Module Analyse de Génomes 2011-2012
Master 2 module FMBS 326 Immunoinformatique

Planning du Module :

Date	Heure	Salle
12/12	9h-12h	TD info TA1Z bat 25
	13h-17h	TD info TA1Z bat 25
13/12	9h-12h	TD info TA1Z bat 25
	13h-17h	TD info TA1Z bat 25
14/12	9h-12h	TD info TA1Z bat 25
	13h-17h	TD info TA1Z bat 25
15/12	9h-12h	TD info TA1Z bat 25
	13h-17h	TD info TA1Z bat 25
16/12	9h-12h	TD info TA1Z bat 25
	13h-16h	TD info TA1Z bat 25

Prévoir:
Une présentation orale (présentation de votre travail de programmation 15-20 minutes) ainsi que des documents à rendre (rapport + scripts sur CD).

Important : Pour chaque site Web que vous visiterez, donnez :

- l'origine des données (source),
- l'organisme (privé ou public) distribuant les données,
- le domaine de spécificité de ce site,
- les liens vers l'extérieur qu'ils fournissent,
- toute autre information intéressante…

PARTIE 1 - Annotation

a/ Quels sont les derniers génomes d'eucaryotes (vertébrés) séquencés ?
Trouver les informations sur le site GOLD (http://www.genomesonline.org) et donner la demarche utilisée.
Vérifier la cohérence des informations dans le site Ensembl.

Nom de l'espece	Taille du génome	Nombre de chromosomes	Publication	Sources

…

b/ La suite du module portera sur l'étude du génome du Pig (*Sus scrofa*) :

Espèce **Pig (*Sus scrofa*)**

Chaîne **T cell Receptor Beta**

Pour la suite vous complèterez le tableau en Annexe 2 en utilisant certaines instructions de l'Annexe 1

c/ Rechercher dans la base de données de références bibliographiques PubMed le ou les articles relatifs aux gènes TRBV, TRBD, TRBJ et TRBC de l'espèce.

d/ A partir de l'article de référence, récupérer les séquences nucléotiques (accessibles sur le site NCBI) et déterminer l'espèce la plus proche phylogénétiquement pour les loci TRB. Trouver les documents et informations allant dans ce sens.

e/ Dans Ensembl donner la localisation chromosomique approximative de la partie du locus TRB trouvé sur le chromosome, son orientation ainsi que les numéros des contigs correspondants.

f/ En vous appuyant sur les exemples de l'Annexe 3 et le protocole en Annexe 5, choisissez une séquence par binôme pour répondre aux questions suivantes :

- Identifier le(s) V-GENE(s), D-GENE(s) et JGENE(s) présent(s). Délimiter (position de debut et position de fin) chaque label constitutif (L-PART1, L-PART2, V-REGION, V-RS, J-REGION, D-REGION.....), les nommer et déterminer la fonctionnalité du (ou des) gène(s)(cf. IMGT® Scientific Chart – Functionality).
- Classer les informations précédentes dans les différents axioms de IMGT-ONTOLOGY (cf. http://www.imgt.org/IMGTindex/ontology.html)
- Fournir une carte de la portion de locus TRB trouvée en vous appuyant sur les représentations des locus présents dans IMGT.

g/ Pour chaque TRBV trouvé, donner le pourcentage d'identité avec les gènes de références de la souris et de l'homme à l'aide d'IMGT/V-QUEST (http://www.imgt.org/IMGT_vquest/share/textes/)

PARTIE 2 - Localisation des différents loci IG et TR de quelques espèces (Annexe 4) :

Indiquer, lorsque l'information est connue, pour chaque locus de chaque espèce, la localisation chromosomique et les positions correspondantes (avec le lien et/ou la référence qui vous ont permis de trouver l'information).
Remarque : Les informations pour l'homme sont vérifiées (source : IMGT Repertoire http://www.imgt.org). Il s'agit, donc pour les autres espèces, de vérifier, de préciser et de compléter le tableau ci-joint.

PARTIE 3 - Programmation.

