	0	0	0	0	0	0	0,002591	0,002591	0	0,002591	0	0
11	seq10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,004425	0	0	0	0	0
12	seq11	0	0	0	0	0,003497	0	0	0	0	0	0	0	0,003497	0	0	0	0	0	0	0
13	seq12	0,003584	0	0	0	0	0	0	0,007168	0,003584	0	0	0	0	0	0	0	0	0	0	0
14	seq13	0	0,003096	0	0	0	0	0	0	0	0	0	0	0	0	0,003096	0	0	0	0	0
15	seq14	0,003584	0	0	0	0	0	0	0,007168	0,007168	0	0	0	0	0	0	0	0	0	0	0
16	seq15	0,003322	0	0	0	0	0,003322	0,003322	0,006645	0	0	0	0,003322	0,003322	0,003322	0,003322	0,003322	0	0	0	0
17	seq16	0	0	0	0	0	0	0	0,004902	0	0	0	0	0	0	0	0	0	0	0	0
18	seq17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	seq18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	seq19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	seq20	0,003367	0	0	0	0	0,003367	0,003367	0,006734	0	0	0	0,003367	0,003367	0,003367	0,003367	0,006734	0	0	0	0
22	seq21	0	0	0	0	0	0,003367	0,003367	0,003367	0	0	0	0	0	0	0,003367	0,003367	0	0	0	0
23	seq22	0,003546	0	0	0	0	0	0	0,003546	0,003546	0	0	0	0	0	0,003546	0	0	0	0	0

Protéines pairs	2- grammes	3- grammes	2-3- grammes	4-grammes	[3-4]-grammes	5-grammes
F1_2	400	6600	7000	23408	30008	28958

Tableau 1.Nombre de descripteurs en fonction de la longueur n de n-grammes

Traitement Datamining

➤ Data-Mining

Terme récent (1995) représentant un mélange d'idées et d'outils provenant de la Statistique, Science de l'information et l'Informatique.

- A évolué vers le data science
- Machine Learning,
- Big Data (explosion des données),
- Formalismes de stockage et de traitement distribués de données (NoSQL, NewSQL, Hadoop, MapReduce ...).

➤ Classification

La classification est la tâche la plus commune de la fouille de données qui semble être une tâche humaine primordiale.

Afin de comprendre notre vie quotidienne, nous sommes constamment obligés à classer, catégoriser et évaluer.

La classification consiste à étudier les caractéristiques d'un nouvel objet pour l'attribuer à une classe prédéfinie.

Le fonctionnement de la classification se décompose en deux phases :

➤ La première étant la phase d'apprentissage :

Dans cette phase, les approches de classification utilisent un jeu d'apprentissage dans lequel tous les objets sont déjà associés aux classes de références connues.

L'algorithme de classification apprend du jeu d'apprentissage et construit un modèle.

➤ La seconde phase est la phase de classification :

Proprement dite, dans laquelle le modèle appris est employé pour classer de nouveaux objets.

➤ **Cross validation**

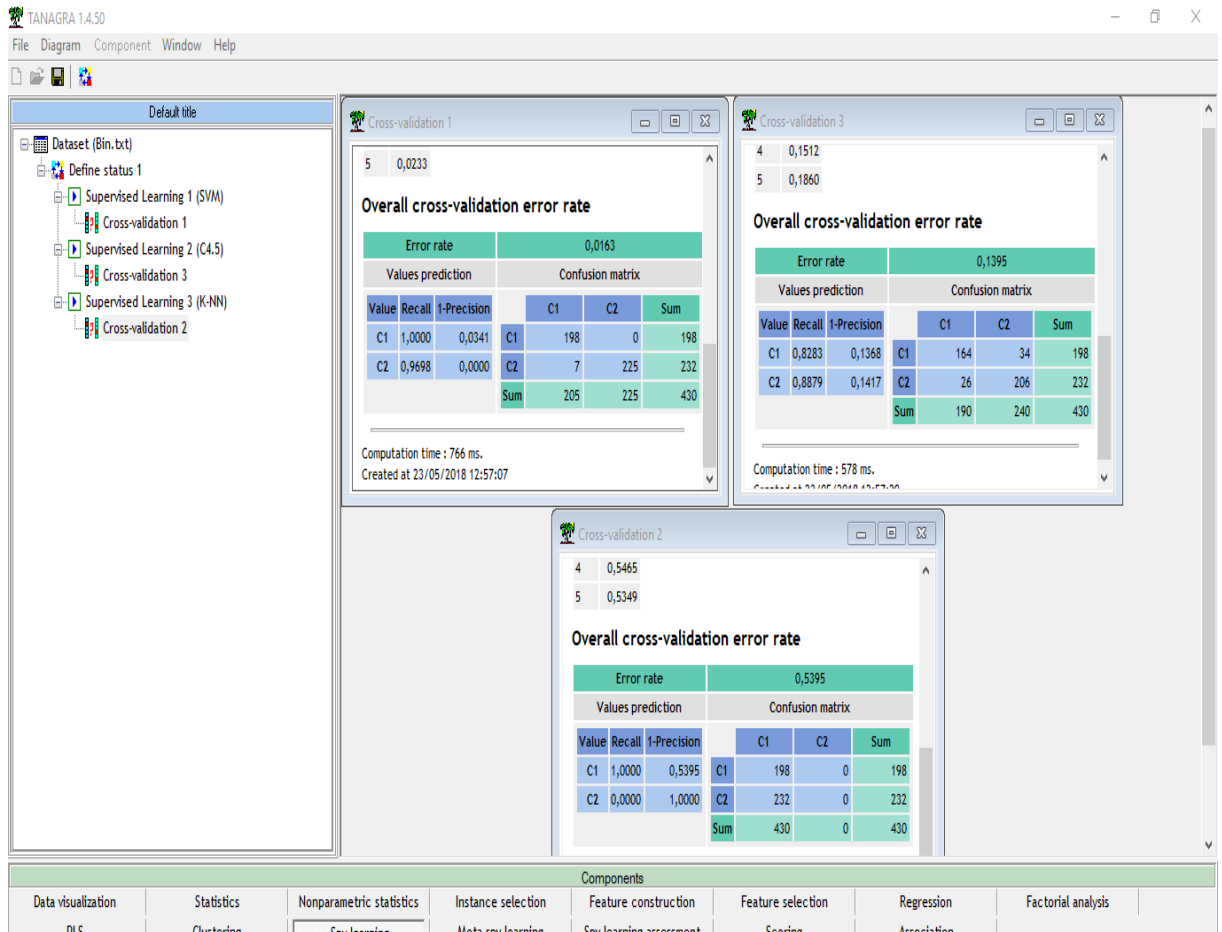
Dans l'apprentissage supervisé, il est généralement accepté de ne pas utiliser le même échantillon pour construire un modèle prédictif et estimer son taux d'erreur. L'erreur obtenue dans ces conditions - appelée taux d'erreur de resubstitution - est (très souvent) trop optimiste, laissant croire que le modèle présentera une excellente performance de prédiction.

Une approche typique consiste à diviser les données en deux parties (approche de retenue): un premier échantillon, le dit échantillon de train est utilisé pour construire le modèle; un second échantillon, dit échantillon d'essai, est utilisé pour mesurer sa performance. Le taux d'erreur mesuré reflète honnêtement le comportement du modèle en généralisation. Malheureusement, sur de petits ensembles de données, cette approche est problématique. En réduisant la quantité de données présentées à l'algorithme d'apprentissage, nous ne pouvons pas apprendre correctement la relation sous-jacente entre les descripteurs et l'attribut de classe. Dans le même temps, la partie consacrée aux tests reste limitée, l'erreur mesurée a une variance élevée.

Nous calculons le taux d'erreur en utilisant un ressemblant méthode. Nous avons choisi une validation croisée 5 X 2 que nous répéter un grand nombre de fois. Cette approche est privilégiée car cela permet d'obtenir des biais et des résultats relativement stables. Dans notre contexte, de fort risque de sur apprentissage, nous cherchons à produire des résultats généraux et à éviter être trop dépendant de notre ensemble de données.

