La taxonomie du NCBI

Présentation détaillée

https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi

1. OBJET

La NCBI Taxonomie est une classification phylogénétique de l'ensemble du règne du vivant servant de référence pour indexer les bases de séquences (génétiques, protéomique, etc.). La ressource est mise à jour quotidiennement en suivant les recommandations de comités internationaux de nomenclature. La version publiée sur le SMT propose un nettoyage automatique du règne Procaryote ne conservant que les taxons pertinents pour l'interopérabilité (exclusion des échantillon environnementaux, etc.) et proposant des alignements vers d'autres référentiels tel que la CIM-11, le NCIT ou LOINC.

2. DESCRIPTION DES DONNEES

2.1. Nombre de concepts

La NCBI mise à disposition sur le SMT présente une base filtrée automatiquement de taxon procaryote. Elle contient dans sa première version 35 113 taxons.

2.2. Domaine(s) couvert(s)

La taxonomie du NCBI permet de décrire les organismes du vivant dans les bases de séquences génomiques, métabolomiques et protéomiques. La NCBI mise à disposition sur le SMT couvre le règne des procaryotes.

2.3. Contenu

La NCBI test que présenté sur le SMT est composé :

- **information terminologiques** : dénominations scientifiques, anciens noms

- **rang et arborescence taxonomique**

- **alignement vers d'autres terminologies biomédicales :** LOINC, CIM-11 et NCIt

Les alignements de la taxonomie du NCBI vers d'autres terminologies biomédicales ont été ajouté à la ressource d'origine et son proposés à des fins de recherche.

3. USAGES

3.1. Cas d'usage

La taxonomie du NCBI est utilisé dans les sciences du vivant et les laboratoires de biologie médicale pour des activités :

- de recherche
- de curation des bases local de taxons

En 2020 la taxonomie du NCBI a été intégrée au cadre d'interopérabilité comme terminologie de référence pour structurer les microorganismes dans les compte-rendu de biologie. A date (juin 2022) il n'est pas requis de l'implémenter dans le CR de Biologie.

3.2. Utilisateurs cibles

Laboratoires de biologie médicale, épidémiologiste, chercheur

4. INTERACTIONS AVEC D'AUTRES TERMINOLOGIES

4.1. Jeux de valeurs

N/A

4.2. Alignements

Tel que proposé sur le SMT, le NCBI est aligné avec trois terminologies :

- Codes LOINC (dans une relation de type « voir aussi ») ce qui permet de lier description du test et l'espace de résultat possible. Cette relation est déduite grâce au descripteur 'analyte' de LOINC qui exprime l'objet du test.
- Codes extension des organismes de la CIM-11 utilisé pour affiner la description d'une pathologie
- Concepts NCIt

Relation	Nb	Commentaire
skos:exactMatch	373	Relation d'alignement exact entre un taxon NCBI et un concept externe
skos:broadMatch	2	Code NCBI décrit un taxon plus XX
skos:narrowMatch	4	
skos:relatedMatch	4 420	Cette relation d'alignement est exclusivement utilisée pour faire le lien entre une analyse LOINC. Ces alignements sont dérivés des alignements exacts entre les alignements LOINC part analyte[1] et NCBI

5. ACCES AUX DONNEES ET OUTIL(S) DISPONIBLE(S)

La taxonomie est consultable sur le site https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi et téléchargeable https://ftp.ncbi.nih.gov/pub/taxonomy/

6. FORMAT ONTOLOGIQUE

Relation	Cardinalité	Type de données	Description
skos:notation	1..1	String	Identifiant unique du taxon

[1] LOINC part : partie descriptive d'un code LOINC permettant de décrire une des 6 dimensions d'un code LOINC. Un part analyte permet de décrire le composant à analyser pour le test.

rdfs:label	2..2	String	Dénomination officielle du taxon. Cette dénomination est typée de manière avec un code langue (en) et sans code langue avec le même contenu
skos:altLabel	0..N	String	Termes alternatifs (ancien nom ou nom usuels) en anglais extrait de la ressource NCBI
https://data.esante.gouv.fr/chu-rouen/NCBI_taxonomy/T_ATT_NCBI_RANK	1..1	String	Rang taxonomique de la ressource exprimé en anglais, extrait du NCBI
rdfs:subClassOf	1..1	URI	Relation hiérarchique entre un taxon et son (unique) parent
skos:exactMatch	0..N	URI	Relation d'équivalence exacte entre le taxon et un concept d'une autre ressource
skos:relatedMatch	0..N	URI	Relation de type « voir aussi » entre le taxon et un concept d'une autre ressource. Cette relation est utilisée pour faire le lien entre code LOINC et taxon NCBI.

