Accepted Manuscript

Proteomic Analysis of Sweet Algerian Apricot Kernels (*Prunus armeniaca* L.) by Combinatorial Peptide Ligand Libraries and LC-MS/MS

Hamida Ghorab, Carmen Lammi, Anna Arnoldi, Zahia Kabouche, Gilda Aiello

Please cite this article as: Ghorab, H., Lammi, C., Arnoldi, A., Kabouche, Z., Aiello, G., Proteomic Analysis of Sweet Algerian Apricot Kernels (*Prunus armeniaca* L.) by Combinatorial Peptide Ligand Libraries and LC-MS/ MS, *Food Chemistry* (2017), doi: http://dx.doi.org/10.1016/j.foodchem.2017.07.054

Proteomic Analysis of Sweet Algerian Apricot Kernels (*Prunus armeniaca* L.) by Combinatorial Peptide Ligand Libraries and LC-MS/MS

Hamida Ghorab[1‡], Carmen Lammi[2], Anna Arnoldi[2*], Zahia Kabouche[1], Gilda Aiello[2‡]

[1]Université des Frères Mentouri-Constantine, Département de Chimie, Laboratoire d'Obtention des Substances Thérapeutiques (LOST), Campus Chaabet-Ersas, 25000 Constantine, Algeria

[2]Department of Pharmaceutical Sciences, University of Milan, via Mangiagalli 25, 20133 Milan, Italy

*Corresponding author: Anna Arnoldi, Department of Pharmaceutical Sciences, University of Milan, via Mangiagalli 25, 20133 Milan, Italy. Tel +390250319372, anna.arnoldi@unimi.it

[‡] HG and GA gave equivalent contributions to this work.

Abbreviations: CPLLs, Combinatorial peptide ligand libraries; **EEP,** eluted enriched proteins; **ACN,** acetonitrile; **FA,** formic acid; **PPI,** protein-protein interaction network; **PDMQ,** Protein Digestion Multi Query.

Abstract

An investigation on the proteome of the sweet kernel of apricot, based on equalisation with combinatorial peptide ligand libraries (CPLLs), SDS-PAGE, nLC-ESI-MS/MS, and database search, permitted identifying 175 proteins. Gene ontology analysis indicated that their main molecular functions are in nucleotide binding (20.9 %), hydrolase activities (10.6 %), kinase activities (7 %), and catalytic activity (5.6%). A protein-protein association network analysis using STRING software permitted to build an interactomic map of all detected proteins, characterised by 34 interactions. In order to forecast the potential health benefits deriving from the consumption of

these proteins, the two most abundant, i.e. Prunin 1 and 2, were enzymatically digested *in silico* predicting 10 and 14 peptides, respectively. Searching their sequences in the database BIOPEP, it was possible to suggest a variety of bioactivities, including dipeptidyl peptidase-IV (DPP-IV) and angiotensin converting enzyme I (ACE) inhibition, glucose uptake stimulation and antioxidant properties.

Keywords: apricot; combinatorial peptide ligand libraries (CPLL); GO analysis; LC-MS/MS; proteomics; prunin.

1. INTRODUCTION

Prunus armeniaca L. (synonym *Armeniaca vulgaris* Lam., subclass *Rosidae*, order *Rosales*, family *Rosaceae*, subfamily *Prunoideae*), widely known as "apricot", is an edible plant widely cultivated for its delicious fruits. The term *armeniaca* is supposed to derive from Armenia, where this plant has been grown since ancient times. Apricot is currently cultivated in many countries with suitable climates. Algeria and Italy are among major world producers, with annual yields equal to 269,308 and 247,146 tonnes, respectively (FAO, 2012).

The fruit is a drupe, with a thin downy skin, enclosing a yellowish-orange flesh (mesocarp) and an inner large, compressed, smooth, woody stone, containing the kernel. Although the fruit is certainly the most important product of this plant, also the kernel has some interest, being a rich source of dietary protein, oil, and fibre (Femenia, Rossello, Mulet, & Canellas, 1995).

Its use in human diet is limited, however, since it contains D(-)-mandelonitrile β-gentiobioside, a toxic bitter cyanogenic glycoside generally named amygdalin (Varsha, Akash, Jasmine, Raj, Jain, & Chaudhary, 2012; Yildirim & Askin, 2010). Its concentration is nonetheless variable and also some non-toxic cultivars derived from a sweet population from China are available (Lu, Lu, Gao, Lu, Lu, & Gao, 1994). Bitter kernels may contain up to 5.5% amygdalin (Yildirim & Askin, 2010), while sweet ones less than 1% (Karsavuran, Charehsaz, Celik, Asma, Yakinci, & Aydin, 2014). Whereas

bitter kernels are exploited as raw material in cosmetic and pharmaceutical applications, sweet ones may be consumed as nuts (Yildirim & Askin, 2010). They are currently eaten as roasted salted titbit and appetisers or ground to a flour used in different food formulations (Ozboy-Ozbas, Seker, & Gokbulut, 2010; Seker, Ozboy-Ozbas, Gokbulut, Ozturk, & Koksel, 2009; Seker, Ozboy-Ozbas, Gokbulut, Ozturk, & Koksel, 2010), often together with almond flour. Often, however, they are discarded by food processing industry, although their use would help maximising available resources and might result in generating innovative foods (Ozcan, 2000).

The oil of apricot kernels has been widely characterised (Turan, Topcu, Karabulut, Vural, & Hayaloglu, 2007; Yildirim, Yildirim, Askin, & Kankaya, 2010) and shown to possess hypocholesterolaemic and antioxidant properties (Kutlu, Durmaz, Ates, & Erdogan, 2009). Other components extensively characterised are sugars, fibre and phytochemicals (Turan, Topcu, Karabulut, Vural, & Hayaloglu, 2007; Yildirim, Yildirim, Askin, & Kankaya, 2010). Literature indicates that this kernel has antioxidant (Durmaz & Alpaslan, 2007) and antimicrobial properties (Lee, Ahn, Kwon, Lee, Kwak, & Min, 2014).

The fruit proteome has been investigated during ripening (D'Ambrosio, Arena, Rocco, Verrillo, Novi, Viscosi, et al., 2013), whereas an extensive characterisation of the kernel proteome is still lacking. Only one recent paper has developed a highly specific competitive enzyme-linked immunosorbent assay to detect the addition of apricot kernels in almond products, based on specific proteins differential analysis (Zhang, Wang, Huang, Lai, Du, Liu, et al., 2016).

The first objective of the present work was to get a comprehensive information on kernel proteome: the identification of minor proteins was improved by the use of spin columns packed with combinatorial peptide ligand libraries (CPLLs), a powerful non-depleting tool for discovering low-abundant proteins (Aiello, Fasoli, Boschin, Lammi, Zanoni, Citterio, et al., 2016; Esteve, D'Amato, Marina, García, Citterio, & Righetti, 2012; Righetti, Fasoli, & Boschetti, 2011). The apricot kernel proteome was then elucidated by mass spectrometry (LC-MS/MS) and searching in suitable databases. The second objective was the classification of the detected proteins based on

physiological function and localisation using bioinformatic tools. Finally, the third objective was the identification of the potential bioactivities of peptides deriving from the digestion of the most abundant proteins through *in silico* enzymatic digestion and search in the BIOPEP database (Minkiewicz, Dziuba, Iwaniak, Dziuba, & Darewicz, 2008).

2. MATERIALS AND METHODS

2.1 Chemicals

All chemicals and reagents were analytical grade. Acetonitrile (ACN), acetic acid, acetone, formic acid (FA), methanol, sodium hydroxide, ammonium bicarbonate (NH_4HCO_3), trichloroacetic acid (TCA), β-mercaptoethanol, dithiothreitol (DTT), iodoacetamide (IAA), urea, sodium chloride (NaCl), trypsin, glycine, tris-(hydroxymethyl)-aminomethane were from Sigma-Aldrich Corporation (St Louis, MO, USA). ProteoMiner™ (CPLL), Laemmli buffer, Precision Plus Protein Standards, 40% acrylamide/bis solution, N,N,N′,N′-tetramethylethylenediamine (TEMED), sodium dodecyl sulphate (SDS), and electrophoresis apparatus for one-dimensional electrophoresis were acquired at Bio-Rad (Hercules, CA, USA).

