Innovation, Science and Economic Development Canada
Canadian Intellectual Property Office

Innovation, Sciences et Développement économique Canada
Office de la propriété intellectuelle du Canada

Office de la propriété intellectuelle du Canada

Listages des séquences biologiques de brevets

Dictionnaire de données

Norme ST.25 de l'OMPI

Version 1.00

2019-01-08

Canada

This document is also available in English under the title *Patent Biological Sequence Listings - Data Dictionary*.

On peut obtenir cette publication sur supports accessibles, sur demande.

Coordonnées
Centre de services à la clientèle
Office de la propriété intellectuelle du Canada
Innovation, Sciences et Développement économique Canada
Place du Portage I
Bureau C229, 2e étage
50, rue Victoria
Gatineau (Québec) K1A 0C9

Téléphone (sans frais) : 1 -866 -997- 1936
ATS : 1 -866 -442 -2476
Téléc. : 819 -953- 2476
ic.contact-contact.ic@canada.ca

Nº au catalogue lu71-4/63-2019F-PDF
ISBN 978-0-660-29227-4

TABLE DES MATIÈRES

1.0 Aperçu des données sur les listages des séquences biologiques canadiennes

Les brevets s'appliquent aux technologies nouvellement créées et aux améliorations apportées aux produits ou aux procédés. La *Loi sur les brevets* décrit en détail ce qui constitue un brevet. Un brevet est un droit accordé par le gouvernement d'empêcher d'autres personnes de fabriquer, d'utiliser ou de vendre votre invention au Canada.

Les données sur les listages des séquences biologiques canadiennes contiennent de l'information sur les nucléotides et les séquences d'acides aminés. La présente section donne un aperçu des renseignements fournis et une description des structures de fichiers.

1.1 Données sur les listages des séquences biologiques

Le listage des séquences biologiques d'un brevet est présenté comme une partie distincte de la description de la demande de brevet ou du brevet octroyé et comme une collecte distincte de données. La structure de fichiers est régie par la norme ST.25 de l'Organisation mondiale de la propriété intellectuelle (OMPI) relative à la présentation du listage des séquences de nucléotides et d'acides aminés dans les demandes de brevets.

Les données sur les listages des séquences biologiques (LSB) se composent du fichier soumis par le demandeur (souvent un fichier texte) et de deux autres types de fichier (.PEP et .SEQ) générés par l'Office. La documentation officielle relative au listage des séquences biologiques dans une demande de brevet demandé ou un brevet octroyé est le dossier fourni par le demandeur. Chaque dossier créé par l'Office est un dossier de travail; il peut être incomplet si le dossier soumis par le client est lui-même incomplet. L'Office exigera un nouveau listage des séquences au client s'il juge que le listage fourni est incomplet.

Il convient également de noter que les demandes entrant dans la phase nationale importent le listage des séquences à partir de la demande internationale présentée à l'Organisation mondiale de la propriété intellectuelle (OMPI). Parfois, les listages des séquences téléchargés sont incomplets.

1.2 Fichiers de données hebdomadaires

Les demandes et les brevets octroyés LSB sont fournies dans un fichier de mise à jour hebdomadaire. Les données sur les brevets sont mises à la disponibilité du public après une période de confidentialité pouvant aller jusqu'à 18 mois après la première date de dépôt de la demande.

Une collecte hebdomadaire LSB contient des fichiers TXT, PEP et SEQ. Cette collecte contient des fichiers mis à jour ainsi que de nouveaux fichiers pour la semaine en cours. La taille du fichier de données hebdomadaires oscille entre 1 Mo et 350 Mo, selon le volume d'activité. La collecte hebdomadaire comprend un rapport indiquant tous les fichiers de la semaine.

Les éléments obligatoires relatifs aux données du fichier fourni par le demandeur comprennent les suivants :

- **des renseignements bibliographiques** (nom du demandeur, titre de l'invention)
- **des renseignements sur le listage des séquences** (nombre de numéros d'identification de séquences, numéro d'identification SEQ, longueur, type, organisme, séquence).

