# Consolidated peptide/protein databases including markers for application III

## Deliverable D9.10

15 May 2023

Author(s): Dalel Askri[1)], Karim Arafah[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<br>Royal Holloway and Bedford New College www.poshbee.eu |
| Deliverable title: | Consolidated peptide/protein databases including markers for application III |
| Deliverable n°: | D9.10 |
| Nature of the deliverable: | Other |
| Dissemination level: | Public |
| WP responsible: | WP9 |
| Lead beneficiary: | CNRS |
| Citation: | Askri, D., Arafah, K., & Bulet, P. (2023). *Consolidated peptide/protein databases including markers for application III* Deliverable D9.10 EU Horizon 2020 PoshBee Project, Grant agreement No. 773921. |
| Due date of deliverable: | Month n°60 |
| Actual submission date: | Month n°60 |

Deliverable status:

| Version | Status | Date | Author(s) |
|---|---|---|---|
| 1.0 | Draft | 15 May 2023 | Dalel Askri[1)], Karim Arafah[1)], & Philippe Bulet[2)]<br>[1)]*Plateforme BioPark d'Archamps, Archamps, France*<br>[2)]*CR, University Grenoble Alpes, IAB Inserm U1209, CNRS UMR 5309, Grenoble, France* |

# Table of contents

Introduction ..... 4

1. Bottom-up proteomics workflow to fill the OSMDBase-1.0 ..... 4
2. MS/MS spectra matching against public protein databases ..... 5
3. Structure of OSMDBase-1.0 ..... 6
4. Conclusion ..... 7
5. References ..... 8

## Introduction

This database, referred to as OSMDBase-1.0, lists the *Osmia* peptides/proteins identified during the off-gel bottom-up proteomics analysis of the *Osmia* haemolymph samples (Table 1) provided by the PoshBee consortium, as part of the experiments of Work Packages (WPs) 3, 5-7, or collected through in-house experiments (infections with pathogens), as part of the experiments conducted within WP9.

OSMDBase-1.0 has been made public. OSMDBase-1.0 is intended to be mined by any researcher looking for specific proteins or interested in cross-referencing their findings with the ones we generated within PoshBee.

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

| WP | Pools |
|---|---|
| 3-Toxicokinetics, toxicodynamics and interactions among agrochemicals | 30 |
| 5-Agrochemical-Nutrition interactions | 72 |
| 6-Agrochemical-pathogen interactions | 09 |
| 7-Effects of chemicals and their interactions with other stressors on bees tested in semi-field and field experiments | 48 |
| 9-Bacterial infection | 37 |

## 1. Bottom-up proteomics workflow to fill the OSMDBase-1.0

Haemolymph is the circulating body fluid in invertebrates, equivalent to human blood. As summarized in Figure 1, haemolymph samples collected from *Osmia bicornis* were regrouped into pools of five 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, Houdelet *et al.,* 2020, Bournonville *et al.,* 2023 and Askri *et al.,* 2023. Briefly, dried samples were suspended in 20 µL of Rapigest 0.1% in 50 mM ammonium bicarbonate (ABC) buffer, and the proteins' cysteine residues were reduced to open disulfide bonds using dithiothreitol and alkylated (blocked) with 4-vinyl-pyridine. The reduced and alkylated proteins were then digested by trypsin.

After an overnight incubation, samples were acidified, centrifuged, and the supernatant transferred into an HPLC autosampler vial. Samples were 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.