2.2 Treatment of apricot kernels

Apricot kernels were purchased in a local market (N'gaous, Batna, Algeria) in July 2015. The sample was extracted according to a literature method (Vita, Lucarotti, Alpi, Balestrini, Mello, Bachi, et al., 2013) with some modifications. Briefly, 0.5 g of kernels were ground to a fine powder using mortar and pestle cooling in ice. The powder was washed with hexane, then extracted with 12 mL of extraction buffer (1 M Urea, 50 mM Tris-HCl pH 8.0, 1% CHAPS, 60 mM DTT) under stirring for 2 h at 4 °C. The homogenate was centrifuged for 30 min at 8,400 g at 4 °C and extracted proteins (supernatant phase) were precipitated using 13% TCA solution containing 0.007% ß-mercaptoethanol in acetone (keeping at -20°C overnight and then at 4 °C for 2 h). The sample was

centrifuged at 8,400 g at 4 °C for 30 min and the pellet was suspended into 5 mL of 50 mM Tris-HCl, pH 8.0, 50 mM NaCl. Part of the supernatant was collected as crude protein extract (**Raw** sample), whereas the other one was submitted to CPLLs incubation (Boschetti & Righetti, 2008). Before the enrichment process, the protein concentration was measured by the Bradford assay using bovine serum albumin as standard. The ProteoMiner enrichment was performed according to the manufacturer protocol. Briefly, the CPLLs spin column containing 50 µL of ProteoMiner beads, stored in a 20% ethanol and 0.5% ACN solution, was conditioned through washing in 1 mL of water followed by 1 mL of phosphate buffered saline (150 mM NaCl, 10 mM NaH_2PO_4). Then, 1 mL of protein solution obtained from the previous extraction was incubated within the column for 2 h at room temperature under gently rocking. In order to remove any excess of non-adsorbed proteins, the CPLL column was washed with 1 mL of PBS buffer for 5 min and then centrifuged at 80 *g*, discarding the solution. To elute bound proteins, the ProteoMiner beads were incubated twice with 100 µL of rehydrated elution reagent (5% acetic acid, 4 M urea, 1% CHAPS), for 15 min at room temperature and centrifuged for 2 min at 80 *g*. The eluted enriched proteins (**EEP**) were collected in a clean tube and stored at -20 °C. All **EEP** solutions were unified and precipitated using chloroform/methanol to remove the SDS excess. The resulting pellets were mixed with Laemmli buffer. Three independent experiments were performed as biological replicates.

2.3 SDS-PAGE analysis

For the SDS–PAGE analysis, 10 µL of **EEP** sample and **Raw** sample were loaded onto a SDS-PAGE gel, composed by a 4% polyacrylamide stacking gel (125 mM Tris-HCl, pH 6.8, 0.1%, m/v, SDS) over a 12% resolving polyacrylamide gel (375 mM Tris-HCl, pH 8.8, 0.1%, m/v, SDS buffer). The cathodic and anodic compartments were filled with Tris-glycine buffer, pH 8.3, containing SDS (0.1%, m/v). Electrophoresis was run at 100 V, until the dye front reached the bottom of the gel and at 150 V until separation end. After electrophoresis, gels were washed and stained/destained using colloidal Coomassie Blue and 7% (v/v) acetic acid in water, respectively.

Images were acquired by the GS800 densitometer and analysed by software Quantity One (Bio-Rad). After this, the two gel lanes corresponding to **EEP** and **Raw** were cut into 11 segments along migration path. Each sample was rinsed with pure water and submitted to standard reduction and alkylation procedures. The reducing solution (150 µL of a 1 mg/mL DTT solution in 50 mM NH_4HCO_3) was added to gel pieces and incubated for 20 min at 56 °C; after its removal, gel pieces were added twice to the washing solution [ACN/50 mM NH_4HCO_3 (1:1)]. The alkylating solution (150 µL, 20 mg/mL IAA solution in 50 mM NH_4HCO_3) was added to gel pieces and incubated in the dark for 45 min. After alkylating solution removal, gel pieces were covered by 50 mM NH_4HCO_3 adding 1 µg of trypsin solution in each sample cooling in ice. After overnight incubation at 37 °C, the resulting tryptic peptides were acidified by adding FA (pH < 3) and the solutions centrifuged at 10,000 rpm at 4 °C for 10 min. The supernatants were then lyophilised using a Speed Vac (Martin Christ), suspended in 20 µL water/ACN (98:2), added with 0.1% FA, and submitted to nano-LC-MS/MS analysis.

2.4 Nano LC-MS/MS analysis

A SL ion trap mass spectrometer (Agilent Technologies, Palo Alto, CA, USA), interfaced to a Chip-nanospray ion source, was operated in data-dependent acquisition mode with the installed Data Analysis software. The reconstituted peptides (4 µL) were loaded onto an enrichment column (Zorbax 300SB-C18, 5 µm pore size) at a flow rate of 4 µL/min for 2 min using isocratic 100% solvent phase (99% water, 1% ACN, and 0.1% FA). After clean-up, the chip valve was switched to separation, conducted in a 43 mm x 75 µm analytical column packed (Zorbax 300SB-C18, 5 µm pore size). The separated peptides were eluted into the mass spectrometer at the constant flow rate of 0.3 µL/min. The LC solvent A was 95% water, 5% ACN, 0.1% FA; solvent B was 5% water, 95% ACN, 0.1% FA. The nano-pump gradient program was as follows: 5% solvent B (0-1 min), 50 % solvent B (1-32 min), 95% solvent B (32-35 min) and back to 5% in 5 min to re-equilibrate the column. The nano-ESI source operated under the following conditions: drying gas temperature 300

°C, flow rate 3 L/min (nitrogen), capillary voltage 1950 V, with endplate offset -500V. Mass spectra were acquired in positive ionisation mode, in the mass range from m/z 300-2200, with target mass 700 m/z, average of 2 spectra, ICC target 30,000 and maximum accumulation time 150 msec. Cationic peptide ions detection was performed by data dependent acquisition AutoMS(n) mode with a dynamic exclusion set at 2 spectra and released after 0.1 min. Each sample was analysed twice. All MS/MS spectra of each duplicate were combined and submitted to database search.

2.5 Protein identification from MS data

Proteins were identified using Spectrum Mill MS Proteomics Workbench (Rev B.04.00, Agilent Technologies, Palo Alto, CA, USA) against the Uniprot-Plants-Viridiplantae database (3,579,823 sequences) obtained from the Uniprot database (version of January, 2016) (The UniProt Consortium, 2017). The parameters used were as follows: carbamidomethylation of cysteine (+57.02 Da) and oxidation of methionine (+15.99 Da) were set as a fixed and variable modification respectively; trypsin was specified as the proteolytic enzyme; 2 missed cleavages were allowed; peptide mass tolerance was set at 1 Da, fragment mass tolerance at 0.7 Da, and peptide charge at +2 and +3. The thresholds used for peptide identification were peptide Local FDR $\leq$ 1%, Score Peak Intensity% $\geq$ 70%, difference of forward and reverse scores $\geq$ 2. Proteins were considered detected when identified by at least two peptides. Supplementary Table S1 reports a list of all identified proteins, MS/MS scores, sequence coverages, and all amino acid sequences of unique recognised peptides.

2.6 Functional categorisation of identified proteins

The identified proteins were categorised based on their cellular localisation, molecular functions or biological processes by using the open access software QuickGo (Binns, Dimmer, Huntley, Barrell, O'Donovan, & Apweiler, 2009). The pathway mapping was performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa, Goto, Kawashima, & Nakaya,

2002). The STRING tool (Search Tool for the Retrieval of Interacting Genes) v.10 software (Szklarczyk, Franceschini, Wyder, Forslund, Heller, Huerta-Cepas, et al., 2015), set on *Arabidopsis thaliana* as reference organism, was used to generate a protein-protein interaction network (PPI). A functional analysis aimed at creating an interaction protein map was conducted subjecting the identified proteins (FASTA sequences IDs), detected in all analysed samples (**Raw** and **EEP**) in STRING. In the interactive network output, proteins are represented by nodes and interactions by connecting continuous and discontinuous lines for direct (physical) or indirect (functional) interactions, respectively. Each connection is supported by at least a literature reference or a canonical information stored in STRING dataset. The confidence value (score) was set to 0.7 (high confidence). The pathways classification was performed after the automatic functional enrichment in STRING, based on information provided by KEGG-Pathway Database.

2.7 *In silico* simulated gastrointestinal digestion of major storage proteins and potential biological activities of generated peptides

The investigation on bioactive peptides was conducted only on Prunin 1 and Prunin 2, the most abundant proteins in this kernel. A prediction of potential bioactive peptides encrypted in these proteins was obtained by combining different *in-silico* enzymatic digestions using the software tool PDMQ - Protein Digestion Multi Query (Haraszi, Tasi, Juhasz & Makai, 2015) in order to simulate gastrointestinal processes: pepsin (pH 1.3) was the first enzyme followed by trypsin and chymotrypsin. All generated peptides were then ranked using the tool PeptideRanker (Mooney, Haslam, Pollastri, & Shields, 2012) in order to evaluate the quality of these results. For ranking these peptides based on the probability of being bioactive, N-to-1 neural algorithm was used to produce a list of probability scores (Supplementary Table S2b). A score higher than 0.5 was considered as indicative for "bioactivity". After such filtering, the peptide sequences were searched against the tool BIOPEP (Minkiewicz, Dziuba, Iwaniak, Dziuba, & Darewicz, 2008).

