From early proteomics developments to protein N-terminal acetylation in plant cells

Willy Vincent Bienvenut

▶ **To cite this version:**

Willy Vincent Bienvenut. From early proteomics developments to protein N-terminal acetylation in plant cells. Genomics [q-bio.GN]. Université Paris-Sud, 2019. tel-04562905

HAL Id: tel-04562905
https://hal.science/tel-04562905

Submitted on 29 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Université de Paris-Sud

HABILITATION A DIRIGER LA RECHERCHE (HDR)

Présenté par Willy V. BIENVENUT
Chargé de Recherche de Classe Normale
Génétique Quantitative et Evolution – Le Moulon
INRA - Université Paris Sud - CNRS – AgroParisTech
Ferme du Moulon, 91190 Gif sur Yvette
France

From early proteomics developments to protein N-terminal acetylation in plant cells

Soutenance prévue le 02/12/2019, salle de l'unité, GQE – le Moulon, devant le jury composé de :

Prof. Amos Bairoch	Professeur à l'université de Genève, Suisse, Rapporteur
Dr. Véronique Santoni	Directrice de recherche à l'IBIMP, Montpellier, Rapporteur
Dr. Myriam Ferro	Directrice de recherche au CEA, Grenoble, Rapporteur
Prof. Catherine Dillmann	Professeur à l'université Paris-Sud, Examinatrice
Dr Michel Zivy	Directeur de recherche à GQE – le Moulon, Gif sur Yvette, Examinateur
Dr Joelle Vinh	Directrice de recherche à l'ENSPCI, Paris, Examinatrice

Summary

1 Curriculum vitae

1.1 Personal information

Willy V. Bienvenut

Born the 10th of August 1969 at Saintes (17, Charente Maritime), French
Senior scientist (CRCN, CNRS), section 23 (Integrative Vegetal Biology)

1.2 Education

1998-2002: Ph. D. at the Geneva University Hospital (HUGe) in the R&D section of the Central Clinical Chemistry Laboratory under the direction of Professor D.H. Hochstrasser

1996-97: DEA (Master2) in Material Science at the INP-ENSC Toulouse

1993-96: Chemist engineering degree at INP-ENSCT (Toulouse, FRANCE)
1995-1996: Erasmus student at the Greenwich University (London, UK) in nutrition and sport science
1993-1996: INP-ENSCT (Toulouse, FRANCE)

1990-92: Technical chemistry degree at the "Institut Universitaire de Technologie" (Poitiers, France)

1.3 Professional experience

Since 06-2019: Senior Scientist at GQE – le Moulon (INRA - Univ. Paris Sud - CNRS – AgroParisTech, Quantitative Genetics and Evolution, Gif sur Yvette)

- Large scale proteomics quantitation using metabolic labelling in plants
- Protein turnover and ½ life determination;
- Influence of biotic/abiotic stress, phenotypes… on protein turnover to improve plant resistance to various stresses…

2016 - 2019: Bioinformatics coordinator for SICaPS facility (I2BC, Gif sur Yvette, France).

- Management of the computing resources (hardware and software);
- Development of dedicated bioinformatics tools;

2012 - 2019: Senior Scientist at CNRS (Institute for Integrative Biology of the Cell, Gif/Yvette)

- Creation of a dedicated database (N-terDB) as a public repository for experimentally characterised mature protein N-termini;
- Development of experimental technology for protein N-terminal enrichment and data processing tools;
- Characterisation of the influence of the N-acetyltransferase in plant resistance to biotic/abiotic stress;
- Identification and characterisation of plastid specific N-acetyltransferase in *A. thaliana*;

2010 - 2012: Researcher in Biology and Mass Spectrometry at CNRS (FRC 3115, Gif/Yvette)

- Management of research projects related to the N-terminal protein acetylation and few collaborative projects developments.
- Management of technical staff associated to the facility and students for temporary trainings

2006 - 2010: Mass spectrometry & Proteomics Core Facility Manager at the "Beatson Institute for Cancer Research", Glasgow.

- Facility management and Scientific project management
- Technical and bio-informatics tools developments

2002 - 2006: Scientific collaborator at the "Protein Analysis Facility", Lausanne University, CH

- Collaborative and international projects development
- Characterisation of N-terminus acetylated proteins in collaboration with the Swiss-Prot team (A. Estreicher and A. Bairoch; SIB, Geneva, CH)

02-2002 to 05-2002: Post-doctoral position at the BPRG, Geneva University, CH

- In charge of the Swiss-2D service for protein identification by mass spectrometry

10-1998 to 01-2002: Ph. D degree at the LCCC/HCUGe, Geneva, CH

- The molecular scanner project;
- Various technological and analytical developments related to MALDI-MS analyses
- β-site for the AB-Sciex "*Proteome analyzer"* (MALDI-TOF/TOF-MS instrument)

1.4 Scientific production

1.4.1 Scientific production overview

35 articles published in peer reviewed journals
3 proceedings paper
1 letter
H index: 20 with an average of 35 citations per article.

35 communication of which 11 as an invited speaker
35 Posters

1.4.2 Articles in Scientific journals (Peer reviewed, selection)

Castrec B, Dian C, Ciccone S, Ebert CL, Bienvenut WV, Le Caer JP, Steyaert JM, Giglione C, Meinnel T. Structural and genomic decoding of human and plant myristoylomes reveals a definitive recognition pattern. *Nat Chem Biol.* **2018;** doi: 10.1038/s41589-018-0077-5.

Bienvenut WV, Giglione C, Meinnel T. SILProNAQ: A convenient approach for proteome-wide analysis of protein N-termini and N-terminal acetylation quantitation. *Meth Mol Biol*, **2017**; 1574:17-34. doi: 10.1007/978-1-4939-6850-3_3

Bienvenut WV, Scarpelli JP, Dumestier J, Meinnel T, Giglione C. EnCOUNTer: A processing tool to discover organelle-targeted proteins N-termini and to quantify associated N-α-acetylation yield. *BMC Bioinformatics*, **2017**; 18(1):182. doi: 10.1186/s12859-017-1595-y.

Linster E, Iwona S, Bienvenut WV, Maple-Grødem J, Myklebust LM, Huber M, Reichelt M, Sticht C, Geir Møller S, Meinnel T, Arnesen T, Giglione C, Hell R and Wirtz M. Proteome imprinting by N-terminal acetylation is a vital hormone-regulated master switch during stress, *Nat Commun*. **2015**; 6:7640. doi: 10.1038/ncomms8640.

Bienvenut WV, Giglione C, Meinnel T. Proteome-wide analysis of the amino terminal status of Escherichia coli proteins at the steady-state and upon deformylation inhibition. *Proteomics* **2015;** 15(14):2503-18. doi: 10.1002/pmic.201500027.

Xu F, Huang Y, Li L, Gannon P, Kapos P, Linster E, Bienvenut W, Polevoda B, Wirtz M, Meinnel T, Hell R, Giglione C, Zhang Y, Chen S, Li X. Two N-terminal acetyltransferases antagonistically regulate the stability of a nod-like receptor in Arabidopsis. *Plant Cell* **2015**; *27(5)*:1547-62. doi: 10.1105/tpc.15.00173.

Dinh TV, Bienvenut WV, Linster E, Feldman-Salit A, Jung VA, Meinnel T, Hell R, Giglione C, Wirtz M. Molecular identification and functional characterization of the first Nα-acetyltransferase in plastids by global acetylome profiling. *Proteomics* **2015;** doi: 10.1002/pmic.201500025.

Wickman GR, Julian L, Mardilovich K, Schumacher S, Munro J, Rath N, Zander SA, Mleczak A, Sumpton D, Morrice N, Bienvenut WV, Olson MF. Blebs produced by actin-myosin contraction during apoptosis release damage-associated molecular pattern proteins before secondary necrosis occurs. *Cell Death Diffe* **2013**; *20(10)*:1293-305.

Bienvenut WV, Sumpton D, Lilla S, Meinnel T, Giglione . Influence of various endogenous and artefactual modifications on large scale proteomics analysis. *Rap. Commun. Mass Spectrom.* **2013;** 27(3):443-50.

Grill B, Chen L, Tulgren E, Baker S, Bienvenut W, Anderson M, Quadroni M, Jin Y, Garner C RAE-1, A novel PHR binding protein, is required for axon termination and synapse formation in C. elegans. *J Neurosci* **2012**; *32(8)*:2628-2636.

Bienvenut WV, Sumpton D, Martinez A, Lilla S, Espagne C, Meinnel T, Giglione C Comparative large-scale characterization of plant vs. mammal proteins reveals similar and idiosyncratic N-alpha acetylation features *Mol. Cell. Proteomics*, **2012**; *11(6)*:M111.015131.

Bienvenut WV, Espagne C, Martinez A, Majeran W, Valot B, Zivy M, Vallon O, Adam Z, Meinnel T Giglione C. Dynamics of post-translational modifications and protein stability in the stroma of *Chlamydomonas reinhardtii* chloroplasts. *Proteomics* **2010**; *11(9)*: 1734-50.

von Kriegsheim A, Baiocchi D, Birtwistle M, Sumpton D Bienvenut WV, Morrice N, Yamada K, Lamond A, Kalna G, Orton R, Gilbert D, Kolch W Cell fate decisions specified by the dynamic ERK interactome. *Nat. Cell Biol.* **2009**, *11(12)*: 1458-64.

Colzani M, Bienvenut W, Faes E, Quadroni M Precursor Ion Scans (PIS) for the targeted detection of stable isotope-labelled peptides. *Rapid Commun Mass Spectrom,* **2009**, *23(22)*:3570-8.

Sumpton D, Bienvenut WV, Artefactual protein modification induced by colloidal Coomassie staining. *Rapid Commun Mass Spectrom* **2009**, *23(10)*:1525-9.

Da Cruz S, Parone PA, Gonzalo P, Bienvenut WV, Tondera D, Jourdain A, Quadroni M, Martinou JC. SLP-2 interacts with prohibitins in the mitochondrial inner membrane and contributes to their stability. *Biochim Biophys Acta*. **2008,** *1783(5)*:904-11

Grill B, Bienvenut WV, Brown HM, Ackley BD, Quadroni M, Jin Y C. elegans RPM-1 regulates axon termination and synaptogenesis through the Rab GEF GLO-4 and the Rab GTPase GLO-1. *Neuron*. **2007,** *55(4)*:587-601.

Vuadens F, Benay C, Crettaz D, Gallot D, Sapin V, Schneider P, Bienvenut WV, Lemery D, Quadroni M, Dastugue B, Tissot JD; Identification of biologic markers of the premature rupture of foetal membranes: proteomic approach. *Proteomics*. **2003**, ***3(8)***:1521-5.

Gattiker A, Bienvenut WV, Bairoch A, Gasteiger E FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification. *Proteomics* **2002**, *2(10)*: 1435-44.

Bienvenut WV, Déon C, Sanchez J-C, Hochstrasser DF; Enhanced protein recovery after electrotransfer using square wave alternating voltage**.** *Analytical Biochemistry* **2002,** *307*: 297-303.

Bienvenut WV, Déon C, Pasquarello C, Campbell J, Sanchez JC, Vestal M, Hochstrasser DH; MALDI MS-MS with high resolution and sensitivity for identification and characterization of proteins *Proteomics* **2002**, *2*: 868-76.

Müller M, Gras R, Appel R D, Bienvenut W V, Hochstrasser D F; Visualization and Analysis of the Molecular Scanner Peptide Mass Spectra. *J. Am. Soc. Mass Spectrom*. **2002**, *13*, 221-31.

Bienvenut W V, Hoogland C, Greco A, Heller M, Gasteiger E, Appel R D, Diaz J-J, Sanchez J-C, Hochstrasser D F; Hydrogen/Deuterium Exchange for Higher Specificity of Protein Identification by Peptide Mass Fingerprinting. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 616-26

Greco A, Bienvenut W, Sanchez J-C, Kindbeiter K, Hochstrasser D, Madjar J-J, Diaz J-J; Identification of ribosome-associated viral and cellular basic proteins during the course of infection with herpes simplex virus type 1. *Proteomics* **2001**, *1*, 454-9.

Gras R, Müller M, Gasteiger E, Gay S, Binz P-A, Bienvenut W V, Hoogland C, Sanchez J-C, Bairoch A, Appel R D, Hochstrasser D F; Improving Protein Identification from Peptide Mass Fingerprinting through a Parameterized Multi-level Scoring Algorithm and an Optimized Peak Detection. *Electrophorersis* **1999:** 20, 3535-50

Binz P-A, Müller M, Walther D, Bienvenut WV, Gras R, Hoogland C, Bouchet G, Gasteiger E, Fabbretti R, Gay S, Palagi P, Wilkins MR, Rouge V, Tonella L, Paesano S, Rossellat G, Karmime A, Bairoch A, Sanchez J-C, Appel RD, Hochstrasser DF; A Molecular Scanner to Highly Automate Proteomic Research and to Display Proteome Images. *Anal. Chem.* **1999**, *71*, 4800-7.

Bienvenut WV, Sanchez J-C, Karmime A, Rouge V, Rose K, Binz P-A, Hochstrasser DF; Toward a Clinical Molecular Scanner for Proteome Research: Parallel Protein Chemical Processing before and during Western-Blot. *Anal. Chem.* **1999**, *71*, 4981-8

1.4.3 Book and Chapters (selection)

Bienvenut WV, Acceleration and Improvement of Protein Identification by Mass Spectrometry. *Springer* Dordrecht, **2005.**

Bienvenut WV, Müller M, Palagi PM, Gasteiger E, Heller M, Jung E, Giron M, Gras R, Gay S, Binz P-A, Hughes GJ, Sanchez J-C, Appel RD, Hochstrasser DF; Mass spectrometry and genomic analysis; Proteomics and mass spectrometry: Some aspects and recent developments*.* JN Housby Ed.; *Kluwer Academic Publishers*, **2001**, pp93-145.

1.4.4 Proceedings (selection)

Bienvenut WV, Barblan J, Quadroni M; Peptide recovery after desalting and concentration steps using Zip Tips. Proceeding of the Swiss Proteomics Society 2002 Congress, Applied Proteomics, Lausanne, 3-5 December 2002, Ed. Fontis Media (ISBN 2-88476-003-2), **2002**, pp 158;

Nadler T, Huang Y, Wagenfeld B, Parker B, Lotti R, Vella GJ, Binz PA, Müller M, Gras R, Hochstrasser DF, Bienvenut W, Sanchez JC, Corthals G, Appel RD; Proteome Analysis by Multi-Dimensional Liquid Chromatography Gel-Electrophoresis and the Molecular Scanner. In *5th Siena 2D electrophoresis meeting,* **2002,** Siena (Italy).

Hochstrasser D, Sanchez JC, Binz PA, Bienvenut W, Appel RD; A clinical molecular scanner to study human proteome complexity. *Novartis Found Symp*. **2000**; 229:33-8; discussion 38-40.

Hochstrasser D, Appel R, Bienvenut W, Binz PA, Chiappe D, Demalte I, Jung E, Sanchez JC; Proteomic complexity and computational aspects in oncology. In *10th Symposium of the Division of Experimental Cancer Research of the German Cancer Society; Journal of Cancer Research and clinical Oncology, supplement to vol. 125, Scientific Proceeding*. **1999**; Heidelberg (Germany).

1.4.5 Patents

Bienvenut WV, Sanchez J-C, Hochstrasser DF; Kit for electroblotting polypeptides separated on an electrophoresis gel. PCT patent application No. 99 01775.8, UK patent application No. 99 07790.1**.**

Bienvenut WV, Hochstrasser D F; Method of identifying polypeptides. US patent application No. 09/107,991, European patent application: CA2244947.

1.4.6 Awards

Riotton Price (2000) of the Geneva Medical University for the article entitle “Toward a Clinical Molecular Scanner for Proteome Research: Parallel Protein Chemical Processing before and during Western-Blot. *Anal. Chem.* 1999, *71*, 4981-8

French Society of Mass Spectrometry Price (2002) for the thesis entitle: ”Accélération et amélioration de l’identification des protéines par spectrométrie de masse”.

