N-terDB user manual

By Willy Bienvenut

I2BC

CNRS, CEA, Université Paris-sud

Contact : N-terDB_Help@i2bc.paris-saclay.fr

Version 0.6

Wednesday, 29 May 2019

Summary

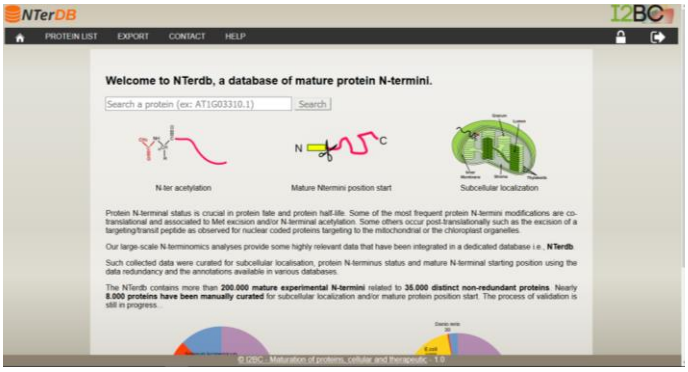
Version 0.6

Wednesday, 29 May 2019

1. Introduction

This document should provide some details and instruction to ease the access, the use and the export of the N-terDB collected, stored and curated data. This document is solely related to the N-terDB. Another document is dedicated to the N-terPred web site (n-terPred.i2bc.paris-saclay.fr) and the associated prediction tools.

2. The N-terDB ressource


NTerDB
I2BC
PROTEIN LIST
EXPORT
CONTACT
HELP
Search a protein (ex: AT1G03310.1)
Search
N
C
N-ter acetylation
Mature Ntermini position start
Subcellular localization
© I2BC - Maturation of proteins, cellular and therapeutic - 1.0

Welcome to NTerdb, a database of mature protein N-termini.

Protein N-terminal status is crucial in protein fate and protein half-life. Some of the most frequent protein N-termini modifications are co-translational and associated to Met excision and/or N-terminal acetylation. Some others occur post-translationally such as the excision of a targeting/transit peptide as observed for nuclear coded proteins targeting to the mitochondrial or the chloroplast organelles.

Our large-scale N-terminomics analyses provide some highly relevant data that have been integrated in a dedicated database i.e., **NTerdb**.

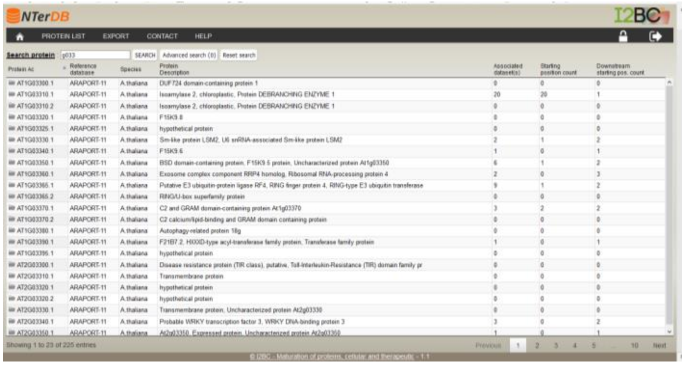
Such collected data were curated for subcellular localisation, protein N-terminus status and mature N-terminal starting position using the data redundancy and the annotations available in various databases.

The NTerdb contains more than **200.000 mature experimental N-termini** related to **35.000 distinct non-redundant proteins**. Nearly **8.000 proteins have been manually curated** for subcellular localization and/or mature protein position start. The process of validation is still in progress.

Figure 1: N-terDB home page available at N-terDB.i2bc.paris-saclay.fr.

The N-terDB web site is accessible at https://N-terDB.i2bc.paris-saclay.fr/. The starting page (Figure 1) could be reached at any time by clicking on the home symbol (⌂) on the top banner strip. This banner line also contains access to the “Protein List”, the “Export”, the “Contact” and the “Help” pages.