2.8 Statistically Analysis

The proteomic analyses were conducted on two independent samples, each injected twice. The statistical analysis of the mass spectrometry identification, carried out following a decoy (reversed) database search, was performed using the MS/MS search option in Spectrum Mill to account for false positives. Peptide scores were compared to those of reversed peptide scores to obtain a delta forward-reverse score. Database Fwd-Rev Score ≥2 and Local False Discovery Rate ≤1% were used.

3. RESULTS AND DISCUSSION

3.1 Characterisation of apricot kernel proteome

In order to obtain an extensive identification of tiny kernel proteins, they were enriched by equalising the most abundant ones using spin columns packed with CPLLs (Aiello, et al., 2016). The workflow adopted included proteome extraction using a strong reducing buffer, pre-fractionation via CPLLs technology, protein separations by SDS-PAGE, and characterisation by MS analysis. Fig. 1A shows the SDS-PAGE profiles of the **Raw** and **EEP** samples. A comparison of these two profiles confirms that CPLLs were very efficient in increasing low-abundant protein detection. The **Raw** lane is characterised by three major protein bands, with molecular weights (MW) equal to 50–75 kDa, 50 kDa and 25kDa, which correspond to Prunin 1 (E3SH28, 63 kDa), Prunin 2 (E3SH29, 53 kDa) and two uncharacterised proteins (M5XS06 and M5XPY4, 25 kDa). Contrariwise, the **EEP** lane exhibits very intense and additional bands, particularly evident in the high MW region corresponding to 100-75 kDa. As revealed by **EEP** eluate profile, the combined use of CPLLs and strong reducing extraction buffer has contributed to increase capture mainly of membrane proteins, as demonstrated by the subsequently mass spectrometry analysis.

Proteins from **Raw** and **EEP** samples were reduced, alkylated, digested, and analysed using nano-LC−MS/MS. It was possible identifying 175 unique gene products with at least two matched

peptides (Supplementary Table S1). The Venn diagram (Fig. 1B) reports the total IDs in terms of identifications obtained by matching MS/MS data against the Uniprot Viridiplantae database. Among identified proteins 44 (25.2 %) and 73 (41.7%) were specific to **Raw** and **EEP** samples, respectively, whereas 58 (33.1%) were identified in both samples (Table 1). Indeed, the CPLLs capture permitted the additional identification of 73 unique gene products that could not be detected via a conventional solubilisation protocol.

Interestingly, 72% of total IDs was attributable to the *Prunus* genus, specifically three of them were assigned to almond (*Prunus dulcis*), whereas the remaining to peach (*Prunus persica*). Only a few proteins were identified by homology to other plant species, such as *Oryza sativa*, *Setaria italica*, *Triticum aestivum*, *Vitis vinifera*, and *A. thaliana*. One of the most recent papers on *Prunus* seed proteome has considered plum and peach (Gonzalez-Garcia, Marina, Garcia, Righetti, & Fasoli, 2016): the analysis permitted to identify 141 unique gene products in plum and 97 in peach, however, only a small percentage of them belonged to the genus *Prunus*, whereas most belonged to *Glycine max*, *Vitis vinifera*, *Zea mays*, and *Populus trichocarpa*.

The most abundant proteins in apricot kernel were Prunin 1 (E3SH28) and Prunin 2 (E3SH29), belonging to the cupin family and associated to 11S globulins. Prunin 1 has been identified as the major component in almond and has been recognised as a main almond allergen (Jin, Albillos, Guo, Howard, Fu, Kothary, et al., 2009). Beyond the nutrient reservoir function, the observed protein profile includes many GTPases involved in vesicle trafficking, cytoskeletal organisation and signal transduction as recently confirmed in peach by another paper (Falchi, Cipriani, Marrazzo, Nonis, Vizzotto, & Ruperti, 2010). We have identified also phosphatidylinositol binding protein, an important lipid binding proteins, playing roles in the stabilisation of membranes, cell wall organisation, seed development and germination (Liu, Zhang, Lu, Zeng, Li, Fu, et al., 2015). In addition, some methyltransferases were recognised. These proteins are connected to the biosynthesis of volatile phenolic derivatives in plants (Lavid, Wang, Shalit, Guterman, Bar,

Beuerle, et al., 2002), as well as involved in the enzymatic methylation of polyphenols, which results in antimicrobial properties.

3.2 Protein functional data analysis

The protein species identified were categorised into specific clusters, using the GO annotation tool, which allows automatic information retrieval from the website available databank. Fig. 2 shows the obtained clusters in which they are involved: biological processes (panel A), molecular functions (panel B), and cellular components (Panel C).

Focusing the attention on biological processes (Fig. 2A), the largest clusters include proteins involved in biological process regulation (12.6%), cellular processes (10.2%), transport activities (8.8%), metabolic processes (8.4%), cellular component organisation (6.9%) and multicellular organisms development (6.4%). Fewer proteins are involved in cellular processes, such as growth factor, response to extracellular stimulus, and in cell-cell signalling. Regarding molecular functions (Fig. 2B), 20.9% proteins are involved in nucleotide binding, 10.6% in hydrolase activities, 7.0% in kinase activities, 5.6% in catalytic activities, and only 2.6% in biosynthetic processes. As for subcellular localisation (Fig. 2C), most detected IDs are localised in membranes (48.4%), nucleus (10.1%), plastids (9.2%), cytoplasm (6.0%), and cell walls (1.4%).

Fig. 3 shows a comparison of the GO for molecular function annotation between **EEP** and **Raw**, based on the proteins uniquely identified in each sample. A 47% increment in the proteome discovery, attributed to minor proteins unrevealed without the CPLLs treatment, was observed for binding activity, whereas a 31% enhancement for catalytic activity.

3.3 Protein biology

The protein classification based on their functional roles highlighted four main categories, transcription regulators, transmembrane transporters, stress-related proteins, and binding activities,

providing a relevant contribution to kernel proteome knowledge during development and maturation.

GRAS domain protein and TFIIB/Zinc finger transcription factor have been identified as transcriptional regulators. Involved in the yield and quality of storage compounds, transcription factors are considered the main protagonists controlling early seed developments as well as genome-wide epigenetic events (Ikeda, 2012). Associated to well-established TFIIB, the mediator complex family, including MD13, has emerged as the most crucial cofactor in RNAP II-mediated transcriptional events due to its role either in growth or developmental processes or biotic and abiotic stress response (Samanta & Thakur, 2015).

The ATP-binding cassette (ABC) transporter family, responsible of adenosine triphosphate (ATP) hydrolysis, one of the largest protein superfamilies in biology represented in all living organisms, is strictly correlated to active transport of a wide variety of substrates through the mitochondrial outer membranes and cytosol (Rea, 2007). The ABC transporter protein superfamily members share a hydrolytic ability useful in a wide array of functions, including DNA repair and RNA translocation.

The voltage dependent anion channel (VDAC) assumes an important role in the regulation of energetic and metabolic functions (Shoshan-Barmatz & Ben-Hail, 2012). It is involved in Ca^{2+} transport across the mitochondrial outer membrane (Bathori, Csordas, Garcia-Perez, Davies, & Hajnoczky, 2006). The regulation of mitochondrial physiology needs an efficient metabolic exchange systems identified here into the Solute Carrier Protein Family, responsible of sugar-phosphate/phosphate exchange.

Pectinesterase (PME) was identified as the major protein involved in cell wall metabolism. The demethylation of cell wall pectins is mediated by pectin methylesterases, whose activity alters cell walls and mediates various physiological and biochemical processes in plants, including elongation growth, water uptake, and fruit ripening (Peaucelle, Braybrook, & Hofte, 2012).

Several proteins related to stress-response, including 17.7 KDa class I heat shock proteins (HSPs) and NB-ARC domain protein, were also detected. Although HSPs were initially identified as being

induced by heat stress, they are overrepresented under a variety of physical and chemical stimuli as well as during plant development and seeds growth (Koo, Park, Kim, Suh, Lee, Lee, et al., 2015; W. Wang, Vinocur, Shoseyov, & Altman, 2004). The possible important roles of HSPs in fruit development and ripening have been recently reported in tomato (Neta-Sharir, Isaacson, Lurie, & Weiss, 2005), apple (A. D. Wang, Tan, Tatsuki, Kasai, Li, Saito, et al., 2009), and apricot (Grimplet, Romieu, Audergon, Marty, Albagnac, Lambert, et al., 2005; Manganaris, Rasori, Bassi, Geuna, Ramina, Tonutti, et al., 2011).