Application des techniques de datamining (SVM, KNN, C4.5) avec Tanagra

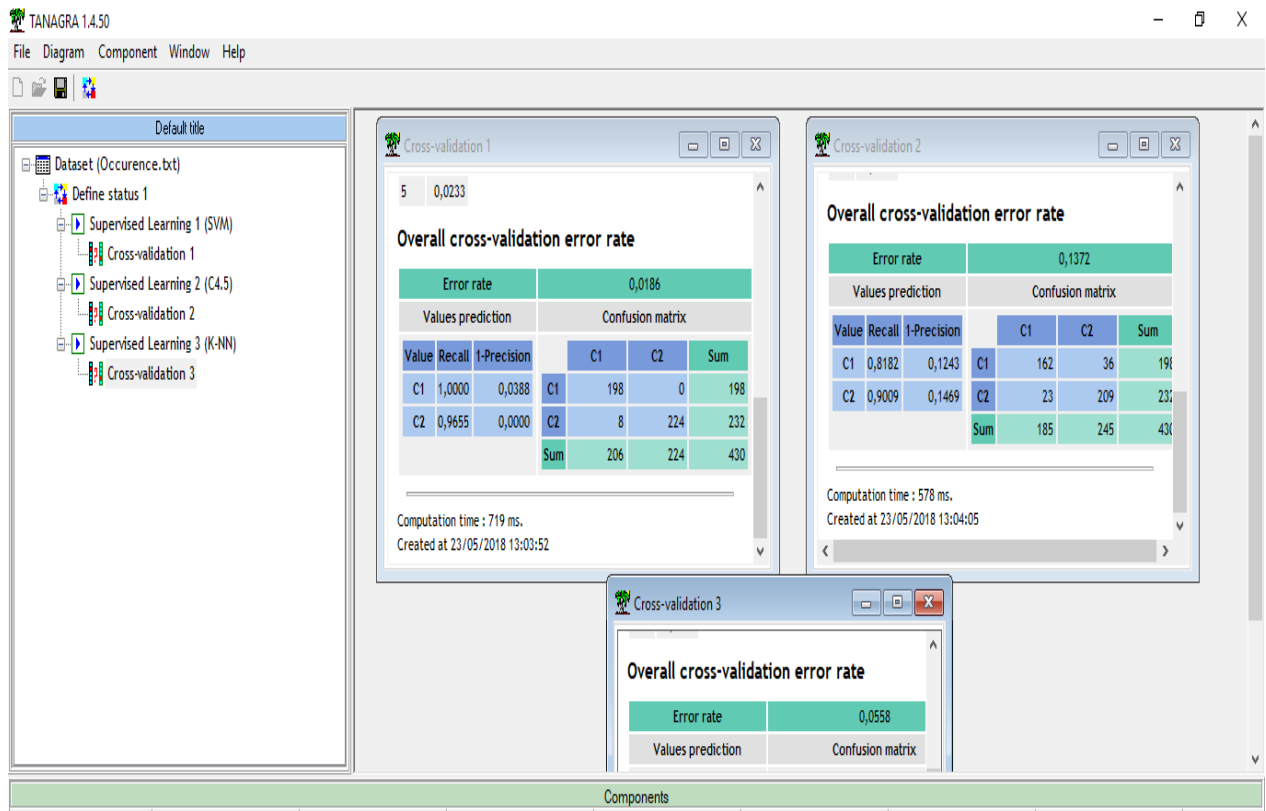
➤ Tableau Booléen(2-grammes)



Algorithme de classification Taux d'erreur

<i>Knn</i>	0.539
<i>C4.5</i>	0.139
SVM	0.016

➤ Tableau Occurrence(2-grammes)