2.1. The N-terDB home page (Figure 1)



Protein Ac	Reference database	Species	Protein Description	Associated dataset(s)	Starting position count	Downstream starting pos. count
AT1G03300.1	ARAPORT-11	A.thaliana	DUF724 domain-containing protein 1	0	0	0
AT1G03310.1	ARAPORT-11	A.thaliana	Isoamylase 2, chloroplastic, Protein DEBRANCHING ENZYME 1	20	20	1
AT1G03310.2	ARAPORT-11	A.thaliana	Isoamylase 2, chloroplastic, Protein DEBRANCHING ENZYME 1	0	0	0
AT1G03320.1	ARAPORT-11	A.thaliana	F15K9.8	0	0	0
AT1G03325.1	ARAPORT-11	A.thaliana	hypothetical protein	0	0	0
AT1G03330.1	ARAPORT-11	A.thaliana	Sm-like protein LSM2, U6 snRNA-associated Sm-like protein LSM2	2	1	2
AT1G03340.1	ARAPORT-11	A.thaliana	F15K9.6	1	0	1
AT1G03350.1	ARAPORT-11	A.thaliana	BSD domain-containing protein, F15K9.5 protein, Uncharacterized protein At1g03350	6	1	2
AT1G03360.1	ARAPORT-11	A.thaliana	Exosome complex component RRP4 homolog, Ribosomal RNA-processing protein 4	2	0	3
AT1G03365.1	ARAPORT-11	A.thaliana	Putative E3 ubiquitin-protein ligase RF4, RING finger protein 4, RING-type E3 ubiquitin transferase	9	1	2
AT1G03365.2	ARAPORT-11	A.thaliana	RING/U-box superfamily protein	0	0	0
AT1G03370.1	ARAPORT-11	A.thaliana	C2 and GRAM domain-containing protein At1g03370	3	2	2
AT1G03370.2	ARAPORT-11	A.thaliana	C2 calcium/lipid-binding and GRAM domain containing protein	0	0	0
AT1G03380.1	ARAPORT-11	A.thaliana	Autophagy-related protein 18g	0	0	0
AT1G03390.1	ARAPORT-11	A.thaliana	F21B7.2, HXXXD-type acyl-transferase family protein, Transferase family protein	1	0	1
AT1G03395.1	ARAPORT-11	A.thaliana	hypothetical protein	0	0	0
AT2G03300.1	ARAPORT-11	A.thaliana	Disease resistance protein (TIR class), putative, Toll-Interleukin-Resistance (TIR) domain family pr	0	0	0
AT2G03310.1	ARAPORT-11	A.thaliana	Transmembrane protein	0	0	0
AT2G03320.1	ARAPORT-11	A.thaliana	hypothetical protein	0	0	0
AT2G03320.2	ARAPORT-11	A.thaliana	hypothetical protein	0	0	0
AT2G03330.1	ARAPORT-11	A.thaliana	Transmembrane protein, Uncharacterized protein At2g03330	0	0	0
AT2G03340.1	ARAPORT-11	A.thaliana	Probable WRKY transcription factor 3, WRKY DNA-binding protein 3	3	0	2
AT2G03350.1	ARAPORT-11	A.thaliana	At2g03350, Expressed protein, Uncharacterized protein At2g03350	1	0	1

Showing 1 to 25 of 225 entries

Figure 2: Protein list obtain for a basic search request using a fragment of A. thaliana protein AC (g033 instead of AT1G03310.1 for example). The list includes the expected protein (AT1G03310.1) but also all ACs fitting with the search criterion (i.e. AT2G033).

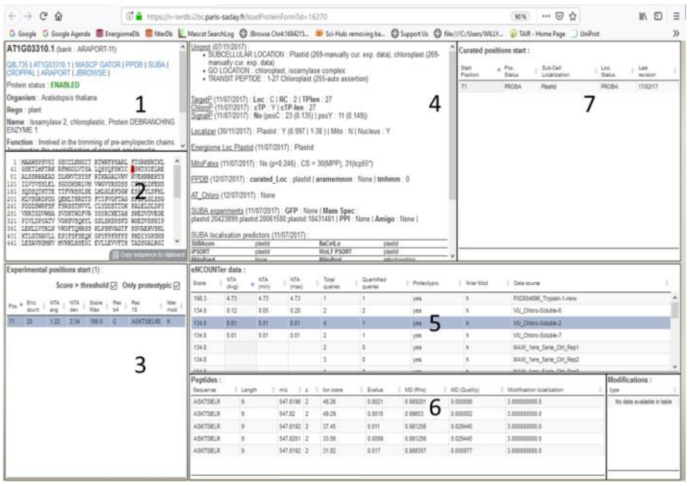
This home page provides basic knowledge of the database contents and some metrics of the presently available data. Basic protein requests based on protein accession identity, or part of it, could be performed from this page. If the protein AC is partial (*e.g.* g033 instead of AT1G03310.1), a list of all protein AC's matching this AC-fragment will be display in a novel window (Figure 2). A description of the columns is available in the next section.

2.2. The "Protein List" Page

The "Protein list" could be reach directly using the top banner strip or if a protein search is performed from the home page. The following chapters describe how to perform protein search (basic and advanced) and the search output format.

2.2.1. Simple protein AC search

Basic search, similarly to the N-terDB home page form, could be performed from this page. As previously defined, the search is solely based on the protein AC (see previous paragraph for details). This value could be the exact protein AC or just part of it (the protein AC's nomenclatures are detailed in section 2.2.3.) and the resulting list is comparable to those obtained from the N-terDB home page. The result table (Figure 2) contains a few columns directly associated to the protein reference database (*e.i.* protein AC, reference database name, species, and protein description) and the collected experimental data (e.g. number of associated datasets, number of starting positions…). Depending of the requested criteria and the size of your screen a few proteins/lines are visible in the starting (or first) page but additional pages could be available by clicking on the link at the bottom right of the page (see Figure 2).



https://n-terdb.i2bc.paris-saclay.fr/loadProteinForm?id=16270

AT1G03310.1 (bank : ARAPORT-11)

Q8L735 | AT1G03310.1 | MASCP GATOR | PPDB | SUBA | CROPPAL | ARAPORT | JBROWSE |

Protein status : ENABLED

Organism : Arabidopsis thaliana

Regn : plant

Name : Isoamylase 2, chloroplastic, Protein DEBRANCHING ENZYME 1

Function : Involved in the trimming of pre-amylopectin chains.

