Consolidated peptide/ protein databases including markers for application I

Deliverable D9.8

21 September 2021

Author(s): Dalel Askri[1)], Karim Arafah[1)], Sébastien Voisin[1)] & Philippe Bulet[2)]

[1)]*Plateforme BioPark d'Archamps, Archamps, France*
[2)]*CR, University Grenoble Alpes, IAB Inserm 1209, CNRS UMR5309, Grenoble, France*

PoshBee

Pan-european assessment, monitoring, and mitigation of stressors on the health of bees

Prepared under contract from the European Commission

Grant agreement No. 773921
EU Horizon 2020 Research and Innovation action

Project acronym:	**PoshBee**
Project full title:	**Pan-european assessment, monitoring, and mitigation of stressors on the health of bee**
Start of the project:	June 2018
Duration:	60 months
Project coordinator:	Professor Mark Brown Royal Holloway and Bedford New College www.poshbee.eu
Deliverable title:	Consolidated peptide/ protein databases including markers for application I
Deliverable n°:	D9.8
Nature of the deliverable:	Other
Dissemination level:	Public
WP responsible:	WP9
Lead beneficiary:	CNRS
Citation:	Askri, D., Arafah, K., Voisin, S. & Bulet, P. (2021). *Consolidated peptide/ protein databases including markers for application I.* Deliverable D9.8 EU Horizon 2020 PoshBee Project, Grant agreement No. 773921.
Due date of deliverable:	Month n°40
Actual submission date:	Month n°40

Deliverable status:

Version	Status	Date	Author(s)
2.0	Draft	21 September 2021	Dalel Askri[1)], Karim Arafah[1)], Sébastien Voisin[1)] & Philippe Bulet[2)] [1)]*Plateforme BioPark d'Archamps, Archamps, France* [2)]*CR, University Grenoble Alpes, IAB Inserm U1209, CNRS UMR 5309, Grenoble, France*

Table of contents

Introduction

This database referred as APIDBase-1.0 lists the *Apis mellifera* proteins identified during the proteomics analysis of the bee haemolymph samples provided by the PoshBee consortium, as part of the experiments of work packages (WP) 3, 5-7 (Figure 1), or collected in our local beehives, as part of the experiments done within WP9.

	Work package titles
WP3	Toxicokinetics, toxicodynamics and interactions among agrochemicals
WP5	Effects of agrochemical-nutrition interactions on bee health in the laboratory
WP6	Effects of agrochemical-pathogen interactions on bee health in the laboratory
WP7	Effects of chemicals and their interactions with other stressors on bees tested in semi-field and field experiments
WP9	OMICS of agrochemical responses in bees

Figure 1: List of the work packages (WP) involved in providing haemolymph samples to WP9.

APIDBase-1.0 has been made public. APIDBase-1.0 is intended to be mined by any researcher looking for specific proteins or interested in cross-referencing their findings with ours. APIDBase-1.0 will be further implemented and updated as more samples are provided by the PoshBee partners and analysed by 10-BIOP.

The list of the identified proteins will be refined. For example, a number of identified proteins have their sequence annotated as hypothetical, low quality, and/or uncharacterized in the reference databases we used during the MS/MS spectra identification step. We are confident of the identification of these proteins in the samples we analysed, and thus could confirm the existence of these hypothetical proteins. We intend to complete APIDBase-1.0 with the description of these proteins that are currently unidentified/putative in the available reference databases (e.g. NCBI, UniProtKB, BeeBase).

Table 1: List of haemolymph pools analysed by 10-BioP and used for APIDBase-1.0

WP	Pools
3-Toxico_Interactions	33
3-Toxico_Interactions	27
3-Toxico_Interactions	12
7-Chem. & Stressors Semi-field, Field exp.	27
5-Agrochemical-Nutrition interactions	28
6-Agrochemical_pathogen interactions	33

1. Bottom-up proteomics workflow to fill the APIDBase-1.0

Haemolymph is the circulating body fluid in invertebrates, equivalent to human blood. As summarized in Figure 2, haemolymph samples collected from *Apis mellifera* were regrouped into pools of five to

eight individual haemolymphs based on the individual Mass Fingerprints generated with MALDI BeeTyping®. The pools were dried by vacuum centrifugation before being analysed by a bottom-up proteomics approach, according to the protocol reported in Masson *et al.*, 2018 and Houdelet *et al.*, 2020. Dried samples are briefly suspended in 20 µL of Rapigest 0.1% in 50 mM ammonium bicarbonate (ABC) buffer. The proteins' cysteine residues are reduced (disulfide bonds are opened) and alkylated (blocked) using dithiothreitol and 4-vinyl-pyridine, respectively. The reduced-alkylated proteins are then digested by trypsin.

