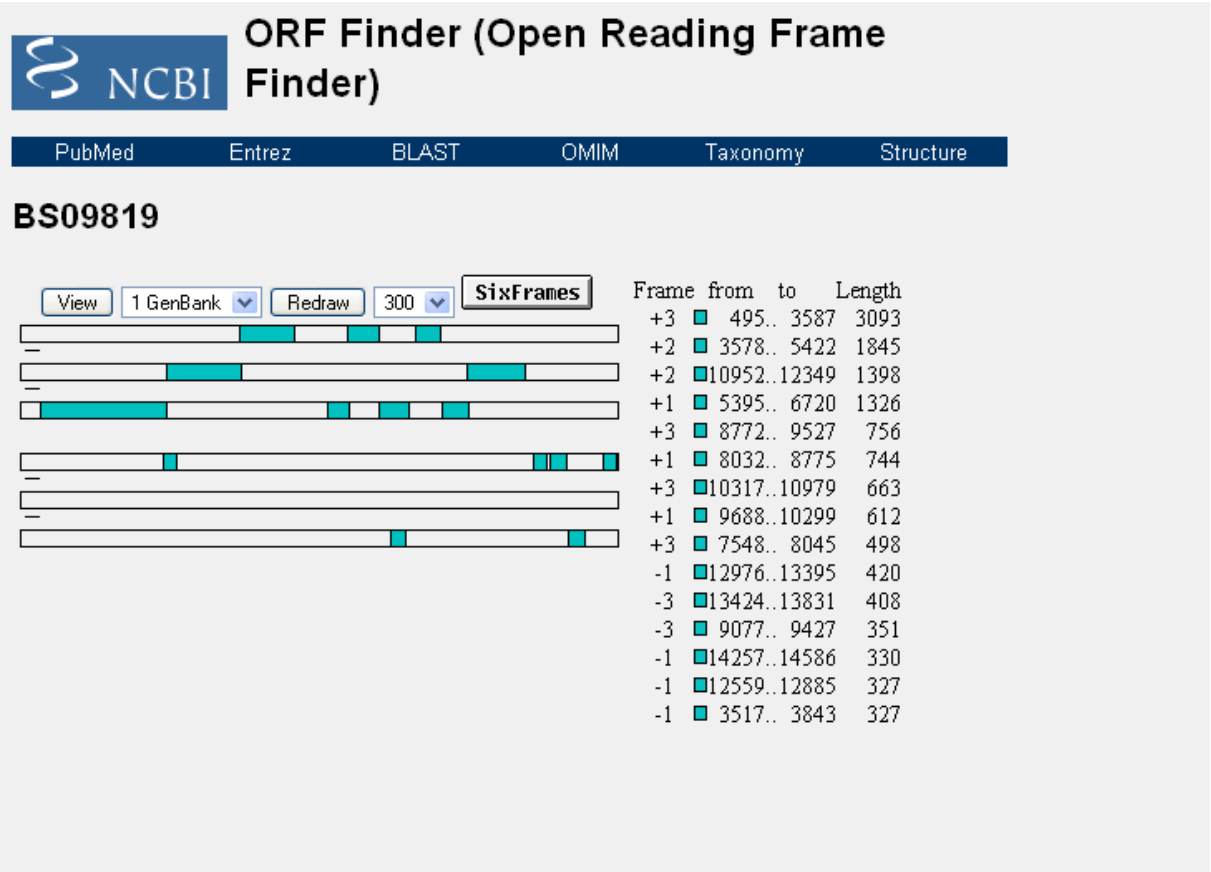
Résultat de la prédiction avec ORFfinder (choix de la taille minimum des ORFs 300pb alors que par défaut 100 pb donc modifier la taille et faire redraw)



Problème d'ORFfinder : codon start ATG simplement alors que chez les bactéries d'autres codons start (GTG et TTG). Problème aussi quand l'ORF est petite. Est-elle vraiment codante ou pas. Donc bien que simple il vaut mieux utiliser des méthodes statistiques.

Donc pour les grandes ORF, le problème va être de trouver le bon codon start et pour les petites de vérifier qu'elles ont les caractéristiques du codant.

Une méthode largement utilisée : GeneMark.