1.3 Description de la structure ST.25

La structure des fichiers est régie par la norme ST.25 de l'Organisation mondiale de la propriété intellectuelle (OMPI) pour la présentation des listages des séquences de nucléotides et d'acides aminés dans les demandes de brevet. Tous les renseignements concernant la norme ST.25 en ce qui concerne la structure des fichiers et les spécifications techniques se trouvent sur le site Web de l'OMPI, à l'adresse suivante :
https://www.wipo.int/export/sites/www/standards/fr/pdf/03-25-01.pdf

Les fichiers de mise à jour hebdomadaire des données sont structurés comme suit :

Fichier ZIP (compressé) hebdomadaire sur les LSB « BSL » :

BSLAAAASS.zip (c.-à-d. BSL201825.zip) contient un dossier pour la semaine d'extraction et un fichier de rapport où AAAA représente l'année et SS représente le numéro de semaine (01 à 52 semaines)

- Répertoire hebdomadaire de la collecte :
  - AAAA-MM-JJ (c.-à-d. 2018-06-23) contient une collection de dossiers intitulée XXXXXXXX (c.-à-d. 02836299) où X représente le numéro de brevet.
    - Répertoire - numéro de brevet : contient les fichiers PEP, SEQ et TXT :
      - Les noms de fichier contiennent un préfixe (CA), le numéro de brevet, la date de production, le type de fichier et la version, tel que CAXXXXXXXXAAAAMMJJ-DNAvXX.PEP (ou .SEQ ou .TXT)
      - Exemples :
        - CA0278203320180618-DNAv03.PEP;
        - CA0278203320180618-DNAv04.SEQ;
        - CA0278203320180618-DNAv03.TXT)

- Fichier de rapport :

- Report.txt contient les dates de l'extraction, le nombre de brevets traités, le nombre de fichiers produits et une liste complète de tous les brevets, ainsi que les noms des fichiers PEP, SEQ et TXT inclus dans la collecte hebdomadaire. Certains renseignements de base concernant les dates de mise à la disponibilité, l'entrée dans la phase nationale et les demandes PCT sont également compris dans le fichier report.txt pour chaque brevet dans la collecte hebdomadaire.

No	Donnée	Identificateur numérique de la norme ST.25	Obligatoire (O), conditionnel (C), ou facultatif (F)	Description
2.0 Listages des séquences biologiques de brevets canadiens				
1.	**Nom du demandeur**	**<110>** *MOUNT SINAI HOSPITAL*	O	Cet identificateur numérique est suivi du nom du demandeur. Il peut s'agir d'une personne ou d'une entreprise. Remarque : Si le nom du demandeur est écrit en caractères autres que ceux de l'alphabet latin, la valeur sera une traduction ou une translittération.
2.	**Titre de l'invention**	**<120>** *METHODS AND COMPOSITIONS FOR MODULATING A STEROID RECEPTOR*	O	Cet identificateur numérique est suivi du titre de l'invention.
3.	**Référence du dossier**	**<130>** *064016-379280*	C	Cet identificateur numérique est suivi du numéro de référence du dossier physique de la demande. Remarque : Cet identificateur numérique s'affiche si le listage des séquences a été fourni à tout moment avant l'attribution d'un numéro de demande.
4.	**Demande de brevet actuelle**	**<140>** *PCT/CA2005/000042*	C	Cet identificateur numérique est suivi du numéro de demande de brevet actuelle. Remarque : Cet identificateur numérique s'affiche si le listage des séquences a été fourni après l'attribution d'un numéro de demande.
5.	**Date de dépôt de la demande actuelle**	**<141>** *2005-01-14*	C	Cet identificateur numérique est suivi de la date de dépôt de la demande actuelle. Remarque : Cet identificateur numérique s'affiche si le listage des séquences a été fourni après l'attribution d'un numéro de demande.
6.	**Demande de brevet antérieure**	**<150>** *60/536,598*	C	Cet identificateur numérique est suivi du numéro de la demande antérieure lorsqu'un listage des séquences est déposé relativement à une demande qui revendique la priorité d'une demande antérieure.
7.	**Date de dépôt de la demande antérieure**	**<151>** *2004-01-15*	C	Cet identificateur numérique est suivi de la date de dépôt de la demande antérieure lorsqu'un listage des séquences est déposé relativement à une demande qui revendique la priorité d'une demande antérieure.
8.	**Nombre de SEQ ID NO (numéros d'identification des séquences)**	**<160>** *22*	O	Cet identificateur numérique est suivi du nombre de séquences (<400>) trouvées dans le document.