1.5 Seminars and talks

1.5.1 Invited speaker (selected)

Bienvenut W.V. The paradigm of N-α acetylation… Journée I2BC Vert, 27th of January 2015, Gif-sur-Yvette, France

Bienvenut W; N-terminal Protein acetylation: A neglected co- and Post-translational modification. Proteomic workshop, 28th of November 2013, Montpellier (France)

Bienvenut WV, Scarpelli J-P, Espagne C, Meinnel T, Giglione C; Large-scale analyses for protein N-terminal modifications: A method of choice for protein maturation characterization. Indo French workshop -Recent trends in Proteomics, 6-8th April 2013, Bangalore (India)

Bienvenut WV, Sumpton D, Martinez A, Lilla S, Espagne C, Meinnel T, Giglione C; Large-scale characterization of plant N-terminal modifications reveals idiosyncratic N-alpha-acetylation features. Conférences Jacques Monod, 3-6 juin 2012, Roscoff (France)

Bienvenut W; SILAC: some advantages and drawbacks on one of the most famous relative quantitation approaches. European BioAlpine Convention and SPS annual meeting, 2008; Genève (Suisse)

Bienvenut WV, Müller M, Pasquarello C, Binz P-A, Corthals G, Sanchez J-C, Hochstrasser DF; Improvement and high throughput protein identification using mass spectrometric techniques: the molecular scanner. 19éme Journée Française de Spectrométrie de Masse, septembre 2002, Chaville (France)

Bienvenut WV, Müller M, Paesano S, Binz P-A, Converset V, Corthals G, Sanchez J-C, Hochstrasser DF; A new method for large-scale proteins identification: the molecular scanner. 7ème colloque de l'association des doctorants du centre de Biophysique moléculaire, May 2002, Orléans (France)

Bienvenut WV, Müller M, Pasquarello C, Paesano S, Binz P-A, Corthals G, Sanchez J-C, Hochstrasser DF; Automated proteome scanner. 17th-20th of September 2002, SIBIOC, Rimini (Italia)

Bienvenut WV, Müller M, Paesano S, Gras R, Binz, PA, Converset V, Déon C, Appel RD, Sanchez JC, Hochstrasser DF; Comprehensive proteome analysis: the molecular scanner in Congrès commun des sociétés de biochimie française et italienne. 2001, Sienna (Italia)

Bienvenut W, Heller M, Converset V, Paesano S, Binz PA, Sanchez JC, Hochstrasser DH; Clinical Molecular Scanner for proteome research. 10th of July 2000, Imperial College, London (UK)

Bienvenut W, Binz PA, Karmime A, Rouge V, Sanchez JC, Hochstrasser DF A Major Step Toward a Clinical Molecular Scanner for Proteome Research: Parallel Protein Chemical Processing followed by MALDI-TOF MS imaging. Congrès commun de la SFEAP et de la BSPR, septembre 1999, Rouen (France)

1.5.2 Posters communication at national and international meeting (selection)

Bienvenut W.V, Charbit P.-A., Scarpelli J.-P., Meinnel T., Giglione C.; eNergiomeDB and N-terPred: novel tools to improve the prediction of plastidic and mitochondrial mature N-termini. 43rd FEBS congress, Prague 7th - 12th of July 2018

Charbit P.-A., Scarpelli J.-P., Meinnel T., Giglione C., Bienvenut W.V. N-terPred: a new generation of protein N-terminus modification prediction tools. N-term Symposium, 2017, 11-13 September 2017, Halle, Germany

Bienvenut W.V., Dumestier J., Scarpelli J.P., Meinnel T., Giglione C. Extraction and statistical validation tool for the quantification of N-terminal acetylation of proteins targeting organelles. SMAP, 30 juin au 2 juillet 2014; Lyon, France

Bienvenut WV, Sumpton D, Martinez A, Lilla S, Espagne C, Meinnel T, Giglione C. Large-scale characterisation of plant *N-alpha-acetylation* reveals idiosyncratic N-alpha-acetylation features. 60th ASMS meeting, 21-24 Mai 2012; Vancouver (ON), Canada

Bienvenut WV, Espagne C, Martinez A, Sumpton D, Lilla S, Majeran W, Meinnel T, Giglione C. Comparative large-scale characterisation of plant *vs.* mammal proteins: Divergences and convergences. SMAP, , 19-22 Septembre 2011; Avignon, France

Bienvenut W, Martinez A, Sumpton D, Lilla S, Meinnel T, Giglione C; N-α-term protein acetylation: a post-translational modification of increasing interest. Congrès de la Société Française d'Electrophorèse et d'Analyse Protéomique ; 06 - 08 Septembre 2010 2010. Marseille, France

Sumpton D, Bienvenut WV; Coomassie Stains: Choices and Concerns; An evaluation of ESI-MS compatibility. 57th ASMS meeting, June 2009, Philadelphia (PA), USA

Lilla S, Bienvenut WV; Fast quality control of recombinant protein: Microwave-assisted sample processing and In-gel digestion. British Society of Proteomics Research meeting, July 2009, Hinxton, UK

Lange E, Lilla S, Sumpton D, Bienvenut, W.; Selectivity and efficiency improvement of relative protein quantification using SILAC European BioAlpine Convention and SPS annual meeting, 2008; Genève, Suisse.

Bienvenut W, Potts A, Quadroni M; Identification by tandem mass spectrometry of N-terminal acetylated proteins in e*ukaryotic* samples. Congrès commun SFCBA 2006, Montpellier, France

Kanor S, Xenarios I, Estreicher A, Quadroni M, Bienvenut WV; Characterization, identification and prediction of N-terminus acetylated proteins. 54th ASMS meeting, June 2006, Seattle (WA), USA)

Bienvenut W, Potts A, Quadroni M; Identification by tandem mass spectrometry of N-terminal acetylated proteins in e*ukaryotic* samples. 53rd ASMS meeting, June 2005. San Antonio (TX), USA

Quadroni M, Walther D, Appel R, Bienvenut W, Potts A; Comparative mass spectrometric analysis of complex peptide mixtures using a novel software tool for two-dimensional data visualization. *Congress of the Swiss Proteomics Society*. 2004. Bern, Switzerland

Bienvenut W, Barblan J, Potts A, Quadroni M; The use of precursor ion scans as survey scans in the LC-MS/MS analysis of peptide mixtures. 51st ASMS meeting, June 2003, Montreal (QC), Canada

Müller M, Gras R, Bienvenut WV, Binz PA, Pasquarello C, Hochstrasser DF, Appel RD; Using Peptide Signal Intensity Distributions for a Better Interpretation of Molecular Scanner Data. *5th Siena 2D electrophoresis meeting*. 2002. Siena, Italy

Bienvenut W, Binz P-A, Sanchez J-C, Hochstrasser DF; Molecular scanner project: Optimisation of the matrix deposition; alpha-cyano-4-hydroxy cinnamic acid solubility in various solvents. 5th Siena Meeting From Genome to Proteome: functional proteomics, Sept. 2-5 2002, Siena, Italia

Müller M, Gras R, Bienvenut WV, Hochstrasser DF, Appel RD; Visualisation and Analysis of Molecular Scanner Peptide Mass Spectra. In *Congress of the Swiss Proteomics Society*. 2001. Geneva, Switzerland

Bienvenut W, Binz P-A, Rouge V, Paesano S, Heller M, Rose K, Sanchez J-C, Hochstrasser D; Toward a clinical molecular scanner for proteome research: parallel protein chemical processing before and during western blot. Journée de la Faculté de Médecine de l'Université de Genève, Jan. 27th 2000, Cartigny, Switzerland

Binz PA, Bienvenut W, Fabbretti R, Gasteiger E, Bairoch A, Appel RD, Sanchez JC, Hochstrasser DF; The "molecular scanner": A highly automated method for protein identification and 2-D PAGE annotation. In *3rd Siena 2D electrophoresis meeting*. 1998. Siena, Italy. (P)

Bienvenut W, Binz P-A, Appel RD, Sanchez J-C, Veuthey J-L, Hochstrasser DF; Towards a molecular scanner for proteome research: parallel protein chemical processing during transblot. Congress of the Swiss Electrophoresis Society, Dec. 3rd 1998, Basle, Switzerland

Bienvenut W, Faupel M, Francotte E, Borredon E. Some new developments of weakly acid and basic acrylamide buffer molecules, Crongrés annuel de la SFE, 30 août – 2 septembre 1995, Paris (France)

1.6 Involvement for the scientific community

1.6.1 Article, Project, Ph. D. review and comities

- Reviews for *Proteomics*, *Journal of Proteomics*, *Bioinformatics, PLOS-One* and various journals of the British Society of Chemistry (*Analyst, Med Chem Comm, Analytical Methods*…)
- Review for the FNR Switzerland, ANR (France), and Czech research agency…
- Part of the thesis comity of Tassadit Ouida (Université de Rouen) as external expert
- Part of various recruitment committees (INRA, CNRS)

1.6.2 Bioinformatics Tools

Swiss-Prot collaborative development (E. Gasteiger/A. Bairoch):

FindPept, IsotopIdent, PeptCutter (http://www.expasy.org/proteomics),

Scientific/technical expert for

SIB (Geneva, CH) associated bio-informatics tools: Peptident, Aldente, Phenyx…
PROFI associated bio-informatics tools: Proline
I2BC / ProNTI group (Gif sur Yvette, France): Termi*N*ator3

Bio-informatic project manager for:

EnCOUNTer: protein N-terminal parsing script for SILProNAQ processed sample; see paragraph 3.2 for description
N-terDB: mature protein N-termini database; see paragraph 3.3 for description
N-TerPred: N-terminal and subcellular localisation tool suite; see paragraph 3.4 for description

1.6.3 Others

Member of the scientific community such as

- French society of Electrophoresis and proteomics analyses (SFEAP)
  - 2012-2016: treasurer
  - Since 2017: vice president
  - 2018: member of the scientific comity of the SFEAP annual meeting
- French Society of Mass Spectrometry (SFSM)
- American society of Mass Spectrometry (ASMS)
- French Society of photosynthesis (SFPhi)

1.7 Teaching and student management

1.7.1 Teaching activity

- **1998-2004**: **Swiss institute of bioinformatics**

Lectures: Methods for protein identification and general protein chemistry

- **2002: Workshop at the SPS congress of Applied Proteomics (3-5/12/2002, Lausanne, Switzerland)**

Mass spectrometry: turning data into results

- **2004: SIB training course (Lausanne, Switzerland)**

Proteomics Using Bioinformatics Tools

- **2003-2004: BSPR (Geneva, Switzerland) and PAF (Lausanne, Switzerland)**

Joined course in “Proteomics and mass spectrometry “

- **2016: Organisation of the 3 weeks practical training in "Fundamental microbiology" master II degree (Université Paris-Saclay, Paris XI**): immersion of the student in professional research laboratory;
- Organisation of the training (dairy, conference, computer resources…)
- Management of a group of 4 students for 3 weeks with a scientific project (i.e. characterisation of the activity of a recombinant acetylase)
- Final evaluation of the student and reports

1.7.2 Staff supervision

Beatson Institute for Cancer Research (BICR, Glasgow, Scotland):

- Managing of a team of researchers: Eva Lange, David Sumpton, Sergio Lilla and Chris Ward
- Providing state of the art proteomics expertise and analytical solutions to the BICR groups and internationally (B. Grill)
- Supervising service-related protein analyses, specific projects development
- Results and developments valorisation through publications, posters and seminars in national and international meetings

Proteomics Analysis Facility (PAF, Lausanne, CH):

- Technician joint supervision: A. Potts and J. Barblan
- Research projects development
- Result a development valorisation through posters and seminars in national and international meetings

R&D laboratory of the LCCC of the University Hospital of Geneva (HUGe):

- Management of an engineer (G. Rosselat), several technicians (A. Karmime, S. Paesano, V. Rouge) and several apprentices (Mark Riesen, Eric Estevez)
- Result and development valorisation through articles, posters and seminars in national and international meetings

1.7.3 Students supervision

Master Students

2017-2018: Melanie Bilong (I2BC-CNRS, Master 2 in Biochemistry and molecular biology)

- Characterisation of cytosolic protein N-terminal acetylation in different species and determination of the influence of specific NatA inhibitors

2015-2016: Mohamed Miali (I2BC-CNRS, Master 2 in Microbiology)

- Characterisation of multiple plastidic N-acetyl transferases in the *A. thaliana* model plants to define the influence of these gene in chloroplast protein maturation mechanisms

2013-2014: Vincent Jung (ISV, Gif, Master 2 in Proteomics)

- Identification and characterisation of the activity for selected chloroplast N-α acetylases

2013-2014: Johan Dumestier (ISV, Gif, Master 2 in Bio-informatics)

- Data processing tool development to extract and validate statistically the data for protein N-terminal acetylation quantitation.

2010-2009: Samuel Croset (BICR, Glasgow, UK, master 2 in Bio-informatics)

- MaRMot, a tool for experimental Multiple Reaction Monitoring (MRM) transitions from design Mascot repository;

2005-2006: Samuel Kanor (PAF, Lausanne, CH, master 2 in bioinformatics):

- Enrichment of proteins N-a-terminal acetylation and their predictions (Master thesis manuscript available at http://www.mpb.unige.ch/reports/rep_Samuel_Kanor.pdf)

2001-2002: Eric Antezana (BPRG/Swiss-Prot, Geneva, CH, master 2 in bioinformatics**)**

- Development of the "IsotopIdent" tool available on the ExPASy (www.expasy.org/proteomics)

2000-2001: Ulrich Wagner (BPRG/Swiss-Prot, Geneva, CH, master 2 in bioinformatics)

- Development of the "PeptideCutter" tool available on the ExPASy (www.expasy.org/proteomics)

1999-2000: Alexander Gatiker (BPRG/Swiss-Prot, Geneva, CH, master 2 in bioinformatics)

- Development of the "FindPept" to tool available on the ExPASy (www.expasy.org/proteomics) and published in a *Proteomics* article (Gattiker, A., et al.; *Proteomics* **2002**; 2: 1435-1444).

Other Students

2015: Gautier Bernal (I2BC, CNRS, Licence Pro)

- Identification and characterisation of the activity for selected chloroplastic N-α acetylases.

1.8 International collaboration and Projects

Collaboration with T Arnesen:

- Validation of Nat inhibitors in *D. rerio using the* SILProNAQ methodology

Project leader of the eNergiome project (ANR programme Blanc Ed. 2013):

- National collaboration for the characterisation of the mature N-terminus of organelles proteins (ANR-Blanc)
- Project coordination involving 3 teams located at Grenoble (T. Rabilloud), Strasbourg (C. Carapito) and Gif/Yvette (W. Bienvenut)

International collaboration with M. Wirtz group (Heidelberg) and L. Xin (Group (Vancouver)

- Protein N-terminal characterization during biotic and abiotic stresses for a better understanding of this modification in cell fate

Long-term collaboration with B. Grill (USA)

- Resulting in three articles in international peer reviewed journals including an article related to the GLO-1 protein in *C. elegans* and an essential protein kinase anchored in the membrane through a Myristoyl group at the protein N-terminus

Long-term collaboration with the Swiss-Prot group (Geneva, Switzerland)

- Experimental data annotation related to protein N-terminal acetylation (almost 1000 entries) especially for *H. sapiens* and *A. thaliana* organisms (Coll. A. Estreicher, M. Tognolli, D. Lieberherr and M. Schneider)

2 Past and present research activity

2.1 Introduction

Although the term "proteome" was proposed quite recently in 1994 [1], the interest for protein isolation and characterisation begins long before. Protein separation using gel electrophoresis (known as SDS-PAGe separation) was originally proposed by Laemmly [2] in 1970. Resolution power of this technique was strongly improved when MacGillivray and Rickwood [3] as well as O'Farrell [4] proposed two-dimensional separation (2-DE) that combines protein isoelectric focussing (IEF) and SDS-PAGe separation. At first, IEF was performed with buffering molecules (Ampholine®) embedded in the polyacrylamide gel matrix. In 1982, Bjellqvist *et al.* [5] improved the resolving power of this method using buffering molecules covalently bound to the gel matrix. Soon after, Görg *et al.* applied such technological improvement to 2DE separation [6]. Nevertheless, one of the major bottleneck of the IEF separation technique was the limited number of buffering molecules (known as "Immobilines"™) used for the preparation of the "immobilised pH gradient" (IPG) strips.

During my technical degree at the "Institut universitaire de technologie" in Poitiers, I performed two temporary trainings at the Ciba-Geigy research centre (Basel, Switzerland) aiming to synthetize new acrylamido-buffering molecules additionally to the commercially available ones. These new molecules were reported for the first time at the annual meeting of the French Electrophoresis Society congress in 1995 (Paris, France). The interest raised by this work led me to start a master project on the separation of basic protein using IEF technology in Prof. D.H. Hochstrasser's team (HUG, Switzerland). In 1997, I started a Ph. D degree associated to the "*Molecular Scanner* project" [7].

2.2 The *Molecular Scanner* project

2.2.1 Introduction

In the 90's, the National Institutes of Health in USA started large-scale sequencing projects to decipher the coding sequence of multiple small genome organisms, such as *E. coli* [8], *C. elegans* or *S. cerevisae*. In 1995, the first genome of a living organism (*H. influenza*) was published by "The Institute of Genomic Research" (TIGR) [9]. Although the translation of the coding sequences was publicly available, it was still unknown if the predicted proteins were actually existing as such and if they were modified. Protein modifications are crucial for activity, but not predictable. At this stage, experimental protein identification was mandatory to confirm and complement the results obtained by sequencing at the nucleotide level. In this context, the development of protein separation techniques and bioinformatics tool dedicated provided the basis of the *Molecular Scanner* project [10].

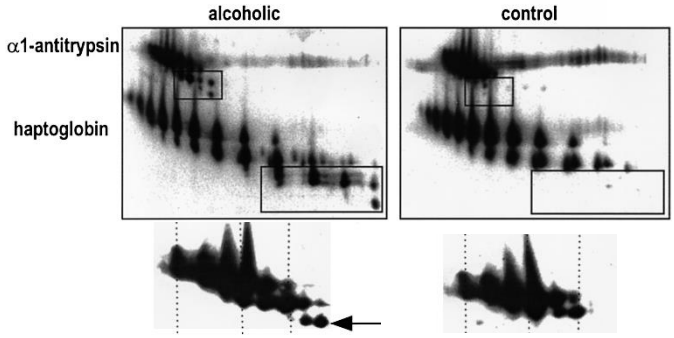


Figure 1: *Patterns of transferrin, a1-antitrypsin and haptoglobin detected by 2-D PAGE of serum proteins from alcoholic and control subjects, followed by immunoblotting and revealed by chemiluminescence; Adapted from [11].*

The original idea of the *Molecular Scanner* project was mainly 2-DE separation orientated. This analytical technique allowed to separate few hundreds to thousands of proteins in a relative short time frame and could be applied to various samples. As an example, Figure 1 shows the changes in serum

protein profile between alcoholic and control individuals [11]. In parallel, image processing and associated bioinformatics tools [12-14] were developed such as MELANIE (Medical Electrophoresis Analysis Interactive Expert) [15] or ExPASy (Expert Protein Analysis System) web resource launched in 1993 [16].

Originally, Edman sequencing was the only technique for protein identification, but this low throughput approach (1 or 2 proteins per day) was not compatible with 2-DE separation (100 to 1000 proteins per gel). In addition, the chemical reactions used in Edman sequencing was not compatible with N-terminally modified protein such as N-terminal acetylation (NTA) [17]. Consequently, higher throughput alternatives were developed such as protein Amino Acid Analysis (AAA) or Peptide Mass Fingerprint (PMF).

In AAA, the proteins were first hydrolysed and the resulting amino acids were quantified after HPLC separation. This method allowed the processing of 50-100 samples a week [18] and protein identification reliability was dramatically improved when AAA was combined with partial Edman sequencing (restricted to the first few residues). Such analytical methods triggered the development of dedicated software tools such as AAcompIdent [19].