Uniprot (07/11/2017) :
- SUBCELLULAR LOCATION : Plastid (269-manually cur. exp. data), chloroplast (269-manually cur. exp. data)
- GO LOCATION : chloroplast, isoamylase complex
- TRANSIT PEPTIDE : 1-27 Chloroplast (255-auto assertion)

TargetP (11/07/2017) : **Loc** : C | **RC** : 2 | **TPlen** : 27
ChloroP (11/07/2017) : **cTP** : Y | **cTP-len** : 27
SignalP (11/07/2017) : **No** (posC : 23 (0.135) | posY : 11 (0.149))

Localizer (30/11/2017) : Plastid : Y (0.997 | 1-38) | Mito : N | Nucleus : Y

Energiome Loc Plastid (11/07/2017) : Plastid

MitoFates (11/07/2017) : No (p=0.246) , CS = 30(MPP), 31(Icp55*)

PPDB (12/07/2017) : **curated_Loc** : plastid | **aramemnon** : None | **tmhmm** : 0

AT_Chloro (12/07/2017) : None

SUBA experiments (11/07/2017) : **GFP** : None | **Mass Spec** : plastid:20423899;plastid:20061580;plastid:18431481 | **PPI** : None | **Amigo** : None |

SUBA localisation predictors (11/07/2017) :

SUBAcon	plastid	BaCelLo	plastid
iPSORT	plastid	WoLF PSORT	plastid
MitoPred	None	MitoProt	mitochondrion

Curated positions start :

Start Position	Pos Status	Sub-Cell Localization	Loc Status	Last revision
71	PROBA	Plastid	PROBA	17/02/17

Experimental positions start (1) :

Score > threshold ☑ Only proteotypic ☑

Pos	Enc count	NTA avg	NTA dev	Score Max	Res b4	Res 10	Nter mod
71	20	1.22	2.34	198.5	C	ASKTSELRE	H

eNCOUNTer data :

Score	NTA (Avg)	NTA (min)	NTA (max)	Total queries	Quantified queries	Proteotypic	N-ter Mod	Data source
198.3	4.73	4.73	4.73	1	1	yes	h	PXD004896_Trypsin-1-new
134.8	0.12	0.05	0.28	2	2	yes	h	VU_Chloro-Soluble-6
134.8	0.01	0.01	0.01	4	1	yes	h	VU_Chloro-Soluble-2
134.8	0.01	0.01	0.01	2	1	yes	h	VU_Chloro-Soluble-7
134.8				2	0	yes	h	MAW_1ere_Serie_Ctrl_Rep1
134.8				3	0	yes	h	MAW_1ere_Serie_Ctrl_Rep2

Peptides :

Sequence	Length	m/z	z	Ion score	Evalue	MD (Rho)	MD (Quality)	Modification localization
ASKTSELR	9	547.8196	2	48.26	0.0021	0.989201	0.000006	3.000000000.0
ASKTSELR	9	547.82	2	49.29	0.0015	0.99853	0.000002	3.000000000.0
ASKTSELR	9	547.8192	2	37.45	0.011	0.981256	0.029445	3.000000000.0
ASKTSELR	9	547.8201	2	33.56	0.0099	0.981256	0.029445	3.000000000.0
ASKTSELR	9	547.8192	2	31.82	0.017	0.988357	0.000877	3.000000000.0

Modifications :

type
No data available in table

Figure 3: The "Protein Page" for the A. thaliana protein AT1G03310.1. This page is subdivided in 7 sub-windows that contain protein details(panel 1, 2 and 4) collected from various references databases (UniProtKB, SUBA, AT_Chloro, PPDB…), the collected experimental data (panel 3, 5 and 6) and the manually curated N-termini (panel 7). Windows 1-7 are more extensively described in section 2.3.)

The first column is associated to the protein AC from its referent database, which is mentioned in the second column, as well as the species in the third column followed by the protein description in the fourth column. Columns 5-7 are related to the experimentally characterised N-termini. Column 5 highlights the number of datasets/projects associated with a single non-redundant protein whereas column 6 stated the non-redundant number of starting position and column 7 the number of downstream starting positions (starting position > Position 2). Data collected for each of the listed protein is accessible by double clicking on the corresponding protein line giving an access to the "protein page" (Figure 3).

2.2.2. Protein "Advanced search" form

A protein "Advanced search" form (Figure 4) is available from the "Protein list" page providing a few more parameters to filter the results. The available parameters are distributed in three distinct classes: "Validation", "Experiments" and "Protein description".

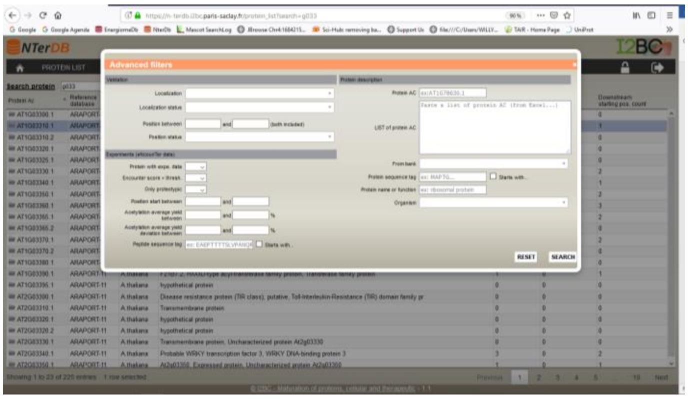


Figure 4: The advance filter table available from the "Protein list" page. This table provides filters on data quality, sub-cellular localisation, sequence…

In the "Validation" section, the filters are associated to the protein subcellular localisation ("Localisation" and "localisation status") and the experimentally characterised starting position ("Position between" and "position status"). If no filters are selected, the entire contains of the N-terDB should be available. Nevertheless, it could be interesting to filter the database available data using different options described below.