But : Obtenir une carte complète du locus IGHV du rat (sur le chromosome 6)

Représentation graphique des gènes V à partir des données fournies (fichier sera distribué).
Automatisation de cette tâche (langage JAVA)

The column description is as follows:

Accession number
Gene name
Gene orientation/direction (1=+ve, -1=-ve)
Gene position (start) in bp
Gene position (stop) in bp
Functionnality

D'après la Charte IMGT®, il existe un code pour la fonctionnalité des gènes (rouge pour un pseudogène, jaune pour un ORF et vert pour un fonctionnel, cf. IMGT® Scientific Chart - V-GENE prototypes with IMGT/LIGM-DB labels).

Remarques :

- La possibilité de réutilisation de l'outil pour d'autres espèces devra être prise en compte.
- La création d'un diagramme de classe sera un plus.
- Chaque allèle ayant sa fonctionnalité, ce serait un plus de permettre au logiciel de gérer plusieurs fonctionnalités pour chaque gène.

Pour vous guider

 IMGT:
http://www.imgt.org
http://www.imgt.org/textes/IMGTScientificChart/RepresentationRules/colorchart.html#LOCUS
http://www.imgt.org/textes/IMGTScientificChart/SequenceDescription/variable.html
http://www.imgt.org/textes/IMGTScientificChart/SequenceDescription/IMGTfunctionality.html
http://www.imgt.org/textes/IMGTScientificChart/Nomenclature/IMGTnomenclature.html
IMGT/V-QUEST :
http://www.imgt.org/IMGT_vquest/share/textes/
 Ensembl :
http://www.ensembl.org
http://www.ensembl.org/Sus_scrofa/index.html
 NCBI:
http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/
http://www.ncbi.nlm.nih.gov/Entrez/
http://www.ncbi.nlm.nih.gov/mapview/

ANNEXE 1 :

Canis familiaris
Protocole pour la localisation des gènes IG Heavy, IG Light et TR

Localisation des loci sur le génome

- **Recherche du chromosome de chaque locus, *via* NCBI, soit par :**
  - Recherche bibliographique dans PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez/)
  - Recherche de gènes dans Entrez Gene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=).

  Saisie de mots clefs précis, exemple: *t cell receptor beta canis lupus*.

  Obtention d'une liste de numéro d'accès (LOC…) correspondant aux gènes recherchés. Dans le résultat de la requête (cf. exemple), nous obtenons la définition du gène, sa **localisation chromosomique,** sa position dans la séquence nucléotidique de référence, et son numéro d'identité (GeneID).

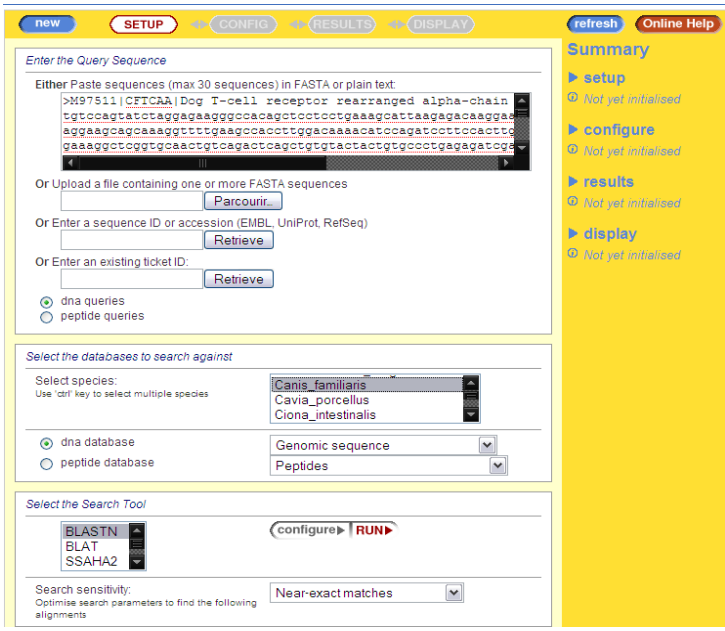
  Exemple de résulats:
  LOC609058 similar to T-cell receptor beta chain V region 86T1 precursor [*Canis lupus familiaris*]
  Chromosome: 16
  Annotation: Chromosome 16, NC_006598.2 (9934782..9935332, complement)
  GeneID: 609058

  Collecte de tous les LOC (séquences de gènes) pour les blaster, éventuellement, contre le génome du chien d'Ensembl, dans le cas où n'avons pas ou peu de séquences dans LIGM-DB.