Genemark analyse la séquence en déplaçant une fenêtre de taille choisie avec un pas lui aussi choisi. Deux sorties sont proposées : uns sortie graphique qui pour chaque fenêtre indique la valeur de la probabilité d'être codée par le cadre correspondant et une sortie textuelle sur lequel un filtre a été appliqué. Si on choisit un seuil de 0.5, seules les régions (entre un codon start et un stop) pour lesquelles la moyenne des probabilités d'être codantes des fenêtres dépassant la valeur seuille de 0.5 seront proposées. On peut avoir des fois des problèmes pour choisir un start car plusieurs possibilités dans une petite région et même proba pour le start dans GenMark (bonne ou mauvaise). Dans ce cas on garde plusieurs choix en espérant trancher après en recherchant la position du Shine-Dalgarno

(Scan_for_matches). Par contre, il faut s'aider du graphique car le start ne peut pas être localisé après que la courbe de la séquence apparaissent comme codante.

Résultats sur les ORF brin direct seulement (ne pas traiter le complémentaire)

ORF	ORFfinder	GM 0.5	GM 0.4	SD (bon start)
ORF1	495	1191	495/528	478-497 -> 495
ORF2	3578	3578	3578	Pas trouvé
ORF3	5395	5395/5434	5395/5434	5380-5397 -> 5395
ORF4	Pas trouvé	Pas trouvé	Pas trouvé	6782-6797 -> 6795
ORF5	7548	7548/7602	7548/7602	7347-7550 -> 7548
ORF6	8032	8032/8059	8032/8059	8047-8061 -> 8059
ORF7	8772	Pas trouvé	Zone d'intérêt 8751-9527	8758-8774 -> 8772
ORF8	9688	Zone d'intérêt 9505-10299	Zone d'intérêt 9505-10299	9515-9531 -> 9529
ORF9	10317	Pas trouvé	10317	10305-10319 -> 10317
ORF10	10952	10952/10970	10952/10970	10954-10972 -> 10970 *
ORF11	Pas trouvé	13974/13983	13974/13983	13969-13985 -> 13983

*Trouvé en baissant le seuil pour la matrice à 30

Zone d'intérêt ORF stop à stop dans laquelle on a détecté une zone codante mais avec le seuil choisi par de start à proposer.

Leur faire faire GenMark.hmm après avant SD (aucune erreur et même ORF4 trouvé).

Prédiction GenMark seuil 0.5

```
Sequence: Thu Nov 15 09:01:00 EST 200
Sequence length: 14616
GC Content:  37.49%
Window length: 120
Window step: 12
Threshold value: 0.500
Matrix: Bacillus subtilis, Thu Oct 27 15:58:26 2005
Matrix author: Dr. Borodovsky Laboratory, School of Biology, Georgia Tech
Matrix order: 5
```

List of Open reading frames predicted as CDSs, shown with alternate starts

(regions from start to stop codon w/ coding function >0.50)

Left end	Right end	DNA Strand	Coding Frame	Avg Prob	Start Prob
1	276	direct	fr 1	0.68	....
28	276	direct	fr 1	0.68	....
112	276	direct	fr 1	0.58	....
1191	3587	direct	fr 3	0.51	0.32
2397	3587	direct	fr 3	0.51	0.00
2589	3587	direct	fr 3	0.59	0.97
2619	3587	direct	fr 3	0.57	0.02
3578	5422	direct	fr 2	0.60	0.98
3623	5422	direct	fr 2	0.61	0.02
3632	5422	direct	fr 2	0.62	0.00
3764	5422	direct	fr 2	0.65	0.04
3914	5422	direct	fr 2	0.71	0.31
4037	5422	direct	fr 2	0.72	0.00
4052	5422	direct	fr 2	0.72	0.01
4082	5422	direct	fr 2	0.71	0.10
5395	6720	direct	fr 1	0.70	0.91
5434	6720	direct	fr 1	0.72	0.96
5443	6720	direct	fr 1	0.72	0.88
5473	6720	direct	fr 1	0.73	0.20
5485	6720	direct	fr 1	0.73	0.01
5578	6720	direct	fr 1	0.72	0.79
7548	8045	direct	fr 3	0.63	0.00
7602	8045	direct	fr 3	0.71	0.19
7668	8045	direct	fr 3	0.81	0.69
7710	8045	direct	fr 3	0.80	0.15
7746	8045	direct	fr 3	0.77	0.04
8032	8775	direct	fr 1	0.70	0.00
8059	8775	direct	fr 1	0.72	0.10
8344	8775	direct	fr 1	0.78	0.66
8395	8775	direct	fr 1	0.75	0.17
8407	8775	direct	fr 1	0.75	0.04
10952	12349	direct	fr 2	0.51	0.59
10970	12349	direct	fr 2	0.52	0.63
11024	12349	direct	fr 2	0.53	0.00
11027	12349	direct	fr 2	0.53	0.00