An alternative method, PMF [20-24], combined protein endoproteolytic digestion (usually with trypsin) followed with the accurate measurement of the resulting peptides by Matrix Assisted Laser Desorption/Ionisation Mass Spectrometry (MALDI-MS) [25, 26]. The endoproteolytic peptides were embedded in α-cyanohydroxycinnamic (ACCA) acid which absorbs the energy from a UV (or IR) laser beam. The laser irradiation vaporized the matrix and promoted the production of peptide ions in gas phase. These singly charged ions [27] were first extracted from the source, accelerated [28] and then analysed using Time of Fight (TOF) analysers. The accurate molecular weight (MW) of the endoproteolytic protein fragments were compared to those generated *in silico* from protein sequences available in databases (DB) such as UniProtKB/Swiss-Prot [29, 30]. This technique offers a convenient and rapid way to identify hundreds of proteins per day.

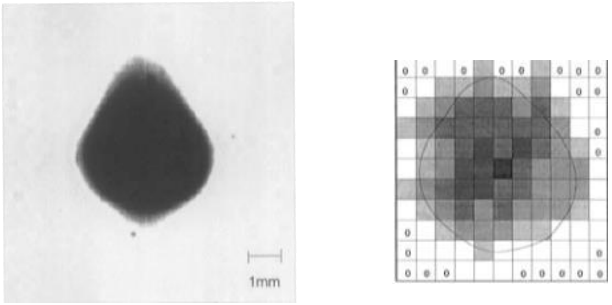


Figure 2: Optical image of the transthyretin (TTR) protein blotted on PVDF membrane after 2-DE separation and image generated from the data acquired by IR-MALDI scanning. For the mass spectrometry experiment, the investigated blot area (TTR spot) was divided into an array of 121 (11 × 11) 500 × 500 µm² squares and the signal intensity averaged over 10 spectra for each square. The contour line of the UV laser densitograph was transferred into the 2D grey-scale histogram for ease of comparison (adapted from [31]).

Our aim within the *Molecular Scanner* project was to use MALDI-MS as an imaging system after 2-DE separation. The MS acquired signal provided data for protein identification by PMF, as well as the 2D localisation of the sample. 2-DE gel were not compatible with MALDI-MS instruments (gel size, thickness, composition…) and the gel-separated proteins had to be first transferred onto a polyvinylidene difluoride (PVDF) membrane. The first imaging attempt of this type was conducted on a MALDI-MS instrument equipped with UV laser [32]. Soon after, IR lasers was used and provided the first MALDI image of a 2-DE separated protein [33] (Figure 2). This approach was able to determine the protein MW, but the measurement was not accurate enough to allow unambiguous identification of

the protein. Then, it appeared that unambiguous protein identification required additional information.

2.2.2 Project development

2DE imaging was a key element in visualizing the complexity of a sample. Originally, the *Molecular Scanner* project was solely based on the MELANIE visualization and differential quantitation software[7]. Nevertheless, such images processing tool was not able to provide protein identity but required additional investigations using AAA, Edman sequencing or PMF. Results of these analyses were manually annotated on the 2-DE image [34]. Then, the revisited *Molecular Scanner* project had the objective to automate proteins identification [35] jointly with 2-DE image annotation [36-38].

2.2.2.1 Sample preparation

MS imagining was proven to be compatible with MALDI-MS analyses [32], however the identification of the proteins was still impossible. Since PMF provided an efficient way to identity proteins, the combination of MALDI-MS imaging and PMF seemed to be the most sensible way to reach our objectives.

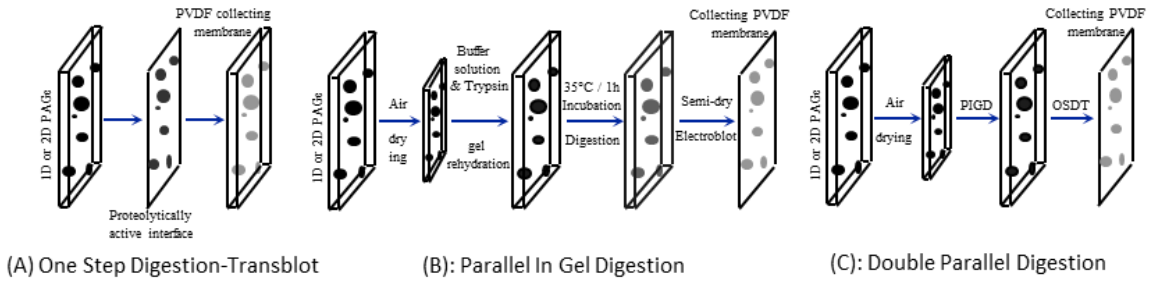


***Figure 3**: Schematic summary of the 3 proposed parallel digestions: (A) OSDT, (B) PIDG and (C) DPD.*

The challenge was to develop a way to endoproteolytically treated 2-DE separated proteins to produce peptides compatible with PMF protein identification. The collected MALDI-MS signal could be used for protein identification by PMF and to create an image based on MALDI-MS signal intensity. One of the major challenges was to conserve the bi-dimensional distribution of the proteins during protein electroblotting combined with the endoproteolytic treatment.

The first attempt to perform large-scale proteins digestion at the gel level was derived from in-gel digestion technique [39] (Figure 3A). The air-dried gel was rehydrated in a Tris-HCl buffer containing trypsin and incubated at 37°C for one hour. This “Parallel In Gel Digestion” (PIGD) provided a proof of concept, but unfortunately the spatial distribution of the proteins was negatively affected. Furthermore, low MW proteins that generate few peptides were generally lost.

In a second approach (Figure 3B), proteins digestion called “One-Step Digestion-Transfer” (OSDT) was performed during the electroblotting step. A hydrophilic membrane with crosslinked trypsin was intercalated between the gel and the collecting membrane. For this approach, a square shape voltage was applied during the electroblot to increase the interaction time between the protein and the immobilized trypsin [36, 40]. Finally, the tryptic peptides were collected on the hydrophobic membrane. Although successful for most proteins, this approach has the drawback of low transfer efficiency for high MW proteins.

In the third option, the “Double Parallel Digestion” (DPD) combined both the PIDG and OSDT (Figure 3C). Fast PIDG (only few minutes of protein digestion) truncated proteins into large polypeptides facilitating the electrotransfer. Then, during the electroblotting step, the large polypeptides were further digested through the trypsin active membranes. This approach provided the best compromise

between transfer efficiency, relevant protein digestion and protein bi-dimensional resolution (Figure 4) [36].

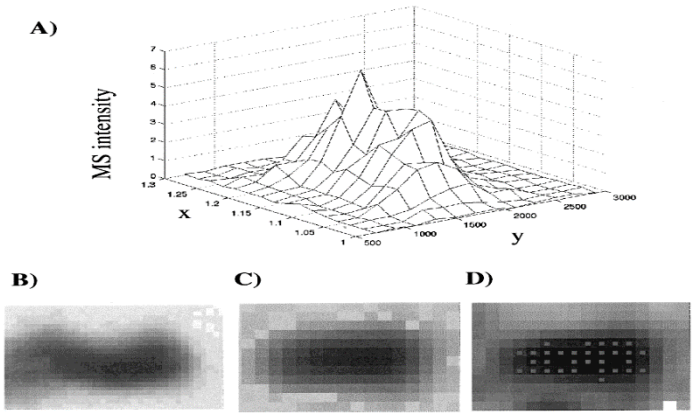


Figure 4: Two-dimensional MS scan of the 1-DE of soybean trypsin inhibitor (ITRA) band blotted on a PVDF membrane (1.1 × 0.9 cm^2) and sprayed with a 10 mg/mL ACCA solution in 70% methanol. An array of 16 × 12 points was defined around the centre of the band. (A) 3-D MS intensity profile. All m/z higher than 1100 Th were considered to create the smoothed image. (B) Amido black-stained image. (C) MS intensity image. (D) MS intensity image, plotted in a logarithmic scale. The white dots represent the positions where the ITRA was unambiguously identified with a minimum of 5 m/z matching values with PeptIdent. (Adapted from [38])

2.2.2.2 Data acquisition and processing

Usually, liquid samples analysed by MALDI-MS are mixed with the appropriate matrix and irradiated to generate ions in gas phase. For the *Molecular Scanner* project, the digested proteins were bound to the PVDF membrane and the peptides should be ionized directly from this surface. Then, the matrix required for the peptide ionisation should be sprayed uniformly on the whole membrane surface then, introduced in the MALDI-MS instrument stuck on the sample plate.

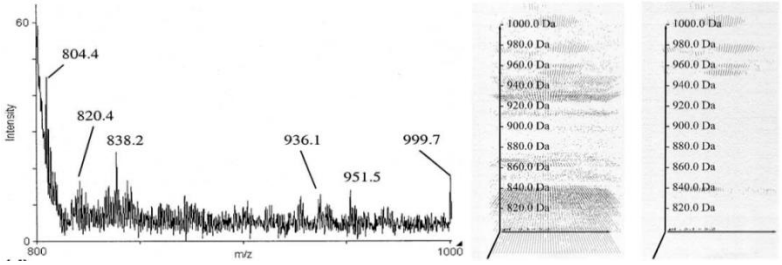
The membrane was scanned automatically and data were collected at each defined position on a defined virtual grid (75 x 500 um steps) [36, 38]. The collected MALDI-MS spectra were used both for protein identification by PMF and for an estimation of the amount of material at each scanned position (Figure 4 and 6). It rapidly appeared that the amount of data generated during each experiment could not be handled manually and required the development of an integrated visualisation tool. A dedicated PMF protein identification tool that could extract automatically the peptide signal from raw MALDI-MS spectra and provide robust protein identification was also required.

2.2.3 The bioinformatics development and imaging tool

Although single and isolated MALDI-MS spectrum could provide enough information for protein identification, the aim of the *Molecular Scanner* was to take advantage of the bi-dimensional distribution of the proteins to improve and strengthen their identification. Then, acquired spectra were used to generate a virtual image (Figure 4 and 5) to generate a grey scale picture based on the collected signal intensity. This visualisation provides an overview of the peptide distribution on the membrane surface. Randomly distributed noise signal on the whole surface of the membrane was interesting to correct the lack of homogeneity of the surface inducing m/z variability. These data were used to perform post-acquisition data recalibration and removed later on during the protein identification by PMF.

Although few PMF identification tools were available [21, 41-43], they appeared to be incompatible with *Molecular Scanner* data. These tools used only few carefully selected m/z values from the MALDI-MS spectrum, whereas the *Molecular Scanner* automated approach provided hundreds of m/z values

that strongly increase the number protein false positive identification. A second drawback was related to the identification of the peptide monoisotopic mass that was not always accurate and frequently assigned to the second (^{13}C) isotope. Finally, most available tools required highly accurate peptide m/z values to avoid erroneous protein identification.



(A) (B) (C)

Figure 5: (A) portion of a spectrum (m/z range of 800-1000 Da) from a collecting membrane obtained from a 2-DE separation of E coli. (B) 3-D representation of the acquired data for the pI range from 5.1 to 5.2 on the x axis and the MW range from 45 to 35 kDa on y axis. Background noise signal is clearly and homogeneously distributed all over the surface such as 804.4 Th peak visible in (A). (C) Remaining signal after background noise suppression highlighting well localized signal such as peak 999.7 Da. (Adapted from [44])

To solve these previously listed drawbacks, SmartIdent [45] was developed to perform peak detection and to develop a relevant scoring scheme allowing the discrimination between relevant peptide signals and background noise. The first part of the algorithm was designed to perform MALDI-MS signal detection and to automatically define a robust peak list even for low intensity or overlapped signals. The second part of the algorithm compared the experimental data to the *in silico* digested proteins available in a selected DB. The algorithm output consists of a list of potential protein hits. Finally, a discriminant score was indicated for each protein to rank them from the most to the less relevant. Few parameters impacted the final scoring scheme and the weight of each of them was defined using a genetics algorithm based on a training dataset. One of these scoring parameters took advantage of the peculiar distribution of the peptide MW errors compared to random matches. This aspect has few noticeable advantages: i) MALDI-MS spectrum did not require to be calibrated before processing and ii) the correlation coefficient of the matched masses *vs.* error deviation was a major element in the final protein identification score.

The development of the *Molecular Scanner* also required a visualisation tool to manage MALDI-MS data and the associated protein identification to provide an interactive image of the sample (Figure 6C) [46]. Since a single protein was not usually identified on a single MALDI-MS scanning position, the clustering methodology took advantage of neighbouring spectra. This approach has the advantage to be based on the filtered MS profile and not on protein identification. Then, an unidentified MS spectrum could be aggregated to a protein/peptide cluster even if the protein was not identified during the PMF identification step. Furthermore, data redundancy could be used to improve and to curate the peak list uploaded for protein identification. This multidimensional visualisation could also be used to characterize unmatched modified peptides. As an example (Figure 6D), the 2-D distribution of the MALDI-MS signal within 2025 ± 0.7 Th. (Thomson) correlated with the phosphoglycerate kinase (PGK or PGK_ECOLI) characterised peptides. Based on mass difference, the FindMod tool [47] proposed a double oxidation of the tryptophan residue (N-formyl kynurenine modification [37]) of the TILWNGPVGVFEFPNFR peptide.

2.2.4 Improvement of protein identification relevance

One of the major drawbacks of PMF protein identification is the lack of information related to the peptide sequence. This could lead from minor peptide assignment ambiguities to major protein identification errors. Since PMF was the main protein identification method used for the *Molecular*

Scanner project few improvements of this technology were investigated to strengthen the final protein identification.

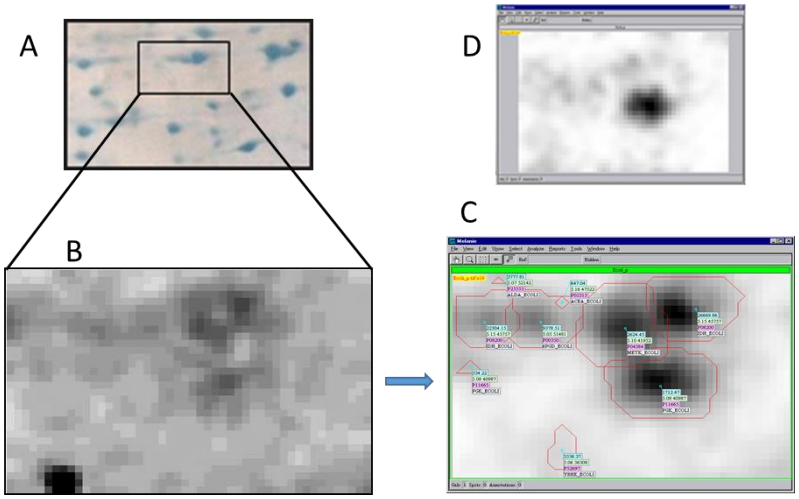


***Figure 6**: (A) 2-DE separation of an E. coli sample blotted on a PVDF membrane and amido black stained. The 2-DE gel was prepared in duplicated and treated by DPD. A 9 x 13 mm area was cut from the collecting membrane (pI from 5.1 to 5.2, MW from 35 to 45 kDa) and scanned every 300 mm by MALDI-MS. (B) The 1536 spectra obtained were automatically treated for peak detection, calibration and used to recreate the MS intensity image. (C) The image is enriched with PFM protein identification results and proteins spot are defined. (D) Image generated with the intensity of MALDI-MS signal within 2025±0.7 Th. The region where these masses are detected corresponds to the PGK_ECOLI spot.*

The easier way to access to the primary sequence of a defined peptide is usually to perform tandem MS (MS^2). At that time, such information was mostly restricted to Electro-Spray Ionisation (ESI) MS instruments, but Applied Biosystems (now ABSciex) provided the first MALDI-MS/MS instrument with real tandem MS capabilities in the early 2000's: the 4700 Proteomics Analyzer™ [37]. The additional data provided by tandem MS analysis were especially relevant for the discrimination of isobaric peptides as observed during the analysis of the bovine creatine kinase protein (UniProKB accession Q9XSC6) where a single MALDI-MS peak fits the molecular mass of two nearly isobaric peptides (GGDDLDPNYVLSSR and LSVEALNSLTGEFK, 1507.7 and 1507.8 Th., respectively). Taking advantage of the tandem MS capacity of this instrument, both sequences were validated with specific MS^2 fragments. Similarly, some peaks observed in the initial MS spectrum cannot be matched to the identified protein. Based on the collected fragments, it was possible to assign the peptide to the correct protein using *de novo* sequencing and to identify an unexpected modification related to multiple oxidation of the tryptophan residues (formyl kynurenine and derivatives) present in these peptides. Unfortunately, the high laser fluence triggered some charging effect. This problem was later solved by coating gold at the surface of the PVDF membrane easing tandem MS acquisition but unfortunately this approach leaded to major signal suppression [48].

An alternative approach was developed using the hydrogen/deuterium exchange on tryptic peptides [49]. Often used to characterise protein surfaces, this approach is based on the property of the amide groups to exchange hydrogen atoms for deuterium atoms. This approach was especially helpful to validate some protein identification based on a limited number of peptides and could ultimately be applied to the PVDF collecting membrane. Nevertheless, this approach requires intensive data processing and is difficult to use with the *Molecular Scanner* approach.

2.2.5 Conclusion

Proteomics developments in the late 90's triggered the development of the *Molecular Scanner* project. This project takes advantage of the development of new technologies such as bioinformatics tools, large-scale data processing, curated protein DB, MS instrumentations… Although these developments were compliant with 2-DE separation but this approach was time-consuming and hence has been progressively replaced by on-line chromatographic separation (such as MudPIT or HILIC…) which were easier to use for large scale protein investigations and compatible with ESI-MS separation.

Nevertheless, some of the work associated to this project was essential in the proteomics field. This is the case for MS-imaging, which is now a technique of choice to screen small molecules (*e.g.* drugs, metabolites), peptides, lipids and proteins in small model animals and various tissues. These previous investigations triggered some technological development such as automated matrix deposition or the development of dedicated instrumentation. On the bioinformatics side, the SmartIdent tool [45] remains a reference in the development of peak detection, data processing and protein identification algorithms with more than 100 citations in the field.