Almost 50 different subcellular localisations are defined in the N-terDB. For some of them, such as mitochondrion or chloroplast, a few sub-organelle localisations are more accurately defined. Subcellular localisation could be defined for a unique localisation, a few selected or all of them. Similarly, the starting position of the mature protein ("Position" min and "position"max) could be defined to restrict the protein N-terminus search area e.g. min = 1 and max =2 to select only the usual protein N-terminus.

Additionally, for these two main parameters, it is possible to define the level of data relevancy. Since the relevancy of the collected data could be variable depending of the protein knowledge (especially for subcellular localisation) or quality of the experimental data (for the N-terminal position), these two parameters are associated with a "quality status ("localization status" and "position status" respectively) to define the strength and le reliability of the localisation and of the experimental data. For the localisation, the defined values are True (clear sub-cellular localisation), Probable (clear localisation with minor inconsistencies), Potential (major inconsistencies in the data available and/or the literature) and Ambiguous (major ambiguities related to the exact subcellular localisation).

For the starting position, a similar rating scale is provided with the additional "Non Proteotypic" (the peptide that match also an alternative protein sequence from the same species [1]) and "False" (if the data quality are not sufficient to validate the observed starting position).

Version 0.6

Wednesday, 29 May 2019

Commenté [1]: Harmoniser localiSation et localiZation. Il semble que localiZation soit plus un terme US.

Commenté [W2R1]:

Table 1: List of the subcellular localization defined for the "advance search" form with the description (column 2) and the number of hits (column 3) for each of them in the N-terDB (April 2019 release)

Defined localisation	**Description of the "localisation"**	**Number of hit in the N-terDB**
Alternative start	Protein with unpredicted N-termini that could be due to alternative start or various splicing variants.	212
Cell junction	Protein found in or associated with a cell junction	8
Cell membrane	Protein involved in the membrane surrounding a cell	240
Cytoplasmic ribosome	protein involved in the cytoplasmic ribosome and translation	260
Cytosol	Protein involved in the cytosol	2464
Endoplasmic reticulum	Protein involved in the endoplasmic reticulum (general)	106
Endosome	Protein involved in the Endosome	44
ER Lumen	Protein involved in the endoplasmic reticulum lumen	5
ER membrane	Protein involved in the endoplasmic reticulum membrane	157
Golgi apparatus	Protein involved in the Golgi apparatus	103
Inner membrane	Protein involved in various internal membranes	143
Inner membrane	Protein involved in the prokaryotic inner cell membrane	142
Intra cellular membrane bound protein	Protein interacting with internal membranes such as membranes from the nucleus, the vacuoles…	12
Intracellular membrane-bounded organelle	Protein involved in an organized structure within the cell and involving single or double bilayer organelle membranes	11
Lysosome	Protein involved in the lysosome organelle	26
Membrane	Protein involved in the other type of membrane associated protein (general)	266
Mito/Chloro	Protein targeting both the mitochondria and the chloroplast	61
Mitochondrion	Protein targeting the mitochondrion (general)	627
Mitochondrion genome	Mitochondrial genome	4
Mitochondrion inner membrane	Protein targeting the mitochondrial inner membrane	106
Mitochondrion inter membrane space	Protein targeting the mitochondrial inter membrane space	14
Mitochondrion membrane	Protein targeting the mitochondrial membrane	20
Mitochondrion outer membrane	Protein targeting the mitochondrial outer membrane	45
Nucleus	Protein located in the nucleus	1065
NXE	Protein with unknown maturation involving the protein N-terminal sequence M[AG]X for the excision of the first two protein residues	9
Outer membrane	Protein involved in the prokaryotic outer cell membrane	19
Periplasm	Protein involved in the space between the inner and outer membrane in Gram-negative bacteria	58

Peroxisome	Protein located in the peroxisome	87
Plastid	Protein localized in the chloroplast (general)	443
Plastid envelope-inner-integral	Protein localized in the chloroplast envelope (inner-integral)	37
Plastid envelope-outer	Protein localized in the Plastid envelope (outer)	12
Plastid envelope-outer-integral	Protein localized in the Plastid envelope (outer-integral)	2
Plastid Genome	Protein associated with the chloroplast genome	38
Plastid Inner membrane	Protein localized in the chloroplast membrane (inner)	35
Plastid membrane	Protein localized in the chloroplast membrane (general)	22
plastid nucleoid	Protein localized in the plastid nucleoid	10
Plastid ribosome	Protein associated to the chloroplast ribosome	30
Plastid stroma	Protein localized in the stroma (general)	396
Plastid thylakoid	Protein localized in the thylakoid (general)	99
Plastid thylakoid-integral	Protein localized in the thylakoid membrane (integral)	41
Plastid Thylakoid Lumen	Protein localized in the thylakoid Lumen	10
Plastid thylakoid lumenal-side peripheral	Protein localized in the thylakoid lumenal-side (peripheral)	41
Plastid thylakoid membrane	Protein localized in the thylakoid membrane	31
Plastid thylakoid stromal-side peripheral	Protein localized in the thylakoid membrane stromal-side peripheral	47
Plastoglobule	Protein associated with the plastoglobule	8
secreted	Secreted protein	117
Unknown	No clear sub-cellular localization	1154
Vacuole	Vacuole membrane proteins	42