After an overnight incubation, samples are acidified, centrifuged, and the supernatant transferred into an HPLC autosampler vial. Samples are separated on a reverse-phase C_{18} capillary column installed on a U3000 nano-HPLC connected to a high-resolution mass spectrometer, a Q-Exactive Orbitrap (all instruments Thermo Scientific). A 155-min long chromatographic method using a linear gradient of acidified acetonitrile was used to separate the peptide digests. The separated peptides were analysed online by the electrospray interface connected to the Q-Exactive Orbitrap for detection and acquisition of MS/MS spectra.

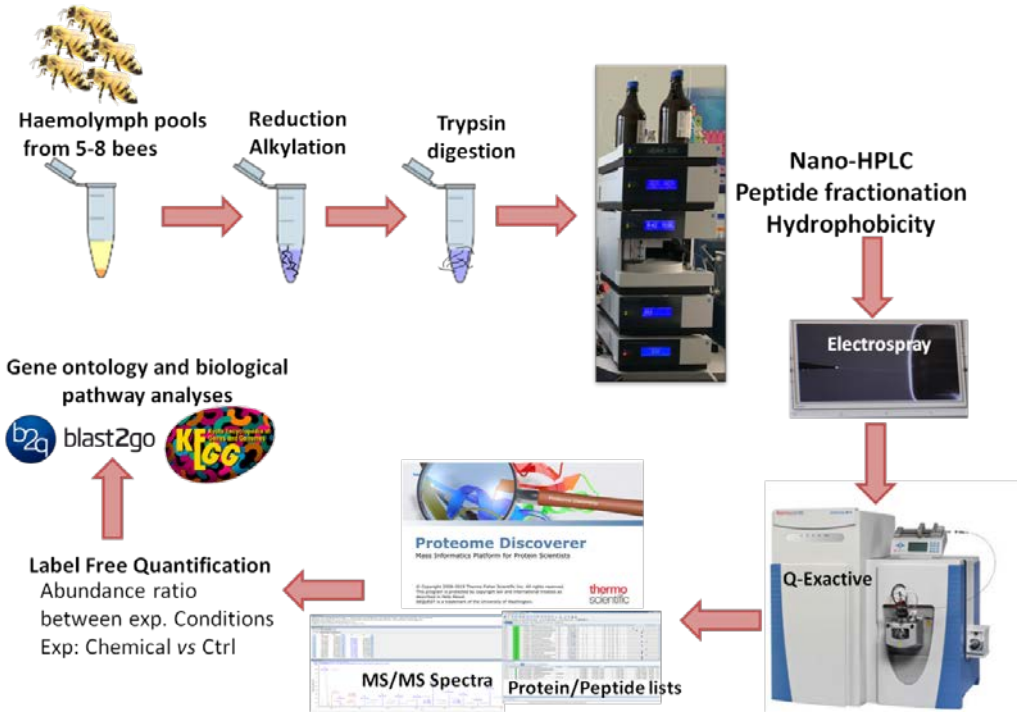

Figure 2: Workflow of bottom-up proteomic analysis of haemolymph samples

2. MS/MS spectra matching against public protein databases

The search algorithm Sequest HT was run by Proteome Discoverer 2.4 (Thermo Fisher Scientific) to match the acquired peptide MS/ MS spectra to a protein sequence database made of entries aggregated from NCBI (https://www.ncbi.nlm.nih.gov/protein) and Uniprot (https://www.uniprot.org/). See Table 2 below for a detailed list of the entries (May 2021 version). The following parameters were used: trypsin digest with two maximum missed cleavages; a tolerance of 10 ppm/0.02 Da for precursors and fragment ions, respectively; cysteine pyridyl-ethylation was set as a fixed modification (4-VP); C-terminal protein amidation, methionine and tryptophan oxidation

were set as variable modifications. The identification confidence was set at a false discovery rate (FDR) of 1%.

