003

The Challenge of Plant Identification in Complex Mixtures: Closely Related Families, Large Proteomes, and Unsequenced Genomes

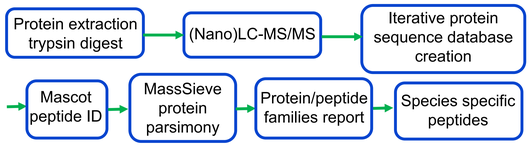
Melinda A. McFarland, Sara M. Handy, Elizabeth Hunter, Christine H. Parker, Ann M. Knolhoff

F.D.A. Center for Food Safety and Applied Nutrition, 5001 Campus Dr., College Park, MD 20740

Introduction

- The market for protein-rich, plant-based foods created from complex mixtures of processed legumes, seeds, fruit, grains, fungi, and vegetables has created a need for methods to identify unknown plant species or unanticipated proteins in complex mixtures of plants.
- Identification of unknown plants is a formidable task due to many closely related species with few known protein sequences, complex genomes, and high sequence homology.
- While food safety efforts have long used LC-MS/MS to identify contaminants and proteins in food, the interface of proteomics and multi-species food analysis remains remarkably complicated.
- Unlike bacterial metaproteomics, informatics solutions to facilitate identification of multiple plant species in a complex mixture have not kept pace.
- We present examples of informatic challenges encountered when proteomics is used to identify closely related seeds, nuts, legumes, and toxic plant contaminants.

Methods



Conclusions

- For multi-species plant mixtures, the protein sequence database should include protein sequences for the most abundant ingredients/plants in your sample to reduce false positive peptide IDs.
- Specificity and accuracy are gained, but sensitivity is sacrificed due to the large size of multi-species plant sequence databases.
- At some point, the number of peptides-to-proteins associations is so large that downstream parsimony-type software crashes.
- Methods to minimize the size of multi-species protein sequence databases are needed, such as food specific or seed specific.
- An automated method to parse species from protein report results would help speed up identification of an unknown plant source in a multi-species plant matrix.
- The presence of proteins from sparsely sequenced plants can be inferred based on similarity to proteins from plants with sequenced genomes from the same phylogenetic family.
    - In these cases, secondary information such as patient symptoms and tissue specific proteins are considered.
- In an outbreak sample, proteomics was able to show that the toxin came from a plant, identified the part of the plant, and narrowed the plant to a phylogenetic family.

1. *BMC Public Health*.**2022**, 22, 623.

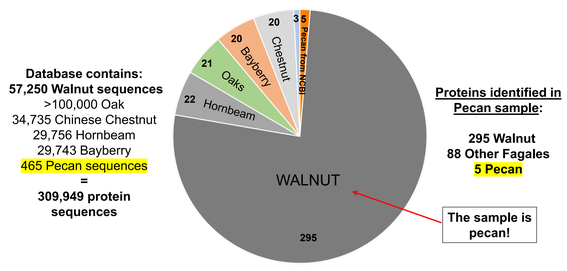
Plant Protein Sequences – Tree Nuts

Tree nut protein sequences in UniProt and NCBI over time

Common Name	Uniprot 2018	NCBI 2018	Uniprot 2019	NCBI 2019	Uniprot 2020	NCBI 2020	S.-Prot 2023	TrEMBL 2023	NCBI 2023	RefSeq 2023
Almond	455	660	50,045	55,146	53,231	88,413	16	53,270	134,164	33,326
Beech nut					59,448	501	3	59,550	2,296	140
Brazil nut	95	208	95	210	95	210	3	92	210	91
Cashew*	95	186	96	188	96	195	1	107	220	84
Chestnut - European	117	147	117	147	127	195	7	189	446	83
Chestnut - Chinese					175	34,735	0	34,348	34,862	83
Coconut	447	953	538	975	542	978	7	22,007	22,752	166
Hazelnut*	473	631	476	636	492	731	13	524	758	77
Hickory nut*					29	305	0	142	479	83
Macadamia nut	99	199	111	211	111	225	7	162	46,917	46,691
Pecan	18	27	134	288	136	465	2	80,176	295,276	54,484
Pistachio	104	211	106	41,431	106	41,431	1	111	41,519	41,299
Walnut - English	45,746	56,234	45,761	56,243	45,776	57,250	4	52,354	88,446	45,959

Table 1. Plant genomes are large and difficult to assemble. Recent years have seen many more plant genomes released to public sequence databases. As multiple cultivars are sequenced, the number of sequences gets prohibitively large (see pecan).

A. Pecan MS/MS data before a pecan genome



B. Pecan MS/MS data after a pecan genome

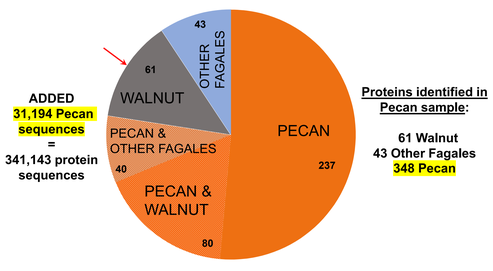


Figure 1. Walnut and pecan are in the same Family and have highly conserved proteins. **A**. If the unknown species is sparsely sequenced, it is easy to wrongly attribute it to a similar species that is better represented in the sequence database. **B**. Addition of protein sequences from the newly sequenced pecan genome to the sequence database in Figure 1A. identifies many more pecan-specific peptides. This translates to reportable proteins that more accurately represent the sample. Note, there are still 61 reportable proteins that are attributed to walnut, not pecan.