3.4 PPI network of apricot kernel

A proteome interactomic map was obtained using the STRING tool for obtaining cross-correlation information. *A. thaliana* was selected as a reference organism, considering the lack of an extensive genome database for *Prunus* specie and the phylogenetic proximity of these species. Fig. 4 shows the PPI network for apricot kernel (*p* value = 0.045) calculated by STRING (confidence score value > 0.7). By removing unconnected proteins, the PPI network contains 34 interactions. Proteins with the highest score values are characterised by the highest connectivity in the network. Supplementary Table S2a lists this latter information and all details, including phylogenetic co-occurrence, genetic neighbourhood and co-expression for each interaction. Among the overall interactions, 14 were endowed high confidence score value (upper to 0.9). The widest interactions involve proteins belonging to RNA-polymerase family protein as well as proteins with transcription regulation and nucleosome positioning activities, such as Chromatin structure-remodelling complex protein SYD.

3.5 Bioactivities of peptides from in silico digestion of prunins

There is now a big interest for bioactive peptides from food sources (Arnoldi, Zanoni, Lammi, & Boschin, 2015) and some previous studies (Gonzalez-Garcia, Marina, & Garcia, 2014; Gonzalez-Garcia, Puchalska, Marina, & Garcia, 2015; Vasquez-Villanueva, Marina, & Garcia, 2015) had given indication that hydrolysates from total protein extracts from kernels of the *Prunus* genus

provide antioxidant activities and angiotensin converting enzyme (ACE) inhibition. The final part of the work was, therefore, dedicated to investigate the possible physiological roles of peptides deriving from Prunin 1 (E3SH28) and Prunin 2 (E3SH29), the major storage proteins in apricot kernel. They were subjected to *in silico* digestion by sequential hydrolysis with pepsin, trypsin, and chymotrypsin. Although useful for obtaining a feasible prediction, this model does not exhaustively represent gastrointestinal digestion, because it does not consider several important aspects, such as peptidase activities, pH variations, and microbiota effects. Only peptides derived from 0 missed cleavage and containing at least 4 amino acid residues were further evaluated, whereas smaller peptides (i.e. di- and tripeptides with MW < 400 Da) were not considered, since they cannot unambiguously belong to a single protein.

To optimise the potential bioactive candidates selection, the predicted peptidome map was ranked by the tool PeptideRanker. A score value was assigned to each peptide using N-to-1 neural network probability: peptides showing score values higher than 0.5 were considered to be potentially bioactive. In a probability range from 0 to 1, predicted values closest to 1 indicate a more confident prediction that the candidate resembles a bioactive peptide (Mooney, Haslam, Pollastri, & Shields, 2012). The total number of predicted peptides after *in silico* digestion was 10 for Prunin 1 and 14 for Prunin 2 (Table 2). Out of these 24 peptides, 7 sequences had high probability to be bioactive, three deriving from Prunin 1 and four from Prunin 2.

The complete peptidome map (24 peptides) was then searched by using the BIOPEP database, including information about known bioactive peptides, in order to associate potential bioactivities to their sequences. This suggested numerous potential bioactivities, including dipeptidyl peptidase-IV (DPP-IV) and angiotensin converting enzyme I (ACE) inhibition, antioxidant properties, and glucose uptake stimulating activity. Most predicted peptides exhibited multifunctional activities, since they contain common structural requirements for each activity (Yea, Ebrahimpour, Hamid, Bakar, Muhammad, & Saari, 2014). Our discussion will be restricted to peptides with high PeptideRanker scores.

All seven peptides were predicted as potential inhibitors of DPP-IV activity. This is a new molecular target correlated with type 2 diabetes development. This ubiquitous enzyme has been shown to cleave and inactivate glucagon-like peptide-1 (GLP-1) and glucose dependent insulinotropic polypeptide (GIP) in the postprandial phase, leading to a loss in their insulinotropic activity (Juillerat-Jeanneret, 2014). In order to exert this activity, a peptide should have a hydrophobic character, a length varying from 2 to 8 amino acid residues, and contain a Pro residue located at the first, second, third, or fourth N-terminal position (Lammi, Zanoni, Arnoldi, & Vistoli, 2016). Based on these considerations, GRPR, NLPIL, NPQGGR, NPSDPQF, SATSPPR, and TPHW seem good candidates as DPP-IV inhibitors. DPP-IV inhibitory peptides have been found in hydrolysates from lupin and soybean protein (Lammi, et al, 2016).

Six peptides were predicted as ACE-inhibitors, i.e. they are potentially hypotensive. The binding of a peptide to the ACE active site is strongly influenced by its C-terminal sequence (Hernandez-Ledesma, Contreras, & Recio, 2011). Structure-activity studies have shown that a positively charged C-terminal residue, such as the ε-amino group of Lys and the guanidine group of Arg, is very important to exert this activity as well as a hydrophobic amino acid residue (aromatic or branched side chain) at least in one of the three C-terminal positions. These peptide motifs are present within peptides GRPR, NPQGGR, NPSDPQF, SATSPPR and TPHW.

Two peptides, SPHW and TPHW, were predicted to be antioxidant for the presence of the fragment *PHW*, which provides this kind of activity. Only peptide NLPIL was predicted as glucose uptake stimulating, owing to the presence of the sequence *IL*. This peptide had been already identified in the sequence of plum kernel proteins (*Prunus domestica* L.) and shown to exert antioxidant capacity *in vitro* (Gonzalez-Garcia, Marina, Garcia, Righetti, & Fasoli, 2016).

An open issue is the stability of these peptides toward hydrolysis by endoproteinases activities that was checked using the open access tool PROSPER (Song, Tan, Perry, Akutsu, Webb, Whisstock, et al., 2012). This process showed that out of these seven peptides only two are susceptible to partial

hydrolysis by endoproteinases: the former, NLPIL, might be cleaved by elastase-2, matrix metalloprotease-9 and cathepsine-K, whereas the latter, NPQGGR, only by cathepsine-K.

However, the intestinal phase of digestion is very difficult to model, since small intestinal brush border membrane contains many amino and carboxyl exopeptidases whose action here has not been considered. This is certainly a main limitation of this approach.

4. CONCLUSIONS

In conclusion, taking advantage of high-throughput technologies, this work has provided innovative information on the proteome of apricot kernel. This represents a major improvement in the knowledge of this food material. Although in general this kernel finds a rare use in human nutrition owing to its bitterness, there are some specific recipes, such as Italian *amaretti*, in which this sensorial characteristic is looked for and carefully modulated precisely through a sapient addition of bitter apricot kernels. Moreover, kernels from sweet varieties may be used without any precaution as roasted salted kernels consumed as titbit or as a flour to be included in different food formulations.

The presence of potentially bioactive peptides encrypted inside its proteins sequences suggests that apricot kernel proteins may offer useful health benefits, increasing even more the interest for this kernel. This fact would help maximising available resources and possibly providing new ingredients to exploit in innovative healthy foods. Finally, considering the challenges in correlating food protein sequences and health, we observe that, in order to facilitate proteome analysis in food chemistry, it would be very useful to develop specific bioinformatics tools including data processing, clustering, dynamics, and integration at various omics levels, and designed to take into account properties, such as taste, technological functions and health promoting features.

ACKNOWLEDGMENTS

References

Aiello, G., Fasoli, E., Boschin, G., Lammi, C., Zanoni, C., Citterio, A., & Arnoldi, A. (2016). Proteomic characterization of hempseed (*Cannabis sativa* L.). *Journal of Proteomics, 147*, 187-196.

Arnoldi, A., Zanoni, C., Lammi, C., & Boschin, G. (2015). The role of grain legumes in the prevention of hypercholesterolemia and hypertension. *Critical Reviews in Plant Sciences, 34*(1-3), 144-168.

Bathori, G., Csordas, G., Garcia-Perez, C., Davies, E., & Hajnoczky, G. (2006). Ca2+-dependent control of the permeability properties of the mitochondrial outer membrane and voltage-dependent anion-selective channel (VDAC). *Journal of Biological Chemistry, 281*(25), 17347-17358.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., & Apweiler, R. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics, 25*(22), 3045-3046.

Boschetti, E., & Righetti, P. G. (2008). The ProteoMiner in the proteomic arena: a non-depleting tool for discovering low-abundance species. *Journal of Proteomics, 71*(3), 255-264.

D'Ambrosio, C., Arena, S., Rocco, M., Verrillo, F., Novi, G., Viscosi, V., Marra, M., & Scaloni, A. (2013). Proteomic analysis of apricot fruit during ripening. *Journal of Proteomics, 78*, 39-57.

Durmaz, G., & Alpaslan, M. (2007). Antioxidant properties of roasted apricot (*Prunus armeniaca* L.) kernel. *Food Chemistry, 100*(3), 1177-1181.