No	Donnée	Identificateur numérique de la norme ST.25	Obligatoire (O), conditionnel (C), ou facultatif (F)	Description
9.	**Logiciel**	**<170>** *PatentIn version 3.3*	F	Cet identificateur numérique est suivi du logiciel utilisé pour créer des listages de séquences à inclure dans les demandes de brevet.
10.	**Numéro d'identification de séquence**	**<210>** *1*	O	Cet identificateur numérique est suivi du numéro d'identification de la séquence correspondant à la séquence suivante (<400>).
11.	**Longueur**	**<211>** *707*	O	Cet identificateur numérique est suivi de la longueur de la séquence (c.-à-d. le nombre de paires de base ou d'acides aminés).
12.	**Type**	**<212>** *PRT*	O	Cet identificateur numérique est suivi du type de molécule séquencée, soit ADN, soit ARN, soit PRT. Remarque : Si une séquence de nucléotides contient à la fois des fragments d'ADN et d'ARN, la valeur qui suit cet identificateur numérique sera celle de l'ADN.
13.	**Organisme**	**<213>** *Homo sapiens*	O	Cet identificateur numérique est suivi du nom de l'espèce ou du nom scientifique de la molécule séquencée. Cette valeur peut aussi être « Séquence artificielle » ou « Non connu ».
14.	**Caractéristique**	**<220>**	C	Cet identificateur numérique est toujours laissé en blanc, mais il est suivi d'autres identificateurs numériques qui comportent une description des points ayant une importance biologique dans la séquence. Remarque : Cet identificateur numérique est utilisé lorsque « n », ou « Xaa », ou une base modifiée ou un acide aminé L modifié ou peu connu est utilisé dans la séquence ou si l'organisme (identificateur numérique <213>) est « Séquence artificielle » ou « Non connu ».
15.	**Nom/clé**	**<221>**	C	Cet identificateur numérique est suivi de la clé de caractérisation qui représente les points ayant une importance biologique dans la séquence. Les clés de caractérisation et leurs définitions se trouvent aux annexes E et F. Remarque : Cet identificateur numérique est utilisé lorsque « n », ou « Xaa », ou une base modifiée ou un acide aminé L modifié ou peu connu est utilisé dans la séquence.

No	Donnée	Identificateur numérique de la norme ST.25	Obligatoire (O), conditionnel (C), ou facultatif (F)	Description
16.	**Emplacement**	**<222>**	C	Cet identificateur numérique est suivi de l'emplacement des points ayant une importance biologique dans la séquence selon la longueur de la séquence. Remarque : Cet identificateur numérique est utilisé lorsque « n », ou « Xaa », ou une base modifiée ou un acide aminé L modifié ou peu connu est utilisé dans la séquence.
17.	**Autres informations**	**<223>**	C	Cet identificateur numérique est suivi de la description des points ayant une importance biologique dans la séquence. Remarque : Cet identificateur numérique est utilisé lorsque « n », ou « Xaa », ou une base modifiée ou un acide aminé L modifié ou peu connu est utilisé dans la séquence ou si l'organisme (identificateur numérique <213>) est « Séquence artificielle » ou « Non connu ».
18.	**Informations concernant la publication**	**<300>**	F	Cet identificateur numérique est toujours laissé en blanc, mais il est suivi des informations concernant la publication.
19.	**Auteurs**	**<301>**	F	Cet identificateur numérique est suivi du nom de l'auteur ou des noms des auteurs de la publication dans le périodique où se trouve le listage des séquences.
20.	**Titre**	**<302>**	F	Cet identificateur numérique est suivi du titre de la publication dans le périodique où se trouve le listage des séquences.
21.	**Périodique**	**<303>**	F	Cet identificateur numérique est suivi du titre du périodique dans lequel se trouve le listage des séquences.
22.	**Volume**	**<304>**	F	Cet identificateur numérique est suivi du volume du périodique dans lequel se trouve le listage des séquences.
23.	**Numéro**	**<305>**	F	Cet identificateur numérique est suivi du numéro du périodique dans lequel se trouve le listage des séquences.
24.	**Pages**	**<306>**	F	Cet identificateur numérique est suivi de la série de pages du périodique dans laquel se trouve le listage des séquences.
25.	**Date**	**<307>**	F	Cet identificateur numérique est suivi de la date de parution du périodique dans lequel le listage des séquences.

No	Donnée	Identificateur numérique de la norme ST.25	Obligatoire (O), conditionnel (C), ou facultatif (F)	Description
26.	**Numéro d'entrée dans la base de données**	**<308>** *P23246*	F	Numéro d'entrée attribué par la base de données, y compris nom de cette base de données.
27.	**Date d'entrée dans la base de données**	**<309>** *2004-06-15*	F	Date d'entrée dans la base de données.
28.	**Numéro du document**	**<310>**	F	Numéro du document, uniquement pour les citations de brevets.
29.	**Date de dépôt**	**<311>**	F	Date de dépôt du document, uniquement pour les citations de brevets.
30.	**Date de publication**	**<312>**	F	Date de publication du document, uniquement pour les citations de brevets.
31.	**Résidus pertinents dans SEQ ID NO : x**	**<313>** *(1)..(707)*	F	Numéros d'identification des séquences (SEQ ID NOS) mentionnés dans la publication citée.
32.	**Séquence**	**<400>** *et Ser Arg Asp Arg Phe Arg Ser Arg Gly Gly Gly Gly Gly Gly Phe* *1 5 10 15* *His Arg Arg Gly Gly Gly Gly Gly Arg Gly Gly Leu His Asp Phe Arg* *20 25 30* *Ser Pro Pro Pro Pro Pro Gly Met Gly Leu Asn Gln Asn Arg Gly Pro Met Gly* *35 40 45*	O	