Identification of an Unknown Toxic Plant Contaminant Without a Genome

OUTBREAK SAMPLES

WFP Super Cereal CSB fortified corn-soy blend

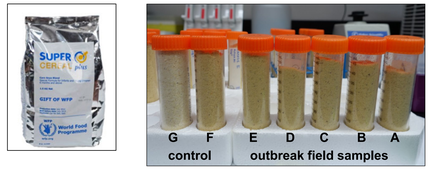


Figure 2. Fortified cereal distributed as food aid was linked to four deaths and 300 illnesses in Uganda. Symptoms included hallucinations, disorientation, and vomiting. A chemical toxin was suspected.[1]

ITERATIVE PROTEIN SEQUENCE DATABASES

1. Protein Toxins

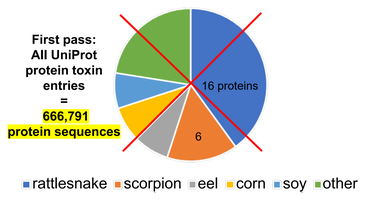


Figure 3. When searched against a protein toxin sequence database, the number of assigned spectra is low at 124 peptides and 40 proteins of at least 2 peptides per protein. It would be exciting if these results were 'real', but they aren't.

2. Corn + Soy + Protein Toxins

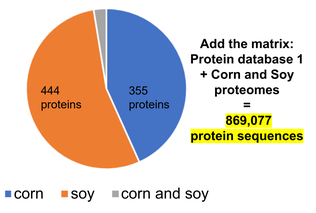


Figure 4. The samples are primarily corn and soy. Corn and soy protein sequences must be included in the search or abundant, high-quality MS/MS spectra may yield false positives. The same data now yields only corn and soy proteins.

Toxin Identified

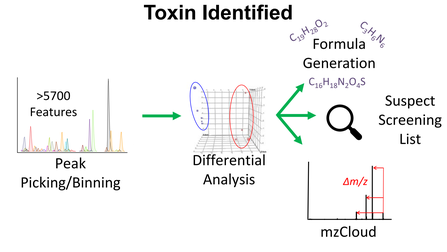


Figure 5. Small molecule non-targeted analysis putatively identified the toxin as tropane alkaloids. LC-HR-MS with analytical standards confirmed the presence of atropine and scopolamine. Both compounds occur naturally in some plants from the Erythroxylaceae, Brassicaceae, and Solanaceae Families.

3. Corn + Soy + All Solanaceae

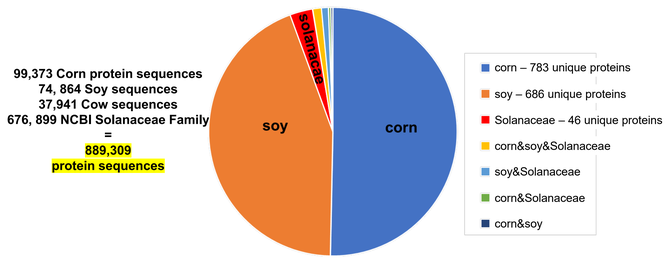


Figure 6. 1,740 proteins were identified, 92% from corn and soy. 136 reportable proteins were from the Solanaceae Family of plants. Of these, 46 proteins from 13 protein families contained only protein entries from the Solanaceae Family and shared no identified peptides with corn or soy. The source of the contaminant was now narrowed to an atropine-containing plant from the Solanaceae Family, not isolated atropine.

Solanaceae Seed Storage Proteins Identified

Protein Description	Sample Name													
Solanaceae Seed Storage Proteins	A		B		C		D		E		F		G	
	Peptides	Hits	Peptides	Hits	Peptides	Hits	Peptides	Hits	Peptides	Hits	Peptides	Hits	Peptides	Hits
11S-GLOBULIN/LEGUMIN FAMILY														
cocosin 1-like [Solanum pennellii]	0	0	**2**	8	**4**	29	0	0	**3**	10	0	0	0	0
PREDICTED: 11S globulin subunit beta-like [Capsicum annuum]	0	0	**1**	4	**4**	30	0	0	**3**	10	0	0	0	0
PREDICTED: legumin A-like [Nicotiana tomentosiformis]	0	0	**3**	9	**6**	35	0	0	**2**	11	0	0	0	0
12S seed storage protein CRA1-like [Solanum lycopersicum]	0	0	**3**	11	**5**	43	0	0	**3**	21	0	0	0	0
11s globulin seed storage protein 2 [Nicotiana attenuata]	**1**	2	**2**	10	**4**	22	0	0	**2**	8	0	0	0	0
VICILIN FAMILY														
PREDICTED: vicilin C72-like [Capsicum annuum]	0	0	**1**	3	**3**	15	0	0	**1**	5	0	0	0	0
vicilin gc72-a [Nicotiana attenuata]	0	0	**1**	6	**2**	11	0	0	**2**	11	0	0	0	0
LIPID TRANSFER PROTEIN														
non-specific lipid-transfer protein A-like [Capsicum annuum]	0	0	0	0	**2**	9	0	0	**1**	6	0	0	0	0

Table 2. Seven of the 13 protein families (15 reportable proteins) are seed storage proteins and were only identified in the atropine-containing samples (red). This implicates the seeds of the plant as the source of contamination.

The lack of a consistent plant species across the seed storage protein IDs suggests that the contaminant species' protein sequences are not well represented in this sequence database. Jimsonweed, an atropine-containing species known to co-harvest with corn and soy has only 408 available protein sequences.

Species Identified

Figure 7. Genome skimming was used to identify the specific plant. Relative spectral counts from seed storage proteins and the amount of Jimsonweed DNA directly correlate with the LC/MS measured alkaloid levels. Current efforts are being made to assess the utility of long contigs from genome skimming data as proxy peptide/protein sequences for sparsely sequenced plants.