2.3. Protein N-terminal acetylation project

2.3.1 Introduction

From the 70's to the 90's, most protein chemistry was devoted to the characterisation of protein primary sequence using Edman sequencing and protein modifications were usually neglected if not considered has a handicap (*e.g.* protein with blocked N-termini) [50].

In late 90's, the development of 2-DE proved that a single gene could be associated to different isoforms [51]. These distinct proteoforms [52] derive from various maturation processes such as initiator methionine (iMet) removal, transit or signal peptide cleavages and a large number of *in vivo* and *in vitro* co- and post–translational modifications [53, 54]. Presently, 644 distinct modifications are annotated in UniProtKB/Swiss-Prot and only a handful of them, including phosphorylation, glycosylation, lysine or N-terminal acetylation (NTA), have been extensively studied by the scientific community.

2.3.2 An overview of the N-terminal acetylation world

mRNA translation by the ribosome is the first step for protein expression, but some additional maturation processes are required to produce functional proteins. Proteins co-translational maturation processes occur at the early stage of protein translation and mainly target protein N-terminus including various proteolytic events *e.g.*, N-terminal methionine excision (NME), N-terminal blocking reactions (*e.g.* NTA) and ligation reactions (*e.g.* N-ubiquitination [55]). After NME [56], protein NTA is probably the second most common protein modification.

If NTA is usually associated with histone, NTA is also a major modification of the N-terminal (or the N-alpha) amino group of the proteins. This modification is catalysed by N-terminal acetyltransferase (Nats) belonging to the superfamily of the "general control non-repressible 5 (GCN5)-related N-acetyltransferases" (GNAT) and requires the transfer of an acetyl moiety from an acetyl-CoA to the N-α/ε-amino group of a protein.

NTA is present in all organisms and the proportion of N-terminal acetylated (NTAed) proteins tends to increase with organism complexity [57, 58]. It affects only a few bacterial proteins [59], 10-20% of the proteome in archaea [60, 61], 40-50% in yeast [62], and 70-90% of human [63] or *A. thaliana* [64] cytosolic proteins. In *Archaea*, experimentally characterised substrates suggest a distinct acetylation pattern targeting preferentially serine and alanine residues uncovered after NME. This type of activity is well conserved in eukaryotes with six N-α-acetyltransferase (Nat) complexes identified,

including N-α-acetyltransferases A (NatA), NatB, NatC [65], NatD [66, 67], NatE [68, 69] and NatF [70] which all exhibit distinct substrate specificity (Figure 7).

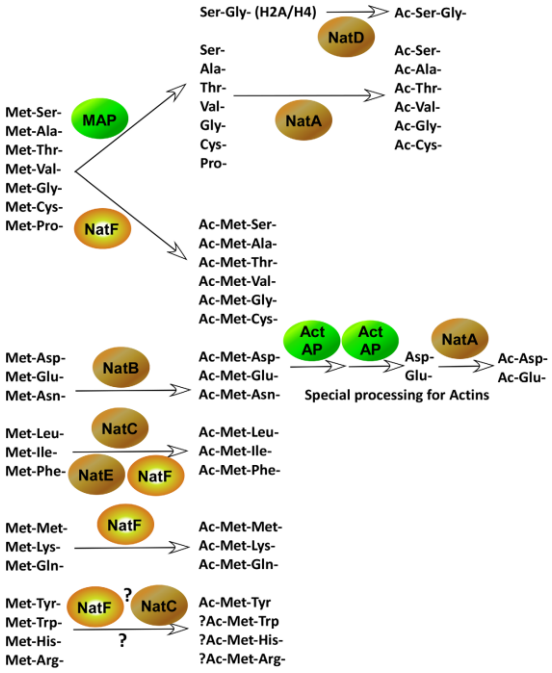


***Figure 7**: The major pathways of protein N-terminal processing in higher eukaryotes (Adapted from [70]).*

The NatA complex combines two subunits, *i.e.* NAA10 (catalytic subunit) and NAA15 (auxiliary subunit), anchored to the ribosome [71]. NatA acts after NME [56] to acetylate predominantly serine, alanine, glycine and threonine residues. This complex can also be trimeric with the addition of the catalytic subunit NAA50 and is then designated as NatE complex [68]. The trimeric complex favours the NTA of NatA substrates, which retain iMet [72]. Homologues of NAA10 and NAA15, respectively NAA11 [73] and NAA16 [74], are present in some vertebrate species (mostly *Euteleostomi* and lower species) due to probable gene duplication. These homologs exhibit similar but lower NatA activity *in vitro* [73]. NatB (NAA20 and NAA25) and NatC (NAA30, NAA35 and NAA38) acetylate polypeptides on iMet with a preference for acidic residues (NatB) or hydrophobic and bulky residues (NatC). NatD (NAA40) targets specifically protein sequences starting with the sequence MSGX and is limited to few candidates, such as histones H2A and H4 [66, 67]. Finally, NatF [70] is anchored to the Golgi apparatus and acetylates specifically transmembrane proteins [75] with broad specificity targeting mainly substrates that have retained iMet [72].

Although *Naa10/Naa15* double knockout in *S. cerevisiae* induces only weak biological defects [50, 76], deletion mutant of both Naa10/Naa15 in higher eukaryotes are not viable [64, 74]. siRNA mediated knockdown of Naa10 or Naa15 impairs cell viability through G0/G1 arrest in eukaryote species [74, 77]. Rope *et al.* [78] identified a NAA10 variant (NP_001243048.1:p.Ser37Pro) in male infants that decreases NatA activity by 60-80% which is lethal. Other missense mutations were identified such as NP_001243049.1:p.Arg110Trp and NP_001243048.1:p.Val107Phe in humans which were identified responsible of severe intellectual disability and cognitive impairment respectively [79]. Gromyko *et al.* reported the influence of Naa10/Naa15 gene knockdown in p53-dependent apoptosis and p53-independent growth [77, 80]. Human NAA10 has also been associated with cancer and cell proliferation [81-83]. Although this modification is involved in various biological processes [82], very little is known about its physiological relevance.

Until recently, NTA was considered as an irreversible modification and N-terminal deacetylase has not been identified so far. However, *E. coli* ribosomal protein L12/L7 exists in two different forms, *i.e.* the

N-terminus acetylated (NTAed) and the non-acetylated counterpart [84]. Gordiyenko *et al.* [85] proposed that the NTA of RL12 was a key elements in stalk ribosomal complex stabilisation. Furthermore, some large-scale studies [86-88] highlight the high frequency of this modification in eukaryotes and confirm the simultaneous presence of both acetylated and non-acetylated proteoforms and the NTA yield appears to be variable. Yi *et al.* suggested that protein NTA yield is metabolically regulated and dependent of the level of acetyl-CoA in the cell [89]. Recently, Varland *et al.* [90] proposed a more complex interaction between the acetyl-CoA level and the NTA yield. It involves Pcl8 (Q08966: cyclin-8, a negative regulation of glycogen synthase GSY2) which was characterised with a lower NTA yield in favourable growing condition (2% glucose vs. 0 % glucose).

As for humans, NTA is vital for plant cytosolic proteins [63, 88] where it plays a major role in stress responses [64, 91]. NTA has also been frequently characterised at the mature protein N-terminus in the chloroplast (Cp) [63, 88, 92]. This modification was not expected in this organelle originating from a bacterial ancestor. Furthermore, NTA of the nucleus-encoded proteins targeting Cp is clearly a posttranslational modification since it occurs after the excision of the Cp transit peptide (cTP), suggesting the existence of plastid-specific Nats.

It is now clear that NTA has essential roles on protein fate. The presence of this modification in the Cp is intriguing compared to the mitochondria where this type of modification remains infrequent. When I first observed this modification during my PhD, it triggered some interesting collaboration with Prof. A. Bairoch and the Swiss-Prot team. Although at the early beginning, my interest for NTA was limited to the submission of experimental results in UniProtKB/Swiss-Prot, I rapidly decided to learn more about it and to focus on the characterization of this modification including the enzymes involved in the NTA process, their substrate specificity and physiological roles.

2.3.3 Development of a mature protein N-terminal enrichment strategy

Although NTA occurs frequently in eukaryotic species such as *H. sapiens* or *A. thaliana*, N-terminally modified peptides represent, at best, a unique tryptic peptide per protein or proteoform. Furthermore, complex biological samples such as tissues or cells contain hundreds to thousands distinct proteoforms

Sample type		Whole *S. lycopersicum* leaf lysate (basic LC-MS/MS analysis; trypsin digestion)	LC-MS/MS analyses of the first 10 SCX fractions after tryptic digestion		
Validation processing		1% FDR on protein (Mascot 2.6)	1% FDR on protein (Mascot 2.6)	1% FDR on protein (Mascot 2.6) Min. 2 peptides per protein	SILProNAQ validated protein N-Termini
Number of identified proteins (Not redundant)		953	1421	687	608
Number of identified peptides	All	7144	15047	N.D.	N.D.
	Not redundant	3319	3366	3439	N.D.
Protein N-term peptides		40	224	139	259
Protein C-term peptides		55	219	34	N.D.
N-termini ratio (vs. characterised proteins)		4%	16%	20%	43%
Fold increase of characterised N-termini		(Reference set)	5.6	3.5	6.5

Table 1: *Number of proteins, peptides and protein N/C-terminal peptides characterised in a tryptic digest of S. lycopersicum leaf lysate (crude mixture) and after an N-terminus peptide enrichment using the SILProNAQ and EnCOUNTer SCX using various validation strategies including data processing.*

over a large dynamic range of concentration (5-10 orders of magnitudes). This high complexity increases the difficulty to target specifically and efficiently protein N-termini. Experimentally, the characterisation of such modified peptides remains rare in complex endoproteolytic peptide mixture although N-terminally acetylated peptides are known to promote MS^2 fragmentation [93] which should ease their characterisation.

In the example shown above (see Table 1), the analysis of *S. lycopersicum* leaf tryptic digest provides only 40 protein N-termini out of thousands of identified proteins (less than 5%). By this approach, I have been able to collect more than 500 distinct proteins N-termini over the years. These data are now available in UniProtKB/Swiss-Prot. To improve protein N-terminal characterisation, a dedicated methodology targeting more efficiently these peptides was first developed at the "Proteomics Analysis Facility" in 2005, when I was working in the biochemistry department, in Lausanne University [94, 95].

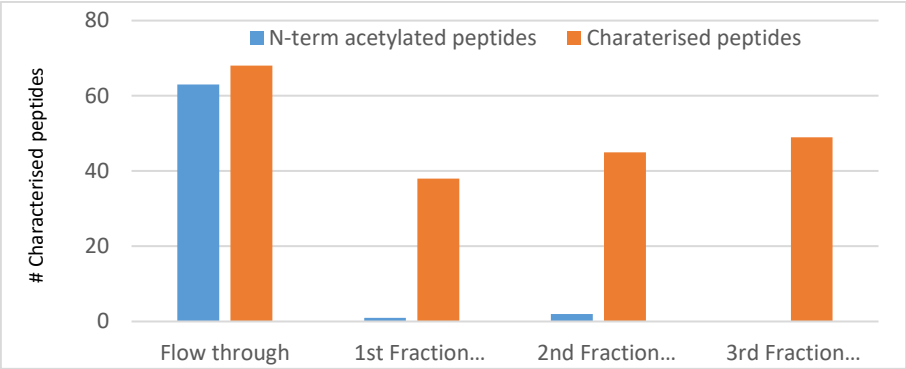


***Figure 8**: Distribution of the NTAed peptides characterised in human CD8 T lymphocytes after SCX separation of 100 µg of Lys-N digested sample using* 20, 90, 500 mM of NaCl *for peptides elution.*

The goal of this development was to segregate the NTAed peptides to favour their characterisation by liquid chromatography (LC) mass spectrometry (MS) analyses. Since this modification decrease the peptide net charge, it is possible to discriminate NTAed peptides using Strong Cation eXchange chromatography. Initially, sample preparations were conducted using Lys-N endopeptidase [95], but trypsin turn out to be a reliable and cost-effective endoproteinase for this approach. The initial experiment conducted using 100 µg of human CD8 T lymphocytes provided the characterisation of more than 60 NTAed peptides/proteins (Figure 8). Most of these N-termini were efficiently segregated in the flow through although few of them were found in the following SCX-elution fractions due to the presence of an arginine or histidine in their sequence.

Although this sample preparation provides an enrichment of NTAed peptides, unmodified N-termini were lost. Moreover, the lack of a relevant negative dataset (free or unmodified N-termini) rapidly appeared to be a major issue in prediction tool development. For a better in silico NTA prediction, it was crucial to retrieve also the unmodified peptides as a negative dataset. Alternative approaches were developed to target specifically free N-termini [96, 97] but, these methods were losing the modified N-termini, which is also a major hurdle in large-scale N-terminomics approach. Then, a new approach which take care of both NTAed and non-acetylated peptides had to be developed.

2.3.4 SILPRroNAQ: A methodology for NTA segregation and Quantification [98]

Since NTA enrichment technique was not able to take care of the non-modified N-termini, a sample preparation step was added to chemically acetylate free amino groups (protein N-termini and Lys residues) with stable isotopically labelled reagent (d_3-acetyl or $^{13}C_2$,d_3-acetyl groups). With these improvements, this method promotes large-scale N-terminomics studies and the characterisation of both NTAed (native or light proteoform) and free N-terminal (heavy labelled proteoform) peptides.

Additionally, the MS signal could be used to determine NTA yield in a similar way as SILAC differential quantification. This step required the development of a dedicated tool, that we called EnCOUNter [99] (Extraction and Calculation Of Unbiased N-Termini) (see section 3.2). This tool is able to provide NTA yield at different protein position which is crucial for nuclear encoded proteins targeting organelles such as mitochondria (Mt) and/or Cp or other proteoforms even at non-canonical positions (out of positions 1-2) such as protein splicing variants or proteins translated from alternative start positions.

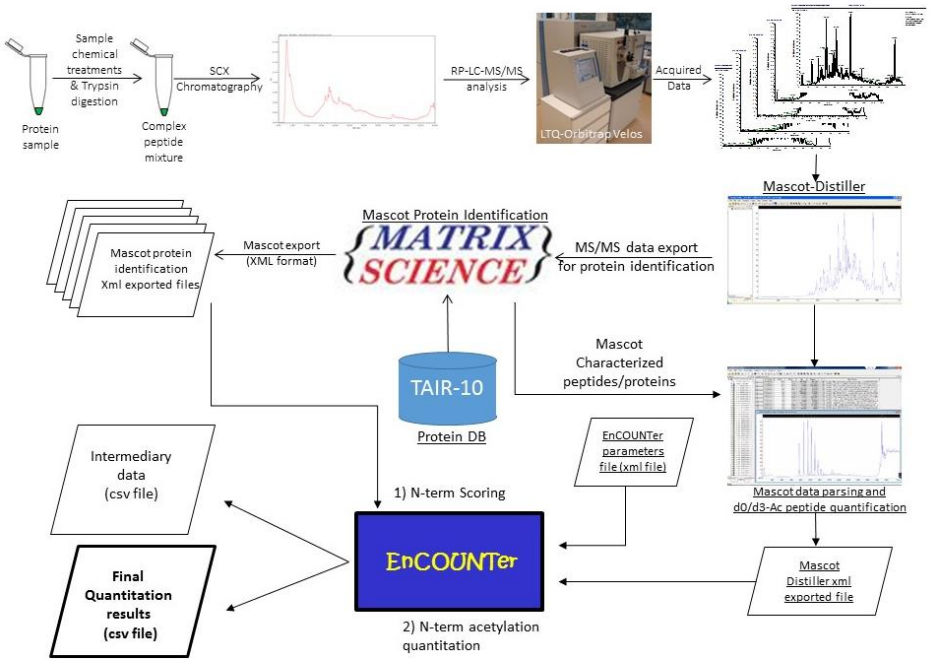


***Figure 9**: Overview of the SILProNAQ processing scheme: from sample to identification of mature N-terminus position and determination of NTA yield. Adapted from [99]*

In summary, the SILProNAQ pathway (Stable Isotope Labelling Protein N-terminal Acetylation Quantitation) combines sample preparation, high resolution MS analyses and post-acquisition data processing (Figure 9). It provides a unique and efficient package able to perform large scale protein N-terminal enrichment and NTA yield quantitation.

2.3.5 Large scale N-terminal investigation in various species

A few N-terminal peptide enrichment techniques by positive or negative selection have been developed over the years. These targeted approaches are frequently based on SCX fractionation [100, 101] including the SILProNAQ approach [98], solid phase extraction [102] or other alternative techniques [103, 104] to unravel the extent and the nature of N-terminal modifications. This approach was applied to several species, including *H. sapiens*, *A. thaliana*, *D. rerio*, *E. coli*, *S. lycopersicum…*

The huge amount of data generated by these studies required a centralized management and storage. The N-terDB (Experimentally characterised N-terminal peptide DB), was created. Currently, 57 projects dealing with various *A. thaliana* samples are available in N-terDB. These projects allowed the characterization of a large number of experimentally validated N-termini. Collected data could be redundant for the most abundant proteins especially for the species well represented in the included project such as *A. thaliana*. The idea was to take advantage of this redundancy to validate the characterized N-termini. Indeed, this approach was especially interesting for downstream N-terminal position (*e.g.* for Mt or Cp nuclear encoded proteins where mature protein N-termini are usually downstream of the predicted protein starting position) that are not always known precisely. This

approach put a spotlight on several interesting observations on the protein N-terminal maturation processes such as:

- Partial iMet excision for the protein N-termini with a proline, valine and threonine residue present at position 2: this type of partial excision was previously described [105] and was further validated,
- N-terminal alanine excision (NAE): this modification occurs when few alanine residues are present at position 2 and at least position 3. Surprisingly, the Met-Ala dipeptide is excised (MA|AX) and the uncovered alanine at position 3 is frequently NTAed; this unusual maturation is preferentially observed for human proteins (few cases observed for Arabidopsis proteins) and reminded the actins maturation process [106],
- Partial protein NTA: numerous protein N-termini were characterized under both N-term version, i.e. free and NTAed; Although N-terminal peptide sequence is directly related to this partial N-terminal acetylation, e.g. inhibitory effect of the lysine residues at position 3 to 5, the observed variability has not been explained so far;
- Multiple transit peptide cleavage positions: although single transit peptide cleavage position is expected for nuclear encoded Cp/Mt proteins, multiple cleavages were characterised experimentally especially for the Cp targeting proteins and confirm the previous observation of Rowland *et al.* [107]; The additional starting positions could be associated to additional Cp maturations that could affect protein stability and half-life [108].