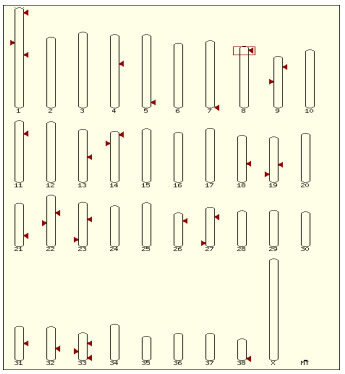
- Recherche sur Ensembl, des Contigs contenant les gènes recherchés :

A partir des séquences de l'espèce *Canis familiaris* présentes dans LIGM-DB, réaliser un BLAST sur Ensembl (http://www.ensembl.org/Multi/blastview).

Collez la séquence en format FASTA.
Sélectionnez l'espèce : Canis_familiaris.



Gestion des résultats :



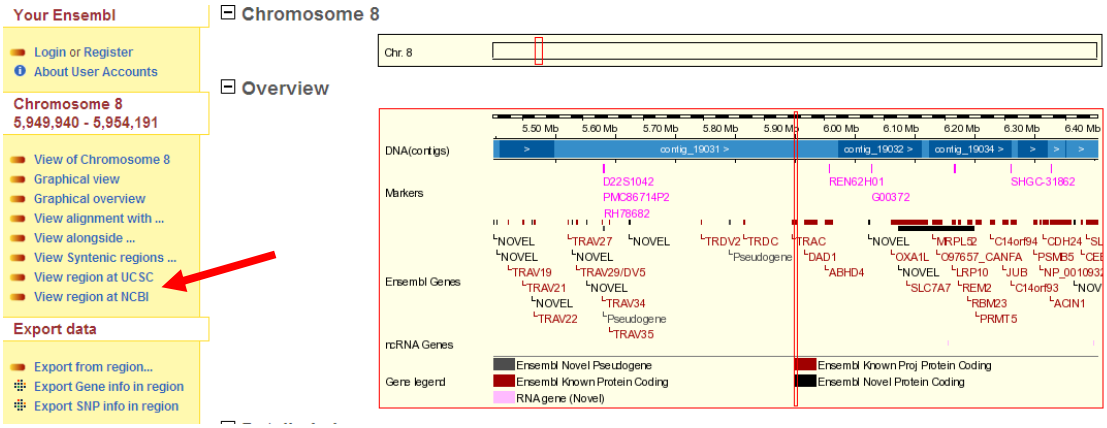
1) Visualiser le meilleur alignement qui apparaît dans un cadre rouge, ce qui permet de vérifier la localisation chromosomique de notre séquence.

2) Visualiser le meilleur alignement, accessible *via* le lien : [A] (alignment)

Links	Query Start	End	Ori	Chromosome Name	Start	End	Ori	Stats Score	E-val	%ID	Length
[A] [S] [G] [C]	246	497	+	Chr:8	5951940	5952191	+	252	6.5e-171	100.00	252
[A] [S] [G] [C]	1	174	+	Chr:8	5309561	5309734	+	170	6.3e-90	99.43	174
[A] [S] [G] [C]	1	171	+	Chr:8	5357376	5357546	+	159	2.3e-83	98.25	171
[A] [S] [G] [C]	1	171	+	Chr:8	5403505	5403675	+	155	5.8e-81	97.66	171
[A] [S] [G] [C]	1	171	+	Chr:8	5400390	5400560	+	139	2.1e-71	95.32	171

Pour cet alignement, déterminer le numéro du contig *via* le lien : [C] (Contig)

3) Détermination du contig grâce à *ContigView* (obtenu par le lien [C]).