11090	12349	direct	fr 2	0.56	0.72
11186	12349	direct	fr 2	0.53	0.00
11201	12349	direct	fr 2	0.53	0.00
11606	12349	direct	fr 2	0.61	0.27
11627	12349	direct	fr 2	0.62	0.03
12559	12906	complement	fr 3	0.63	0.15
12559	12894	complement	fr 3	0.66	0.19
12559	12885	complement	fr 3	0.68	0.16
12559	12870	complement	fr 3	0.70	0.66
12559	12699	complement	fr 3	0.54	0.01
12559	12696	complement	fr 3	0.50	0.01
12976	13395	complement	fr 3	0.75	0.58
12976	13341	complement	fr 3	0.80	0.25
12976	13326	complement	fr 3	0.79	0.01
12976	13206	complement	fr 3	0.69	0.51
13424	13852	complement	fr 1	0.75	0.96
13424	13840	complement	fr 1	0.78	0.97
13424	13831	complement	fr 1	0.80	1.00
13424	13642	complement	fr 1	0.72	0.03
13424	13615	complement	fr 1	0.68	0.01
13974	14216	direct	fr 3	0.53	0.13
13983	14216	direct	fr 3	0.56	0.26
14031	14216	direct	fr 3	0.64	0.84
14257	14586	complement	fr 3	0.74	....
14257	14427	complement	fr 3	0.56	0.10
14257	14529	complement	fr 3	0.71	....

List of Regions of interest

(regions from stop to stop codon w/ a signal in between)

LEnd	REnd	Strand	Frame
1	276	direct	fr 1
486	3587	direct	fr 3
3569	5422	direct	fr 2
5392	6720	direct	fr 1
7542	8045	direct	fr 3
8026	8775	direct	fr 1
9077	9499	complement	fr 1
9505	10299	direct	fr 1
10302	10979	direct	fr 3
10943	12349	direct	fr 2
12268	12543	complement	fr 3
12559	12978	complement	fr 3
12976	13416	complement	fr 3
13424	13861	complement	fr 1
13947	14216	direct	fr 3
14257	14616	complement	fr 3

Prédiction GenMark seuil 0.4

```
Sequence: Thu Nov 15 08:58:34 EST 2007
Sequence length: 14616
GC Content:  37.49%
Window length: 120
Window step: 12
Threshold value: 0.400
Matrix: Bacillus subtilis, Thu Oct 27 15:58:26 2005
Matrix author: Dr. Borodovsky Laboratory, School of Biology, Georgia Tech
Matrix order: 5
```

List of Open reading frames predicted as CDSs, shown with alternate starts

(regions from start to stop codon w/ coding function >0.40)

Left end	Right end	DNA Strand	Coding Frame	Avg Prob	Start Prob
1	276	direct	fr 1	0.68	....
28	276	direct	fr 1	0.68	....
112	276	direct	fr 1	0.58	....
495	3587	direct	fr 3	0.43	0.03
528	3587	direct	fr 3	0.43	0.24
1191	3587	direct	fr 3	0.51	0.32
1356	3587	direct	fr 3	0.48	0.02
1467	3587	direct	fr 3	0.47	0.13
3578	5422	direct	fr 2	0.60	0.98
3623	5422	direct	fr 2	0.61	0.02
3632	5422	direct	fr 2	0.62	0.00
3764	5422	direct	fr 2	0.65	0.04
3914	5422	direct	fr 2	0.71	0.31
4037	5422	direct	fr 2	0.72	0.00
4052	5422	direct	fr 2	0.72	0.01
4082	5422	direct	fr 2	0.71	0.10
5395	6720	direct	fr 1	0.70	0.91
5434	6720	direct	fr 1	0.72	0.96
5443	6720	direct	fr 1	0.72	0.88
5473	6720	direct	fr 1	0.73	0.20
5485	6720	direct	fr 1	0.73	0.01
5578	6720	direct	fr 1	0.72	0.79
7548	8045	direct	fr 3	0.63	0.00
7602	8045	direct	fr 3	0.71	0.19
7668	8045	direct	fr 3	0.81	0.69
7710	8045	direct	fr 3	0.80	0.15
7746	8045	direct	fr 3	0.77	0.04
8032	8775	direct	fr 1	0.70	0.00
8059	8775	direct	fr 1	0.72	0.10
8344	8775	direct	fr 1	0.78	0.66
8395	8775	direct	fr 1	0.75	0.17
8407	8775	direct	fr 1	0.75	0.04

