eNergiome and N-TerPred: novel tools to improve the prediction of plastidic and mitochondrial mature N-termini,

Willy Vincent V Bienvenut, P-A Charbit, J-P Scarpelli, T. Meinnel, C. Giglione

eNergiome and N-TerPred:

novel tools to improve the prediction of plastidic and mitochondrial mature N-termini,

W V Bienvenut, P-A Charbit, J-P Scarpelli, T Meinnel, C Giglione,
Institute for Integrative Biology of the Cell, Paris-Saclay University, France

Project overview

Protein N-terminal maturation is essential for protein activity, sub-cellular location and half-life…
Huge interest to know the exact status and position of proteins' N-termini but…

Main issues:

Experimental data are not available for all proteins…
Prediction tools are available but not always reliable…

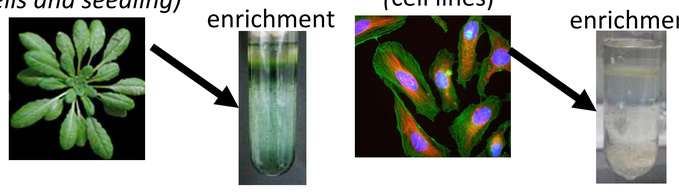
Objectives :

Collecting experimental data : the eNergiomeDB.
Data validation: based on data redundancy and manual validation
Datamining for transit peptide cleavage site prediction.

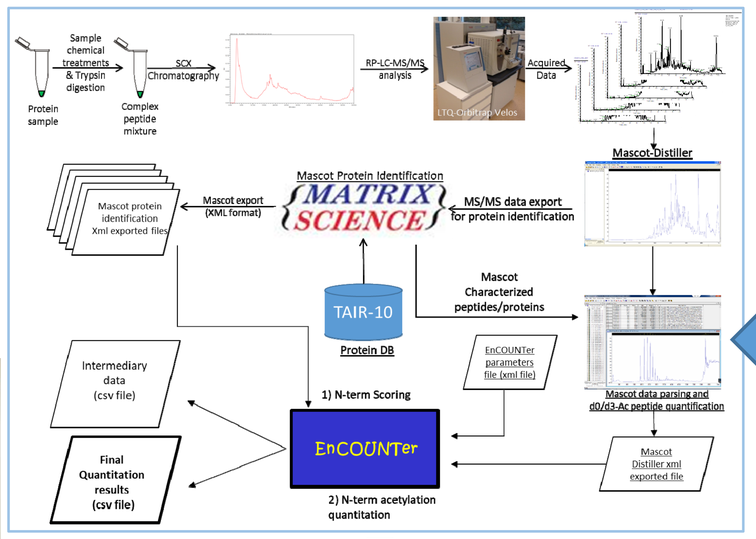
Deliverable :

N-TerPred tool suite:
protein mature N-termini and N-terminal transit peptide prediction tool

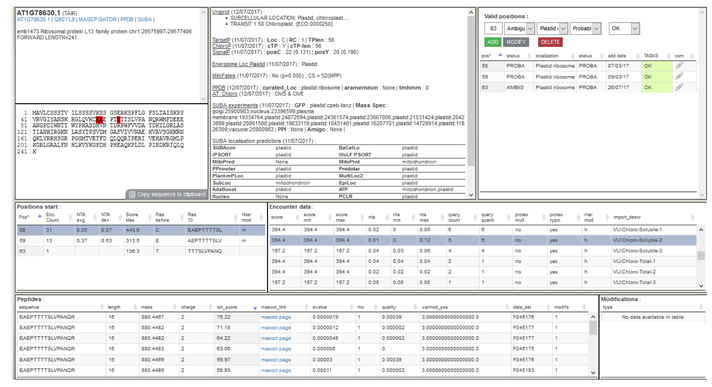
A. thaliana (cells and seedling) → Chloroplast enrichment → *H. sapiens* (cell lines) → Mitochondria enrichment



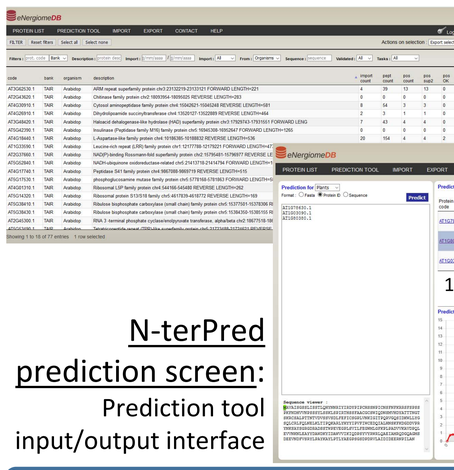
Sample preparation and data processing [1-2]



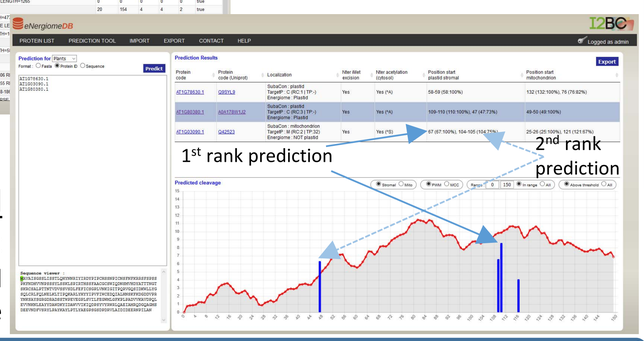
Data reprocessing :
Data collected from repositories
e.g. PXD002069 & PXD002690 [3-5]



eNergiomeDB protein screen: collected data (experimental data, prediction…) and validated mature N-termini.