Annexe A – Liste des nucléotides

Symbole	Signification	Origine de la désignation
a	a	adénine
g	g	guanine
c	c	cytosine
t	t	thymine
u	u	uracile
r	g ou a	purine
y	t/u ou c	pyrimidine
m	a ou c	amino
k	g ou t/u	keto
s	g ou c	interactions fortes (liaisons 3 H)
w	a ou t/u	interactions faibles (liaisons 2 H)
b	g ou c ou t/u	autre que a
d	a ou g ou t/u	autre que c
h	a ou c ou t/u	autre que g
v	a ou g ou c	autre que t et u
n	a ou g ou c ou t/u, non connu ou autre	n'importe lequel

Annexe B – Liste des nucléotides modifiés

Symbole	Signification
ac4c	4-acétylcytidine
chm5u	5-(carboxyhydroxyméthyl)uridine
cm	2'-O-méthylcytidine
cmnm5s2u	5-carboxyméthylaminométhyl-2-thiouridine
cmnm5u	5-carboxyméthylaminométhyluridine
d	dihydrouridine
fm	2'-O-méthylpseudouridine
gal q	bêta, D-galactosylquéuosine
gm	2'-O-méthylguanosine
i	inosine
i6a	N6-isopentényladénosine
m1a	1-méthyladénosine
m1f	1-méthylpseudouridine
m1g	1-méthylguanosine
m1i	1-méthylinosine
m22g	2,2-diméthylguanosine
m2a	2-méthyladénosine
m2g	2-méthylguanosine
m3c	3-méthylcytidine
m5c	5-méthylcytidine
m6a	N6-méthyladénosine
m7g	7-méthylguanosine
mam5u	5-méthylaminométhyluridine
mam5s2u	5-méthoxyaminométhyl-2-thiouridine
man q	bêta, D-mannosylquéuosine
mcm5s2u	5-méthoxycarbonylméthyl-2-thiouridine
mcm5u	5-méthoxycarbonylméthyluridine
mo5u	5-méthoxyuridine
ms2i6a	2-méthylthio-N6-isopentényladénosine
ms2t6a	N-((9-bêta-D-ribofuranosyl-2-méthylthiopurine-6-yl) carbamoyl) thréonine
mt6a	N-((9-bêta-D-ribofuranosylpurine-6-yl)N-méthylcarbamoyl) thréonine
mv	ester méthylé d'uridine 5 oxy-acétique acide
o5u	acide d'uridine 5 oxy-acétique

osyw	wybutoxosine
p	pseudouridine
q	quéuosine
s2c	2-thiocytidine
s2t	5-méthyl-2-thiouridine
s2u	2-thiouridine
s4u	4-thiouridine
t	5-méthyluridine
t6a	N-((9-bêta-D-ribofuranosylpurine-6-yl)-carbamoyl) thréonine
tm	2'-O-méthyl-5-méthyluridine
um	2'-O-méthyluridine
yw	wybutosine
x	3-(3-amino-3-carboxypropyl)uridine, (acp3)u

Annexe C – Liste des acides aminés

Symbole	Signification
Ala	alanine
Cys	cystéine
Asp	acide aspartique
Glu	acide glutamique
Phe	phénylalanine
Gly	glycine
His	histidine
Ile	isoleucine
Lys	lysine
Leu	leucine
Met	méthionine
Asn	asparagine
Pro	proline
Gln	glutamine
Arg	arginine
Ser	sérine
Thr	thréonine
Val	valine
Trp	tryptophane
Tyr	tyrosine
Asx	Asp ou Asn
Glx	Glu ou Gln
Xaa	non connu ou autre