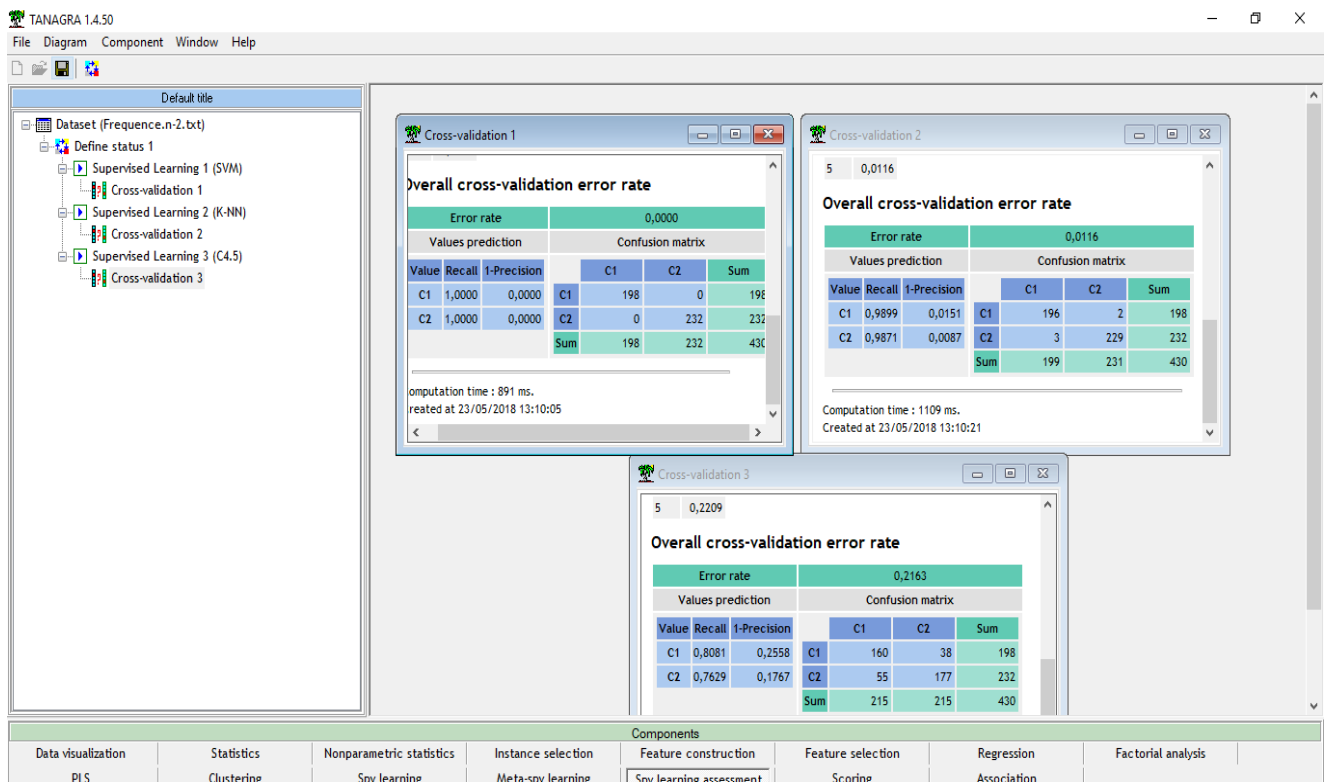
Algorithme de classification

Knn 0.0558

C4.5 0.137

SVM **0.018**

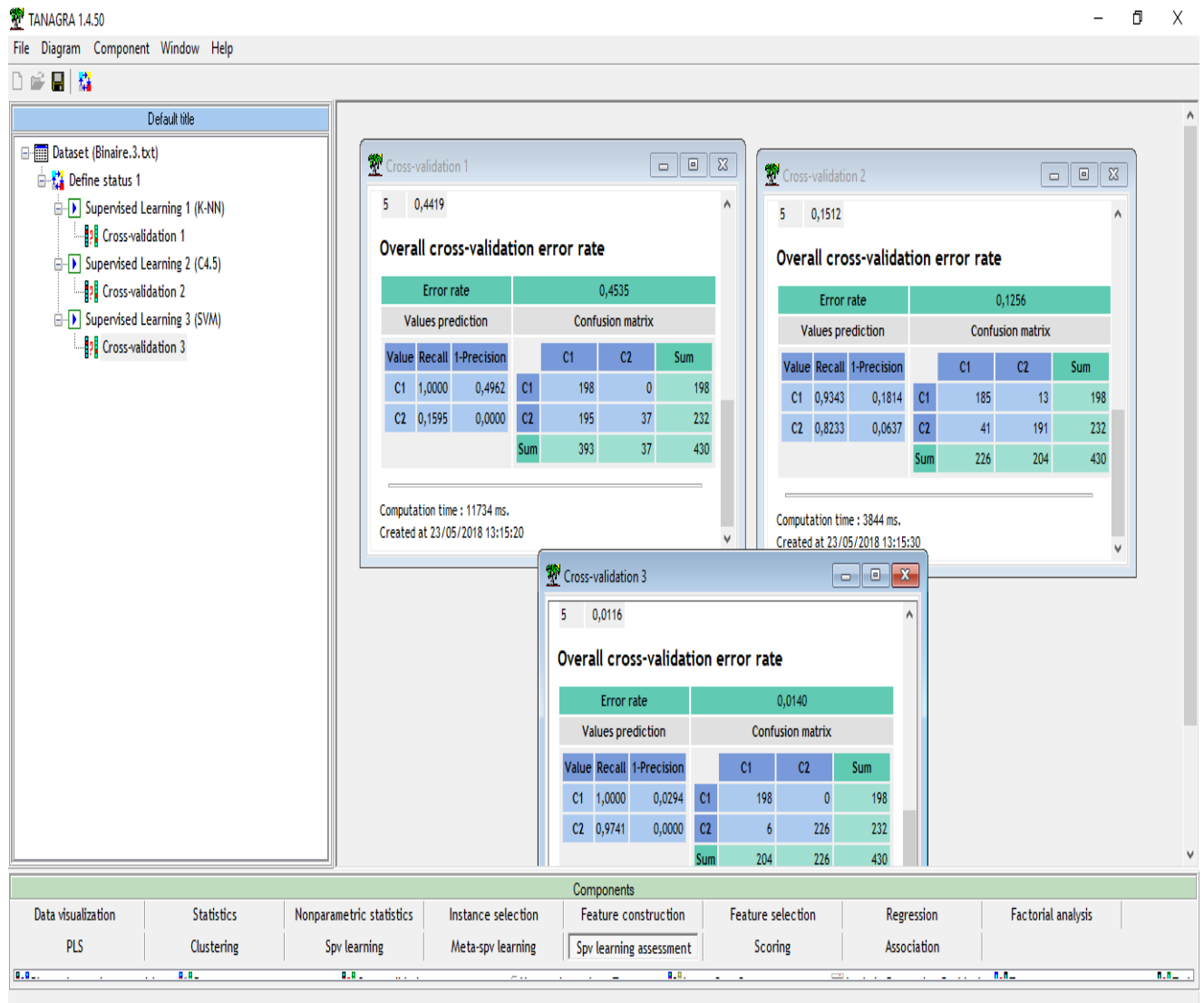
➤ Tableau Fréquences(2-grammes)



Algorithme de classification

Algorithme de classification	Taux d'erreur
<i>Knn</i>	0.0116
<i>C4.5</i>	0.2163
SVM	0.00

➤ Tableau Booléen(3-grammes)

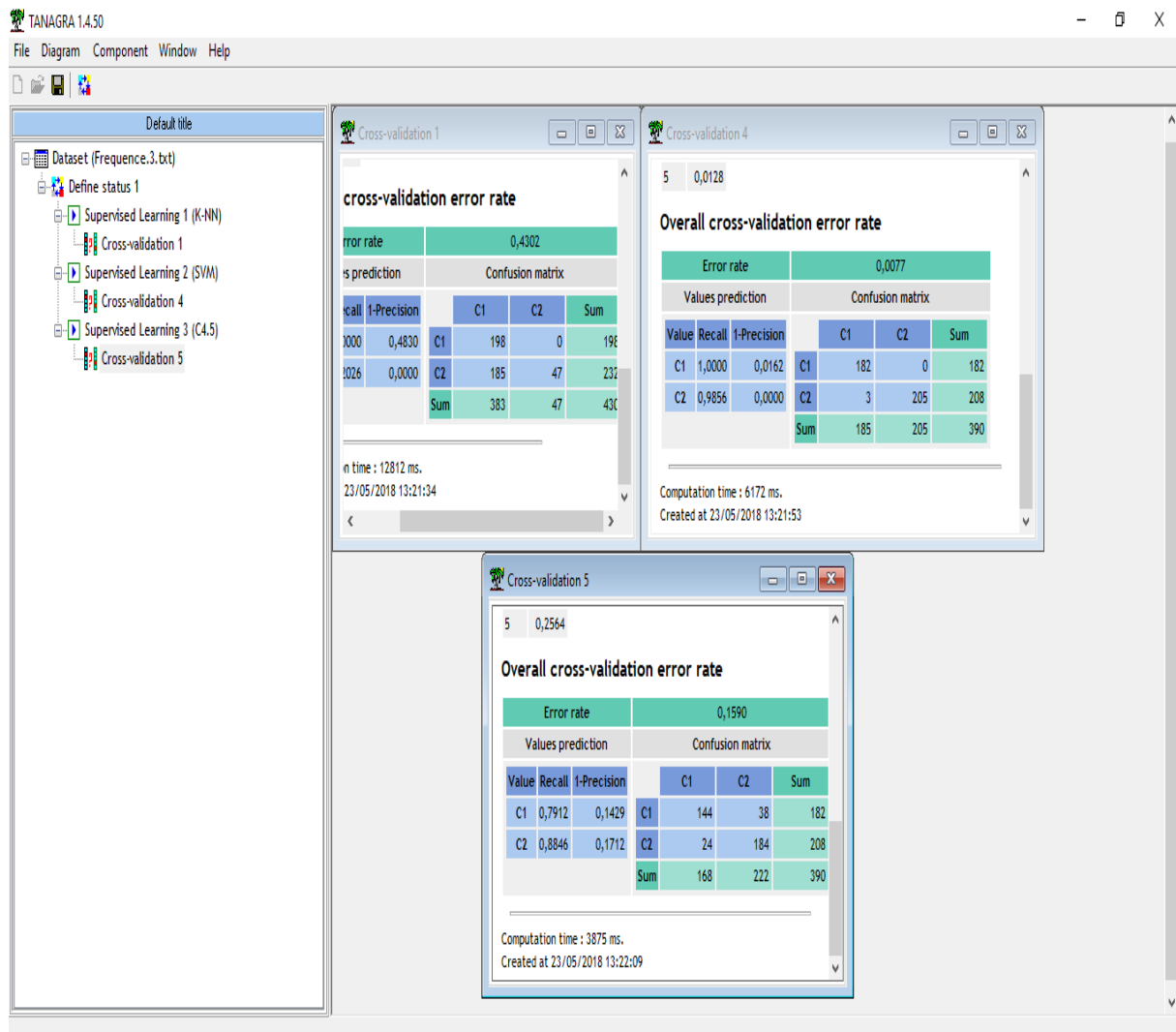


Algorithme de classification

Taux d'erreur

<i>Knn</i>	0.453
<i>C4.5</i>	0.125
SVM	0.014

➤ Tableau Fréquence(3-grammes)