Esteve, C., D'Amato, A., Marina, M. L., García, M. C., Citterio, A., & Righetti, P. G. (2012). Identification of olive (*Olea europaea*) seed and pulp proteins by nLC-MS/MS via combinatorial peptide ligand libraries. *Journal of Proteomics, 75*(8), 2396-2403.

Falchi, R., Cipriani, G., Marrazzo, T., Nonis, A., Vizzotto, G., & Ruperti, B. (2010). Identification and differential expression dynamics of peach small GTPases encoding genes during fruit development and ripening. *Journal of Experimental Botany, 61*(10), 2829-2842.

FAO (2012). www.fao.org (2012 data)

Femenia, A., Rossello, C., Mulet, A., & Canellas, J. (1995). Chemical composition of bitter and sweet apricot kernels. *Journal of Agricultural and Food Chemistry, 43*(2), 356-361.

Gonzalez-Garcia, E., Marina, M. L., & Garcia, M. (2014). Plum (*Prunus Domestica* L.) by-product as a new and cheap source of bioactive peptides: Extraction method and peptides characterization. *Journal of Functional Foods, 11*, 428-437.

Gonzalez-Garcia, E., Marina, M. L., Garcia, M. C., Righetti, P. G., & Fasoli, E. (2016). Identification of plum and peach seed proteins by nLC-MS/MS via combinatorial peptide ligand libraries. *Journal of Proteomics, 148*, 105-112.

Gonzalez-Garcia, E., Puchalska, P., Marina, M. L., & Garcia, M. C. (2015). Fractionation and identification of antioxidant and angiotensin-converting enzyme-inhibitory peptides obtained from plum (*Prunus domestica* L.) stones. *Journal of Functional Foods, 19*, 376-384.

Grimplet, J., Romieu, C., Audergon, J. M., Marty, I., Albagnac, G., Lambert, P., Bouchet, J. P., & Terrier, N. (2005). Transcriptomic study of apricot fruit (*Prunus armeniaca*) ripening among 13 006 expressed sequence tags. *Physiologia Plantarum, 125*(3), 281-292.

Haraszi, R., Tasi, C., Juhasz, A., & Makai, S. (2015) PDMQ - Protein Digestion Multi Query software tool to perform in silico digestion of protein/peptide sequences. bioRxiv 014019, doi.org/10.1101/014019

Hernandez-Ledesma, B., Contreras, M. D., & Recio, I. (2011). Antihypertensive peptides: Production, bioavailability and incorporation into foods. *Advances in Colloid and Interface Science, 165*(1), 23-35.

Ikeda, Y. (2012). Plant Imprinted Genes Identified by Genome-wide Approaches and Their Regulatory Mechanisms. *Plant and Cell Physiology, 53*(5), 809-816.

Jin, T. C., Albillos, S. M., Guo, F., Howard, A., Fu, T. J., Kothary, M. H., & Zhang, Y. Z. (2009). Crystal Structure of Prunin-1, a Major Component of the Almond (Prunus dulcis) Allergen Amandin. *Journal of Agricultural and Food Chemistry, 57*(18), 8643-8651.

Juillerat-Jeanneret, L. (2014). Dipeptidyl peptidase IV and its inhibitors: therapeutics for type 2 diabetes and what else? *Journal of Medicinal Chemistry, 57*(6), 2197-2212.

Kanehisa, M., Goto, S., Kawashima, S., & Nakaya, A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Research, 30*(1), 42-46.

Karsavuran, N., Charehsaz, M., Celik, H., Asma, B. M., Yakinci, C., & Aydin, A. (2014). Amygdalin in bitter and sweet seeds of apricots. *Toxicological and Environmental Chemistry, 96*(10), 1564-1570.

Koo, H. J., Park, S. M., Kim, K. P., Suh, M. C., Lee, M. O., Lee, S. K., Xia, X. L., & Hong, C. B. (2015). Small Heat Shock Proteins Can Release Light Dependence of Tobacco Seed during Germination. *Plant Physiology, 167*(3), 1030-+.

Kutlu, T., Durmaz, G., Ates, B., & Erdogan, A. (2009). Protective effect of dietary apricot kernel oil supplementation on cholesterol levels and antioxidant status of liver in hypercholesteremic rats. *Journal of Food Agriculture & Environment, 7*(3-4), 61-65.

Lammi, C., Zanoni, C., Arnoldi, A., & Vistoli, G. (2016). Peptides Derived from Soy and Lupin Protein as Dipeptidyl-Peptidase IV Inhibitors: In Vitro Biochemical Screening and in Silico Molecular Modeling Study. *Journal of Agricultural and Food Chemistry, 64*(51), 9601-9606.

Lavid, N., Wang, J., Shalit, M., Guterman, I., Bar, E., Beuerle, T., Menda, N., Shafir, S., Zamir, D., Adam, Z., Vainstein, A., Weiss, D., Pichersky, E., & Lewinsohn, E. (2002). O-methyltransferases involved in the biosynthesis of volatile phenolic derivatives in rose petals. *Plant Physiol, 129*(4), 1899-1907.

Lee, H.-h., Ahn, J.-H., Kwon, A.-R., Lee, E. S., Kwak, J.-H., & Min, Y.-H. (2014). Chemical Composition and Antimicrobial Activity of the Essential Oil of Apricot Seed. *Phytotherapy Research, 28*(12), 1867-1872.

Liu, F., Zhang, X., Lu, C., Zeng, X., Li, Y., Fu, D., & Wu, G. (2015). Non-specific lipid transfer proteins in plants: presenting new advances and an integrated functional analysis. *Journal of Experimental Botany, 66*(19), 5663-5681.

Lu, Y., Lu, Z., Gao, S., Lu, Y. M., Lu, Z. R., & Gao, S. Z. (1994). A study on the evolution relationship and classification of apricots via peroxidase isozyme zymogram analysis. *Acta Agriculturae Boreali-Sinica, 9*(4), 69-74.

Manganaris, G. A., Rasori, A., Bassi, D., Geuna, F., Ramina, A., Tonutti, P., & Bonghi, C. (2011). Comparative transcript profiling of apricot (*Prunus armeniaca* L.) fruit development and on-tree ripening. *Tree Genetics & Genomes, 7*(3), 609-616.

Minkiewicz, P., Dziuba, J., Iwaniak, A., Dziuba, M., & Darewicz, M. (2008). BIOPEP database and other programs for processing bioactive peptide sequences. *Journal of AOAC International, 91*(4), 965-980.

Mooney, C., Haslam, N. J., Pollastri, G., & Shields, D. C. (2012). Towards the Improved Discovery and Design of Functional Peptides: Common Features of Diverse Classes Permit Generalized Prediction of Bioactivity. *Plos One, 7*(10).

Neta-Sharir, I., Isaacson, T., Lurie, S., & Weiss, D. (2005). Dual role for tomato heat shock protein 21: protecting photosystem II from oxidative stress and promoting color changes during fruit maturation. *Plant Cell, 17*(6), 1829-1838.

Ozboy-Ozbas, O., Seker, I. T., & Gokbulut, I. (2010). Effects of resistant starch, apricot kernel flour, and fiber-rich fruit powders on low-fat cookie quality. *Food Science and Biotechnology, 19*(4), 979-986.

Ozcan, M. (2000). Composition of some apricot (*Prunus armeniaca* L.) kernels grown in Turkey. *Acta Alimentaria, 29*(3), 289-293.

Peaucelle, A., Braybrook, S., & Hofte, H. (2012). Cell wall mechanics and growth control in plants: the role of pectins revisited. *Frontiers in Plant Science, 3.*

Rea, P. A. (2007). Plant ATP-Binding cassette transporters. *Annual Review of Plant Biology, 58*, 347-375.

Righetti, P. G., Fasoli, E., & Boschetti, E. (2011). Combinatorial peptide ligand libraries: The conquest of the 'hidden proteome' advances at great strides. *Electrophoresis, 32*(9), 960-966.

Samanta, S., & Thakur, J. K. (2015). Importance of Mediator complex in the regulation and integration of diverse signaling pathways in plants. *Frontiers in Plant Science, 6.*

Seker, I. T., Ozboy-Ozbas, O., Gokbulut, I., Ozturk, S., & Koksel, H. (2009). Effects of fiber-rich apple and apricot powders on cookie quality. *Food Science and Biotechnology, 18*(4), 948-953.

Seker, I. T., Ozboy-Ozbas, O., Gokbulut, I., Ozturk, S., & Koksel, H. (2010). Utilization of apricot kernel flour as fat replacer in cookies. *Journal of Food Processing and Preservation, 34*(1), 15-26.

Shoshan-Barmatz, V., & Ben-Hail, D. (2012). VDAC, a multi-functional mitochondrial protein as a pharmacological target. *Mitochondrion, 12*(1), 24-34.