eNergiomeDB main screen:
Filtered protein list selection



N-terPred prediction screen:
Prediction tool input/output interface

EnCOUNTer xml export file → Literature data (manual annotation)

https://N-TerPred.i2bc.paris-saclay.fr/

eNergiomeDB

eNergiomeDB overview

More than 10000 proteins/entries:
3000 *H. sapiens*, 6000 *A. thaliana*,
700 *S. lycopersicum*, 500 *E. coli*

More than 8500 distinct N-termini :
4500 at protein N-terminus (Pos 1-2)
of which 176/117 *Hs/At* mitochondrial proteins
and 100 plastidic proteins

4000 downstream mature N-termini including:
280/400 Mitochondrial N-term (*Hs* /*At*)
2300 plastidic (*At*) N-term

1700 N-termini used for prediction tool training :
125/225 mitochondrial (*H sapiens/A. thaliana*)
900 plastidic (*A. thaliana*)

2300 N-term quantified for Acetylation yield :
1230 at protein N-term (Pos 1&2)
1114 downstream of the protein N-term (Pos >2)

other prediction tools

Data mining based for :
* Initial Met excision prediction,
* N-terminal protein acetylation status and yield…

Plastidic localisation prediction

Randomized dataset using 1400 proteins per training round
(700 plastid/700 cytoplasmic candidates)
70% Train | 30% Test

Data mining using k-nearest neighbors Network
combining few predictors results (TargetP, WolfPSort, Predotar)

Localization Prediction Results

Plastid localization : True/False

	eNergiome-Loc (train)	eNergiome-Loc (Test)	TargetP (Test)	SubaCon (Test)
Accuracy	95.0%	**94.5%**	91.3%	94.3%
Sensitivity	93.2%	**91.7%**	90.2%	89.8%
Specificity	96.5%	97.4%	92.5%	**98.9%**
Matthews Corr. Coef.	0.90	**0.89**	0.83	**0.89**
FDR	3.5%	2.6%	7.5%	**1.1%**

Conclusion:

eNergiomeDB:
Manually Curated N-terminome data

N-TerPred toolbox:
A powerful prediction suite for
* Subcellular localization
* Transit Peptide length
(Mitochondrial/Plastid)
* N-term Met excision
* N-term Acetylation

Better N-TerPred reliability
for transit peptide prediction vs. TargetP

cTP prediction: 91% vs. 63%
Subcellular localization: 95% vs. 91.%

Combined predictions:
88% vs. 75%

Mitochondria transit peptide cleavage site

Chloroplast transit peptide cleavage site prediction

900 stromal mature N-term
70% Train | 30% Test
370 Mitochondrial mature N-term

Datamining using Position Weight Matrix

Transit Peptide length Prediction Results

2 levels of confidence defined: Top-Pred & Extended *

		N-terPred				ChloroP			
		Train dataset		Test dataset		Train dataset		Test dataset	
		Hits	%	Hits	%	Hits	%	Hits	%
Total hits		*371*	-	*160*	-	*371*	-	*160*	-
Top Pred *	**True**	**315**	**85%**	**104**	**65%**	204	55%	66	41%
	False	56	15%	56	35%	167	45%	94	59%
*Extended Pred ** *	*True*	***336***	***91%***	***115***	***72%***	*235*	*63%*	*86*	*54%*
	False	58	16%	58	36%	58	16%	58	36%

Combined Predictions LocPred/N-TerPred vs TargetP/ChloroP

	LocPred + N-TerPred		ChloroP (Loc + cTP)		TargetP (Loc + cTP)	
	Top Pred *	Extended Pred *	Top Pred *	Extended Pred *	Top Pred *	Extended Pred *
Global prediction	**85.0%**	**87.6%**	71.8%	74.6%	79.4%	82.2%
Sensitivity	**0.95**	**0.95**	0.92	0.93	0.86	0.88
Specificity	**0.82**	1.00	0.67	1.00	0.78	2.00
Accuracy	**0.85**	0.96	0.72	0.96	0.79	1.96
Matthews Corr. Coef.	**0.67**	**0.72**	0.46	0.51	0.52	0.59

*** Top-Pred**: Exp cTP within ± 2 residues of the 1st ranked prediction position.
* ***Extended*** **:** Exp cTP within ± 2 residues of the 1st or 2nd ranked prediction positions

References:

SILProNAQ sample processing:
[1] Bienvenut *et al. Methods Mol Biol* **2017,** *1574*: 17-34
EnCOUNTer parsing tool:
[2] Bienvenut *et al. BMC Bioinformatics* **2017,** *18(1)*:182
SILProNAQ/EnCOUNTer applications:
[3] Bienvenut *et al. Proteomics* **2015,** *15(14):* 2503-18
[4] Lindser *et al. Nat Commun*. **2015,** *6*: 7640
[5] Dinh *et al. Proteomics* **2015,** *15(14):* 2426-35