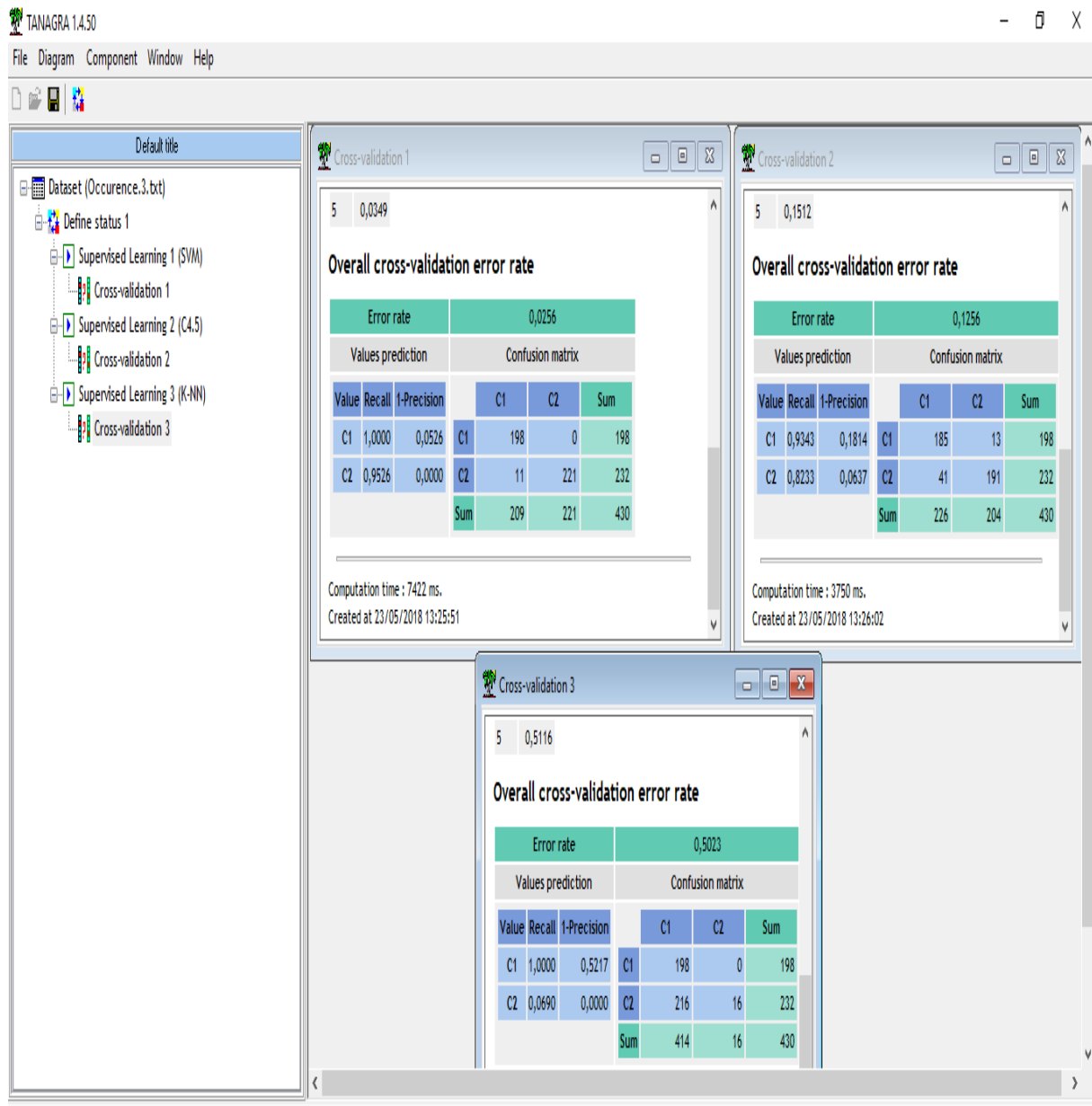
Algorithme de classification *Taux d'erreur*

Knn 0.430

C4.5 0.159

SVM **0.007**

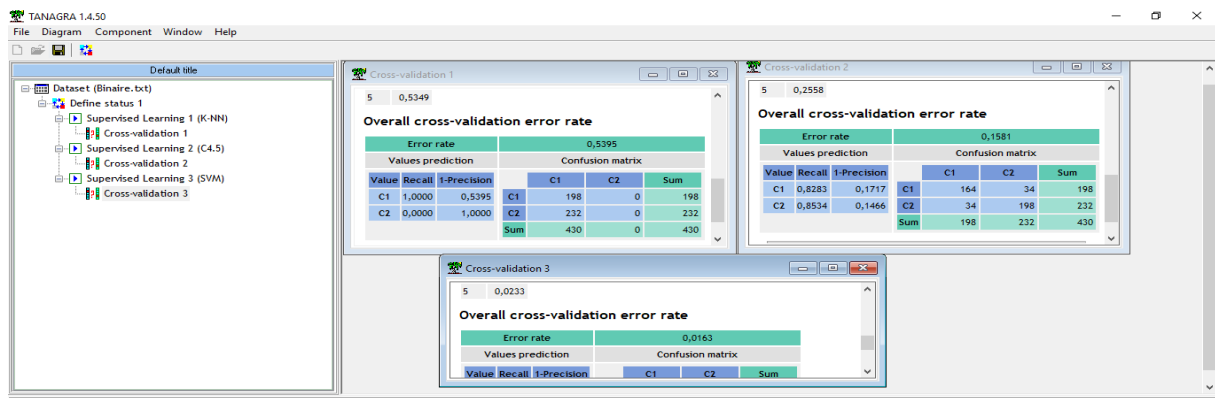
➤ Tableau Occurrences(3-grammes)



Algorithme de classification Taux d'erreur

<i>Knn</i>	0.5023
<i>C4.5</i>	0.1256
SVM	0.0256

➤ Tableau Booléen([2-3]-grammes)



Algorithme de classification Taux d'erreur

Knn 0.539

C4.5 0.158

SVM 0.016

➤ Tableau fréquences ([2-3]-grammes)



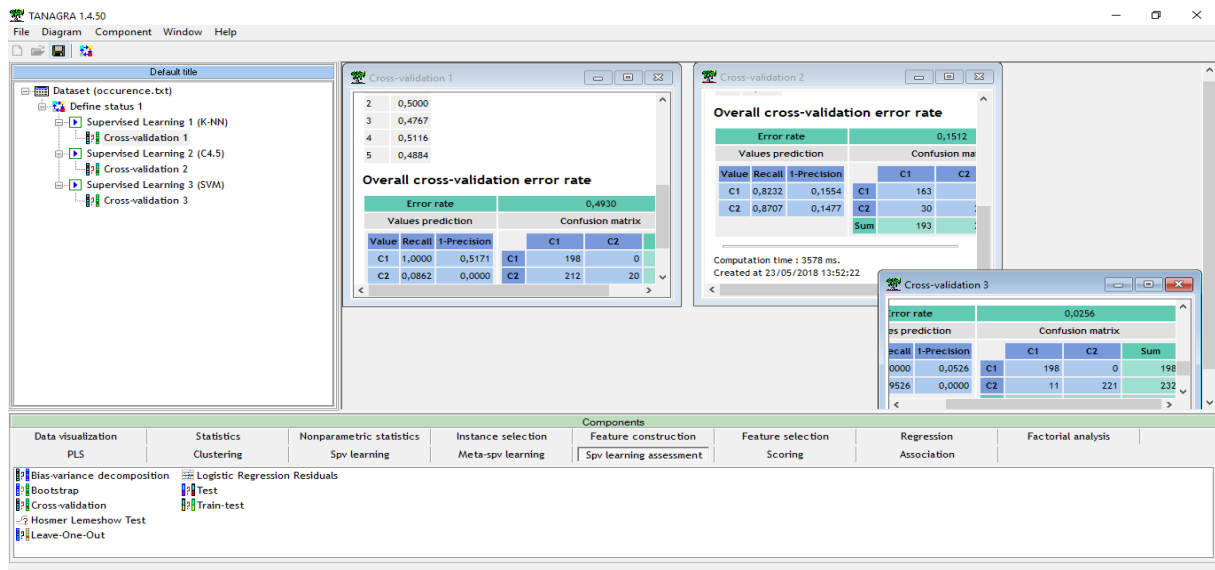
Algorithme de classification Taux d'erreur

Knn 0.369

C4.5 0.204

SVM 0.007

➤ Tableau occurrences([2-3]-grammes)