Song, J., Tan, H., Perry, A. J., Akutsu, T., Webb, G. I., Whisstock, J. C., & Pike, R. N. (2012). PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. *PLoS One, 7*(11), e50300.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., & von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research, 43*(Database issue), D447-452.

The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research, 45*(D1), D158-D169.

Turan, S., Topcu, A., Karabulut, I., Vural, H., & Hayaloglu, A. A. (2007). Fatty acid, triacylglycerol, phytosterol, and tocopherol variations in kernel oil of Malatya apricots from Turkey. *Journal of Agricultural and Food Chemistry, 55*(26), 10787-10794.

Varsha, R., Akash, J., Jasmine, C., Raj, V., Jain, A., & Chaudhary, J. (2012). *Prunus armeniaca* (Apricot): an overview. *Journal of Pharmacy Research, 5*(8), 3964-3966.

Vasquez-Villanueva, R., Marina, M. L., & Garcia, M. C. (2015). Revalorization of a peach (*Prunus persica* (L.) Batsch) byproduct: Extraction and characterization of ACE-inhibitory peptides from peach stones. *Journal of Functional Foods, 18*, 137-146.

Vita, F., Lucarotti, V., Alpi, E., Balestrini, R., Mello, A., Bachi, A., Alessio, M., & Alpi, A. (2013). Proteins from Tuber magnatum Pico fruiting bodies naturally grown in different areas of Italy. *Proteome Science, 11*.

Wang, A. D., Tan, D. M., Tatsuki, M., Kasai, A., Li, T. Z., Saito, H., & Harada, T. (2009). Molecular mechanism of distinct ripening profiles in 'Fuji' apple fruit and its early maturing sports. *Postharvest Biology and Technology, 52*(1), 38-43.

Wang, W., Vinocur, B., Shoseyov, O., & Altman, A. (2004). Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends in Plant Science, 9*(5), 244-252.

Yea, C. S., Ebrahimpour, A., Hamid, A. A., Bakar, J., Muhammad, K., & Saari, N. (2014). Winged bean [*Psophorcarpus tetragonolobus* (L.) DC] seeds as an underutilised plant source of bifunctional proteolysate and biopeptides. *Food & Function, 5*(5), 1007-1016.

Yildirim, F. A., & Askin, M. A. (2010). Variability of amygdalin content in seeds of sweet and bitter apricot cultivars in Turkey. *African Journal of Biotechnology, 9*(39), 6522-6524.

Yildirim, F. A., Yildirim, A. N., Askin, M. A., & Kankaya, A. (2010). Total oil, fatty acid composition and tocopherol content in kernels of several bitter and sweet apricot (*Prunus armeniaca* Batsch) cultivars from Turkey. *Journal of Food Agriculture & Environment, 8*(3-4), 196-201.

Zhang, S., Wang, S., Huang, J., Lai, X., Du, Y., Liu, X., Li, B., Feng, R., & Yang, G. (2016). High-specificity quantification method for almond-by-products, based on differential proteomic analysis. *Food Chemistry, 194*, 522-528.

Legends of Figures

Fig. 1. A) SDS-PAGE profiles of **Raw** sample *versus* **EEP** sample; Mr = molecular mass ladder; staining with micellar Coomassie blue. B) Venn diagrams of all identified species in **Raw** and **EEP** samples against Uniprot_Viridiplantae database. In total, 44 proteins were identified exclusively in **Raw**, 73 only in **EEP**, and 58 in both samples. The squares indicate the numbers of IDs assigned to *Prunus*.

Fig. 2. Gene Ontology (GO) analysis of identified gene products. Pie graphs of (A) biological processes, (B) molecular functions, and (C) cellular components show the percentage of proteins in each functional category.

Fig. 3. Molecular Function GO Term Annotation Comparison obtained plotting the unique ID entries for **Raw** and **EEP**, respectively.

Fig. 4. Protein-Protein interaction network obtained by STRING software (p-value = 0.045). Different colored edges represent the existence of different types of evidence. A green line indicates neighborhood evidence; a blue line, gene-co-occurrence; a yellow line, textmining; a purple line, experimental evidence.

A)

B)

A)
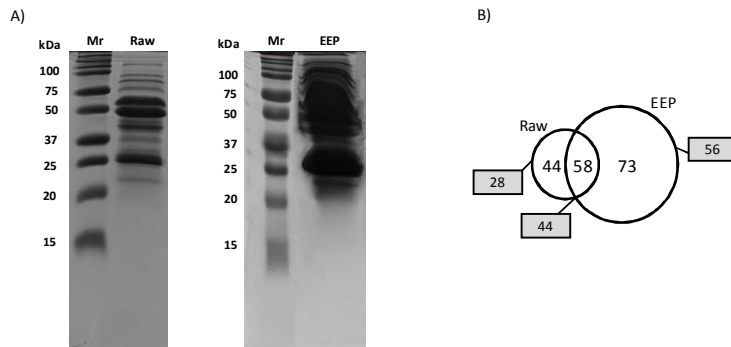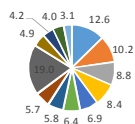
- biological_process
- cellular process
- transport
- metabolic process
- cellular component organization
- multicellular organism development
- biosynthetic process
- nucleobase-containing compound metabolic process
- other*
- reproduction
- cellular protein modification process
- response to stress
- carbohydrate metabolic process

B)
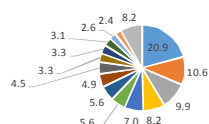
- nucleotide binding
- hydrolase activity
- binding
- transferase activity
- kinase activity
- catalytic activity
- DNA binding
- transporter activity
- cellular protein modification process
- cellular process
- molecular_function
- protein binding
- biosynthetic process
- nucleobase-containing compound metabolic process
- other*

C)

1.38
1.38
2.3
3.23
3.23
3.23
5.07
5.99
9.22
10.14
6.44
48.39

- membrane
- nucleus
- plastid
- cytoplasm
- intracellular
- cytoskeleton
- cytosol
- Golgi apparatus
- endosome
- nucleoplasm
- cell wall
- other*

Molecular Function GO Term Annotation Comparison
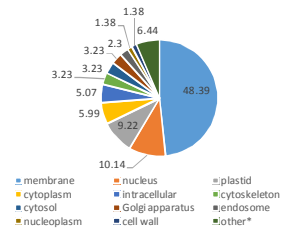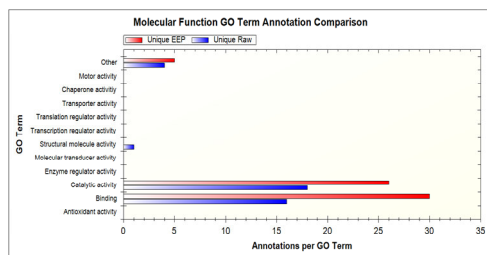
Table 1. Identified proteins in *P. armeniaca* kernels: comparison of **EEP** and **Ra**w compositions [a]