Annexe D – Liste des acides aminés modifiés ou peu connus

Symbole	Signification
Aad	acide 2-aminoadipique
bAad	acide 3-aminoadipique
bAla	bêta-alanine, bêta-acide aminopropionique
Abu	acide 2-aminobutyrique
4Abu	acide 4-aminobutyrique, acide pipéridinique
Acp	acide 6-aminocaproïque
Ahe	acide 2-aminoheptanoïque
Aib	acide 2-aminoisobutyrique
bAib	acide 3-aminoisobutyrique
Apm	acide 2-aminopimélique
Dbu	acide 2,4-diaminobutyrique
Des	desmosine
Dpm	acide 2,2-diaminopimélique
Dpr	acide 2,3-diaminopropionique
EtGly	N-éthylglycine
EtAsn	N-éthylasparagine
Hyl	hydroxylysine
aHyl	allo-hydroxylysine
3Hyp	3-hydroxyproline
4Hyp	4-hydroxyproline
Ide	isodesmosine
aIle	allo-isoleucine
MeGly	N-méthylglycine, sarcosine
MeIle	N-méthylisoleucine
MeLys	6-N-méthyllysine
MeVal	N-méthylvaline
Nva	norvaline
Nle	norleucine
Orn	ornithine

Annexe E – Liste des clés de caractérisation concernant les séquences de nucléotides

Clé	Description
allele	individu ou souche apparenté contenant des formes stables différentes d'un même gène, qui se distingue de la séquence présentée à cet emplacement (et possiblement à d'autres emplacements)
attenuator	1) région d'ADN où se produit une régulation de la terminaison de la transcription qui contrôle l'expression de certains opérons bactériens; 2) segment de séquence situé entre le promoteur et le premier gène de structure, qui provoque une terminaison partielle de la transcription
C_region	région constante de la chaîne lourde et de la chaîne légère de l'immunoglobuline et des chaînes alpha, bêta et gamma du récepteur d'un lymphocyte T; comprend un ou plusieurs exons, selon la chaîne
CAAT_signal	séquence CAAT; partie d'une séquence conservée située environ 75 paires de bases en amont du site d'initiation des unités de transcription eucaryotes, qui peut jouer un rôle dans la fixation de l'ARN polymérase; consensus=GG (C ou T) CAATCT
CDS	séquence codante (« coding sequence »); séquence de nucléotides correspondant à celle des acides aminés dans une protéine (l'emplacement comprend le codon d'arrêt); contient la traduction conceptuelle des acides aminés
conflict	les déterminations indépendantes de la « même » séquence diffèrent sur ce site ou dans cette région
D-loop	boucle de déplacement (« displacement loop »); région au sein de l'ADN mitochondrial où une petite partie d'ARN est appariée à un brin d'ADN, entraînant le déplacement du brin original d'ADN dans cette région; désigne aussi le déplacement d'une région d'ADN double brin sous l'effet d'un envahisseur simple brin dans la réaction catalysée par la protéine RecA
D-segment	segment de diversité (« diversity segment ») de la chaîne lourde de l'immunoglobuline et de la chaîne bêta du récepteur d'un lymphocyte T
enhancer	séquence en cis entraînant l'utilisation accrue de (certains) promoteurs eucaryotes et dont l'action s'exerce quelle que soit l'orientation et l'emplacement (en amont ou en aval) par rapport au promoteur
exon	région du génome codant une partie de l'ARN messager épissé; peut contenir la région 5'UTR, tous les CDS et la région 3'UTR
GC_signal	séquence GC; région conservée riche en GC, située en amont du site d'initiation des unités de transcription eucaryotes, qui peut prendre la forme de copies multiples et se produire dans les deux sens; consensus=GGGCGG
gene	région présentant un intérêt biologique, identifiée comme étant un gène et à laquelle un nom a été attribué
iDNA	AND intercalaire; ADN éliminé par l'un des types de recombinaison
intron	segment d'ADN qui est transcrit, puis éliminé par aboutement des séquences (exons) situées de part et d'autre
J_segment	segment de jonction (« joining segment ») de la chaîne légère et de la chaîne lourde de l'immunoglobuline, ainsi que des chaînes alpha, bêta et gamma du récepteur d'un lymphocyte T
LTR	répétition terminale longue (« long terminal repeat »); séquence directement répétée aux deux extrémités d'une séquence définie, du type de celle que l'on trouve dans les rétrovirus
mat_peptide	séquence codante d'un peptide mature ou d'une protéine mature; séquence codante du peptide ou de la protéine à l'état mature ou final, qui suit la modification post-traductionnelle; l'emplacement ne comprend pas le codon d'arrêt (contrairement au CDS correspondant)
misc_binding	site d'un acide nucléique fixant, par covalence ou non, un autre fragment de molécule, qui ne peut être décrit par aucune autre clé de fixation (primer_bind ou protein_bind)
misc_difference	séquence de caractérisation différente de celle qui est présentée dans l'entrée et ne pouvant pas être décrite par une autre clé de différence (conflict, unsure, old_sequence, mutation, variation, allele ou modified_base)