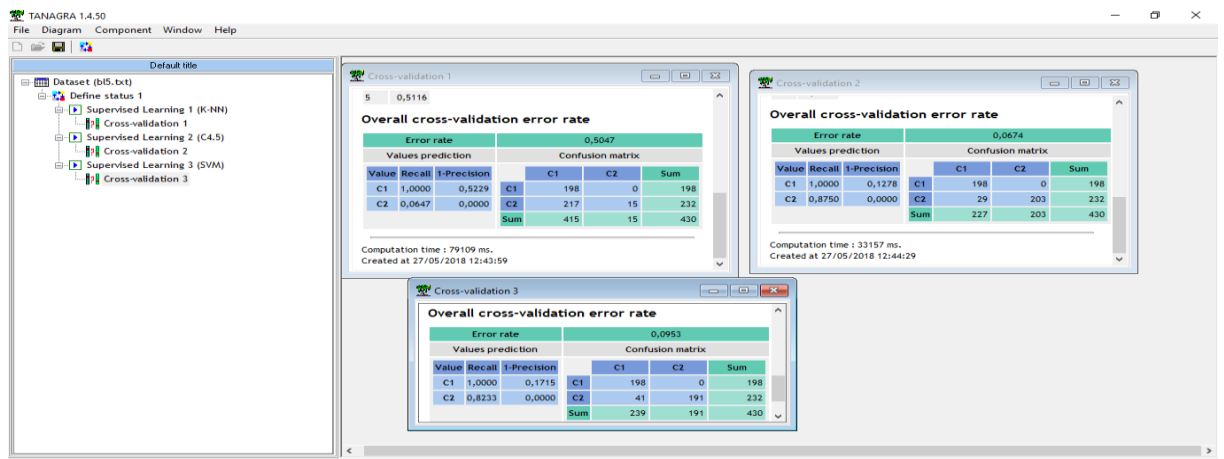
Algorithme de classification Taux d'erreur

Knn 0.493

C4.5 0.151

SVM 0.025

➤ Tableau Booléen (4-grammes)



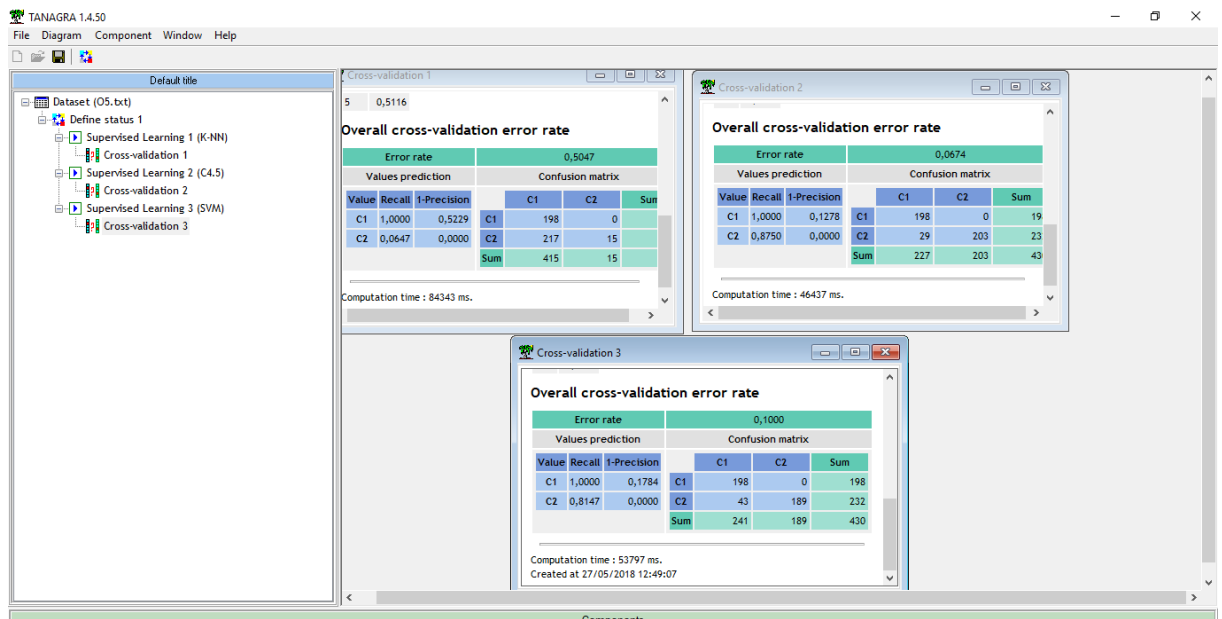
Algorithme de classification Taux d'erreur

Knn 0.504

C4.5 0.064

SVM 0.095

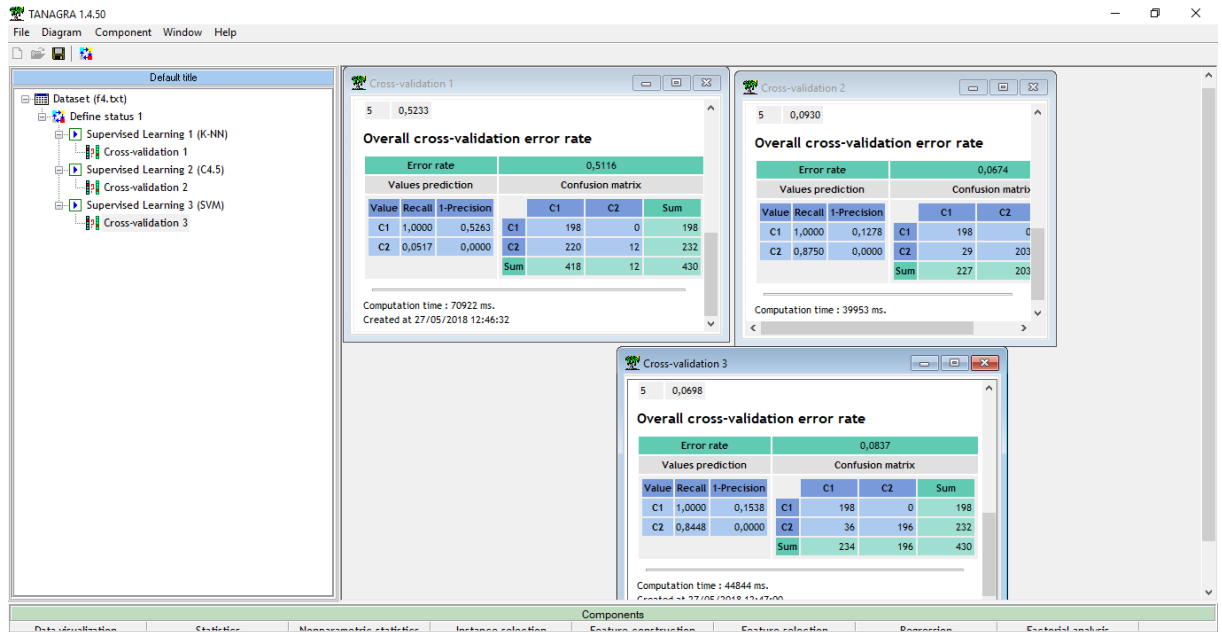
➤ Tableau Occurrences (4-grammes)



Algorithme de classification Taux d'erreur

Knn	0.504
C4.5	0.067
SVM	0.100

➤ Tableau Fréquences(4-grammes)