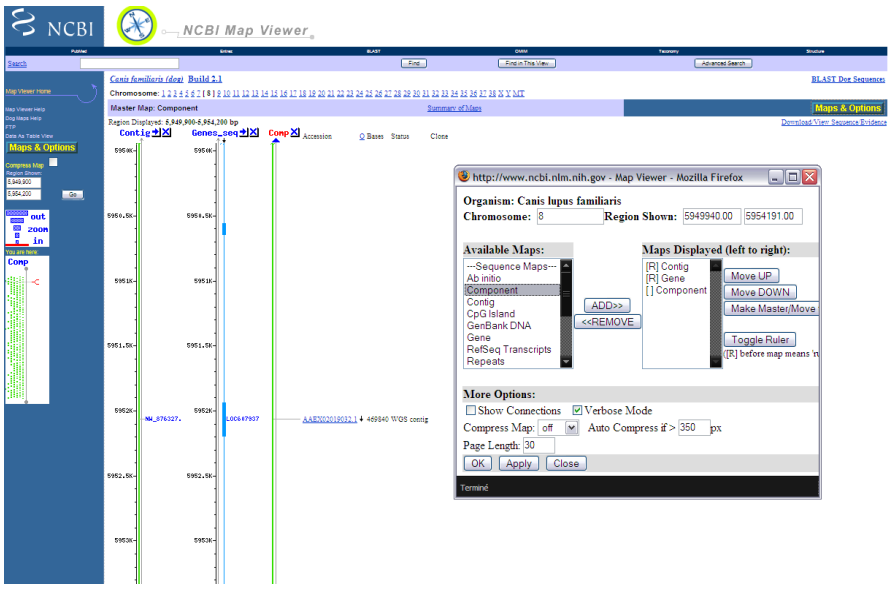
Sur la page *ContigView*, le contig de l'alignement est indiqué par un trait rouge.

Obtenir le numéro d'accès (NCBI) du contig :

Sur la page *ContigView*, dans le menu de gauche, choisir View region at NCBI. (cf. flèche rouge)

Remarque: Au préalable, il est nécessaire de centrer sur la région contenant le contig.

Cliquer sur le contig et choisir dans le menu qui s'affiche "Centre on the contig". Ainsi, nous obtiendrons **uniquement le numéro d'accès** du contig cible sur Map Viewer.



La page de Map Viewer s'ouvre. A droite, dans le menu Maps & Options, choisir d'afficher Component.
S'affichent, alors, sur la carte les numéros d'accès :

- du contig du NCBI (colonne Contig: séquence regroupant les components)
- du ou des gènes contenus dans la séquence (colonne Genes_seq)
- de la séquence nucléotidique correspondant au contig d'Ensembl (colonne Comp).

Les étapes suivantes sont :

Récupération des séquences d'Ensembl: On récupère le format fasta des séquences d'Ensembl pour les coller dans le champ texte d'un blast par exemple.

BLAST sur les bases d'IMGT :
Sur le site d'IMGT®, nous pouvons effectuer un BLAST contre la base de données IMGT/LIGM-DB : http://www.imgt.org/blast/blast.html.
Ainsi il est possible de comparer nos séquences trouvées avec ce qui existe dans la base, mais des bases peuvent être faites (comme une base de données ne regroupant que les séquences de chien par exemple).

Localiser la région chromosomique du locus cible:
Récupérer les séquences nucléotidiques des contigs adjacents (en amont et en aval) du premier contig identifié comme contenant des gènes d'intérêts.
Pour chaque contig, réaliser le protocole précédemment décrit:

- Réaliser un BLAST sur les bases IMGT,
- Obtenir le numéro d'accès NCBI (si des gènes d'intérêt y sont localisés).

Synthèse des résultats :
Dans un tableau, noter :

- l'espèce, la race
- le chromosome,
- la position du locus sur le chromosome
- la taille du locus en Kb
- le nombre de paires de bases analysées (pour estimer si la recherche couvre une assez grande longueur chromosomique, en comparant avec les locus respectifs de l'homme)
- le nom des contig d'Ensembl, et le numéro d'accès du NCBI
- la position de chaque séquence sur le chromosome de 5' en 3'
- la taille des gaps entre chaque séquence
- les résultats du Blast pour déterminer la présence ou non de gènes d'intérêt.