In the “Experiments” section, a few optional filters could be applied directly to the experimental data collected in the N-terDB. Then, filters could be applied to select only:

- The “Protein with experimental data”: only the protein associated with experimental data will be considered,
- The peptide with an “EnCOUTer score higher than the threshold” (the threshold score is defined automatically by EnCOUNter during the training phase, see ref [2] for details). This score defines the most relevant N-termini. Nevertheless, in some case, it could be useful to visualize also N-termini with an EnCOUNter score below the threshold…
- If the “Peptide are proteotypic” or not considering it (in any case if the peptide is not proteotypic, the NTA yield is not defined due to the origin of the peptide that could be shared between different proteins),
- With a “position start between” a “minimum” and a “maximum” sequence position in the protein sequence,

- With an “Acetylation average yield between” a minimum and a maximum value,
- With an “Acetylation average yield deviation between” a minimum and a maximum value,
- Protein with an experimentally characterised N-termini “peptide sequence tag”. It could be a sequence tag present in the protein core sequence or the characterised N-termini when “Starts with…” is selected.

In the “Protein description” section, the “Protein AC” zone allows to perform a protein search based on the protein AC (even incomplete or fractional). It is also possible to call for a list of protein using the “List of protein AC” section (one protein AC per line). The bank(s) of reference could be selected e.g. UniprotKB, ARAPORT…but also, the organism(s). The “protein Name or function” could be used as a text mining on the protein name and general description present in the protein fasta title.

Finally, it is possible to define a “sequence” that could be defined to be N-terminal (select the option “Starts with…”) or internal depending of the option selected.

2.2.3. Protein AC format used in N-terDB

A few different resources are used in the N-terDB. This include species-specific database such as ARAPORT for *A. thaliana* (https://www.araport.org/) or more general database such as UniProtKB for the *H. sapiens* reference proteome provided on the UniProtKB web site (https://www.uniprot.org/). Nevertheless, the whole UniProtKB is not available in N-terDB but only the reference proteome provided by UniProtKB for the *H. sapiens, A. thaliana, S. lycopersicum, D. rerio, S. cerevisae, E. coli.* The list of the available species depends primarily of the availability of experimental data.

Arabidopsis thaliana samples are primarily associated to the ARAPORT-11 database with the conventional TAIR AC format *i.e.,* ATXGYYYYY.Z (with X= 1-5, M or C; Y = 0-9 and Z >= 1). The *S. lycopersicum* samples are primarily associated to the Solgenomics reference database (https://solgenomics.net/) with the following format: SolycXXgYYYYYY.Z.K (with XX = 01-12; Y = 0-9; Z = 1-2 and K = 1).

For the other species (e.g. *E. coli*, *D. rerio*, *H. sapiens*…), the UniProtKB reference proteomes are used associated to the UniProtKB accession numbers. The UniProtKB accession numbers are also available when it is possible for the *S. lycopersicum* and *A. thaliana* samples.

Internally, the N-terDB define its own accession number, which is purely numerical. This type of Ac’s could be used especially to retrieve the experimental data directly using an internet link. For example, the N-terDB AC for the AT1G03310.1 is “16270” and could be call directly with the following link: https://N-terDB.i2bc.paris-saclay.fr/loadProteinForm?id=16270.

An exchange table is available in the help section of the N-terDB web page to connect the N-terDB Ac’s to the other Ac’s used in this resource.

2.3. The “Protein window”

From the “Protein list”, it is possible to access all information stored in the N-terDB for a unique protein by double clicking on therelated protein line. The “Protein window (Figure 3) is subdivided in 7 sub-windows described below.

2.3.1. Sub-window 1: Protein AC and description

This window resumes the protein identity using the AC defined in the reference database that could be different depending of the species. It contains a few interactive links including a link to the reference database and a few additional resources such as CROPAL, SUBA, NeXtprot… These links are dependant of the species and are listed in Table 1.

Table 1: List of the reference databases and the resources associated to the species presently available in the N-terDB

Species	**Reference database**	**Other resources available**	**Prediction tools used**
Arabidopsis thaliana	*ARAPORT-11*[3]	TAIR, UniProtKB[4], MASCP_GATOR[5], PPDB[6], SUBA[7], CROPPAL[8], AT_CHLORO[9], JBROWSE	TargetP[10], ChloroP[11], SignalP[12], Localizer[13], MitoFates[14]
Danio rerio	UniProtKB[4]		TargetP[10]/SignalP[12], MitoFates[14]
Escherichia coli	UniProtKB[4]		TargetP[10]/SignalP[12], MitoFates[14]
Homo sapiens	UniProtKB[4]	NeXtprot[15], PeptideAtlas[16], *h*Blat[17]	TargetP[10]/SignalP[12], MitoFates[14]
Saccharomyces cerevisae	UniProtKB[4]		TargetP[10]/SignalP[12], MitoFates[14]
Solanum lycopersicum	Solgenomics[18]	CROPPAL [8]	TargetP[10], ChloroP[11], SignalP[12], Localizer[13], MitoFates[14]

Commenté [WB3]: A préciser

2.3.2. Sub-window 2: Protein sequence

The sequence of the protein displayed in this window is the protein sequence found in the reference database. This sequence could be copied manually to the clipboard if required simply by clicking on the button located at the bottom right of the window "copy sequence to clipboard".