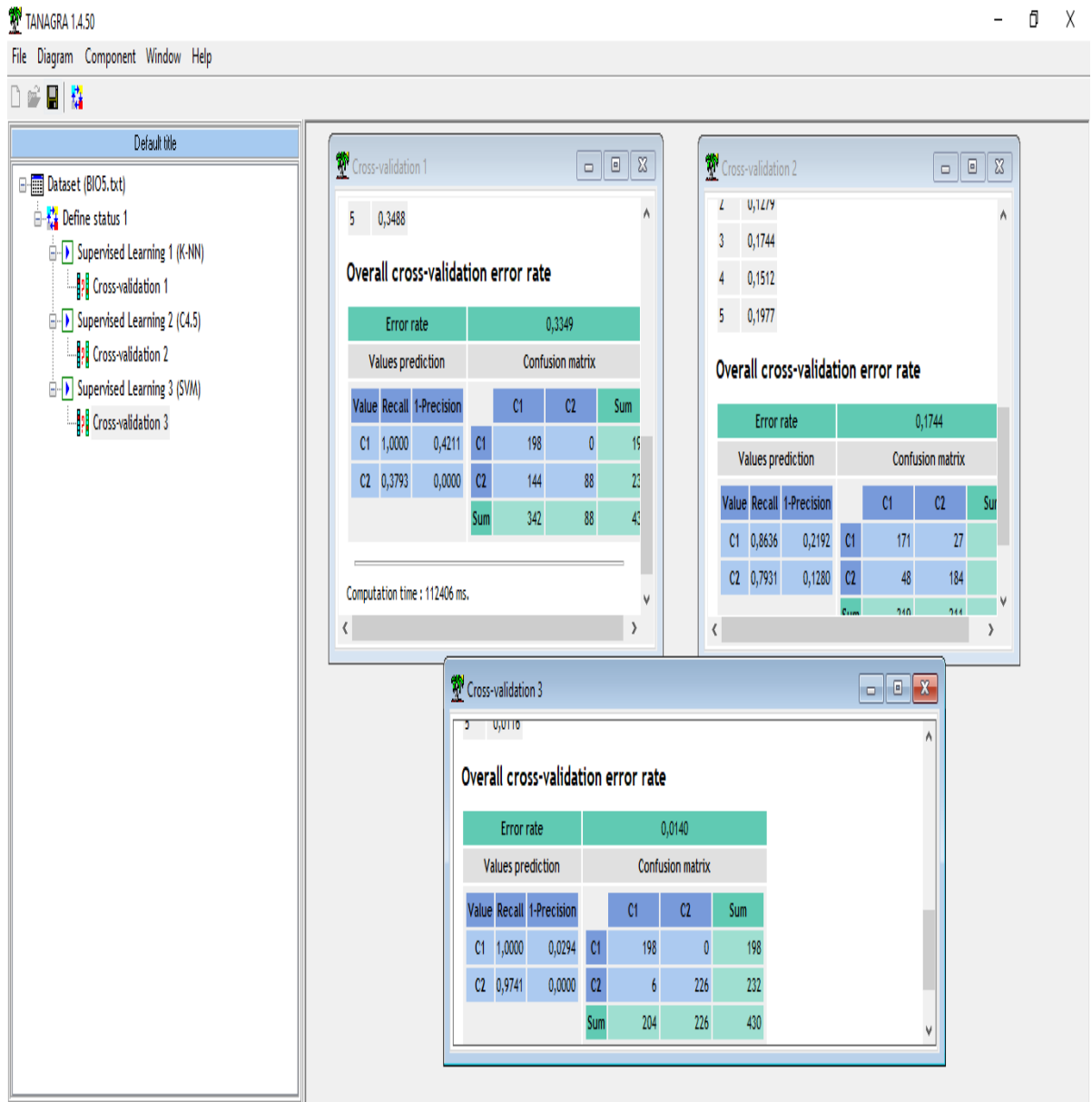
Algorithme de classification Taux d'erreur

Knn 0.511

C4.5 0.067

SVM 0.083

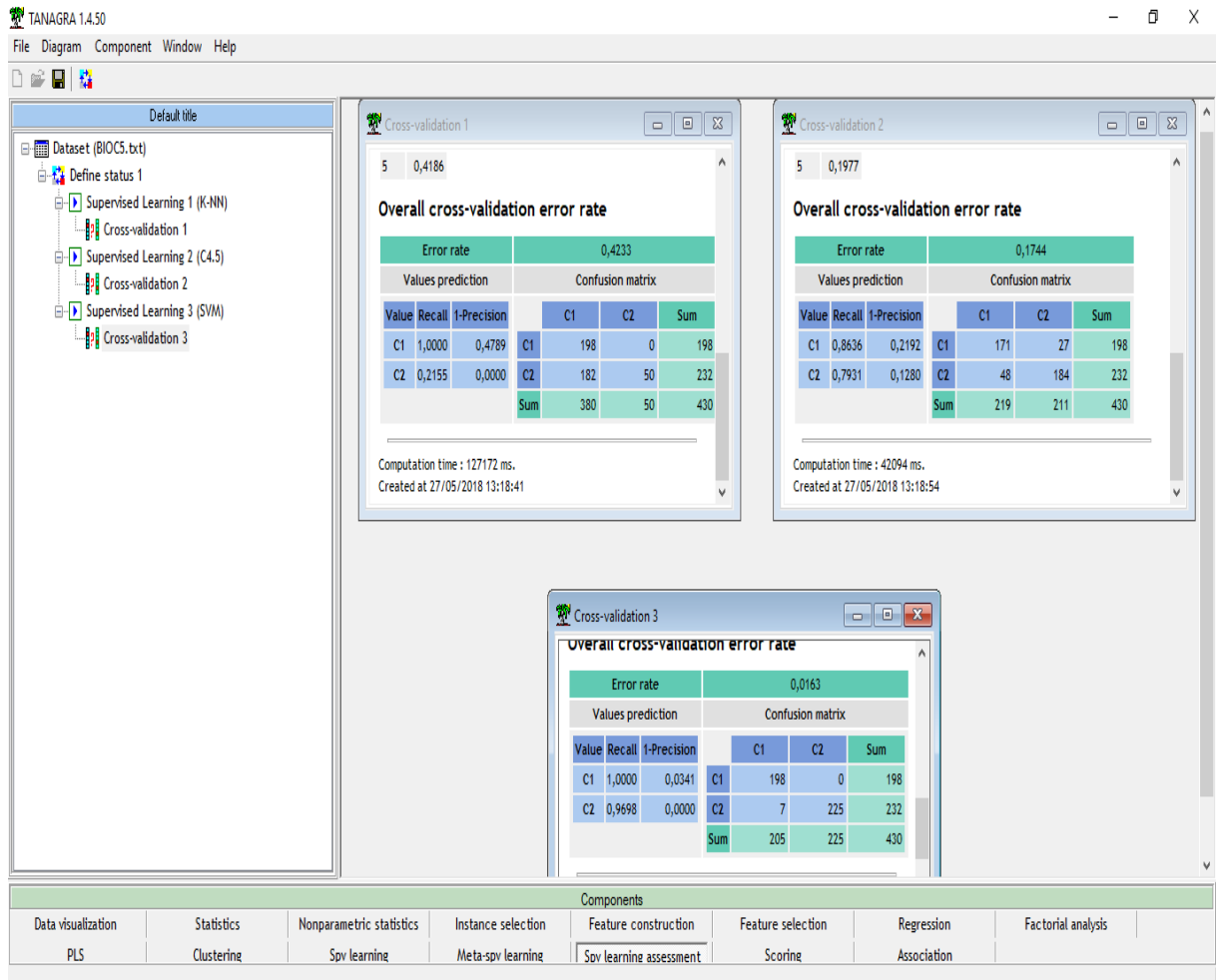
➤ Tableau Booléen ([3-4]-grammes)