ANNEXE 2 :
A remplir

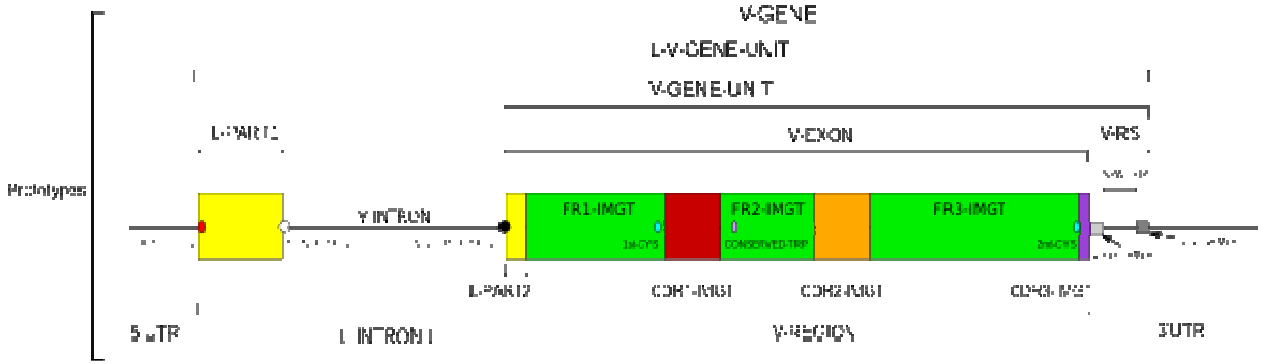
Nom du gène	Fonctionnalité	EMBL Accession number	Ensembl Contig	Localisation chromosomique	Position dans la séquence	% Homologie avec l'espèce la plus proche	Sens

ANNEXE 3 - Exemples d'annotation

V-GENE

>X62106.0|HSVI2|*Homo sapiens* VI-2 gene for immunoglobulin heavy chain

```
tgagagctcc gttcctcacc atggactgga cctggaggat cctcttcttg gtggcagcag      60
ccacaggtaa gaggctccct agtcccagtg atgagaaaga gattgagtcc agtccaggga     120
gatctcatcc acttctgtgt tctctccaca ggagcccact cccaggtgca gctggtgcag     180
tctggggctg aggtgaagaa gcctggggcc tcagtgaagg tctcctgcaa ggcttctgga     240
tacaccttca ccggctacta tatgcactgg gtgcgacagg cccctggaca agggcttgag     300
tggatgggat ggatcaaccc taacagtggt ggcacaaact atgcacagaa gtttcagggc     360
agggtcacca tgaccaggga cacgtccatc agcacagcct acatggagct gagcaggctg     420
agatctgacg acacggccgt gtattactgt gcgagagaca cagtgtgaaa acccacatcc     480
tgagggtgtc agaaacccaa gggaggaggc ag
```

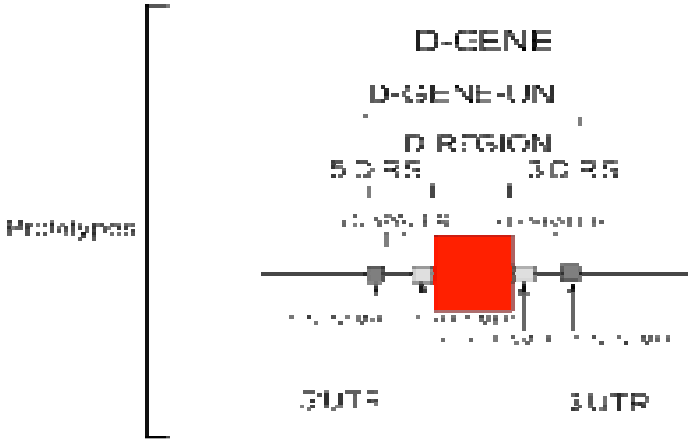


V-GENE (ProteinDisplay *Mus musculus*)

IGHV gene	FR1-IMGT (1-26)	CDR1-IMGT (27-38)	FR2-IMGT (39-55)	CDR2-IMGT (56-65)	FR3-IMGT (66-104)	CDR3-IMGT (105-115)
	1 10 20	30	40 50	60	70 80 90 100	110
AC073561,IGHV1-4	QVQLQQSGA.ELARPGASVKMSCKAS	GYTFTSYT....	MHWVKQRPGQGLEWIGY	INPSSGYT..	KYNQKFK.DKATLTADKSSSTAYMQLSSLTSEDSAVYYC	AR
AC090843,IGHV1-5	EVQLQQSGT.VLARPGASVKMSCKTS	GYTFTSYW....	MHWVKQRPGQGLEWIGA	IYPGNSDT..	SYNQKFK.GKAKLTAVTSASTAYMELSSLTNEDSAVYYC	TR