10317	10979	direct	fr 3	0.46	0.98
10952	12349	direct	fr 2	0.51	0.59
10970	12349	direct	fr 2	0.52	0.63
11024	12349	direct	fr 2	0.53	0.00
11027	12349	direct	fr 2	0.53	0.00
11090	12349	direct	fr 2	0.56	0.72
11186	12349	direct	fr 2	0.53	0.00
11201	12349	direct	fr 2	0.53	0.00
11606	12349	direct	fr 2	0.61	0.27
11627	12349	direct	fr 2	0.62	0.03
12559	12906	complement	fr 3	0.63	0.15
12559	12894	complement	fr 3	0.66	0.19
12559	12885	complement	fr 3	0.68	0.16
12559	12870	complement	fr 3	0.70	0.66
12559	12699	complement	fr 3	0.54	0.01
12559	12696	complement	fr 3	0.50	0.01
12976	13395	complement	fr 3	0.75	0.58
12976	13341	complement	fr 3	0.80	0.25
12976	13326	complement	fr 3	0.79	0.01
12976	13206	complement	fr 3	0.69	0.51
13424	13852	complement	fr 1	0.75	0.96
13424	13840	complement	fr 1	0.78	0.97
13424	13831	complement	fr 1	0.80	1.00
13424	13642	complement	fr 1	0.72	0.03
13424	13615	complement	fr 1	0.68	0.01
13974	14216	direct	fr 3	0.53	0.13
13983	14216	direct	fr 3	0.56	0.26
14031	14216	direct	fr 3	0.64	0.84
14103	14216	direct	fr 3	0.46	0.17
14257	14586	complement	fr 3	0.74	....
14257	14427	complement	fr 3	0.56	0.10
14257	14529	complement	fr 3	0.71	....

List of Regions of interest

(regions from stop to stop codon w/ a signal in between)

LEnd	REnd	Strand	Frame
1	276	direct	fr 1
486	3587	direct	fr 3
3569	5422	direct	fr 2
5392	6720	direct	fr 1
7542	8045	direct	fr 3
8026	8775	direct	fr 1
8751	9527	direct	fr 3
9077	9499	complement	fr 1
9505	10299	direct	fr 1
10302	10979	direct	fr 3
10943	12349	direct	fr 2
12268	12543	complement	fr 3
12559	12978	complement	fr 3

```
    12976     13416  complement  fr 3
    13424     13861  complement  fr 1
    13947     14216  direct      fr 3
    14257     14616  complement  fr 3
```

Prédiction GenMark.hmm

Parse predicted by GeneMark.hmm 2.4

```
GeneMark.hmm PROKARYOTIC (Version 2.5a)
Model organism: Bacillus_subtilis
Thu Nov 15 09:29:47 2007

Predicted genes
   Gene    Strand    LeftEnd    RightEnd       Gene     Class
    #                                        Length
    1        +          <1         276         276         1
    2        +         495        3587        3093         1
    3        +        3578        5422        1845         1
    4        +        5395        6720        1326         2
    5        +        6795        6965         171         2
    6        +        7548        8045         498         1
    7        +        8059        8775         717         1
    8        +        8772        9527         756         2
    9        +        9529       10299         771         2
   10        +       10317       10979         663         1
   11        +       10970       12349        1380         1
   12        -       12559       12885         327         1
   13        -       12976       13395         420         1
   14        -       13424       13831         408         1
   15        +       13983       14216         234         1
   16        -       14257      >14616         360         1
```

Recherche du RBS (Shine-Dalgarno)

Résultats en recherchant le motif : GGAGG 5...12 DTG

BS09819:[5380,5397]: ggagg aaggcagtaa atg
BS09819:[6600,6618]: ggagg aagcagtacct ttg
BS09819:[6782,6797]: ggagg tgaccaat atg
BS09819:[7116,7130]: ggagg aacacaa ttg
BS09819:[9515,9531]: ggagg cgatgtaag gtg

Résultats en recherchant le motif : GGAGG[1,0,0] 5...12 DTG (une base de différence dans le motif GGAGG).