6.1. Processus de filtre automatique

La taxonomie NCBI telle que proposée sur le site du *National Center for Biotechnology Information* contient plus de 2 000 0000 de taxons tout règne confondu

(Figure 1-A) dont seul 677 000 correspondent à des vrais taxons ; le reste étant gérés pour les activités de classifications de séquences (Figure 1 – B).

Dans le règne procaryote sur 6% (30 000/500 000) on effectivement un intérêt à être conserver pour l'interopérabilité. C'est pourquoi en amont la version du NCBI mise à disposition sur le SMT a été filtré automatiquement. L'extraction des taxons procaryote comme suit :

1. Extraction de la branche procaryote

2. Nettoyage de nœuds feuilles[2] en utilisant une succession de 6 filtres utilisant le rang taxonomie (suppression des « strain » ou « no rank ») et des patrons de libellé principal (une espèce avec plus de deux caractères ou contenant un caractère numérique).

Les méthodes de filtre automatique ont permis d'extraire 35 113 taxons (proches des 30 000 vrais taxons du NCBI). La méthode de nettoyage mise en place élimine les taxons sans descendant, cela garantie donc de conserver l'arborescence d'origine de la taxonomie NCBI. Le Tableau 1 présente le résultat de la méthode de filtre par rang taxonomique.

A : Contenu brute de la taxonomie (2021-11)

New Taxonomy Nodes: 2021/11/30 and before

Ranks:	higher taxa	genus	species	lower taxa	total
Archaea	725	225	12,633	361	13,944
Bacteria	8,488	4,561	467,793	45,671	526,512
Eukaryota	65,618	94,231	1,400,283	40,045	1,600,177
Fungi	6,199	7,023	161,520	4,042	178,784
Metazoa	46,838	67,065	994,752	18,159	1,126,814
Viridiplantae	8,279	16,350	198,085	15,567	238,281
Viruses	1,738	2,195	45,225	174,632	223,790
All taxa	76,603	101,214	1,943,767	260,728	2,382,311

Dates: 2021/11/30 and before

B : Contenu filtré sur les 'vrai' taxons (2021-11)

New Taxonomy Nodes: 2021/11/30 and before

Ranks:	higher taxa	genus	species	lower taxa	total
Archaea	226	225	809	0	1,260
Bacteria	1,588	4,558	23,614	923	30,683
Eukaryota	21,721	94,101	485,307	34,182	635,311
Fungi	1,517	7,023	51,931	1,493	61,964
Metazoa	14,880	66,935	253,983	17,118	352,916
Viridiplantae	3,652	16,350	165,490	15,208	200,700
Viruses	569	2,171	7,175	7	9,922
All taxa	24,123	101,057	516,891	35,112	677,183

Dates: 2021/11/30 and before

Taxa: Customize Use Default

Filters: ☑ exclude unclassified ☑ exclude uncultured ☑ exclude informal nam

Figure 1 Etat des lieux du contenu de la taxonomie du NCBI en novembre 2021[3]

[2] nœud feuille : nœud d'un arbre n'étant parent d'aucun autre nœud.

[3] https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=statistics&period=&from=&to=

Rang taxonomique	Nombre
superkingdom	**1**
phylum	**167**
subphylum	1
class	**110**
subclass	5
order	**256**
suborder	7
family	**646**
subfamily	1
tribe	2
no rank	*2944*
clade	128

Rang taxonomique	Nombre
genus	**4495**
subgenus	1
species group	92
species subgroup	26
species	**24064**
subspecies	**637**

Rang taxonomique	Nombre
forma specialis	522
isolate	504
serotype	247
serogroup	138
strain	77
varietas	25
biotype	7
pathogroup	5
forma	4

Tableau 1 : Contenu de la ressource mise à disposition sur le SMT.

7. LICENCE

L'usage de la taxonomie du NCBI est soumis à une licence CC0 1.0 Universel (https://creativecommons.org/publicdomain/zero/1.0/legalcode.fr)

8. SUPPORT

Vous avez des remarques ? Contactez ans-terminologies@sante.gouv.