2.3.3. Sub-window 3: Experimental position start

The experimental data are listed in this window and sorted by peptide N-terminal position (labelled "Pos"). Each line (corresponding to a unique starting position) could contain the data from few different projects. The number of projects sharing the same starting position is define in the second column labelled "Enc. Count" (*i.e.* EnCOUNTer counts).

This table also include the average N-terminus acetylation yield ("NTA avg") and the standard deviation ("NTA dev") if more than one value is available in N-terDB which is continuously updated if additional data are provided. The EnCOUNTer score, which is used to discriminate the most relevant N-termini, is available in the "Score Max" column. Finally, the last two columns provide the residue upstream ("Res b4") of the cleavage position and the 10 first residues ("Res 10") after le cleavage position. Finally, the last column resumes the N-terminal modification determined from the MS/MS data. The value could be "H" (not acetylated), "J" (partially acetylated) or "L" (N-terminal acetylated) depending of the characterised modification group identified at the N-terminal side of the identified peptides. Such modification is defined for each starting position. As an example, if all peptides are characterised with heavy-labelled acetyl group (d3-Ac), the final modification will be "not modified" or "H". Similarly, if all peptides are characterised with an endogenous acetyl group (Ac), the final modification will be "N-terminally acetylated" or "L". If both "L" and "H" are characterised, the final labelling will be "partially N-terminally acetylated" or "J".

The results are displayed only if the EnCOUNter score is higher than the threshold and/or the characterised peptide is proteotypic. Nevertheless, by unchecking the “Score>threshold” or “Only proteotypic” options, the filtered data become available. Detail of the experimental data is shown by clicking on the line of interest and the associated data become available in the sub-window 5 (EnCOUNTer data window).

2.3.4. Sub-window 4: external prediction tools results

This windows provides as many information as possible from other sources to strengthen the subcellular localisation and/or the starting position including the manually curated UniProtKB/Swiss-Prot[4] and other resources associated to experimental data such as AT_Chloro[9], PPDB[6] or PeptideAtlas[16] depending of the species (see Table 1 for more details). Additionally, results of selected prediction tools are available including the TargetP tool suite [10,11,19] (ChloroP result is only available for plant species), Localizer[13], MitoFates[14]…

2.3.5. Sub-window 5: EnCOUNTer data

This window lists a few details and the origin of the data provided in the sub-window 3. The table provides information related to the original and EnCOUNTer processed data:

- Column 1: the “ENCOUNT score”: this is the score used by EnCOUNter to discriminate internal peptides *vs.* protein N-terminal peptides. If the score in lower than the EnCOUNTer determined threshold, the NTA yield will not be calculated;
- Column 2: NTA (Av); this is the average N-terminal acetylation yield. This value is dynamically updated depending of the available data.
- Column 3 and 4: NTA (Min) and NTA (Max): since the NTA yield available in this section is directly linked to the ratio of the heavy isoform over the light isoform (d3-Ac/Ac of H/L), the NTA is highly variable and it is not possible to use arithmetic mean but logarithmic mean. Then, NTA yield bounds are defined as minimum and maximum NTA yield.
- Column 5. “N-ter Mod”: although the NTA yield is determine from the survey MS scan, the characterised peptides could also supply a qualitative information related to the identified N-terminal group (endogenous NTA or chemical d3-acetyl group). The letter in this column resume the peptide modification, i.e. “h”: d3-Ac (or not acetylated in vivo), “l”: endogenous NTA and “j”: partial NTA. Some of the data available in the N-terDB were prepared with the dN-top approach [20], which is compatible with the EnCOUNTer processing. In this case, it is not possible to determine the NTA yield and the N-terminal modification of the peptide. Then the column remains empty.
- Column 6: “Proteotypic”: This column highlights if the considered peptide could match few different proteins (see Ref [1] for explanation). If this is not the case, the characterised peptide will be considered as “Proteotypic”. By the way, this investigation consider only protein from unique species and do not considered protein variants (from alternative start or splicing variant) as distinct proteins.
- Column 7: “Total number of query(ies)”: some projects aggregates a few raw MS files. Then, the same peptide could be found more a few times. This column highlights the number of times that the considered peptides have been characterised in the considered project.
- Column 8: “Quantified query(ies)”: this is the number of peptides that have been used by EnCOUNTer to determine the NTA yield. If there is more than 1 spectrum used for NTA yield determination, the NTA yield (column 2) should be associated with a min/max value (columns 3-4). At best, the value in column 7 and 8 are equal but usually value in column 8 is lower than in column 7.