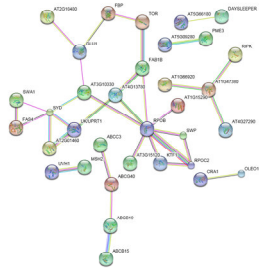| UniProt accession [b]	Taxonomy	Protein description [c]	Summed MS/MS Search score	% aa	Protein MW (kDa)	P unique	EEP	Ra
		Carbon Metabolism						
M5X7A6	PRUPE	Starch synthase, chloroplastic/amyloplastic	11.51	5.4	715.99	2		x
M5VX90	PRUPE	Fructose catabolic process	12.09	10	375.69	2		x
M5WKH0	PRUPE	Tricarboxylic acid cycle	16.65	11.3	469.45	2		x
		Stress Related and Environmental Response Proteins						
M5XL25	PRUPE	Small heat shock protein (HSP20)	27.50	15.5	173.81	2	x	x
O82011	SOLPE	17.7 kDa class I heat shock protein	20.20	16.8	176.85	2	x	x
M5X1C6	PRUPE	Small heat shock protein (HSP20)	15.20	14.5	183.45	2	x	x
M5XPY4	PRUPE	Response to stress	14.74	10	240.60	2	x	x
M5WG38	PRUPE	Response to oxidative stress	12.40	9.7	384.07	2	x	
M5XKD9	PRUPE	NB-ARC domain	14.26	4.9	109.79	2		x
M5WH03	PRUPE	Plant hormone signal transduction	15.68	6.3	767.31	2	x	
M5X2S0	PRUPE	Plant-pathogen interaction	20.35	8	592.80	2	x	x
M5XA26	PRUPE	Pectinesterase	16.68	3.4	640.13	2	x	x
M5W532	PRUPE	Dirigent-like protein	13.43	34	20.78	2		x
A5BMZ0	VITVI	Putative uncharacterized protein	17.80	5.1	135.59	2	x	x
		Oxidoreductive proteins						
D8SML7	SELML	Cytochrome P450	21.63	9.1	538.96	2	x	
M5VT70	PRUPE	Glutamine amidotransferases class-II	15.30	2.5	178.79	2	x	
M5WFA0	PRUPE	Cytochrome P450	13.93	7.4	540.28	2		x
M5VNJ0	PRUPE	Alcohol dehydrogenase GroES-like domain	13.91	14.3	353.18	2		x
M5WZ08	PRUPE	Short chain dehydrogenase	12.65	13.9	375.32	2		x
M5XNX5	PRUPE	Cytochrome P450/ E-class, group I	12.08	11.1	590.05	2		x
		Nucleotide Binding						
M5VPJ6	PRUPE	ABC transporter	12.79	7.6	692.18	2		x
M5X1Y2	PRUPE	Spliceosome	17.45	3.0	136.11	2		x
A0A0K9QEM4	SPIOL	Basic-leucine zipper domain	17.03	5.6	783.79	2		x
B2XWN9	FAGEA	DNA-directed RNA polymerase subunit beta	16.47	2.1	155.64	2	x	
M5XRM5	PRUPE	Phosphoribulokinase	16.66	6.3	104.95	2	x	x
M5X306	PRUPE	Triphosphate hydrolase activity	16.45	1.8	207.33	2	x	
M5VVC5	PRUPE	mRNA surveillance pathway	16.19	1.1	425.57	2	x	x
M0TFM1	MUSAM	Serine/threonine/dual specificity protein kinase	16.31	8.5	561.70	2		x
A0A0D9Y5E4	ORYZ	ATP binding. D-mannose binding lectin	16.55	3	193.23	2	x	x
A9RFQ7	PHYPA	Myb-like DNA-binding domain	16.04	2.4	265.05	2	x	
M5X879	PRUPE	NB-ARC domain/TIR domain/NACHT domain	17.09	2.3	116.47	2	x	x
M5Y288	PRUPE	F-box domain	16.81	8.4	469.80	2	x	x
M5X7F6	PRUPE	NB-ARC domain/Leucine Rich repeats/ATPase domain	16.71	2.8	166.99	2	x	x
M5X3N3	PRUPE	Phosphotransferase enzyme family	14.18	4.8	94.55	2	x	x
M5VXN8	PRUPE	SET domain/Tesmin/TSO1-like CXC domain	14.17	5.2	100.69	2	x	
M5XPK4	PRUPE	Cation transporting ATPase,	14.16	4.3	114.09	2	x	x
M5XIB1	PRUPE	Snf2-ATP coupling, chromatin remodelling complex	16.02	1.3	326.26	2	x	x
M5X9G3	PRUPE	Protein tyrosine kinase/PAN-like domain	15.82	6	832.15	2	x	x
M5WJX8	PRUPE	Auxin response factor	15.78	5.2	80.07	2	x	x
M5WL01	PRUPE	Phospholipid-translocating ATPase	15.69	6.8	81.86	2	x	
M5WQM8	PRUPE	Phosphatidylinositol signaling system	15.09	2.5	204.24	2	x	
M5WXM8	PRUPE	NAC domain	13.69	4.7	967.98	2	x	x
M5X2H9	PRUPE	tRNA synthetases class	15.46	4.3	91.41	2	x	
M5VX87	PRUPE	Uncharacterized protein	12.51	7	66.15	2		x
A0A078IK57	BRANA	ATPase family	19.07	3.4	130.69	2	x	x
K3XUR4	SETIT	Phosphatidylinositol-4-phosphate 5-Kinase	18.69	1.9	204.14	2	x	x
A8HM74	CHLRE	Kinesin motor domain/K-box region	17.23	10.3	54.06	2	x	
M5X7V1	PRUPE	Protein serine/threonine kinase activity	17.67	8.7	51.44	2		x
V4V803	ROSI	Protein kinase domain/Leucine Rich Repeat	17.13	5.6	100.34	2		x
C1MUM4	MICPC	Helicase conserved C-terminal domain	17.97	5.2	112.94	2	x	
		Transcription factor activity						
A0A0E0EBP5	ORYZ	Transcription factor TFIIB repeat	16.03	6.7	69.96	2	x	
F2DTK8	HORVD	GRAS domain family	15.38	7.4	74.12	2	x	
M5VWQ0	PRUPE	Anaphase-promoting complex subunit 4 WD40 domain	16.16	4.7	124.87	2	x	
M5VYG0	PRUPE	Mediator complex subunit 13 C-terminal	11.04	2.8	206.67	2		x
M5W6G9	PRUPE	Mediator complex subunit MED14	15.85	2	188.73	2	x	x
A0A087SNJ1	AUXPR	RNA-induced silencing complex	17.44	5.7	94.44	2		x
F4IHS2	ARATH	Chromatin structure-remodeling complex protein SYD	15.97	0.7	391.97	2		x

M5XKW1	PRUPE	PB1 domain	13.91	3.6	106.72	2	x	x
A0A087H600	ARAAL	U3 ribonucleoprotein	18.83	7.1	59.37	2	x	x
Transferase activity								
M5VJV5	PRUPE	Hexosyltransferase	16.42	11.1	43.36	2	x	
M5WDX6	PRUPE	MT-A70	14.99	8.7	72.92	2	x	x
M5W9P4	PRUPE	Glycosyltransferase family	14.70	7.3	59.077	2	x	
M5WEC8	PRUPE	Putative S-adenosyl-L-methionine-dependent methyltransferase	14.62	7.1	70.40	2	x	x
M5XEZ7	PRUPE	Early transcription elongation factor of RNA pol II	13.88	1.9	161.24	2	x	x
M5VUM2	PRUPE	16S rRNA methyltransferase	12.61	11.7	43.57	2		x
M5WAL1	PRUPE	Glycosyltransferase family 29	11.26	7	49.68	2		x
M5WMV1	PRUPE	Ubiquitin ligase	19.31	7.1	60.91	2	x	x
Transporter activity								
M5X572	PRUPE	ABC_membrane	31.46	8.1	137.73	4	x	x
M5XIX7	PRUPE	Mito_carr	26.46	23.4	34.54	3		x
R0HKY1	BRAS	Substrate-specific transmembrane transporter activity	17.20	10.4	55.13	2		x
M5WAE5	PRUPE	Membrane insertase	14.08	8.1	60.05	2	x	
M5XM22	PRUPE	RNA transport	13.03	14.7	19.06	2	x	x
M5WUT7	PRUPE	ABC2_membrane	19.58	4.3	110.56	2		x
M5WRV7	PRUPE	Voltage gated chloride channel	18.30	5	84.04	2		x
A0A059AD39	EUCGR	ABC transporter	17.45	4.9	136.39	2		x
A0A078G4Y7	BRANA	ABC transporter	15.68	4.6	137.29	2		x
S8CX84	LAMI	Nucleoporin autopeptidase	17.02	5.1	114.76	2		x
M5VYB7	PRUPE	Proton-dependent oligopeptide transporter family	17.33	8.7	63.98	2	x	
M5WDR1	PRUPE	Uncharacterized protein	14.41	7.5	71.64	2	x	
M5X292	PRUPE	Sugar transporter	14.27	6.1	55.94	2	x	
M5X4E2	PRUPE	Endocytosis	15.03	10.9	44.71	2	x	
M5WPM9	PRUPE	ABC transporter	13.64	2.1	165.79	2	x	
M5VMX6	PRUPE	Major Facilitator Superfamily	13.48	13.5	48.16	2		x
M5XKR7	PRUPE	Cation transmembrane transporter activity	17.90	5.6	98.39	2		x
A0A0D2NP98	CHLO	Autophagy-related protein 3	20.15	14.1	39.30	2	x	x
Kinase activity								
M8BR74	AEGTA	Casein kinase I isoform delta-like protein	15.85	3.6	145.93	2	x	
M5VG74	PRUPE	Protein kinase domain	14.43	3.8	148.66	2		x
M5WQP2	PRUPE	Protein tyrosine kinase	14.91	4.9	114.65	2		x
M5X8G0	PRUPE	Protein kinase domain	13.55	9.4	67.11	2		x
M5X2Q8	PRUPE	Di-glucose binding within endoplasmic reticulum	13.52	6.1	98.79	2		x
M5W5T2	PRUPE	Phosphatidylinositol 3- and 4-kinase	11.66	2	280.37	2		x
Hydrolase activity								
M5WI23	PRUPE	Aspartic-type endopeptidase activity	31.00	8.8	47.59	2	x	
M5X2H5	PRUPE	Alpha-1,2-Mannosidase	15.15	9.3	70.64	2	x	
A0A0D2V8W9	GOSRA	Dynamin GTPase effector domain	18.59	7.8	78.06	2	x	
M5XVK6	PRUPE	Lipase/Hydrolase	17.99	8.2	69.94	2		x
M5XBV0	PRUPE	Protein serine/threonine phosphatase activity	15.03	9.7	47.49	2	x	x
M5WYB9	PRUPE	Metal dependent phosphohydrolases	12.94	6.1	80.54	2	x	x
M5VWB9	PRUPE	Exonuclease/phosphatase	13.86	13.1	57.98	2		x
M5XKG8	PRUPE	Hydroxyisourate hydrolase activity	14.68	15.4	36.92	2	x	
A0A0D2RN96	GOSRA	Pectate lyase	15.99	16.6	42.62	2	x	
Binding activity								
A0A077S2P4	WHEAT	IQ calmodulin-binding motif	20.01	1.5	259.90	2	x	x
V4KUD8	EUTSA	Fatty-acyl-CoA binding	17.32	45	101.54	2	x	x
C1ECC7	MICSR	Zinc ion binding	17.42	6.7	52.82	2	x	x
M5XLQ2	PRUPE	BAH domain /Agenet domain	17.47	8.1	73.47	2		x
M5VWP7	PRUPE	DYW family of nucleic acid deaminases	17.33	5.1	86.51	2		x
M5X7A2	PRUPE	GTPase activity	17.27	4.3	72.88	2		x
M5X8P0	PRUPE	Phosphatidylinositol 3-phosphate binding	14.05	4.5	112.99	2	x	x
M5X715	PRUPE	Glycosyltransferase	13.97	3	110.87	2	x	x
E3W0S0	ROSI	DNA-directed RNA polymerase subunit beta	13.41	4.2	121.82	2		x
M5VPV2	PRUPE	Regulator of chromosome condensation	10.77	4.6	121.38	2		x
M5WQZ5	PRUPE	Chromatin binding	13.31	3	174.52	2	x	x
M5W7U6	PRUPE	DYW family of nucleic acid deaminases	16.60	5.8	75.17	2		x
M5X1G9	PRUPE	Oleosin	21.23	12.8	15.61	2		x
M5XB96	PRUPE	Mitochondrial ribosomal protein subunit L20	17.82	5.8	83.16	2	x	x
M5VWL5	PRUPE	Phosphatidylinositol binding	18.94	5.7	75.78	2		x
M5X9Z6	PRUPE	Proline-binding domain	15.44	3.2	180.67	2	x	
M5WNU7	PRUPE	Plant-specific actin-binding protein	12.94	6.1	80.54	2	x	x
M5W984	PRUPE	Mismatch repair	27.38	7.8	106.68	3	x	x
M5WS62	PRUPE	Uridine kinase	19.15	7.9	53.94	2		x