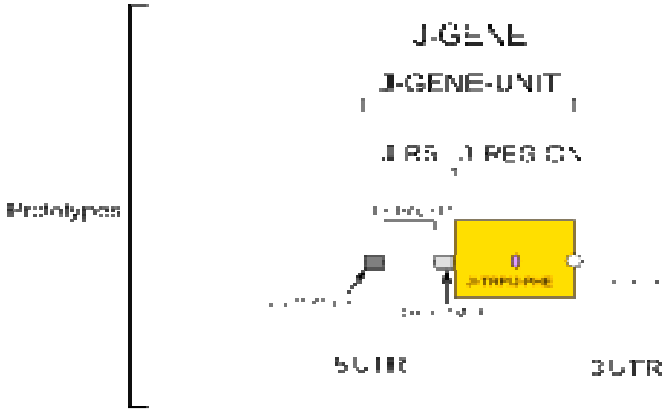
D-GENE

>J00256
ccagccgcagggttttttgg tcaactcaaa ac ctgtgc taactgggga cacagtg att cc caccct aaaccatgctcccccggg



J-GENE

>J00256
accccgggct gtgggtttct gtgcccctgg ctcagggctg actcaccgtg gctgaatact
tccagcactg gggccagggc accctggtca ccgtctcctc aggtgagtct gctgtactgg
ggatagcggg gagccatgtg tactgggcca agcaagggct ttggcttcag



ANNEXE 4 :

A remplir (quand ceci est possible)

	Human ***Homo sapiens***	**Mouse** ***Mus musculus***	**Rat** ***Rattus norvegicus***	**Cat** ***Felix catus***	**Dog** ***Canis familiaris***	**Zebrafish** ***Danio rerio***
IGH	14q32.33	12F1-F2				
IGK	2p11.2					
IGL	22q11.2					
TRA	14q11.2					
TRB	7q34					
TRG	7p14					
TRD	14q11.2		15p13			

ANNEXE 5 : Utilisation du logiciel NTI Vector

Vous travaillez sur la **version DEMO** du logiciel, certaines fonctions sont limitées.

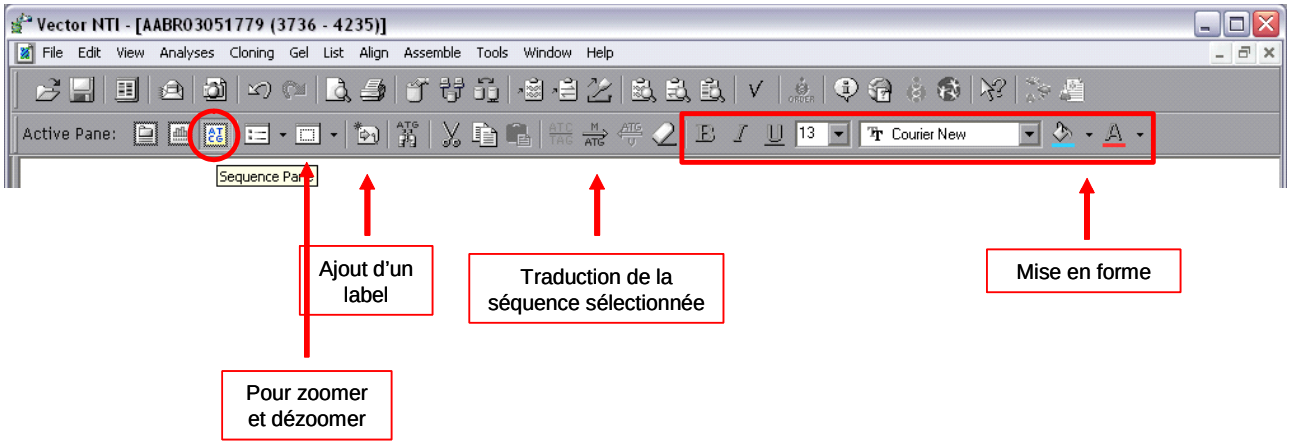
- Lancez le logiciel : Menu démarrer / Invitrogen / Vector NTI Advance 10 / **Vector NTI**
- Allez chercher la séquence à étudier dans GenBank : Tools / Open / Retrieve DNA-RNA by GenBank NID ...
- Entrez le numéro d'accession de votre séquence (AAEX...)
- Après chargement de la séquence, décochez la visualisation des sites de restriction (View / Display Setup / décochez "Restriction Map"), changez le code de traduction (View / Display Setup / Sequence Setup / choisissez "1-Letter AA Code"), et n'affichez que le brin sens (View / Display Setup / Sequence Setup / cochez "Prefer Single-Stranded Display").