2.3.6. Sub-window 6: MS/MS data and spectra

This sub-window provides data associated to the characterized peptides. It includes:

- The characterized peptide sequence in column 1,
- The peptide length in column 2,
- The peptide m/z in column 3,
- The peptide charge (z=) in column 4,
- The identification Mascot score in column 5,
- The peptide e-Value in column 6,
- Two Mascot distiller quality coefficients (“Rho” and “Quality” coefficients; for more details about these coefficient see Ref. [2]) used for N-terminal acetylation quantification in column 7 and 8,
- The positions and the modifications characterised on the peptide are available in column 9 using the Mascot coding scheme: Y.XXXXXXX.Z with X,Y,Z =1-9; 0= no modification, Y: N-terminus modification and Z: C-terminus modification, X: associated to the modification at a defined position in the peptide sequence). The peptide-associated modification could be made visible in the “Modifications” area at the right of the characterised peptide list by clicking on the selected line.

The MS/MS spectrum should be made available soon.

2.3.7. Sub-window 7: Curated start positions list

This subsection provides the curated starting positions. Each line details one unique starting position:

- The starting position considered call “Start Position”,
- The defined status for the starting position called “Pos. Status”,
- The protein “Sub-cell localization,”
- The defined status for the subcellular localization called “Loc. Status”,
- The date of the last modification.

The status for the starting position and the Sub-cell localisation are defined within

- True: the parameter agrees and was previously defined in alternative reference databases.
- Probable: the parameter appears highly probable compared to data available in the literature, prediction tools, reference databases…
- Potential: the parameter appears to be relevant but experimental data remain scarce as well as reference databases annotations.
- Ambiguous: although the considered parameter appears to be valuable, some experimental data, database annotation, tool prediction… does not fit at all the N-terDB data.
- Non Proteotypic: only used for the “Start Position” when the associated peptide is not proteotypic [1]
- False (not used for the Sub-cell localisation): the MS/MS spectra is of poor quality and the defined position is most likely a false “Start position”.

Although the quality criteria applied to the “starting position” and the “Sub-cellular localisation” is defined from the collected information at the time of data validation, these values could be subject to modification depending of additional data and/or reference databases collected information.

2.4. The “Export” Page “

The export page is dedicated to the experimental data selection and their export in an *.xlsx file (xml format should be available soon). The selection of the data uses a similar research form detailed in the “protein list advanced search” form described in section 2.2.2. selected data could be export using two different formats:

- By protein,
- By peptide.

For the “protein export format”, each line of the excel table correspond to one single protein whereas in the peptide export format, each line corresponds to a single non-redundant curated starting position. Then, for the peptide export a few line could be related to the same protein but different starting positions.

A few parameters could be selected and included in the final xlsx file:

- protein_N-terDB_identifier: N-terDB unique identifier code
- protein_accession: the reference database accession number
- protein_uniprot_code: The UniprotKB accession number (same as the protein_accession if UniProtKB is the reference database)
- organism_species: the sample species
- import_counts: numbers of projects where this peptide/protein was found
- peptide_counts: numbers of time the same peptide/protein was characterised
- validation_position_start: the experimentally characterised “starting position” (not available for the “protein” format)
- validation_position_start_status: the “starting position” validation status (not available for the “protein” format)
- validation_subcell_localization: the defined sub-cellular localisation (not available for the “protein” format)
- validation_subcell_localization_status: the “sub-cellular localisation“ validation status (not available for the “protein” format)
- encounter_count: number of occurrences of “peptide/protein (depending of the format selected) in the N-terDB
- encounter_score: the EnCOUNTer score used to discriminate N-terminal peptides (see ref. [2] for a description of the EnCOUNTer score calculation, not available for the “protein” format)
- encounter_acetylation_yield_avg: Average N-terminus acetylation yield (not available for the “protein” format)
- encounter_acetylation_yield_deviation: N-terminus acetylation yield standard-deviation (if more than 1 value is available; not available for the “protein” format)
- encounter_acetylation_light_count: number of times the considered peptide was characterised with an endogenous N-terminal acetylation group (not available for the “protein” format)
- encounter_acetylation_heavy_count: number of times the considered peptide was characterised with an heavy d3-acetylated N-terminal group (not available for the “protein” format)
- encounter_acetylation_heavy_light_count: number of times the considered peptide was characterised both with endogenous and heavy d3-acetylated N-terminal groups (not available for the “protein” format)

- encounter_acetylation_summary: N-terminus status considering the characterised N-terminal groups: H for not modified of free N-terminus, L for N-terminal acetylated peptide or J for partial N-terminal acetylation (not available for the "protein" format).
- prediction_suba_localization: Sub-cellular localisation provided by SUBA [7]
- prediction_energiome_plastid_loc: Sub-cellular localisation provided by Localizer [13]
- prediction_targetp_localization : Sub-cellular localisation provided by TargetP [10]
- prediction_targetp_clivage: Transit peptide length(mitochondria or plastid) provided by TargetP [10]
- prediction_chlorop_localization: plastid localisation (binary answer: Y/N) provided by TargetP [10]
- prediction_chlorop_clivage: chloroplast transit peptide length provided by TargetP [10]
- prediction_energiome_plastid_loc: Chloroplast subcellular localisation using N-terPred (binary result: Y/N)
- protein_description: description of the considered protein
- protein_sequence: whole sequence of the considered protein

An xml file should be available soon with similar option and data.