M5XRX7	PRUPE	RNA binding activity	18.11	2.6	202.43	2		x
M5XS06	PRUPE	CTLH/CRA C-terminal to LisH motif domain	15.82	17.6	24.98	2	x	x
DNA metabolic process								
M5X0E1	PRUPE	Nucleotide excision repair activity	23.58	6.6	117.19	3	x	
Q5NAA4	ORYSJ	Helicase-like protein	21.38	2.4	188.53	2	x	
Storage Proteins								
E3SH28	PRUDU	Prunin 1	132.8	25.7	63.39	9	x	x
E3SH29	PRUDU	Prunin 2	96.65	12.5	57.23	6	x	x
M5Y3W2	PRUPE	Nutrient reservoir protein	43.03	5.6	95.67	3	x	x
Cytoscheleton organization and multicellular component development								
M5XNW4	PRUPE	Microtubule-associated protein 7	13.91	3.6	106.72	2	x	x
I6UTH7	SORBI	Tan1	19.28	5.9	38.34	2	x	x
Translation activity								
M5WYZ2	PRUPE	Translation initiation factor eIF3	15.77	3.5	185.56	2		x
A0A022Q5F9	ERYGU	Ribosomal protein S18	19.00	14.7	30.513	2	x	
M5Y4G5	PRUPE	Zinc finger PHD-type	18.14	3.7	168.10	2	x	x
Other								
M5Y6F3	PRUPE	Band_7	17.34	11.3	47.52	2		x
M5VPQ9	PRUPE	HAUS augmin-like complex subunit 3	15.19	6.4	70.02	2	x	x
A9YTJ2	PRUDU	F-box associated interaction domain	14.22	16.6	31.78	2	x	x
M5VVG0	PRUPE	Retrotransposon gag protein	12.45	6	71.06	2	x	x
M5XVF3	PRUPE	Sieve element occlusion (SEO) protein	15.24	6.2	91.56	2		x

a) Proteins with unknown function are listed in Supplementary Table S1

b) Protein ID, according to the Uniprot database.

c) Protein description, according to the KEGG database

Table 2. Predicted bioactive peptides generated from Prunin 1 (E3SH28) and Prunin 2 (E3SH29) by *in silico* digestion. The best candidates (Peptide Ranker score values higher than 0.5) are labelled in bold.

Protein Ids	Peptide	Score	Potential bioactive sequence	Activity
E3SH28				
	GRPR	0.78	GRP, PR, RP, GR	ACE inhibitor
			RP	DPP IV inhibitor
	GVIQQAGNQG	0.13	AG, GV, QG	ACE inhibitor
			AG, GV, IQ, NQ, QA, QG, QQ, VI	DPP IV inhibitor
	IVPQNH	0.15	VP, PQ,	ACE inhibitor
			IV	Glucose uptake stimulating peptide
			VP, NH, PQ, QN	DPP IV inhibitor
	NLPIL	0.62	IL	Glucose uptake stimulating peptide
			LP, IL, NL,PI	DPP IV inhibitor
	NVNAH	0.08	AH	ACE inhibitor
			AH	Antioxidative
			AH, NA, NV, VN	DPP IV inhibitor
	QGQNDNR	0.23	GQ, QG,	ACE inhibitor
			GQ	Neuropeptide inhibitor
			GQ, DN, ND, NR,QG, QN	DPP IV inhibitor
	QQGEQGRPGQH	0.21	PG,	Prolyl endopeptidase inhibitor
			GRP, RP, GR, GQ, GE, QG, PG	ACE inhibitor
			PG	Antithrombotic
			PG	Peptide regulating the stomach mucosal membrane activity
			GQ	Neuropeptide inhibitor
			GQ, RP, GE, PG, QG, QH, QQ	DPP IV inhibitor
	SPHW	0.91	PH	ACE inhibitor
			PHW	Antioxidant
			SP, HW, PH	DPP IV inhibitor
	SPQNQCQ	0.38	PQ	ACE inhibitor
			SP, NQ, PQ, QN	DPP IV inhibitor
	VSSDH	0.09	VS	DPP IV inhibitor
E3SH29				
	AVITQASNEG	0.06	EG, TQ, AV	ACE inhibitor
			AS, AV, EG, NE, QA, TQ, VI	DPP IV inhibitor
	EPDNH	0.16	EP, DN, NH	DPP IV inhibitor
	GQNK	0.14	GQ, NK	ACE inhibitor
			GQ	Neuropeptide inhibitor
			GQ, QN	DPP IV inhibitor
	IPQNH	0.26	IP, PQ	ACE inhibitor
			IP, NH, PQ, QN	DPP IV inhibitor
	IQSEAGVTES	0.05	AG, GV, EA, TE	ACE inhibitor
			SE	Stimulating vasoactive substance release
			AG, ES, GV, IQ, QS,TE, VT	ACE inhibitor
	NPQGGR	0.57	GR, GG, QG, PQ	ACE inhibitor
			NP, GG, PQ, QG	DPP IV inhibitor
	NPQQQGR	0.31	GR, QG, PQ	ACE inhibitor
			NP, PQ, QG, QQ	DPP IV inhibitor
	NPSDPQF	0.76	PQ	ACE inhibitor

		NP, DP, PQ, PS, QF	DPP IV inhibitor
QGQDDNR	0.27	GQ, QG	ACE inhibitor
		GQ	Neuropeptide inhibitor
		GQ, DN, NR, QD, QG	DPP IV inhibitor
RPSR	0.41	RP	ACE inhibitor
		RP, PS	DPP IV inhibitor
SATSPPR	0.55	PR, PP	ACE inhibitor
		PP, SP, AT, TS	DPP IV inhibitor
TNANA	0.06	NA, TN	DPP IV inhibitor
TPHW	0.75	PH	ACE inhibitor
		PHW	Antioxidant
		TP, HW, PH	DPP IV inhibitor
VAVS	0.04	VAV, AV,	ACE inhibitor
		VA, AV, VS	DPP IV inhibitor

- CPLL's, LC-MS and Bioinformatics for apricot kernel proteome and peptide discovery
- Low abundance protein identification
- In total 175 proteins characterised by searching in *Viridiplantae* databases
- Potentially bioactive peptides encrypted in main storage proteins