Our DB contains so far 126 distinct projects in 6 different species. A total of 9'668 proteins associated to 12'738 distinct non-redundant N-termini were manually validated and annotated of which more than 7'000 are associated to the NTA yield.

2.4 Identification and characterisation of N-acetyltransferases

2.4.1 Global Acetylation Profiling (GAP) test.

NTA catalysed by ribosome-attached Nats is among the most common protein modification in the eukaryotic cytosol with more than 80% of NTAed proteins. Experimental approaches confirm the presence of NTAed substrates in few species such as *H. Sapiens* [63], *A. thaliana* [63], *C. reinhardtii* [88], *D. melanogaster* [109], *S. cerevisae* [86] or *D. rerio* [110]. Using an *in silico* approach, probable eukaryote Nats orthologues including plants[63] were identified highlighting strong evolutionary conservation (Figure 10).

Despite NTA is wide spread in eukaryote species and NatA complex is known to be essential in human [111, 112] and *A. thaliana* [64]its general/global significance is still unclear. NTA has been shown to affect stability of several yeast proteins. However, due to the stable expression of Nats, the absence of N-deacetylases, and its co-translational occurrence, NTA was supposed to be a static and unregulated protein modification.

Recently, a collaborative work uncovered specific and essential functions of cytosolic NTA for development, biosynthetic pathways and stress responses in plants. We have showed that NTA is dynamically regulated in plants upon stress [64] and identified new interaction partners of Nats involved in regulating the acetylation activity [91]. Our results establish cytosolic NTA as an important cellular surveillance mechanism during stress and contribute to the regulation of global protein turnover in plants.

An *in vitro* test using a peptide library was set to determine the specificities of few recombinant Nat [113]. Although successful with NAA10 and NAA50, this approach was time-consuming and we decided

to develop an alternative approach. Our method took advantage of the low number of NTAed proteins in *E. coli* [59]. We have expressed recombinant Nats in *E. coli* and looked for bacterial proteins exhibiting a modified N-terminus (neo-NTAed substrates) using the SILProNAQ sample preparation. This approach, the Global Acetylome Profiling (GAP) assay, was successfully tested with *At*Naa10 [114].

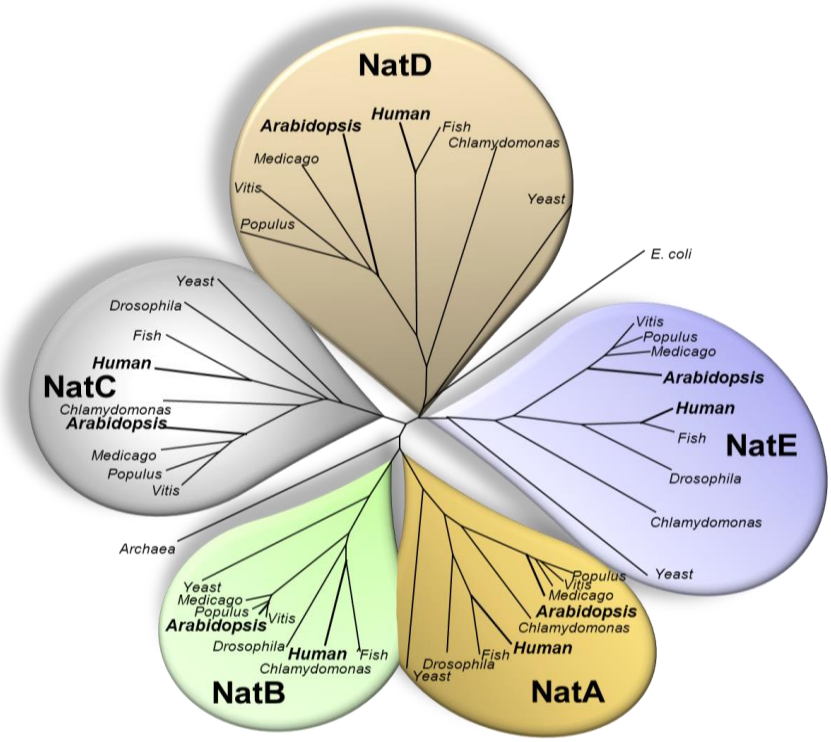


***Figure 10**: Phylogenetic tree including a few Nat A/B/C/D and F catalytic subunits orthologues from plant (V. vinifera, P. trichocarpa, M. truncatula and A. thaliana), algae (C. reinhardtii), and few bony vertebrates (H. sapiens and D. rerio) compared to RimI from E. coli.*

2.4.2 Chloroplastic N-acetyltransferase

Although NTA is a frequently occurring modification in the cytosol, this modification is not expected in organelles of endosymbiotic origin such as Mt. However approximately 25-30 % of the Cp proteins are modified by NTA [63, 92]. The underlying mechanisms are not yet understood and the associated Nats are not identified. Nats are not encoded in the chloroplastic genome, hence they should origin from the nuclear genome and be targeted to the Cp compartment. We therefore performed *in silico* investigation to identify putative *A. thaliana* Nat genes that would i) match the Prosite GNAT-associated profiles (PS51186) and ii) be predicted to have a chloroplastic subcellular location. Out of the dozen candidates, most appear to be active using the GAP test (Table 2) and seven were confirmed to be located specifically in the chloroplast.

Since cytosolic Nat were clearly involved in Lys acetylation (KAT) such as *h*NAA10 [115, 116], it was interesting to determine if the Cp Nats also expressed this activity (collaborative work with Dr I. Finkemeier, Universität Münster, Germany). To make a long story short, we could show that the Cp contains several acetyltransferases with dual KAT/Nat activity (Table 2).

Since cytosolic NatA and NatB were clearly involved in the biotic and abiotic stress responses, this is questioning if these CpNats has an implication in stress modulation in the Cp which are prime targets for stress factors such as drought and excessive light. Then, the seven newly identified CpNats

candidates will be the primary subject of a more extensive biochemical characterisation (ERACaps project: *KATNat: Elucidating the multifaceted functions of protein acetyltransferases in plant stress response and regulation of metabolism*).

Potential Cp Nats	Subcellular Localisation	Nat activity	KAT activity
CpGNAT01	Cp	XX	XX
CpGNAT02	Cp	XX	XX
CpGNAT03	Cp	XX	XX
CpGNAT05	Cp		X
CpGNAT06	Cp	X	XX
CpGNAT08	Cytosol	X	XX
CpGNAT09	Nucleus/Cytosol	X	X
CpGNAT11	Cp	X	XX
CpGNAT12	Cytosol		XX
CpGNAT13	Cp	X	X

***Table 2**: List of potential A. thaliana Cp KAT/Nat enzymes with their NTA and Kat activity (XX: main, X: medium) and subcellular localisation.*

3 When the Wet lab meet the Dry lab

3.1 Introduction

During my Ph. D, the development of the *Molecular Scanner* project required the development of a bioinformatics infrastructure to process raw data. Although I never have been involved in program coding, I was strongly associated to the creation of algorithms related to MS-data processing, protein identification [45] and multi-layered MS-based gel imaging [44]. Additionally, I was involved in the development of few other tools namely IsotopIdent, PeptideCutter and FindPept [117] which are still available on the ExPASy web server (https://www.expasy.org/).

In 2006, I proposed a master project for the experimental characterization of protein N-termini. The collected data were used to develop a NTA prediction tool in collaboration with Dr. I. Xenarios using a Support Vector Machine (SVM) approach [95, 118]. The prediction tool highlighted the inhibitory effect of the Pro residue at position P'2 as expected [119] but also at P'4. Interestingly we also observed the negative influence of the lysine residue at position P'2, P'3 and P'4, an effect that has not been reported earlier.

This initial project highlighted the strong interest of highly curated experimental data for datamining approach and to provide reliable prediction tool. Unfortunately, the amount of experimental data provides during this investigation was still limited and the resulting prediction tool was still suffering some inaccuracies. This drawback could be solved if additional experimental data were collected allowing prediction tool improvement.

3.2 EnCOUNTer: A processing tool to parse large-scale analyses results

The development of the SILProNAQ methodology provided a convenient way to target both type of protein N-termini, *i.e.* free N-termini and NTAed. While the characterisation of cytoplasmic N-termini was simple (usually located at position 1-2 of the protein sequence), the location of novel N-termini generated for instance by the cleavage of chloroplast or mitochondria transit peptides (so-called *downstream N-termini*) were much more difficult to determine.

Therefore, it was crucial to determine the accurate NTA yield to better understand organelle N-terminal maturation processes and their potential involvement in stress resistance mechanisms [64, 91]. Since no tool was available to perform this task, we developed a novel pathway to collect non-redundant N-termini and to determine their respective NTA yield.

Such data processing took advantage of Mascot Distiller (MD) software (Matrix Science, London, UK) that provided raw MS signal extraction and quantification. Pre-processed data were exported in an xml file and further refined using the EnCOUNTer parsing script [99]. This script was developed to discriminate mature protein N-termini and to determine the NTA yield.

To this end, the EnCOUTer script included a scoring scheme to rank and highlight mature protein N-termini. The main scoring coefficient (or scoring matrix) was defined using a manually validated training dataset to sort out protein downstream N-termini. Few other scoring coefficients were added to improve the discrimination power of this approach including data redundancy, protein sequence around the neo-starting position, N-terminal modification, etc. Finally, a threshold score was defined depending of the selected parameters applied to the training dataset. The defined scoring scheme was applied to the samples prepared with the SILProNAQ method to extract the most relevant N-termini at any positions of the protein sequence.

Additionally, to the harvest mature starting positions, the EnCOUTer script could determine the NTA yield for each distinct N-termini. This calculation was based on MD pre-quantitation. Since raw signals could be variable in quality, we used the MD parameters (user defined parameters) to favour the quantitation of the most reliable N-termini. Finally, a list of each N-terminus, starting position and NTA yield was provided for each unique starting position even when several distinct positions were characterised for a single protein. This tool was able to collect a few hundreds to thousands of protein N-termini per experiment [99].

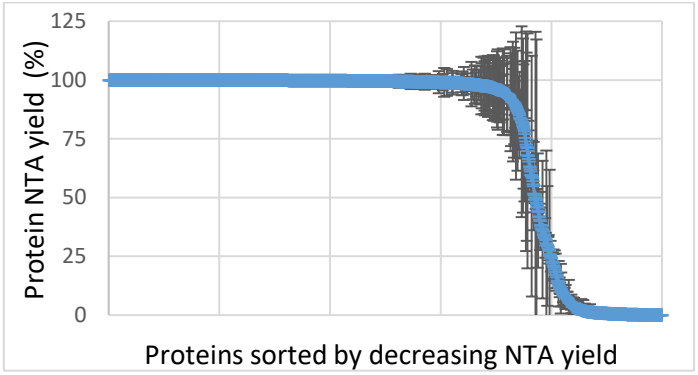


***Figure 11**: NTA yield distribution of the 1'500 cytosolic proteins at position 1/2 available in the N-terDB (True & Probable confidence level) from different species including H. sapiens, A. thaliana, S. lycopersicum.*

3.3 The N-terDB

The recent development of a SILProNAQ methodology allowed the identification of numerous experimentally characterized N-termini. However, when analyzed individually, the collected data from a single experiment remained partial and did not offer a clear and global overview of the N-terminome extent. We then decided to create a dedicated DB to collect and aggregate data provided by EnCOUTer. This allowed us to view all results at a time and to validate those occurring in independent experiments, based on the rationale that real signals should be reproducible compared to random noise. Each experimentally characterized NAT position and subcellular location was tagged with a confidence level (True, Probable, Potential, Ambiguous and False). "True and probable" ranked hits were the most

relevant data that could be used later for data mining approach and the development of new prediction tools. "Potential and Ambiguous" results could also be used with caution.

The N-terDB provides a unique resource of experimentally characterised mature protein N-termini. When raw data quality was sufficient (MS raw data quality), this information is complemented with NTA yield. So far, 11'000 N-termini are available of which 4'000 with a quantified NTA yield. Figure 11 shows the distribution of the NTA yield for all protein N-termini at position 1/2 available in the N-terDB. Note that half of the experimentally characterised protein N-termini are located downstream of the predicted N-terminus (mature protein N-termini at position higher than position 2).

N-terDB provides a unique resource of manually validated N-termini for Mt and Cp mature proteins. For example, the chloroplast cytochrome b6-f complex iron-sulphur subunit (AT4G03280.1) is associated to multiple mature starting positions of which 3 have distinct NTA yield (Figure 12). These curated data are now available on a publicly available website: https://N-terDB.i2bc.paris-saclay.fr

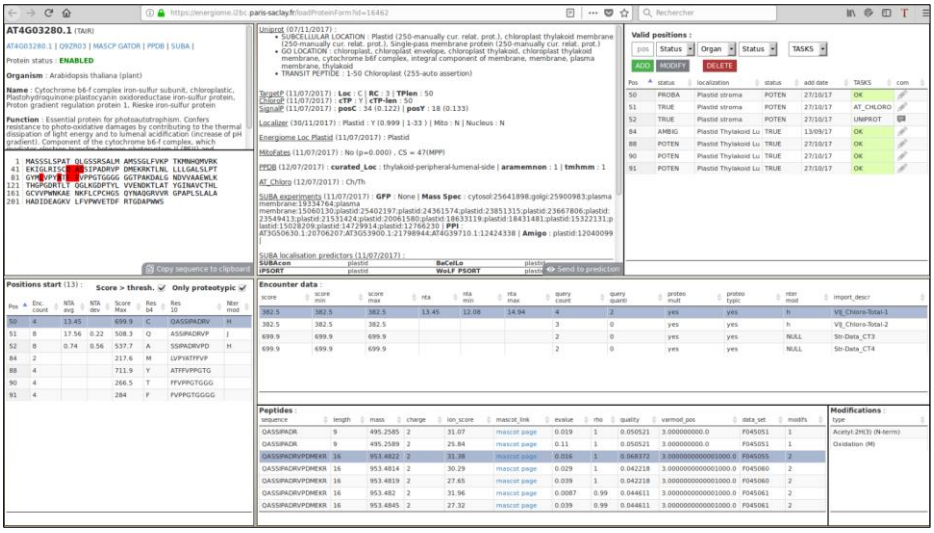


***Figure 12**: N-terDB "protein-form" interface of the "Cytochrome b6-f complex iron-sulphur subunit" protein (ARAPORT11 Ac: AT4G03280.1) and web links, information collected from other reference DB, predictors results and manually validated starting position and subcellular localisation are reported.*

3.4 N-terPred: valorisation of the data collected in N-terDB

3.4.1 Cytosolic N-terminal modification

The proteins expressed in the cytosol are mainly subject to the co-translational NME and NTA. The experimentally characterised N-termini are a material of choice to improve software tools dedicated to predict modifications of cytosolic N-termini.

A few different approaches based on machine learning algorithms and traditional decision tree classifiers were tested to predict NME and NTA status. This work was conducted with the majority of the data available in the N-terDB for *H. sapiens* and *A. thaliana.* The best predictions were obtained with a decision tree classifier.

3.4.1.1 Cytosolic NME

Excision of the iMet is one of the most frequently observed N-terminal modification. The performance of the newly defined rules for the prediction tools N-terPred (NME) and TermiNator3 [120] were determined for *H. sapiens*, *A. thaliana*, *D. rerio, S. lycopersicum* and *S. cerevisiae*. The prediction results are summarized in Table 3.

Both, Termi*N*ator3 and N-terPred (NME) performed similarly for the prediction of the iMet excision (NME) with N-terPred (NME) being slightly better. The false positive and negative rates are usually below 1 % except for the *H. sapiens* and *E. coli* where they were slightly higher mainly due to partial iMet excision. Since the predictions were only based on complete iMet excision, partial NME increase the false positive rate.

		Sensitivity	Specificity	Accuracy	False positive rate	False negative rate	MCC
H. sapiens	Termi*N*ator3	99,1%	87,4%	96,1%	4,2%	0,9%	0,896
	N-terPred	99,3%	89,0%	96,6%	3,7%	0,7%	**0,911**
A. thaliana	Termi*N*ator3	99,7%	99,1%	99,5%	0,5%	0,3%	0,989
	N-terPred	99,7%	99,4%	99,6%	0,3%	0,3%	**0,992**
E. coli	*Terminator3*	93,4%	88,5%	91,2%	9,4%	6,6%	0,822
	N-terPred	97,1%	85,8%	91,9%	11,1%	2,9%	0,840

Table 3: *Results obtained for various species using TermiNator3 tool and N-terPred for NME predictions. The table summarizes a few statistical indicators such as sensitivity, specificity, accuracy, the false positive /negative rate and the Matthew Correlation Coefficient (MCC).*

Partial iMet excision tends to occur more frequently in proteins with threonine, valine or proline residues located at position 2 [56]. However, distant residues may also influence the efficiency of the excision. To improve NME prediction, three classes (retaining iMet, NME and partial NME) should be defined instead of two likes it done at the moment (retaining iMet and NME). Unfortunately, the current dataset (48 partially excised iMet) were not sufficient to populate the partial NME class and to determine these distant influences.

Matthew's Correlation Coefficient (MCC) was used as a prediction indicator and was higher than 0.9 for most of the results (including *D. rerio* and *S lycopersicum*, data not shown) which indicates a very good correlation between the experimental data and our predictions. These results show that substrate specificity is strongly conserved between species and both NME prediction tools (Termi*N*ator3 and N-terPred (NME/NTA)) provide highly reliable predictions (accuracy, specificity and sensitivity are frequently close to 99%). Additional data are required to improve partial iMet excision.