Beaucoup parmi lesquels:
BS09819:[478,497] : ggggg ctatagactaat atg
BS09819:[5380,5397] : ggagg aaggcagtaa atg
BS09819:[6782,6797] : ggagg tgaccaat atg

BS09819:[7537,7550] : tgagg tgatgt atg
BS09819:[8047,8061] : ggaga gtgtgac atg
BS09819:[8758,8774] : gaagg tgtaaaaag atg
BS09819:[9515,9531] : ggagg cgatgtaag gtg
BS09819:[10302,10319]: tgagg gggccgggaa atg

Résultats en recherchant le motif avec la matrice consensus : {(-23, -53, 20, -33), (-22, -34, 20, -46), (14, -46, -7, -15), (-27, -52, 19, -17), (-4, -17, 14, -10)} >44 5...12 DTG

On génère aussi des faux positifs mais moins que lorsque l'on relâche les contraintes sur le motif consensus. Motifs prédits intéressants:

BS09819:[478,497] : ggggg ctatagactaat atg
BS09819:[5380,5397] : ggagg aaggcagtaa atg
BS09819:[6782,6797] : ggagg tgaccaat atg
BS09819:[8047,8061] : ggaga gtgtgac atg
BS09819:[8758,8774] : gaagg tgtaaaaag atg
BS09819:[9515,9531] : ggagg cgatgtaag gtg
BS09819:[10305,10319]: ggggg ccgggaa atg

Annexe 1 : Ensemble des résultats en recherchant le motif : GGAGG[1,0,0] 5...12 DTG
BS09819:[10039,10056]: ggagc aggcctgctc atg
BS09819:[10115,10128]: ggggg attaga ttg
BS09819:[10275,10294]: ggaag gcagaaaaacgt ttg
BS09819:[10302,10319]: tgagg gggccgggaa atg
BS09819:[10841,10854]: cgagg agatat atg
BS09819:[10859,10874]: agagg ggaatgct ttg
BS09819:[11810,11828]: ggagc aggaaatcaaa ttg
BS09819:[11890,11909]: ggaga aagcatctgtag atg
BS09819:[12275,12290]: ggaga acttactc ttg
BS09819:[12668,12682]: ggagt ttcaggg atg
BS09819:[12924,12937]: cgagg gaaggt atg
BS09819:[12985,13001]: agagg ttcgttttt ttg
BS09819:[13216,13228]: gtagg gaact ttg
BS09819:[13256,13270]: gcagg ttttggc gtg
BS09819:[13300,13315]: ggtgg cggagctg atg
BS09819:[13596,13612]: ggagc cctagcttc atg
BS09819:[13688,13700]: ggtgg caggt ttg
BS09819:[13895,13908]: gtagg ctggaa atg
BS09819:[13969,13985]: ggaag gtgaggaaa gtg
BS09819:[14174,14193]: tgagg atctttttact atg
BS09819:[14450,14467]: tgagg cagctgttca atg

BS09819:[1884,1899] : gcagg gaaaacat ttg
BS09819:[1900,1916] : ggaga ttttctcat atg
BS09819:[193,210] : ggagc agatttgata atg
BS09819:[3026,3042] : ggaga tagaacgat atg
BS09819:[3342,3357] : ggcgg tatgttat atg
BS09819:[3778,3791] : agagg tttcaa ttg
BS09819:[435,451] : agagg gatgtagaa ttg
BS09819:[4715,4733] : ggaga acgtgtagcca ttg
BS09819:[478,497] : ggggg ctatagactaat atg
BS09819:[4798,4813] : cgagg gcgacatt ttg
BS09819:[5054,5072] : ggaag acagctatccg gtg
BS09819:[5113,5125] : ggaag cttct ttg
BS09819:[5291,5309] : gtagg aataggcacac atg
BS09819:[5380,5397] : ggagg aaggcagtaa atg
BS09819:[5704,5719] : ggagc agctggca ttg
BS09819:[5866,5884] : gaagg agttagcggaa ttg
BS09819:[5923,5938] : ggaga tttgttaa ttg
BS09819:[5986,6005] : ggaga aaaagttcctgg atg
BS09819:[6307,6319] : ggaag acctg gtg
BS09819:[646,660] : gcagg acactat atg
BS09819:[6600,6618] : ggagg aagcagtacct ttg
BS09819:[6636,6652] : ggatg gagctgtag gtg
BS09819:[6782,6797] : ggagg tgaccaat atg
BS09819:[6894,6909] : ggatg tgtaactg gtg
BS09819:[7116,7130] : ggagg aacacaa ttg
BS09819:[7537,7550] : tgagg tgatgt atg
BS09819:[7712,7725] : ggaga gcctga ttg
BS09819:[7865,7877] : ggaag aagaa ttg
BS09819:[8047,8061] : ggaga gtgtgac atg
BS09819:[8437,8454] : gcagg ccaattttcg atg
BS09819:[8456,8473] : ggatg aggcagcggc ttg
BS09819:[8589,8604] : ggaag gtataacg gtg
BS09819:[8715,8730] : ggagc acttgttt atg
BS09819:[8758,8774] : gaagg tgtaaaaag atg
BS09819:[8812,8831] : ggacg ttttccagaaaa atg
BS09819:[8911,8930] : ggtgg tctattatattt atg
BS09819:[9262,9279] : ggacg ctagttctcg gtg
BS09819:[9304,9319] : ggtgg ttcattcc ttg
BS09819:[9515,9531] : ggagg cgatgtaag gtg
BS09819:[9649,9663] : ggaag agatcaa gtg
BS09819:[9690,9706] : ggaag ctatcgcag ttg
BS09819:[97,114] : ggggg gctcggggat atg
BS09819:[9763,9778] : gcagg gaaatttc atg
BS09819:[9846,9862] : ggaga tttgctcta ttg
BS09819:[9943,9961] : ggtgg tgcctggctgg ttg