- Afin de pouvoir enregistrer votre travail, vous devez, après importation de votre séquence, **faire une modification** (par exemple ajouter un label), puis **fermer le programme** NTI Vector. Il vous demandera s'il doit enregistrer les modifications. Acceptez l'enregistrement par défaut, la séquence est enregistrée à son nom (AAEX…) dans la base Database DNA/RNAs.

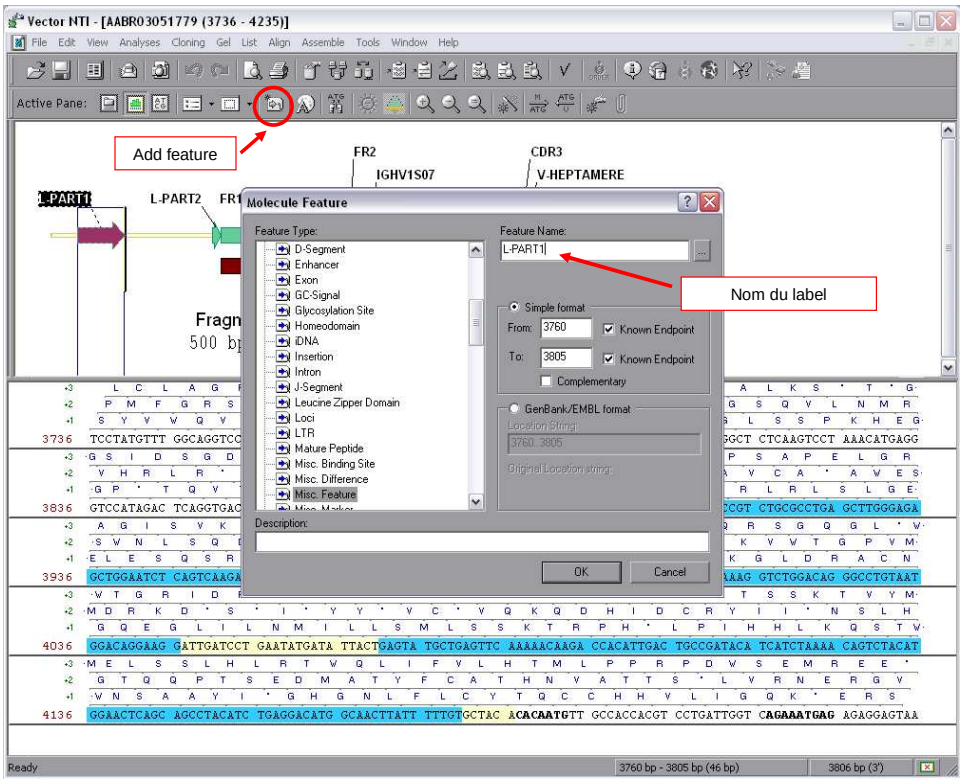
- Vous pouvez maintenant travailler sur votre séquence : ouvrez le fichier enregistré dans la base locale.

La mise en forme (couleurs, gras, italique, …) et les traductions ne peuvent pas être enregistrées. Pensez à ajouter des "**features**" pour ne pas perdre votre travail lors de la fermeture du fichier.

Principales fonctionnalités de la barre d'outils «Sequence pane » :



Ajout d'un label :



Détermination des positions et de la taille des labels :