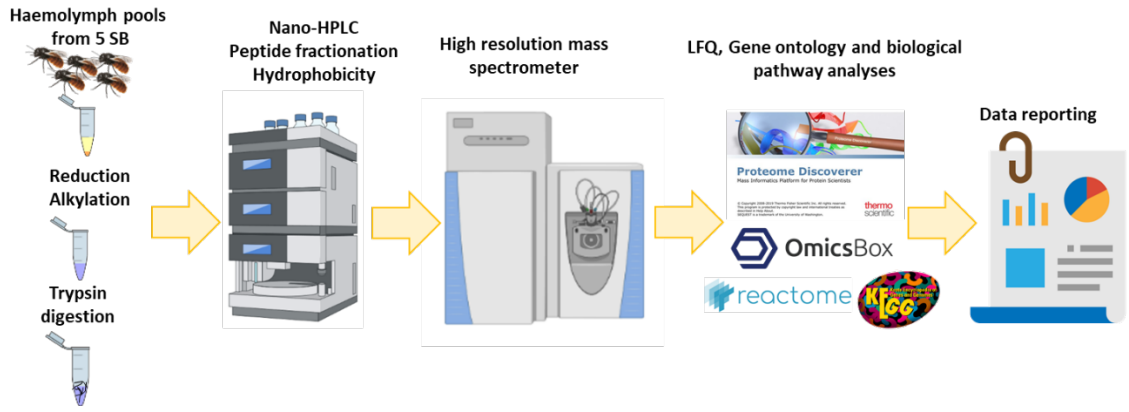


**Figure 1: Workflow of the off-gel bottom-up proteomic analysis of *Osmia* haemolymph samples.**
SB: Solitary bees, HPLC: High performance liquid chromatography, LFQ: Label free quantification

## 2. MS/MS spectra matching against public protein databases

The search algorithm Sequest HT was run by Proteome Discoverer 3.0 (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 UniProtKB (https://www.uniprot.org). Table 2 below details the list of the entries (April 2022 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 oxidations 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 OSMDBase-1.0

The protein identification reports generated as described above for each analysed haemolymph pool were aggregated together. Redundant entries with the same accession numbers were removed. Different entries corresponding to different isoforms of the same protein were kept. For OSMDBase-1.0, the protein lists were restricted to those belonging to *Osmia*. This merged database contains all identified proteins, with no distinction between the experimental condition or quantification data. OSMDBase-1.0 contains **643 accession numbers**, corresponding to **568 proteins** identified in *Osmia* haemolymph pools. Only 13 proteins are referenced as *Osmia* proteins. The rest of the proteins were found by homology in other species of bees *Apis* and *Bombus*. Indeed, in UniProtKB (accessed on 04/05/2023) only 47 entries are registered as *Osmia* proteins and almost 50% of these entries were identified in our samples (13 distinct proteins).

The list of the identified proteins will be refined as a number of identified proteins have their sequence annotated as hypothetical, low quality, and/or uncharacterized (not annotated) in the reference databases we used during the MS/MS spectra identification step. The availability of the *Osmia* genome may serve this issue. We intend to complete OSMDBase-1.0 with the description of these proteins that are currently unidentified/putative in the available reference databases (e.g. NCBI, UniProtKB, BeeBase).

**Table 3: List of the 13 distinct proteins identified as *Osmia* in our *Osmia* haemolymph samples**

| |
|---|
| Cytochrome c oxidase subunit 2 |
| Cytochrome P450 mono-oxygenase |
| Elongation factor 1-alpha (Fragment) |
| Mitochondrial phospholipid hydroperoxide glutathione peroxidase 2 (Fragment) |
| Odorant-binding protein 3 |
| Odorant-binding protein 4 |
| Odorant-binding protein 5 |
| Osmin |
| Peroxiredoxin-6-like protein (Fragment) |
| Serine protease inhibitor |
| Superoxide dismutase (Fragment) |
| Superoxide dismutase 1 (Fragment) |
| Vitellogenin |

The content of each column in the OSMDBase-1.0 is described below.

**The names in bold are the column headers:**

**Accession**: Reference code of the protein entry into the original protein sequence database. Entries of type P81463 or A0A6P3UBV8 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).

**CalcpI:** 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:** Genetic information in the protein entry.

## 4. Conclusion

By this database OSMDBase v1.0, we provide potential biomarkers issued from multi-species studies in experimental conditions covering different stressors in different environments (semi-field and laboratory) in relationship with *Osmia* health. According to the literature, this has never been done

before. All this information has now been made publicly available to serve the scientific community working on bee health and pollinator sustainability.

## 5. References

Askri D, Straw EA, Arafah K, Voisin SN, Bocquet M, Brown MJF, Bulet P. Parasite and Pesticide Impacts on the Bumblebee (*Bombus terrestris*) Haemolymph Proteome. Int J Mol Sci. 2023 Mar 11;24(6):5384. doi: 10.3390/ijms24065384. PMID: 36982462; PMCID: PMC10049270.

Bournonville L, Askri D, Arafah K, Voisin SN, Bocquet M, Bulet P. Unraveling the *Bombus terrestris* Hemolymph, an Indicator of the Immune Response to Microbial Infections, through Complementary Mass Spectrometry Approaches. Int J Mol Sci. 2023 Feb 28;24(5):4658. doi: 10.3390/ijms24054658. PMID: 36902086; PMCID: PMC10003634.

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

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