misc_feature	région présentant un intérêt biologique, qui ne peut pas être décrite par une autre clé de caractérisation; nouvelle caractéristique ou caractéristique rare
misc_recomb	site de toute recombinaison généralisée, spécifique d'un site ou réplicative, où se produit la cassure et la réunion de l'ADN double brin et qui ne peut pas être décrite par une autre clé de recombinaison (iDNA ou virion) ou par un autre qualificateur de clé source (/insertion_seq, /transposon, /proviral)
misc_RNA	tout transcrit ou produit de l'ARN qui ne peut pas être défini par une autre clé de l'ARN (prim_transcript, precursor_RNA, mRNA, 5'clip, 3'clip, 5'UTR, 3'UTR, exon, CDS, sig_peptide, transit_peptide, mat_peptide, intron, polyA_site, rRNA, tRNA, scRNA ou snRNA)
misc_signal	toute région contenant un signal qui commande ou modifie une fonction ou l'expression d'un gène, qui ne peut pas être décrite par une autre clé de signal (promoter, CAAT_signal, TATA_signal, – 35_signal, – 10_signal, GC_signal, RBS, polyA_signal, enhancer, attenuator, terminator ou rep_origin)
misc_structure	toute structure ou conformation secondaire ou tertiaire qui ne peut pas être décrite par une autre clé de structure (stem_loop et D-loop)
modified_base	le nucléotide indiqué est un nucléotide modifié, qui doit être remplacé par la molécule indiquée (donnée dans la valeur qualificative mod_base)
mRNA	ARN messager; comprend la région non traduite en 5' (5'UTR), les séquences codantes (CDS, exon) et la région non traduite en 3' (3'UTR)
mutation	la souche apparentée présente un changement brusque et transmissible dans la séquence, à cet emplacement
N_region	des nucléotides supplémentaires sont insérés entre des segments d'immunoglobuline réarrangés
old_sequence	la séquence présentée est la version modifiée d'une ancienne séquence à cet emplacement
polyA_signal	site de reconnaissance indispensable à la coupure d'un transcrit par endonucléase, suivie d'une polyadénylation; consensus=AATAAA
polyA_site	site d'un transcrit auquel sont ajoutés des résidus d'adénine par polyadénylation post-transcriptionnelle
precursor_RNA	ARN précurseur, c'est-à-dire tout type d'ARN qui n'est pas encore mature; peut comprendre la région coupée en 5' (5'clip), la région non traduite en 5' (5'UTR), les séquences codantes (CDS, exon), les séquences intercalaires (intron), la région non traduite en 3' (3'UTR) et la région coupée en 3' (3'clip)
prim_transcript	transcrit primaire (initial, non remanié); comprend la région coupée en 5' (5'clip), la région non traduite en 5' (5'UTR), les séquences codantes (CDS, exon), les séquences intercalaires (intron), la région non traduite en 3' (3'UTR) et la région coupée en 3' (3'clip)
primer_bind	site de fixation non covalent pour amorces dans l'initiation de la réplication, de la transcription ou de la transcription inverse; comprend les sites pour les éléments de synthèse, par exemple les amorces de l'amplification en chaîne par polymérase (PCR)
promoter	région d'une molécule d'ADN jouant un rôle dans la fixation de l'ARN polymérase en vue de l'initiation de la transcription
protein_bind	site de fixation non covalent des protéines sur un acide nucléique
RBS	site de fixation du ribosome (« ribosome binding site »)
repeat_region	région du génome contenant des unités de répétition
repeat_unit	unité d'un élément de répétition
rep_origin	origine de la réplication; site de départ pour la duplication d'un acide nucléique en vue de l'obtention de deux copies identiques
rRNA	ARN ribosomique mature; molécule d'ARN de la particule ribonucléoprotéique (ribosome) qui assemble les acides aminés en protéines
S_region	région de commutation (« switch region ») des chaînes lourdes de l'immunoglobuline; joue un rôle dans le réarrangement de la chaîne lourde de l'ADN, qui conduit à l'expression d'une classe d'immunoglobuline