Algorithme de classification Taux d'erreur

<i>Knn</i>	0.334
<i>C4.5</i>	0.174
SVM	0.014

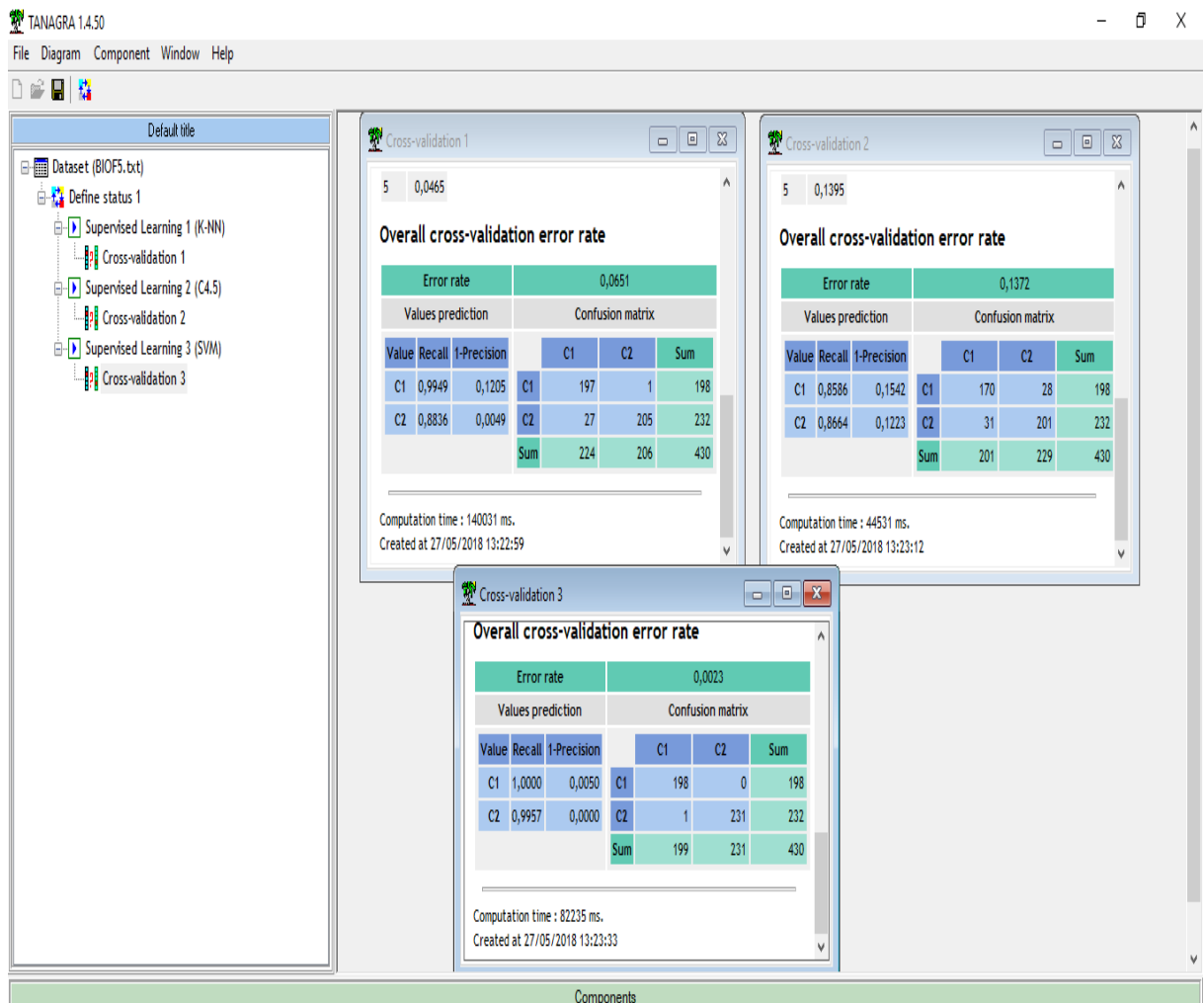
➤ Tableau Occurrences ([3-4]-grammes)



Algorithme de classification *Taux d'erreur*

<i>Knn</i>	0.423
C4.5	0.174
<i>SVM</i>	0.0163

➤ Tableau Fréquences([3-4]-grammes)



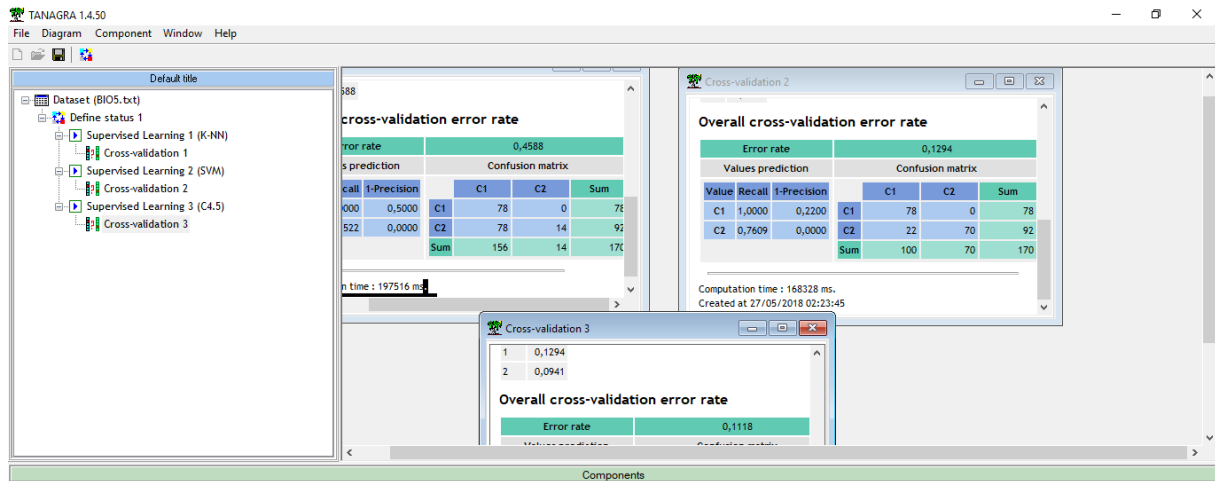
Algorithme de classification *Taux d'erreur*

Knn 0.065

C4.5 0.137

SVM **0.002**

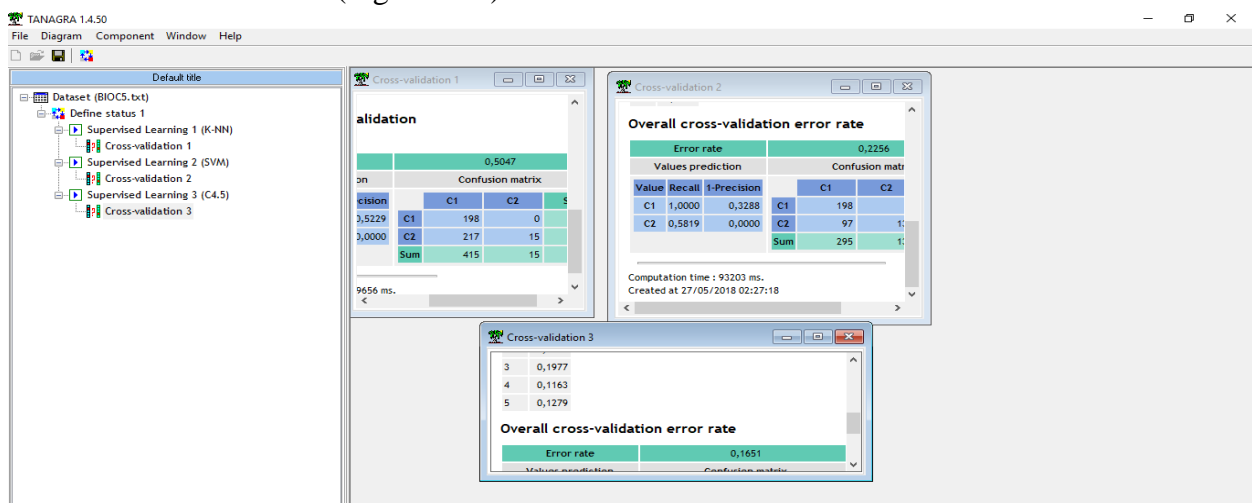
➤ Tableau Booléen(5-grammes)