3.4.1.2 Cytosolic NTA

Experimentally collected data confirmed the widespread occurrence of this modification with a continuous distribution of the NTA yield from not-modified to fully-acetylated N-termini (Figure13). 70-75% of the N-termini are fully acetylated whereas 15-20% are not modified. Some 10-15% were partially acetylated with variable NTA yield.

Although protein NTA yield prediction was our main target, prediction result rapidly highlighted an under-sampling issue. Although the first residues of the polypeptide chain were critical for NTA, inhibitory effects of distant residues are known such as proline at P'2 or lysine up to position P'4. For NatA and considering only the first four residues of the N-terminus (residue at P'1 are restricted to alanine, cysteine, glycine, proline, serine, threonine and valine), there was more than 50'000 possible combinations compared to few thousand N-termini experimentally characterized and quantified for NTA yield. As a result, it was only possible to define a new class of modification called "partial NTA" in addition to the NTAed and the free N-termini (FNT). This class was not considered when comparing N-terPred (NTA) and TermiNator3 results. NTA at position 1 and 2 were analysed independently and the result obtained for *A. thaliana* and *H. sapiens* are shown in Table 4. The N-terPred (NTA) tool clearly

outperformed Termi*N*ator3 for both positions 1 and 2. This result was expected for position 1 since Termi*N*ator3 does not predict NatC substrates while N-terPred (NTA) does. The prediction accuracy is also slightly improved for the NatA substrates due to the implementation of inhibitory effects, for instance in the presence of lysine (position P'2 to P'4) at the protein N-terminal sequence. These effects were previously not considered. N-terPred (NTA) was also tested successfully on *D. rerio*, *S. lycopersicum or S. cerevisiae*.

Species	Start Pos	Prediction tool	Sensitivity	Specificity	Accuracy	False pos. rate	False neg. rate	MCC
H. sapiens	Pos 1	Termi*N*ator3	80,2%	100,0%	84,0%	0,0%	19,8%	0,661
		N-terPred	99,1%	67,9%	93,3%	6,9%	0,9%	**0,768**
	Pos 2	Termi*N*ator3	97,1%	81,3%	93,1%	6,2%	2,9%	0,813
		N-terPred	99,6%	89,5%	97,0%	3,4%	0,4%	**0,921**
A. thaliana	Pos 1	Termi*N*ator3	85,5%	94,6%	86,4%	0,6%	14,5%	0,559
		N-terPred	99,3%	77,8%	97,3%	2,2%	0,7%	**0,829**
	Pos 2	Termi*N*ator3	94,4%	74,2%	89,3%	8,5%	5,6%	0,710
		N-terPred	98,8%	82,4%	94,7%	5,5%	1,2%	**0,856**
S. cerevisae	Pos 1	Termi*N*ator3	89,9%	91,8%	90,5%	4,3%	10,1%	0,795
		N-terPred	99,3%	90,5%	96,4%	4,5%	0,7%	0,919
	Pos 2	Termi*N*ator3	88,8%	95,4%	91,1%	2,8%	11,2%	0,819
		N-terPred	100,0%	68,8%	89,0%	14,5%	0,0%	0,767

***Table 4**: NTA prediction for A. thaliana, H. sapiens and S. cerevisiae using TermiNator3 and N-terPred (NTA). The table summarizes different statistical indicators and the Matthew Correlation Coefficient (MCC: +1 perfect prediction; 0 random prediction; -1 total disagreement) to rate the prediction power of the software tools.*

N-terPred (NTA) prediction tool is highly reliable for all eukaryotic species and outperforms Termi*N*ator3.

3.4.2 Subcellular location prediction

Although most proteins are expressed in the cytosol, some are relocated in specific organelles. This is the case for most Mt or Cp proteins. The TP that targets the protein to the correct location is often excised during proteins translocation. Our SILProNAQ investigation provided the characterisation of relevant mature neo-N-termini for nuclear encoded proteins. Although numerous prediction tools are currently available, their reliability is often questioned.

Thanks to N-terDB curated data, we could test the reliability of some of these TP prediction tools. We also used these data to develop a subcellular location prediction tool specifically for nuclear-encoded Cp and Mt proteins.

3.4.2.1 Chloroplast subcellular location

Since experimental investigations to determine the accurate protein subcellular location are scarce, prediction tools are often used to overcome this gap especially if the considered species are poorly annotated. This is the case for nuclear encoded protein targeting Cp in plant species. Some of the available tools are based solely on protein sequences, *e.g.* TargetP [121], Predotar [122] or WolfPsort [123] whereas SUBA (Subcellular Location Database for Arabidopsis) [124] combines the results of a few prediction tools and literature data mining.
Our idea was to provide a predictor solely based on the protein sequence. To achieve this goal, we tested a few data mining techniques including the SVM (Support Vector Machine) or the logistic regression to combine the prediction results of different existing scripts (Predotar,

ChloroP, WolfPsort, BaCelLo, etc.). The best prediction result was reached using a combination of 3 predictors (TargetP, Predotar, and WolfPsort)

Prediction tool	Sensitivity	Specificity	Accuracy	False Positive Rate	False Negative Rate	MCC
SUBA	92,6%	97,4%	95,5%	4,1%	7,4%	0,906
N-terPred	93,3%	95,8%	94,8%	6,4%	6,7%	0,892
TargetP	90,1%	92,4%	91,5%	11,4%	9,9%	0,823
Predotar	81,4%	96,9%	90,8%	5,5%	18,6%	0,808
Wolf Psort	87,1%	80,4%	83,1%	25,5%	12,9%	0,662

***Table5**: Chloroplast subcellular location predicted by several programs using a composite dataset which includes 690 chloroplastic, 164 mitochondrial and 888 cytosolic proteins from A. thaliana.*

We also took advantage of the curated data available in N-terDB to test and compare the efficiency of few frequently used protein location prediction tools (Table 5) including TargetP, Predotar, and WolfPsort, the web resource SUBA and our new prediction tool that we called N-terPred (Loc). SUBA clearly provided the best result for protein location prediction followed by our new prediction tools. Since N-terPred (Loc) was solely based on protein sequences, it displayed a clear advantage compared to SUBA which was solely dedicated to *A. thaliana* and dependant of published data.
Finally, TargetP and Predotar which we some one of the most frequently used software for protein location prediction suffered high false positive or negative rates whereas WolfPsort had both high false positive and false negative rate.

3.4.2.2 Mitochondria subcellular location

As for Cp protein, Mt location is frequently determined by dedicated predictions tools. Results obtained with commonly used software tools are shown in Table 6. Although SUBA remained the best predictor, it was penalized by a high rate of false positives (>26%). In view of these results, it appeared that there is a need for the development of a new bioinformatics tool for optimal mitochondrial prediction.

Prediction tool	Sensitivity	Specificity	Accuracy	False Positive Rate	False Negative Rate	MCC
SUBA	95,1%	96,4%	96,3%	26,8%	4,9%	0,816
MultiLoc2	86,9%	95,5%	94,7%	33,8%	13,1%	0,731
TargetP	83,5%	94,3%	93,3%	39,6%	16,5%	0,675
EpiLoc	54,4%	98,5%	94,4%	21,6%	45,6%	0,625
WolfPsort	54,4%	96,6%	92,7%	38,3%	45,6%	0,539

***Table 6**: Mitochondrial subcellular location obtained with commonly used prediction tools using the same composite dataset used for chloroplast protein prediction (see in Table 5).*

Our investigations are under progress but the low number of mitochondrial proteins identified remains a bottleneck. Other sources of curated data have to be investigated.

3.4.3 TP cleavage site

The exact cleavage position of mitochondrial TP and chloroplastic TP was determined experimentally for few dozen to hundreds of proteins. However, the starting position of these organelle mature proteins remains unknown or subject to prediction tools like TargetP/ChloroP, MitoProt II or Localizer. Data available in N-terDB were first used to test the existing script then to develop our own predictors for stromal, thylakoid and mitochondrial location.

3.4.3.1 N-terPred (xTP): prediction of the cleavage site of the transit peptides

The prediction of the transit peptide cleavage position is provided by few prediction tools including TargetP/ChloroP/SignalP tools suite [125-127], Localizer [128], iPSORT [129] or MitoProt II [130]. TargetP/ChloroP is usually considered to be one of the best prediction software (data not shown) with a success rate of 50% for stromal proteins (see next paragraph). We decided to improve this TP prediction rate and developed a new scoring approach based on the experimental data collected in N-terDB.

To this end, we used a pattern recognition approach call the Position Weight Matrix (PWM) [131] that was initially used to identify translation initiation sites in *E. coli* genomic sequences. This approach required a training step using a reference dataset that contain experimentally characterized TP starting positions and a negative dataset populated with random protein sequences. These sequences were used to define a "position probability matrix" which was finally converted to a "position weight matrix" (PWM). The PWM was used to determine a score for each position using a sliding window on the tested protein sequences to identify potential TP sites. Score higher than 0 highlighted probable TP and the position with the highest score was considered to be the TP cleavage position.

The accuracy of the prediction was strongly dependent of the quality of the training dataset. Then, only the most relevant collected TP cleavage positions (True and Probable rated) were retained for the Cp and the Mt datasets. In the case of the Cp dataset, proteins targeting the thylakoid that are known to undergo a double transit and "signal" peptide cleavage [132] were excluded. We also defined a specific dataset for the thylakoid-lumen proteins to predict the double transit/signal [132] peptides cleavage. We only selected protein targeting the thylakoid-lumen and protein with an experimental cTP in agreement with the ChloroP prediction were also removed from this dataset.

Finally, four datasets were defined related to human mTP, Arabidopsis mTP, cTP and lumenTP. These datasets provide the associated PWM and their prediction performance were described in the following paragraphs.

3.4.3.2 N-terPred (cTP): Chloroplast-stroma TPs

Figure 13 shows an overview of the ChloroP and Localizer predictions compared to N-terDB (plastid). ChloroP correctly predicted cTP length for only 17% of the proteins. The "positive hit" strongly increases if the prediction result is considered to be within ± 1 (44 %) or ± 2 residues (52%). The results obtained with Localizer were worse with only 14% of correctly predicted cTP length within ± 2 residues.

With more than 40% of exact predictions (71% within ± 2 residues), N-TerPred (cTP) clearly outperformed ChloroP (or TargetP). It was also successfully tested on *S. lycopersicum* (data not shown).

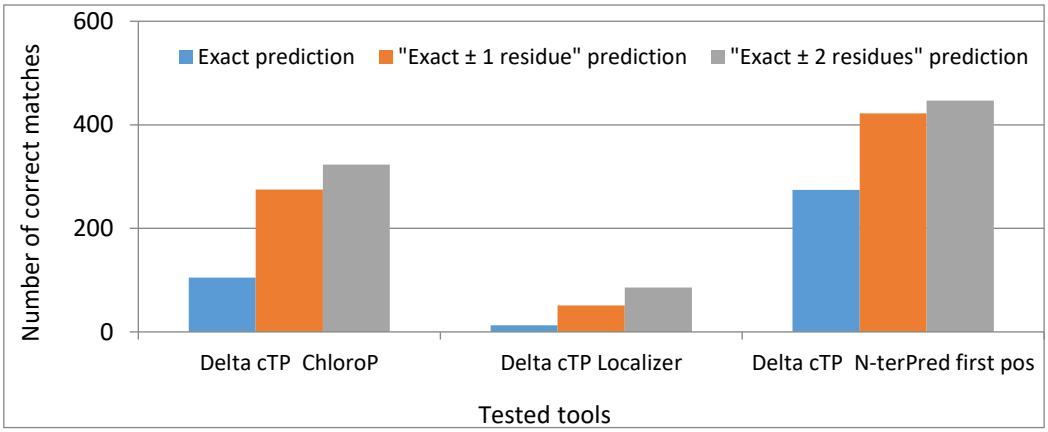


***Figure 13**: Comparison between cTP predictions and experimental data curated in N-terDB.*

3.4.3.3 N-terPred (lumenTP): Chloroplast-thylakoid TP

LumenP [133] was a predictor associated to TargetP tool suite to identify proteins targeting to the lumen of plant chloroplast. It also predicted the cleavage sites but this tool is not anymore available. Since N-terDB provided identification of luminal proteins with unambiguous starting position pattern (Figure 14), we used these data to create a scoring matrix dedicated to predict the luminal-specific TP cleavage sites.

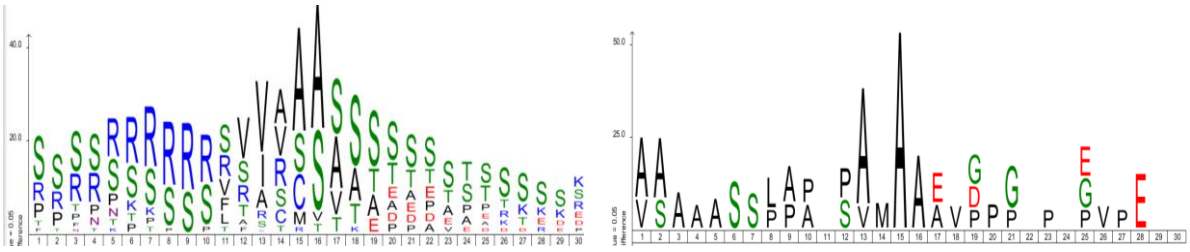


***Figure 14:** N-terminal sequence profile of A) stromal and B) thylakoid/lumen proteins. Position 16 corresponds to the first residue of the mature protein (cleavage between pos. 15-16) after transit peptide excision.*

Our new tool, called N-terPred (LumenTP), correctly predicts over 60% of TP cleavage sites (within ± 2 residues) compared to 10% using ChloroP (data not shown). It is at the moment the sole software able to accurately predict this thylakoid protein maturation event.

3.4.3.4 N-terPred (mTP): Mitochondria transit peptide

Our investigation also considered the nuclear-encoded protein targeting the Mt. Since mitochondrial TP (mTP) maturation differs between plants and mammals, we had to provide distinct scoring matrixes depending of the species (Table 7).

Species	*H. sapiens*		*A. thaliana*	
Dataset	TargetP	N-terPred	TargetP	N-terPred
False	82	57	86	26
Exact ± 2 residues	102	109	52	93
Total hits	184	184	138	138
% True	55%	59%	38%	67%

***Table 7**: Prediction results for Mt targeting proteins for H. sapiens and A. thaliana*

For human Mt proteins, TargetP and N-terPred gave similar predictions with 50-60% correct predictions (within ± 2 residues). This is slightly different for *A. thaliana* Mt proteins. In this case, over two thirds of correct predictions were obtained with N-terPred for only one third with TargetP. This difference is likely due to the lower number of plant Mt proteins included in the training dataset used during the development of TargetP.

3.4.4 Other targets

Data collected in the N-terDB also contain numerous proteins subjected to post–translation maturation, such as signal or pro-peptide cleavage. It appears that he prediction tool SignalP is highly accurate but some erroneous positions can be detected (data not shown). It would be interesting to use N-terDB collected data to improve this type of prediction in the future. Other protein maturation events, such as N-Alanine Excision (NAE) that was frequently observed experimentally remains uncharacterised so far. Unfortunately, this modification has only been characterised experimentally and is not associated so far to any biological interest.

4 Future Work

Proteostasis is defined as the processes required maintaining the equilibrium between protein synthesis and degradation. Processes that regulate protein synthesis and degradation in cell are constantly balanced to respond to cellular development and environmental stresses. This is a key element in the molecular mechanisms required in cell cycle and survival [134]. These mechanisms maintain cell vital functions through tight regulations from gene transcription to post-translational proteins modifications. At steady state, the delicate equilibrium between protein synthesis and protein decay is fundamental for an efficient biological activity.

It has been shown repeatedly that the abundance of proteins only moderately correlates with that of transcripts [135, 136]. This weak correlation may be due to different processes occurring downstream of transcription, from transcript translatability to protein turnover. While the analysis of protein abundance variations according to different factors is the common objective of comparative proteomics, few analyses have attempted to explore the impact of protein turnover on protein abundances and their regulation.

The general aim of this new project is to study the turnover of proteins in plants. The primary objective is to develop a method for the large-scale analysis of protein half-life (PHL) and turnover rate (PTR). This approach will be used:

- To study the relationship between protein N-terminal sequence and PTR,
- To quantify the impact of PTR on the amounts of protein in different situations such as cellular compartment, plant development, responses to environmental variations
- To analyse the influence of genetic variability on protein turnover and measure its contribution to the variation of protein abundances.

4.1 Large-scale determination of protein turnover and half-life

Large scale proteomics approach using pulse-chase SILAC (Stable Isotope Labelling by Amino acids in Cell culture) on cultured cells provided a method of choice to study cell proteostasis [137]. However, this approach is not compatible with autotroph species despite it has been used with seeds occasionally [138]. Indeed, soon after seed germination, plant seedlings acquire an autotrophic capability and are not anymore dependent of AAs supplied. Alternatively, it has been proposed to use inorganic sources of ^{15}N and/or ^{13}C isotopes [139] instead of stable-isotopes labelled amino acids. Full

^{15}N metabolic labelling of the proteins appeared to be difficult to reach [140] while, partial ^{15}N labelling was used few times for this type of investigations but required extensive data processing. Unfortunately, the currently available data processing scripts requires intensive pre- and post-data processing and are highly specific to the parameters of the developing laboratory [141]. The first goal of this project is to provide an alternative to the pulse-chase SILAC approach able to determine large scale PTR/PHL which will be compatible with plant to sort out the previously detailed biological questions.

This development first requires the optimisation of the experimental conditions to perform ^{15}N pulse labelling for plants. At first, *A. thaliana* will be used for the methodological development. This model plant is easy to grow in laboratory or culture chambers and its genome is correctly annotated and well documented. Later, this approach should be adapted to other plant including maize which is of direct interest at the GQE unit where I am now working. The required dedicated processing module able to convert raw LC-MS data to PTR and associated values will be integrated to our currently existing raw data processing tools *i.e.* MassChroQ [142], to provide statistically relevant result for protein turnover rate, protein half-life, constants of protein synthesis (Ks) and degradation (Kd) at the proteome scale.