Annexe II : Ensemble des résultats en recherchant le motif avec la matrice consensus

BS09819:[10039,10056]: ggagc aggcctgctc atg
BS09819:[10081,10094]: gggga tgctat atg
BS09819:[10115,10128]: ggggg attaga ttg
BS09819:[10276,10294]: gaagg cagaaaaacgt ttg
BS09819:[10305,10319]: ggggg ccgggaa atg
BS09819:[10862,10874]: gggga atgct ttg
BS09819:[11810,11828]: ggagc aggaaatcaaa ttg
BS09819:[11890,11909]: ggaga aagcatctgtag atg
BS09819:[12275,12290]: ggaga acttactc ttg
BS09819:[12668,12682]: ggagt ttcaggg atg
BS09819:[12929,12945]: gaagg tatgcaagc gtg
BS09819:[13300,13315]: ggtgg cggagctg atg
BS09819:[13596,13612]: ggagc cctagcttc atg
BS09819:[13688,13700]: ggtgg caggt ttg
BS09819:[13970,13985]: gaagg tgaggaaa gtg
BS09819:[1900,1916] : ggaga ttttctcat atg
BS09819:[193,210] : ggagc agatttgata atg
BS09819:[3026,3042] : ggaga tagaacgat atg
BS09819:[439,451] : ggatg tagaa ttg
BS09819:[4715,4733] : ggaga acgtgtagcca ttg
BS09819:[478,497] : ggggg ctatagactaat atg
BS09819:[5380,5397] : ggagg aaggcagtaa atg
BS09819:[5704,5719] : ggagc agctggca ttg
BS09819:[5866,5884] : gaagg agttagcggaa ttg
BS09819:[5923,5938] : ggaga tttgttaa ttg
BS09819:[5986,6005] : ggaga aaaagttcctgg atg
BS09819:[6600,6618] : ggagg aagcagtacct ttg
BS09819:[6636,6652] : ggatg gagctgtag gtg
BS09819:[6782,6797] : ggagg tgaccaat atg
BS09819:[6894,6909] : ggatg tgtaactg gtg
BS09819:[7116,7130] : ggagg aacacaa ttg
BS09819:[7712,7725] : ggaga gcctga ttg
BS09819:[8047,8061] : ggaga gtgtgac atg
BS09819:[8454,8473] : gggga tgaggcagcggc ttg
BS09819:[8590,8604] : gaagg tataacg gtg
BS09819:[8715,8730] : ggagc acttgttt atg
BS09819:[8758,8774] : gaagg tgtaaaaag atg
BS09819:[8911,8930] : ggtgg tctattatattt atg
BS09819:[9226,9239] : gggga tttaca atg
BS09819:[9304,9319] : ggtgg ttcattcc ttg
BS09819:[9515,9531] : ggagg cgatgtaag gtg
BS09819:[97,114] : ggggg gctcggggat atg

BS09819:[9846,9862] : ggaga tttgctcta ttg
BS09819:[9943,9961] : ggtgg tgcctggctgg ttg