Algorithme de classification Taux d'erreur

<i>Knn</i>	0.458
C4.5	0.111
<i>SVM</i>	0.129

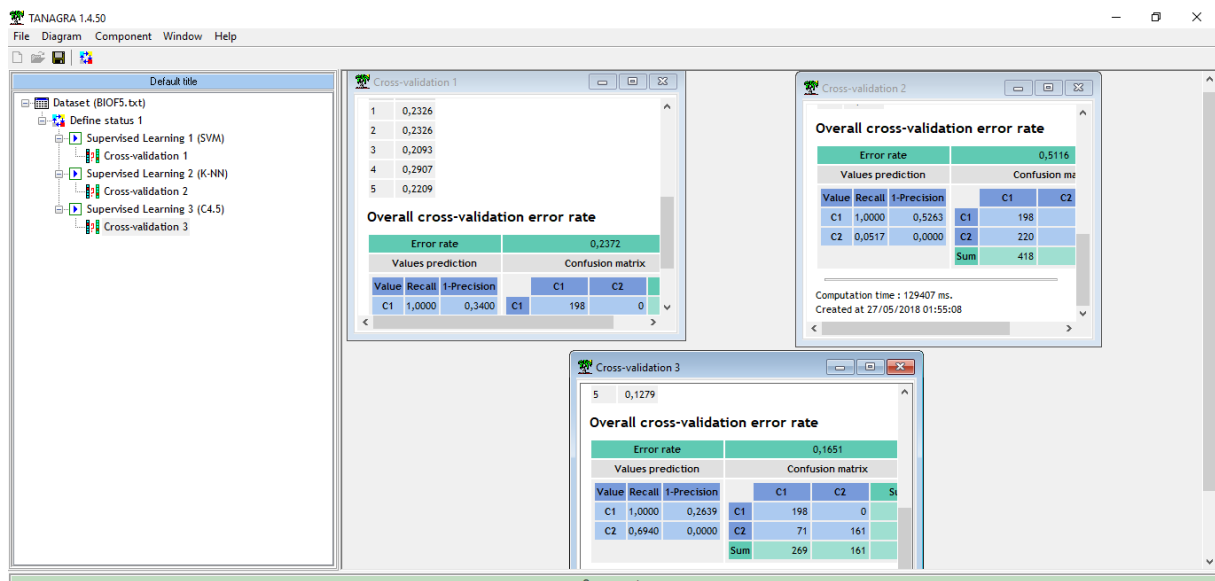
➤ Tableau Occurrences (5-grammes)



Algorithme de classification *Taux d'erreur*

<i>Knn</i>	0.504
C4.5	0.165
<i>SVM</i>	0.225

➤ Tableau Fréquences(5-grammes)



Algorithme de classification *Taux d'erreur*

<i>Knn</i>	0.511
<i>C4.5</i>	0.165
SVM	0.237

<i>Protéines pairs</i>	<i>[2]-grams</i>	<i>[3]-grams</i>	<i>[2-3]- grams</i>	<i>[4]- grammes</i>	<i>[3-4]- grammes</i>	<i>[5]- grammes</i>
<i>Occurrence</i>	0.055	0.502	0.539	0.504	0.423	0.504
<i>Booléen</i>	0.539	0.453	0.369	0.504	0.334	0.458
<i>Fréquences</i>	0.011	0.430	0.369	0.511	0.065	0.511

Tableau 2.Résultats détaillés pour l'algorithme KNN

<i>Protéines pairs</i>	<i>[2]-grams</i>	<i>[3]-grams</i>	<i>[2-3]- grams</i>	<i>[4]- grammes</i>	<i>[3-4]- grammes</i>	<i>[5]- grammes</i>
<i>Occurrence</i>	0.018	0.025	0.025	0.100	0.016	0.225
<i>Booléen</i>	0.016	0.014	0.016	0.095	0.014	0.129
<i>Fréquences</i>	0.0	0.007	0.007	0.083	0.002	0.237

Tableau 3.Résultats détaillés pour l'algorithme SVM

➔ **Meilleure résultat 2 et [3-4]-grammes Tableau Fréquences.**

⇒ **Meilleure Algorithme de classification SVM**

<i>Protéines pairs</i>	<i>[2]- grammes</i>	<i>[3]- grammes</i>	<i>[2-3]- grammes</i>	<i>[4]- grammes</i>	<i>[3-4]- grammes</i>	<i>[5]- grammes</i>
<i>Occurrence</i>	0.137	0.125	0.151	0.067	0.174	0.165
<i>Booléen</i>	0.139	0.125	0.158	0.064	0.174	0.111
<i>Fréquences</i>	0.216	0.159	0.204	0.067	0.137	0.165

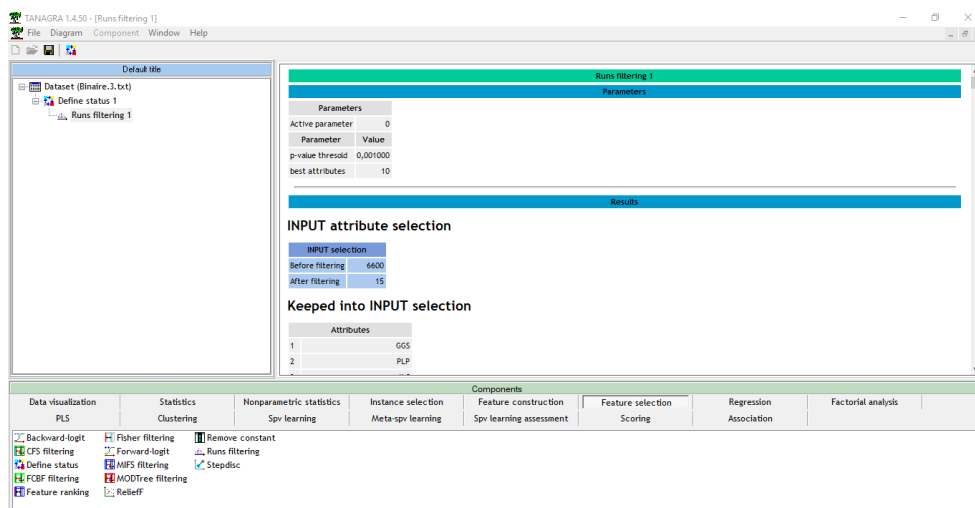
Tableau 4.Résultats détaillés pour l'algorithme C4.5

⇒ **Meilleure Algorithme de classification SVM et C4.5**

Sélection d'attributs

La sélection d'attributs consiste à réduire l'ensemble des attributs considérés et peut augmenter la précision d'un algorithme de regroupement/classification, améliorer la qualité des données, réduire le temps de calcul et/ou l'espace mémoire, en réduisant la dimension de l'espace des attributs par l'élimination des attributs redondants, non pertinents ou bruités.

Sélection d'attributs Avec Tanagra



N-grammes = 3

Fisher filtering

Nombre d'attributs avant sélection= 6600

Nombre d'attributs après sélection= 251

Logiciels et langages utilisés

Langage R

R est un langage de programmation et un logiciel libre dédié aux statistiques et à la Science des données soutenu par la R Foundation for Statistical Computing. R fait partie De la liste des paquets GNU et est écrit en C (langage), Fortran et R.

Le langage R est largement utilisé par les statisticiens et les data miner pour le Développement de logiciels statistiques et l'analyse des données.

Logiciel Tanagra :

TANAGRA est un logiciel gratuit de DATA MINING destiné à l'enseignement et à la recherche. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données.

TANAGRA est un projet ouvert au sens qu'il est possible à tout chercheur d'accéder au code et d'ajouter ses propres algorithmes pour peu qu'il respecte la licence de distribution du logiciel.

L'objectif principal du projet TANAGRA est d'offrir aux chercheurs et aux étudiants une **plate-forme de Data Mining** facile d'accès, respectant les standards des logiciels du domaine, notamment en matière d'interface et de mode de fonctionnement, et permettant de **mener des études** sur des données réelles et/ou synthétiques.

Conclusion

Le travail réalisé dans ce document vise à présenter un processus optimal de classification des protéines.

Le processus tirer le maximum d'avantages, d'une part, en identifiant le meilleur couple de n-grammes et Classificateur. D'un autre côté, les processus de classification doit être réalisé dans un délai raisonnable. Par conséquent, nous avons varié la taille de n-grammes entre {2,3,2-3,3-4 ,4 ,5}, puis, évalué chaque type de n-grammes avec divers classificateurs, tels que KNN, Arbre de décision et SVM. Les résultats montrent que le couple optimal n'est pas le même habituellement. Mais le SVM et les classificateurs d'arbre de décision donnent les meilleurs résultats Respectivement avec 2 grammes ,3 grammes, [3-4]-grammes.