Table 2: Organisms added in the protein sequence database used for matching MS analysis

Organism(s)	**Database**	**Entries**
"Hymenoptera"[Organism]	NCBI	1,532,988
Nosema	NCBI	24,258
bee[All Fields] AND virus[All Fields]	NCBI	5,529
invertebrate iridescent virus	NCBI	3,878
Crithidia OR *Lotmaria*	NCBI	32,221
Aethina tumida	NCBI	20,214
Tropilaelaps	NCBI	59,574
Varroa	NCBI	14,703
Pediococcus acidilactici	NCBI	321,381
Serratia marcescens	Uniprot	75,531
Paenibacillus larvae	Uniprot	28,521
Paenibacillus alvei	Uniprot	18,384
Enterococcus faecalis	Uniprot	188,012
Melissococcus plutonius	Uniprot	3,610
Ascosphaera apis	Uniprot	6,492
Aspergillus fumigatus	Uniprot	78,738
Aspergillus flavus	Uniprot	64,428
Aspergillus niger	Uniprot	81,151
Saccharibacter	Uniprot	3,960
Spiroplasma	Uniprot	44,259
Bifidobacterium	Uniprot	505,200
Lactobacillus	Uniprot	1,239,866

3. Structure of APIDBase-1.0

The protein identification reports generated as described above for each analysed haemolymph pool were aggregated together and the protein entries sorted by accession numbers. Redundant entries with the same accession numbers were removed; however, different entries corresponding to different isoforms of the same protein were kept. For this first version of APIDBase-1.0, the protein list was restricted to those belonging to *Apis mellifera* (or one of its sub-species), the first species investigated within PoshBee in WP9. This merged database contains all identified proteins, with no distinction of experimental conditions or quantification data.

The APIDBase-1.0 contains **1934 accession numbers**, corresponding to **1581 distinct proteins**. In the following section, we describe the content of each column in the APIDBase-1.0. The names in bold are the column headers:

Accession: Reference code of the protein entry into the original protein sequence database. Entries of type P35581 or A0A088ADL8 are from UniprotKB, other entry types are from NCBI.

Description: The description of that protein in the UniprotKB or NCBI database.

Species: The organism to which that protein belongs.

NbAAs; MW [kDa]: The number of amino acids (**NbAAs**) and the molecular weight in kilodaltons (**MW [kDa]**) of the full protein sequence. *Caution! The sequence used for these calculations is the full protein sequence deduced from the precursor form in the original UniprotKB/NCBI database entry.* As mentioned for the coverage, to have a corresponding molecular mass, additional calculation needs to be conducted (e.g., deduction of 2 Da per cysteine pairing, and/or elimination of the molecular mass of the signal peptide if predicted by SignalP-5.0 server, and/or the molecular mass of a pro-domain predicted by ProP1.0 Server).

calc. pI: Calculated isoelectric point of the full protein sequence. *Caution! The sequence used for this calculation is the full protein sequence in the original UniprotKB/NCBI database entry, based on the full genomic sequence*.

Biological Process, Cellular Component; Molecular Function, GO Accessions: Gene Ontology (GO) terms recorded in the protein entry. http://geneontology.org/.

Pfam IDs: Pfam protein domains recorded in the entry database. http://pfam.xfam.org/.

Entrez Gene ID; Gene Symbol; Chromosome; Ensembl Gene ID: Genetic information in the protein entry.

KEGG Pathway Accessions; KEGG Pathways; NbProteinPathwayGroups: KEGG information recorded in the protein entry. https://www.genome.jp/kegg/.

4. References

Masson V, Arafah K, Voisin S, Bulet P. (2018) Comparative Proteomics Studies of Insect Cuticle by Tandem Mass Spectrometry: Application of a Novel Proteomics Approach to the Pea Aphid Cuticular Proteins. Proteomics. 2018 Feb;18(3-4). doi: 10.1002/pmic.201700368. Epub 2018 Feb 2. PMID: 29327416

Houdelet C, Sinpoo C, Chantaphanwattana T, Voisin SN, Bocquet M, Chantawannakul P, Bulet P. (2021) Proteomics of Anatomical Sections of the Gut of *Nosema*-Infected Western Honeybee (*Apis mellifera*) Reveals Different Early Responses to *Nosema* spp. Isolates. J Proteome Res. 2021 Jan 1;20(1):804-817. doi: 10.1021/acs.jproteome.0c00658. Epub 2020 Dec 11. PMID: 33305956