2.5. The "Contact" Page

This page includes the most relevant people to contact for help. It is strongly recommended to used the generic Email address: N-terDB@i2bc.paris-saclay.fr.

Commenté [WB4]: Demander à Jean-Pierre de la creer.

2.6. The "Help" Page

The help page contains most of the document related to the N-terDB such as posters and related N-terDB articles. You could also find the up to date user manual.

3. References

(1) Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; Kuster, B.; Aebersold, R. *Nat Biotechnol* **2007**, *25*, 125-131.
(2) Bienvenut, W. V.; Scarpelli, J. P.; Dumestier, J.; Meinnel, T.; Giglione, C. *BMC Bioinformatics* **2017**, *18*, 182.
(3) Krishnakumar, V.; Hanlon, M. R.; Contrino, S.; Ferlanti, E. S.; Karamycheva, S.; Kim, M.; Rosen, B. D.; Cheng, C. Y.; Moreira, W.; Mock, S. A.; Stubbs, J.; Sullivan, J. M.; Krampis, K.; Miller, J. R.; Micklem, G.; Vaughn, M.; Town, C. D. *Nucleic Acids Res* **2015**, *43*, D1003-1009.
(4) UniProt Consortium, T. *Nucleic Acids Res* **2018**, *46*, 2699.
(5) Joshi, H. J.; Hirsch-Hoffmann, M.; Baerenfaller, K.; Gruissem, W.; Baginsky, S.; Schmidt, R.; Schulze, W. X.; Sun, Q.; van Wijk, K. J.; Egelhofer, V.; Wienkoop, S.; Weckwerth, W.; Bruley, C.; Rolland, N.; Toyoda, T.; Nakagami, H.; Jones, A. M.; Briggs, S. P.; Castleden, I.; Tanz, S. K., et al. *Plant Physiol* **2011**, *155*, 259-270.
(6) Sun, Q.; Zybailov, B.; Majeran, W.; Friso, G.; Olinares, P. D.; van Wijk, K. J. *Nucleic Acids Res* **2009**, *37*, D969-974.
(7) Hooper, C. M.; Castleden, I. R.; Tanz, S. K.; Aryamanesh, N.; Millar, A. H. *Nucleic Acids Res* **2017**, *45*, D1064-D1074.
(8) Hooper, C. M.; Castleden, I. R.; Aryamanesh, N.; Jacoby, R. P.; Millar, A. H. *Plant Cell Physiol* **2016**, *57*, e9.
(9) Salvi, D.; Bournais, S.; Moyet, L.; Bouchnak, I.; Kuntz, M.; Bruley, C.; Rolland, N. *Methods Mol Biol* **2018**, *1829*, 395-406.
(10) Emanuelsson, O.; Brunak, S.; von Heijne, G.; Nielsen, H. *Nat Protoc* **2007**, *2*, 953-971.
(11) Emanuelsson, O.; Nielsen, H.; von Heijne, G. *Protein Sci* **1999**, *8*, 978-984.

Version 0.6

Wednesday, 29 May 2019

(12) Almagro Armenteros, J. J.; Tsirigos, K. D.; Sonderby, C. K.; Petersen, T. N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. *Nat Biotechnol* **2019**, *37*, 420-423.
(13) Sperschneider, J.; Catanzariti, A. M.; DeBoer, K.; Petre, B.; Gardiner, D. M.; Singh, K. B.; Dodds, P. N.; Taylor, J. M. *Sci Rep* **2017**, *7*, 44598.
(14) Fukasawa, Y.; Tsuji, J.; Fu, S. C.; Tomii, K.; Horton, P.; Imai, K. *Mol Cell Proteomics* **2015**, *14*, 1113-1126.
(15) Gaudet, P.; Michel, P. A.; Zahn-Zabal, M.; Britan, A.; Cusin, I.; Domagalski, M.; Duek, P. D.; Gateau, A.; Gleizes, A.; Hinard, V.; Rech de Laval, V.; Lin, J.; Nikitin, F.; Schaeffer, M.; Teixeira, D.; Lane, L.; Bairoch, A. *Nucleic Acids Res* **2017**, *45*, D177-D182.
(16) Farrah, T.; Deutsch, E. W.; Omenn, G. S.; Sun, Z.; Watts, J. D.; Yamamoto, T.; Shteynberg, D.; Harris, M. M.; Moritz, R. L. *J Proteome Res* **2014**, *13*, 60-75.
(17) Kent, W. J. *Genome Res* **2002**, *12*, 656-664.
(18) Foerster, H.; Bombarely, A.; Battey, J. N. D.; Sierro, N.; Ivanov, N. V.; Mueller, L. A. *Database (Oxford)* **2018**, *2018*.
(19) Petersen, T. N.; Brunak, S.; von Heijne, G.; Nielsen, H. *Nat Methods* **2011**, *8*, 785-786.
(20) Bertaccini, D.; Vaca, S.; Carapito, C.; Arsene-Ploetze, F.; Van Dorsselaer, A.; Schaeffer-Reiss, C. *J Proteome Res* **2013**, *12*, 3063-3070.