4.2 Biological investigations

The N-end rules dogma strongly associates protein N-terminal residues to stability and PHLs [143]. Although, N-terminal acetylation is known to influence protein faith during stress events [64, 91], the direct influence of protein N-terminal modification such as protein N-terminal acetylation remains to be defined. The development of a reliable and large-scale method to determine PHLs in *A. thaliana* and maize will provide a unique opportunity to decipher the influence of protein sequence and especially protein N-terminal modifications in PTRs and PHLs.

The influence of protein N-terminal modification *vs.* PHL will be extensively investigated and I will take advantage of the data available in N-terDB. This database provides manually curated mature protein N-termini and associated protein N-terminal acetylation yield. It should be possible to determine more precisely the influence of protein N-terminal acetylation to protein PTRs and the influence of this modification in plant stress response. The knowledge acquired on *A. thaliana* will be useful to our investigate on maize of which has a more complex genome.
Additionally, the GQE institute is directly interested in the influence of genetic variability in the adaptation to the environment. Maize is one of the favourite model and my hosting team recently performed a GWAS (Genome Wide Association Study) analysis of protein amounts in maize (254 genotypes analysed [144]) and several hotspots of protein QTLs were detected. These results offer the possibility to choose genotypes showing a large variability of protein abundance, to study the possible contribution of PTR to this variation, and to look accordingly to possible candidate genes underlying protein QTL hotspots.

Additional collaborative investigation will take advantage of this approach to investigate a few biological subjects where the determination of the PTRs is a key element to better understand seeds germination and storage proteins reuptake at low temperature (L. Rajjou, IJPB, Versailles) or plant cell wall response to environmental stresses (E. Jamet, LRSV, Toulouse).

5 Conclusions

From the early beginning of the proteomic to the present situation, large-scale protein identification and characterisation remains a constant challenge in term of instrument sensitivity/accuracy and data processing. My work was continuously aimed at the improvement of these two points through my participation in the development of a new instrument and data processing scrips.

If originally my Ph. D. project was mainly methodological, the evolution of my interest led me to better understanding of the presence of NTA in eukaryotic cells, including, more recently, plants.

The novelty of this approach applied to plants is expected to be as fruitful as the large-scale PTRs conducted against cultured cells and should provide a new area in plant proteomics. The associated biological investigations should benefit of my expertise previously acquired on protein N-terminal acetylation and should improve our knowledge on the existing links between protein turnover and few parameters such as protein N-terminal status, genetic variability or environment variations.

6 References

1. Wilkins M: **Proteomics data mining**. *Expert Rev Proteomics* 2009, **6**(6):599-603.
2. Laemmli UK: **Cleavage of structural proteins during the assembly of the head of bacteriophage T4**. *Nature* 1970, **227**(259):680-685.
3. MacGillivray AJ, Rickwood D: **The heterogeneity of mouse-chromatin nonhistone proteins as evidenced by two-dimensional polyacrylamide-gel electrophoresis and ion-exchange chromatography**. *Eur J Biochem* 1974, **41**(1):181-190.
4. O'Farrell PH: **High resolution two-dimensional electrophoresis of proteins**. *J Biol Chem* 1975, **250**(10):4007-4021.
5. Bjellqvist B, Ek P, Righetti P, Gianazza E, Gorg A, Westermeir R, Postel W: *J Biochem Biophys* 1982, **6**: 317-339.
6. Westermeier R, Postel W, Weser J, Gorg A: **High-resolution two-dimensional electrophoresis with isoelectric focusing in immobilized pH gradients**. *J Biochem Biophys Methods* 1983, **8**(4):321-330.
7. Hochstrasser DF, Appel RD, Vargas R, Perrier R, Vurlod JF, Ravier F, Pasquali C, Funk M, Pellegrini C, Muller AF *et al*: **A clinical molecular scanner: the Melanie project**. *MD Comput* 1991, **8**(2):85-91.
8. Link AJ, Robison K, Church GM: **Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12**. *Electrophoresis* 1997, **18**(8):1259-1313.
9. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al*: **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd**. *Science* 1995, **269**(5223):496-512.
10. Hochstrasser DF: **Proteome in perspective**. *Clin Chem Lab Med* 1998, **36**(11):825-836.
11. Gravel P, Walzer C, Aubry C, Balant LP, Yersin B, Hochstrasser DF, Guimon J: **New alterations of serum glycoproteins in alcoholic and cirrhotic patients revealed by high resolution two-dimensional gel electrophoresis**. *Biochem Biophys Res Commun* 1996, **220**(1):78-85.
12. Cottrell J: **Protein identification by peptide mass fingerprint (Review)**. *Peptide Research* 1994, **7(3)**:115-124.
13. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A: **rotein Identification and Analysis Tools on the ExPASy Server.** In: *The Proteomics Protocols Handbook.* Edited by Walker JM. Totowa, N.J.: Humana Press; 2005: 571-607.
14. Zhang W, Chait BT: **ProFound: an expert system for protein identification using mass spectrometric peptide mapping information**. *Anal Chem* 2000, **72**(11):2482-2489.
15. Appel RD, Hochstrasser DF, Funk M, Vargas JR, Pellegrini C, Muller AF, Scherrer JR: **The MELANIE project: from a biopsy to automatic protein map interpretation by computer**. *Electrophoresis* 1991, **12**(10):722-735.
16. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A: **ExPASy: The proteomics server for in-depth protein knowledge and analysis**. *Nucleic Acids Res* 2003, **31**(13):3784-3788.
17. Hughes GJ, Frutiger S, Paquet N, Ravier F, Pasquali C, Sanchez JC, James R, Tissot JD, Bjellqvist B, Hochstrasser DF: **Plasma protein map: an update by microsequencing**. *Electrophoresis* 1992, **13**(9-10):707-714.
18. Golaz O, Wilkins MR, Sanchez JC, Appel RD, Hochstrasser DF, Williams KL: **Identification of proteins by their amino acid composition: an evaluation of the method**. *Electrophoresis* 1996, **17**(3):573-579.
19. Wilkins MR, Ou K, Appel RD, Sanchez JC, Yan JX, Golaz O, Farnsworth V, Cartier P, Hochstrasser DF, Williams KL *et al*: **Rapid protein identification using N-terminal "sequence tag" and amino acid analysis**. *Biochem Biophys Res Commun* 1996, **221**(3):609-613.
20. James P, Quadroni M, Carafoli E, Gonnet G: **Protein identification by mass profile fingerprinting**. *Biochem Biophys Res Commun* 1993, **195**(1):58-64.
21. Pappin D, Hojrup P, Bleasby A: **Rapid identification of proteins by petide mass fingerprint**. *Curr Biol* 1993, **3(6)**:327-332.
22. Mann M, Hojrup P, Roepstorff P: **Use of mass spectrometric molecular weight information to identify proteins in sequence databases**. *Biol Mass Spectrom* 1993, **22**:338-345.
23. Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C: **Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases**. *Proceedings of the National Academy of Sciences of the United States of America* 1993, **90**(11):5011-5015.
24. Yates JR, III, Speicher S, Griffin PR, Hunkapiller T: **Peptide mass maps: A highly informative approach to protein identification**. *Anal Biochem* 1993, **214**:397-408.

25. Karas M, Hillenkamp F: **Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons**. *Anal Chem* 1988, **60**(20):2299-2301.
26. Hillenkamp F, Karas M: **Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization**. *Methods Enzymol* 1990, **193**:280-295.
27. Karas M, Gluckmann M, Schafer J: **Ionization in matrix-assisted laser desorption/ionization: singly charged molecular ions are the lucky survivors**. *J Mass Spectrom* 2000, **35**(1):1-12.
28. Takach E, Hines W, Patterson D, Juhasz P, Falick A, Vestal M, Martin S: **Accurate mass measurements using MALDI-TOF with delayed extraction**. *J Prot Chem* 1997, **16**(5):363-369.
29. Bairoch A, Boeckmann B: **The SWISS-PROT protein sequence data bank: current status**. *Nucleic Acids Res* 1994, **22**(17):3578-3580.
30. UniProt Consortium T: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2018, **46**(5):2699.
31. Schleuder D, Hillenkamp F, Strupat K: **IR-MALDI-mass analysis of electroblotted proteins directly from the membrane: comparison of different membranes, application to on-membrane digestion, and protein identification by database searching**. *Anal Chem* 1999, **71**(15):3238-3247.
32. Eckerskorn C, Strupat K, Karas M, Hillenkamp F, Lottspeich F: **Mass spectrometric analysis of blotted proteins after gel electrophoretic separation by matrix-assisted laser desorption/ionization**. *Electrophoresis* 1992, **13**(9-10):664-665.
33. Eckerskorn C, Strupat K, Schleuder D, Hochstrasser D, Sanchez JC, Lottspeich F, Hillenkamp F: **Analysis of proteins by direct-scanning infrared-MALDI mass spectrometry after 2D-PAGE separation and electroblotting**. *Anal Chem* 1997, **69**(15):2888-2892.
34. Tonella L, Walsh BJ, Sanchez JC, Ou K, Wilkins MR, Tyler M, Frutiger S, Gooley AA, Pescaru I, Appel RD *et al*: **'98 Escherichia coli SWISS-2DPAGE database update**. *Electrophoresis* 1998, **19**(11):1960-1971.
35. Hochstrasser D, Sanchez JC, Binz PA, Bienvenut W, Appel RD: **A clinical molecular scanner to study human proteome complexity**. *Novartis Found Symp* 2000, **229**:33-38; discussion 38-40.
36. Bienvenut WV, Sanchez JC, Karmime A, Rouge V, Rose K, Binz PA, Hochstrasser DF: **Toward a clinical molecular scanner for proteome research: parallel protein chemical processing before and during western blot**. *Anal Chem* 1999, **71**(21):4800-4807.
37. Bienvenut WV, Deon C, Pasquarello C, Campbell JM, Sanchez JC, Vestal ML, Hochstrasser DF: **Matrix-assisted laser desorption/ionization-tandem mass spectrometry with high resolution and sensitivity for identification and characterization of proteins**. *Proteomics* 2002, **2**(7):868-876.
38. Binz PA, Muller M, Walther D, Bienvenut WV, Gras R, Hoogland C, Bouchet G, Gasteiger E, Fabbretti R, Gay S *et al*: **A molecular scanner to automate proteomic research and to display proteome images**. *Anal Chem* 1999, **71**(21):4981-4988.
39. Rosenfeld J, Capdevielle J, Guillemot J, Ferrara P: **In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis**. *Analytical Biochemistry* 1992, **203**(1):173-179.
40. Bienvenut WV, Deon C, Sanchez JC, Hochstrasser DF: **Enhanced protein recovery after electrotransfer using square wave alternating voltage**. *Anal Biochem* 2002, **307**(2):297-303.
41. Wilkins M, Gooley A. In: *Proteome Research: New Frontiers in functional genomics.* Berlin: Springer-Verlag; 1997: 35-64.
42. Clauser KR, Baker P, Burlingame AL: **Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching**. *Anal Chem* 1999, **71**(14):2871-2882.
43. Zhang Z, McElvain J: **Improvements in protein identification by MALDI-TOF MS peptide mapping**. *Anal Chem* 2000, **72**:2337-2350.
44. Muller M, Gras R, Appel RD, Bienvenut WV, Hochstrasser DF: **Visualization and analysis of molecular scanner peptide mass spectra**. *J Am Soc Mass Spectrom* 2002, **13**(3):221-231.
45. Gras R, Muller M, Gasteiger E, Gay S, Binz PA, Bienvenut W, Hoogland C, Sanchez JC, Bairoch A, Hochstrasser DF *et al*: **Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection**. *Electrophoresis* 1999, **20**(18):3535-3550.
46. Muller M, Gras R, Binz PA, Hochstrasser DF, Appel RD: **Molecular scanner experiment with human plasma: improving protein identification by using intensity distributions of matching peptide masses**. *Proteomics* 2002, **2**(10):1413-1425.
47. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF: **Protein identification and analysis tools in the ExPASy server**. *Methods Mol Biol* 1999, **112**:531-552.
48. Scherl A, Zimmermann-Ivol CG, Di Dio J, Vaezzadeh AR, Binz PA, Amez-Droz M, Cochard R, Sanchez JC, Gluckmann M, Hochstrasser DF: **Gold coating of non-conductive membranes before matrix-assisted laser desorption/ionization tandem mass spectrometric analysis prevents charging effect**. *Rapid Commun Mass Spectrom* 2005, **19**(5):605-610.
49. Bienvenut WV, Hoogland C, Greco A, Heller M, Gasteiger E, Appel RD, Diaz JJ, Sanchez JC, Hochstrasser DF: **Hydrogen/deuterium exchange for higher specificity of protein identification by peptide mass fingerprinting**. *Rapid Commun Mass Spectrom* 2002, **16**(6):616-626.
50. Perrot M, Massoni A, Boucherie H: **Sequence requirements for Nalpha-terminal acetylation of yeast proteins by NatA**. *Yeast* 2008, **25**(7):513-527.

51. Sarto C, Deon C, Doro G, Hochstrasser DF, Mocarelli P, Sanchez JC: **Contribution of proteomics to the molecular analysis of renal cell carcinoma with an emphasis on manganese superoxide dismutase**. *Proteomics* 2001, **1**(10):1288-1294.
52. Smith LM, Kelleher NL, Consortium for Top Down P: **Proteoform: a single term describing protein complexity**. *Nat Methods* 2013, **10**(3):186-187.
53. Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, Binz PA, Ou K, Sanchez JC, Bairoch A, Williams KL *et al*: **High-throughput mass spectrometric discovery of protein post-translational modifications**. *J Mol Biol* 1999, **289**(3):645-657.
54. Banks RE, Dunn MJ, Hochstrasser DF, Sanchez JC, Blackstock W, Pappin DJ, Selby PJ: **Proteomics: new perspectives, new biomedical opportunities**. *Lancet* 2000, **356**(9243):1749-1756.
55. Ciechanover A, Ben-Saadon R: **N-terminal ubiquitination: more protein substrates join in**. *Trends Cell Biol* 2004, **14**(3):103-106.
56. Frottin F, Martinez A, Peynot P, Mitra S, Holz RC, Giglione C, Meinnel T: **The proteomics of N-terminal methionine cleavage**. *Mol Cell Proteomics* 2006, **5**(12):2336-2349.
57. Soppa J: **Protein acetylation in archaea, bacteria, and eukaryotes**. *Archaea* 2010, **2010**.
58. Drazic A, Myklebust LM, Ree R, Arnesen T: **The world of protein acetylation**. *Biochim Biophys Acta* 2016, **1864**(10):1372-1401.
59. Bienvenut WV, Giglione C, Meinnel T: **Proteome-wide analysis of the amino terminal status of Escherichia coli proteins at the steady-state and upon deformylation inhibition**. *Proteomics* 2015.
60. Falb M, Aivaliotis M, Garcia-Rizo C, Bisle B, Tebbe A, Klein C, Konstantinidis K, Siedler F, Pfeiffer F, Oesterhelt D: **Archaeal N-terminal protein maturation commonly involves N-terminal acetylation: a large-scale proteomics survey**. *J Mol Biol* 2006, **362**(5):915-924.
61. Aivaliotis M, Gevaert K, Falb M, Tebbe A, Konstantinidis K, Bisle B, Klein C, Martens L, Staes A, Timmerman E *et al*: **Large-scale identification of N-terminal peptides in the halophilic archaea Halobacterium salinarum and Natronomonas pharaonis**. *J Proteome Res* 2007, **6**(6):2195-2204.
62. Mullen JR, Kayne PS, Moerschell RP, Tsunasawa S, Gribskov M, Colavito-Shepanski M, Grunstein M, Sherman F, Sternglanz R: **Identification and characterization of genes and mutants for an N-terminal acetyltransferase from yeast**. *Embo J* 1989, **8**(7):2067-2075.
63. Bienvenut WV, Sumpton D, Martinez A, Lilla S, Espagne C, Meinnel T, Giglione C: **Comparative large scale characterization of plant versus mammal proteins reveals similar and idiosyncratic N-alpha-acetylation features**. *Mol Cell Proteomics* 2012, **11**(6):M111 015131.
64. Linster E, Stephan I, Bienvenut WV, Maple-Grodem J, Myklebust LM, Huber M, Reichelt M, Sticht C, Geir Moller S, Meinnel T *et al*: **Downregulation of N-terminal acetylation triggers ABA-mediated drought responses in Arabidopsis**. *Nat Commun* 2015, **6**:7640.
65. Polevoda B, Arnesen T, Sherman F: **A synopsis of eukaryotic Nalpha-terminal acetyltransferases: nomenclature, subunits and substrates**. *BMC Proc* 2009, **3 Suppl 6**:S2.
66. Hole K, Van Damme P, Dalva M, Aksnes H, Glomnes N, Varhaug JE, Lillehaug JR, Gevaert K, Arnesen T: **The human N-alpha-acetyltransferase 40 (hNaa40p/hNatD) is conserved from yeast and N-terminally acetylates histones H2A and H4**. *PLoS One* 2011, **6**(9):e24713.
67. Song OK, Wang X, Waterborg JH, Sternglanz R: **An Nalpha-acetyltransferase responsible for acetylation of the N-terminal residues of histones H4 and H2A**. *J Biol Chem* 2003, **278**(40):38109-38112.
68. Starheim KK, Gromyko D, Velde R, Varhaug JE, Arnesen T: **Composition and biological significance of the human Nalpha-terminal acetyltransferases**. *BMC Proc* 2009, **3 Suppl 6**:S3.
69. Liszczak G, Arnesen T, Marmorstein R: **Structure of a ternary Naa50p (NAT5/SAN) N-terminal acetyltransferase complex reveals the molecular basis for substrate-specific acetylation**. *J Biol Chem* 2011, **286**(42):37002-37010.
70. Van Damme P, Hole K, Pimenta-Marques A, Helsens K, Vandekerckhove J, Martinho RG, Gevaert K, Arnesen T: **NatF Contributes to an Evolutionary Shift in Protein N-Terminal Acetylation and Is Important for Normal Chromosome Segregation**. *PLoS Genet* 2011, **7**(7):e1002169.
71. Gautschi M, Just S, Mun A, Ross S, Rucknagel P, Dubaquie Y, Ehrenhofer-Murray A, Rospert S: **The yeast N(alpha)-acetyltransferase NatA is quantitatively anchored to the ribosome and interacts with nascent polypeptides**. *Mol Cell Biol* 2003, **23**(20):7403-7414.
72. Van Damme P, Hole K, Gevaert K, Arnesen T: **N-terminal acetylome analysis reveals the specificity of Naa50 (Nat5) and suggests a kinetic competition between N-terminal acetyltransferases and methionine aminopeptidases**. *Proteomics* 2015.
73. Arnesen T, Betts MJ, Pendino F, Liberles DA, Anderson D, Caro J, Kong X, Varhaug JE, Lillehaug JR: **Characterization of hARD2, a processed hARD1 gene duplicate, encoding a human protein N-alpha-acetyltransferase**. *BMC Biochem* 2006, **7**:13.
74. Arnesen T, Gromyko D, Kagabo D, Betts MJ, Starheim KK, Varhaug JE, Anderson D, Lillehaug JR: **A novel human NatA Nalpha-terminal acetyltransferase complex: hNaa16p-hNaa10p (hNat2-hArd1)**. *BMC Biochem* 2009, **10**:15.
75. Aksnes H, Goris M, Stromland O, Drazic A, Waheed Q, Reuter N, Arnesen T: **Molecular Determinants of the N-Terminal Acetyltransferase Naa60 Anchoring to the Golgi Membrane**. *J Biol Chem* 2017.
76. Park EC, Szostak JW: **ARD1 and NAT1 proteins form a complex that has N-terminal acetyltransferase activity**. *EMBO J* 1992, **11**(6):2087-2093.