	différente à partir du même lymphocyte B
satellite	nombreuses séquences répétées en tandem (identiques ou apparentées) d'une courte unité de répétition de base; nombre d'entre elles ont une composition de base ou une propriété différente de la moyenne du génome, qui leur permet d'être séparées du reste de l'ADN génomique (bande principale)
scRNA	petit ARN cytoplasmique (« small cytoplasmic RNA »); l'une des nombreuses petites molécules d'ARN cytoplasmique présentes dans le cytoplasme et (parfois) dans le noyau d'un eucaryote
sig_peptide	séquence codante d'un peptide-signal; séquence codante d'un N-terminal de protéine sécrétée; ce domaine joue un rôle dans l'intégration du polypeptide naissant dans la membrane; séquence leader
snRNA	petit ARN nucléaire (« small nuclear RNA »); l'une des nombreuses petites espèces d'ARN confinées au noyau; plusieurs snRNA jouent un rôle dans l'excision-épissage ou dans d'autres réactions de maturation moléculaire de l'ARN
source	permet d'identifier la source biologique de l'intervalle de séquence indiqué; cette clé est obligatoire; chaque entrée doit comporter, au minimum, une clé source unique couvrant la séquence tout entière; il est possible d'utiliser plus d'une clé source par séquence
stem_loop	épingle à cheveux; région d'une double hélice formée par l'appariement de bases entre des séquences contiguës (inversées) complémentaires appartenant à un même brin d'ARN ou d'ADN
STS	site de séquence étiqueté; séquence d'ADN courte et unique caractérisant un point de repère de la cartographie du génome et qui peut être détectée moyennant une amplification en chaîne par polymérase (PCR); la carte d'une région du génome peut être dressée par détermination de l'ordre d'une série de STS
TATA_signal	séquence TATA; séquence de Goldberg-Hogness; heptamère conservé, riche en A et T, situé environ 25 paires de bases en amont du site d'initiation de chaque unité transcrite par l'ARN polymérase II des eucaryotes, qui peut jouer un rôle dans le positionnement de l'enzyme aux fins d'une initiation correcte; consensus=TATA(A ou T)A(A ou T)
terminator	séquence d'ADN située soit à l'extrémité du transcrit, soit à côté d'un promoteur qui conduit l'ARN polymérase à terminer la transcription; peut aussi être le site de fixation d'un répresseur
transit_peptide	séquence codante d'un peptide-transit; séquence codante d'un N-terminal de protéine d'un organite codée par le noyau; ce domaine joue un rôle dans l'importation post-traductionnelle de la protéine dans l'organite
tRNA	ARN de transfert mature; courte molécule d'ARN (75-85 bases) qui permet la traduction d'une séquence d'acides nucléiques en une séquence d'acides aminés
unsure	l'auteur n'est pas certain de l'exactitude de la séquence dans cette région
V_region	région variable de la chaîne légère et de la chaîne lourde de l'immunoglobuline et des chaînes alpha, bêta et gamma du récepteur d'un lymphocyte T; codes applicables à la portion variable du terminal amino; peut être composée des segments suivants : V_segments, D_segments, N_régions et J_segments
V_segment	segment variable de la chaîne légère et de la chaîne lourde de l'immunoglobuline et des chaînes alpha, bêta et gamma du récepteur d'un lymphocyte T; codes applicables à la plus grande partie de la région variable (V_region) et aux quelques acides aminés du peptide leader qui subsistent
variation	souche apparentée contenant des mutations stables du même gène (par exemple : RFLP, polymorphismes, etc.) qui diffèrent de la séquence présentée à cet emplacement (et éventuellement à d'autres emplacements)
3'clip	région d'un transcrit précurseur située en position 3', qui est coupée durant la maturation moléculaire
3'UTR	région en position 3' à l'extrémité d'un transcrit mature (qui suit le codon de terminaison) qui n'est pas traduite en protéine
5'clip	région en position 5' d'un transcrit précurseur qui est coupée durant la maturation moléculaire
5'UTR	région en position 5' située à l'extrémité d'un transcrit mature (avant le codon d'initiation) qui n'est pas traduite en protéine
-10_signal	séquence de pribnow; région conservée située à environ 10 paires de bases en amont du site d'initiation des unités de transcription bactériennes, qui peut jouer un rôle dans la fixation de l'ARN polymérase;

	consensus=TAtAaT
-35_signal	hexamère conservé situé à environ 35 paires de bases en amont du site d'initiation des unités de transcription bactériennes; consensus=TTGACa [ ] ou TGTTGACA [ ]

Annexe F – Liste des clés de caractérisation concernant les séquences de protéines