77. Arnesen T, Gromyko D, Pendino F, Ryningen A, Varhaug JE, Lillehaug JR: **Induction of apoptosis in human cells by RNAi-mediated knockdown of hARD1 and NATH, components of the protein N-alpha-acetyltransferase complex**. *Oncogene* 2006, **25**(31):4350-4360.
78. Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, Swensen JJ, Johnson WE, Moore B, Huff CD, Bird LM *et al*: **Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency**. *Am J Hum Genet* 2011, **89**(1):28-43.
79. Saunier C, Stove SI, Popp B, Gerard B, Blenski M, AhMew N, de Bie C, Goldenberg P, Isidor B, Keren B *et al*: **Expanding the Phenotype Associated with NAA10-Related N-Terminal Acetylation Deficiency**. *Hum Mutat* 2016, **37**(8):755-764.
80. Gromyko D, Arnesen T, Ryningen A, Varhaug JE, Lillehaug JR: **Depletion of the human Nalpha-terminal acetyltransferase A induces p53-dependent apoptosis and p53-independent growth inhibition**. *Int J Cancer* 2010, **127**(12):2777-2789.
81. Lee CF, Ou DS, Lee SB, Chang LH, Lin RK, Li YS, Upadhyay AK, Cheng X, Wang YC, Hsu HS *et al*: **hNaa10p contributes to tumorigenesis by facilitating DNMT1-mediated tumor suppressor gene silencing**. *J Clin Invest* 2010, **120**(8):2920-2930.
82. Lee MN, Kweon HY, Oh GT: **N-alpha-acetyltransferase 10 (NAA10) in development: the role of NAA10**. *Exp Mol Med* 2018, **50**(7):87.
83. Wang Z, Guo J, Li Y, Bavarva JH, Qian C, Brahimi-Horn MC, Tan D, Liu W: **Inactivation of androgen-induced regulator ARD1 inhibits androgen receptor acetylation and prostate tumorigenesis**. *Proc Natl Acad Sci U S A* 2012, **109**(8):3053-3058.
84. Nesterchuk MV, Sergiev PV, Dontsova OA: **Posttranslational Modifications of Ribosomal Proteins in Escherichia coli**. *Acta Naturae* 2011, **3**(2):22-33.
85. Gordiyenko Y, Deroo S, Zhou M, Videler H, Robinson CV: **Acetylation of L12 increases interactions in the Escherichia coli ribosomal stalk complex**. *J Mol Biol* 2008, **380**(2):404-414.
86. Arnesen T, Van Damme P, Polevoda B, Helsens K, Evjenth R, Colaert N, Varhaug JE, Vandekerckhove J, Lillehaug JR, Sherman F *et al*: **Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans**. *Proc Natl Acad Sci U S A* 2009, **106**(20):8157-8162.
87. Bienvenut WV, Martinez A, Sumpton D, Lilla S, Meinnel T, Giglione C: **α-N protein acetylation: A N-terminal modification of increasing interest.** In: *27éme Congrès de la SFEAP: 6-8 septembre 2010 2010; Marseille*. Marseille: SFEAP; 2010: 60.
88. Bienvenut WV, Espagne C, Martinez A, Majeran W, Valot B, Zivy M, Vallon O, Adam Z, Meinnel T, Giglione C: **Dynamics of post-translational modifications and protein stability in the stroma of Chlamydomonas reinhardtii chloroplasts**. *Proteomics* 2011, **11**(9):1734-1750.
89. Yi CH, Pan H, Seebacher J, Jang IH, Hyberts SG, Heffron GJ, Vander Heiden MG, Yang R, Li F, Locasale JW *et al*: **Metabolic regulation of protein N-alpha-acetylation by Bcl-xL promotes cell survival**. *Cell* 2011, **146**(4):607-620.
90. Varland S, Aksnes H, Kryuchkov F, Impens F, Van Haver D, Jonckheere V, Ziegler M, Gevaert K, Van Damme P, Arnesen T: **N-terminal Acetylation Levels Are Maintained During Acetyl-CoA Deficiency in Saccharomyces cerevisiae**. *Mol Cell Proteomics* 2018, **17**(12):2308-2323.
91. Xu F, Huang Y, Li L, Gannon P, Linster E, Huber M, Kapos P, Bienvenut W, Polevoda B, Meinnel T *et al*: **Two N-Terminal Acetyltransferases Antagonistically Regulate the Stability of a Nod-Like Receptor in Arabidopsis**. *Plant Cell* 2015.
92. Zybailov B, Rutschow H, Friso G, Rudella A, Emanuelsson O, Sun Q, van Wijk KJ: **Sorting signals, N-terminal modifications and abundance of the chloroplast proteome**. *PLoS One* 2008, **3**(4):e1994.
93. Sonsmann G, Romer A, Schomburg D: **Investigation of the influence of charge derivatization on the fragmentation of multiply protonated peptides**. *J Am Soc Mass Spectrom* 2002, **13**(1):47-58.
94. Bienvenut WV, Estreicher A, Potts A, Quadroni M: **Identification by tandem mass spectrometry of N-terminal acetylated proteins in eukaryotic samples.** In: *1er Symposium de Chimie et Biologie Analytiques: 26-29 Septembre 2005 2005; Corum de Montpellier*. 115.
95. Kanor-Kudaya S: **Enrichment of N-alpha Terminal Acetylation proteins and their predictions**. Geneva: Geneva university; 2006.
96. Kleifeld O, Doucet A, auf dem Keller U, Prudova A, Schilling O, Kainthan RK, Starr AE, Foster LJ, Kizhakkedathu JN, Overall CM: **Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products**. *Nat Biotechnol* 2010, **28**(3):281-288.
97. Bertaccini D, Vaca S, Carapito C, Arsene-Ploetze F, Van Dorsselaer A, Schaeffer-Reiss C: **An improved stable isotope N-terminal labeling approach with light/heavy TMPP to automate proteogenomics data validation: dN-TOP**. *J Proteome Res* 2013, **12**(6):3063-3070.
98. Bienvenut WV, Giglione C, Meinnel T: **SILProNAQ: A Convenient Approach for Proteome-Wide Analysis of Protein N-Termini and N-Terminal Acetylation Quantitation**. *Methods Mol Biol* 2017, **1574**:17-34.
99. Bienvenut WV, Scarpelli JP, Dumestier J, Meinnel T, Giglione C: **EnCOUNTer: a parsing tool to uncover the mature N-terminus of organelle-targeted proteins in complex samples**. *BMC Bioinformatics* 2017, **18**(1):182.
100. Mohammed S, Heck A, Jr.: **Strong cation exchange (SCX) based analytical methods for the targeted analysis of protein post-translational modifications**. *Curr Opin Biotechnol* 2010, **22**(1):9-16.

101. Helbig AO, Gauci S, Raijmakers R, van Breukelen B, Slijper M, Mohammed S, Heck AJ: **Profiling of N-acetylated protein termini provides in-depth insights into the N-terminal nature of the proteome**. *Mol Cell Proteomics* 2010, **9**(5):928-939.
102. Kleifeld O, Doucet A, Prudova A, auf dem Keller U, Gioia M, Kizhakkedathu JN, Overall CM: **Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates**. *Nat Protoc* 2011, **6**(10):1578-1611.
103. Li L, Yan G, Zhang X: **Isolation of acetylated and free N-terminal peptides from proteomic samples based on tresyl-functionalized microspheres**. *Talanta* 2015, **144**:122-128.
104. Zhang X, Ye J, Hojrup P: **A proteomics approach to study in vivo protein N(alpha)-modifications**. *J Proteomics* 2009, **73**(2):240-251.
105. Frottin F, Espagne C, Traverso JA, Mauve C, Valot B, Lelarge-Trouverie C, Zivy M, Noctor G, Meinnel T, Giglione C: **Cotranslational proteolysis dominates glutathione homeostasis to support proper growth and development**. *Plant Cell* 2009, **21**(10):3296-3314.
106. Cook RK, Sheff DR, Rubenstein PA: **Unusual metabolism of the yeast actin amino terminus**. *J Biol Chem* 1991, **266**(25):16825-16833.
107. Rowland E, Kim J, Bhuiyan NH, van Wijk KJ: **The Arabidopsis Chloroplast Stromal N-Terminome: Complexities of Amino-Terminal Protein Maturation and Stability**. *Plant Physiol* 2015, **169**(3):1881-1896.
108. Bouchnak I, van Wijk KJ: **N-Degron Pathways in Plastids**. *Trends Plant Sci* 2019, **24**(10):917-926.
109. Goetze S, Qeli E, Mosimann C, Staes A, Gerrits B, Roschitzki B, Mohanty S, Niederer EM, Laczko E, Timmerman E *et al*: **Identification and functional characterization of N-terminally acetylated proteins in Drosophila melanogaster**. *PLoS Biol* 2009, **7**(11):e1000236.
110. Ree RM, Myklebust LM, Thiel P, Foyn H, Fladmark KE, Arnesen T: **The N-terminal acetyltransferase Naa10 is essential for zebrafish development**. *Biosci Rep* 2015.
111. Popp B, Stove SI, Endele S, Myklebust LM, Hoyer J, Sticht H, Azzarello-Burri S, Rauch A, Arnesen T, Reis A: **De novo missense mutations in the NAA10 gene cause severe non-syndromic developmental delay in males and females**. *Eur J Hum Genet* 2014.
112. Ree R, Geithus AS, Torring PM, Sorensen KP, Damkjaer M, study DDD, Lynch SA, Arnesen T: **A novel NAA10 p.(R83H) variant with impaired acetyltransferase activity identified in two boys with ID and microcephaly**. *BMC Med Genet* 2019, **20**(1):101.
113. Van Damme P, Evjenth R, Foyn H, Demeyer K, De Bock PJ, Lillehaug JR, Vandekerckhove J, Arnesen T, Gevaert K: **Proteome-derived peptide libraries allow detailed analysis of the substrate specificities of N{alpha}-acetyltransferases and point to hNaa10p as the posttranslational actin N{alpha}-acetyltransferase**. *Mol Cell Proteomics* 2011.
114. Dinh TV, Bienvenut WV, Linster E, Feldman-Salit A, Jung VA, Meinnel T, Hell R, Giglione C, Wirtz M: **Molecular identification and functional characterization of the first Nalpha-acetyltransferase in plastids by global acetylome profiling**. *Proteomics* 2015.
115. Jeong JW, Bae MK, Ahn MY, Kim SH, Sohn TK, Bae MH, Yoo MA, Song EJ, Lee KJ, Kim KW: **Regulation and destabilization of HIF-1alpha by ARD1-mediated acetylation**. *Cell* 2002, **111**(5):709-720.
116. Dorfel MJ, Lyon GJ: **The biological functions of Naa10 - From amino-terminal acetylation to human disease**. *Gene* 2015, **567**(2):103-131.
117. Gattiker A, Bienvenut WV, Bairoch A, Gasteiger E: **FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification**. *Proteomics* 2002, **2**(10):1435-1444.
118. Bienvenut WV, Kanor S, Estreicher A, Quadroni M, Xenarios I: **Characterisation, identification and prediction of N-terminus acetylated proteins.** In: *54th ASMS Conference on Mass Spectrometry and Allied Topics: 2006; Seattle, WA*. ASMS.
119. Arnesen T: **Towards a functional understanding of protein N-terminal acetylation**. *PLoS Biol* 2011, **9**(5):e1001074.
120. Martinez A, Traverso JA, Valot B, Ferro M, Espagne C, Ephritikhine G, Zivy M, Giglione C, Meinnel T: **Extent of N-terminal modifications in cytosolic proteins from eukaryotes**. *Proteomics* 2008, **8**(14):2809-2831.
121. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence**. *J Mol Biol* 2000, **300**(4):1005-1016.
122. Small I, Peeters N, Legeai F, Lurin C: **Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences**. *Proteomics* 2004, **4**(6):1581-1590.
123. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W585-587.
124. Hooper CM, Castleden IR, Tanz SK, Aryamanesh N, Millar AH: **SUBA4: the interactive data analysis centre for Arabidopsis subcellular protein locations**. *Nucleic Acids Res* 2017, **45**(D1):D1064-D1074.
125. Emanuelsson O, Nielsen H, von Heijne G: **ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites**. *Protein Sci* 1999, **8**(5):978-984.
126. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: **Locating proteins in the cell using TargetP, SignalP and related tools**. *Nat Protoc* 2007, **2**(4):953-971.

127. Almagro Armenteros JJ, Tsirigos KD, Sonderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H: **SignalP 5.0 improves signal peptide predictions using deep neural networks**. *Nat Biotechnol* 2019, **37**(4):420-423.
128. Sperschneider J, Catanzariti AM, DeBoer K, Petre B, Gardiner DM, Singh KB, Dodds PN, Taylor JM: **LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell**. *Sci Rep* 2017, **7**:44598.
129. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S: **Extensive feature detection of N-terminal protein sorting signals**. *Bioinformatics* 2002, **18**(2):298-305.
130. Claros MG, Vincens P: **Computational method to predict mitochondrially imported proteins and their targeting sequences**. *Eur J Biochem* 1996, **241**(3):779-786.
131. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A: **Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli**. *Nucleic Acids Res* 1982, **10**(9):2997-3011.
132. Gomez SM, Bil KY, Aguilera R, Nishio JN, Faull KF, Whitelegge JP: **Transit peptide cleavage sites of integral thylakoid membrane proteins**. *Mol Cell Proteomics* 2003, **2**(10):1068-1085.
133. Westerlund I, Von Heijne G, Emanuelsson O: **LumenP--a neural network predictor for protein localization in the thylakoid lumen**. *Protein Sci* 2003, **12**(10):2360-2366.
134. Hartl FU: **Cellular Homeostasis and Aging**. *Annu Rev Biochem* 2016, **85**:1-4.
135. Belouah I, Nazaret C, Petriacq P, Prigent S, Benard C, Mengin V, Blein-Nicolas M, Denton AK, Balliau T, Auge S *et al*: **Modeling Protein Destiny in Developing Fruit**. *Plant Physiol* 2019, **180**(3):1709-1724.
136. Jamet E, Roujol D, San-Clemente H, Irshad M, Soubigou-Taconnat L, Renou JP, Pont-Lezica R: **Cell wall biogenesis of Arabidopsis thaliana elongating cells: transcriptomics complements proteomics**. *BMC Genomics* 2009, **10**:505.
137. Fierro-Monti I, Racle J, Hernandez C, Waridel P, Hatzimanikatis V, Quadroni M: **A novel pulse-chase SILAC strategy measures changes in protein decay and synthesis rates induced by perturbation of proteostasis with an Hsp90 inhibitor**. *PLoS One* 2013, **8**(11):e80423.
138. Lewandowska D, ten Have S, Hodge K, Tillemans V, Lamond AI, Brown JW: **Plant SILAC: stable-isotope labelling with amino acids of arabidopsis seedlings for quantitative proteomics**. *PLoS One* 2013, **8**(8):e72207.
139. Snijders AP, de Vos MG, Wright PC: **Novel approach for peptide quantitation and sequencing based on 15N and 13C metabolic labeling**. *J Proteome Res* 2005, **4**(2):578-585.
140. Bindschedler LV, Palmblad M, Cramer R: **Hydroponic isotope labelling of entire plants (HILEP) for quantitative plant proteomics; an oxidative stress case study**. *Phytochemistry* 2008, **69**(10):1962-1972.
141. MacCoss MJ, Wu CC, Liu H, Sadygov R, Yates JR, 3rd: **A correlation algorithm for the automated quantitative analysis of shotgun proteomics data**. *Anal Chem* 2003, **75**(24):6912-6921.
142. Valot B, Langella O, Nano E, Zivy M: **MassChroQ: a versatile tool for mass spectrometry quantification**. *Proteomics* 2011, **11**(17):3572-3577.
143. Nguyen KT, Mun SH, Lee CS, Hwang CS: **Control of protein degradation by N-terminal acetylation and the N-end rule pathway**. *Exp Mol Med* 2018, **50**(7):91.
144. Blein-Nicolas M, Negro SS, Balliau T, Welcker C, Bosquet LC, Nicolas SD, Charcosset A, Zivy M: **Integrating proteomics and genomics into systems genetics provides novel insights into the mechanisms of drought tolerance in maize**. *bioRxiv* 2019:636514.