Clé	Description
CONFLICT	séquences différentes selon divers documents
VARIANT	les auteurs signalent qu'il existe des variants de la séquence
VARSPLIC	description des variants de la séquence produites par une excision-épissage différentielle
MUTAGEN	site modifié à titre expérimental
MOD_RES	modification post-traductionnelle d'un résidu
ACETYLATION	N-terminal ou autre
AMIDATION	en général, au C-terminal d'un peptide mature actif
BLOCKED	groupe de blocage N- ou C-terminal indéterminé
FORMYLATION	de la méthionine N-terminale
GAMMA-CARBOXYGLUTAMIC ACID HYDROXYLATION	de l'asparagine, de l'acide aspartique, de la proline ou de la lysine
METHYLATION	en général, de la lysine ou de l'arginine
PHOSPHORYLATION	de la sérine, de la thréonine, de la tyrosine, de l'acide aspartique ou de l'histidine
PYRROLIDONE CARBOXYLIC ACID	glutamate N-terminal ayant formé un lactame cyclique interne
SULFATATION	en général, de la tyrosine
LIPID	liaison covalente d'un fragment lipidique
MYRISTATE	groupe myristate rattaché, par une liaison amide, au résidu N-terminal glycine de la forme mature d'une protéine ou à un résidu de lysine interne
PALMITATE	groupe palmitate lié, par une liaison thioether, à un résidu de cystéine ou, par une liaison ester, à un résidu de sérine ou de thréonine
FARNESYL	groupe farnésol rattaché, par une liaison thioether, à un résidu de cystéine
GERANYL-GERANYL	groupe géranyl-géranyl rattaché, par une liaison thioether, à un résidu de cystéine
GPI-ANCHOR	groupe glycosyl-phosphatidylinositol (GPI) lié au groupe alpha carboxyl du résidu C-terminal de la forme mature d'une protéine
N-ACYL DIGLYCERIDE	cystéine N-terminale de la forme mature d'une lipoprotéine procaryote assortie d'un acide gras amidé et d'un groupe glycérile auquel deux acides gras sont rattachés par des liaisons ester
DISULFID	liaison disulfure; les extrémités « FROM » et « TO » représentent les deux résidus qui sont liés par une liaison disulfure intrachaîne; si les extrémités « FROM » à « TO » sont identiques, la liaison disulfure est une liaison interchaîne et le champ de description donne la nature de la liaison réticulée
THIOLEST	liaison thiolester; les extrémités « FROM » et « TO » représentent les deux résidus liés par la liaison thiolester
THIOETH	liaison thio-éther; les extrémités « FROM » et « TO » représentent les deux résidus liés par la liaison thio-éther
CARBOHYD	site de glycosylation; la nature de l'hydrate de carbone (lorsqu'elle est connue) est donnée dans le champ de description

METAL	site de fixation pour un ion métallique; le champ de description donne la nature du métal
BINDING	site de fixation pour tout groupe chimique (coenzyme, groupe prosthétique, etc.); la nature chimique du groupe est donnée dans le champ de description
SIGNAL	séquence-signal (prépeptide)
TRANSIT	peptide-transit (mitochondrial, chloroplastique ou destiné à un micro-organisme)
PROPEP	propeptide
CHAIN	chaîne polypeptidique dans la protéine mature
PEPTIDE	peptide actif libéré
DOMAIN	domaine d'intérêt dans la séquence; la nature de ce domaine est donnée dans le champ de description
CA_BIND	région de fixation du calcium
DNA_BIND	région de fixation de l'ADN
NP_BIND	région de fixation d'un phosphate nucléotidique; la nature du phosphate nucléotidique est donnée dans le champ de description
TRANSMEM	région transmembranaire
ZN_FING	région d'un doigt de zinc
SIMILAR	similitude avec une autre séquence protéique; des informations détaillées sur cette séquence figurent dans le champ de description
REPEAT	répétition de séquence interne
HELIX	structure secondaire – Hélices, telles que les hélices alpha, les hélices 3-10 ou les hélices pi
STRAND	structure secondaire – Brin bêta, tel que le brin bêta à liaison hydrogénée ou le résidu dans un brin isolé à pont bêta
TURN	structure secondaire – Virages, tels que le virage à liaison H (virage 3, virage 4 ou virage 5)
ACT_SITE	acides aminés jouant un rôle dans l'activité de l'enzyme
SITE	tout autre site présentant un intérêt dans la séquence
INIT_MET	la séquence commence par un initiateur méthionine
NON_TER	le résidu situé à une extrémité de la séquence n'est pas le résidu terminal; appliqué à la position 1, cela signifie que la première position n'est pas la position N-terminale de la molécule complète; s'il est appliqué à la dernière position, cela signifie que cette position n'est pas la position C-terminale de la molécule complète; il n'y a pas de champ de description pour cette clé
NON_CONS	résidus non consécutifs; indique que deux résidus dans une séquence ne sont pas consécutifs et qu'il existe un certain nombre de résidus non séquencés entre eux
UNSURE	zones d'incertitude dans la séquence; sert à décrire les régions d'une séquence pour lesquelles les auteurs ne sont pas sûrs de la définition