TP 370 The Challenge of Plant Identification in Complex Mixtures: Closely Related Families, Large Proteomes, and Unsequenced Genomes

Melinda A. McFarland[1]; Sara M. Handy[1]; Elizabeth Hunter[1]; Christine H. Parker[1]; Ann M. Knolhoff[1]

[1]FDA-CFSAN, College Park, MD

Introduction

Identification of unknown plant proteins is a formidable task due to a large number of species with few known protein sequences, complex genomes, and high sequence homology. While food safety efforts have long used MS to identify contaminants in food and identification of proteins by LC-MS/MS has become routine, the interface of proteomics and multi-species food analysis remains remarkably complicated. Interest in plant-based protein-rich food continues to grow, but informatics solutions to facilitate the identification of multiple plant species in a complex mixture have not kept pace. This is especially true when dealing with unknown plant contaminants. We present two examples of informatic challenges encountered in proteomics identification of closely related seeds, legumes, and toxic plant contaminants.

Methods

Tryptic digests of a variety of plant-based samples, including tree nuts and corn and soy blends, were analyzed by nano-LC-MS/MS on an Orbitrap Elite or Lumos mass spectrometer coupled to a nanoAquity UPLC. Data files were searched using Mascot against custom multi-species protein sequence databases created from available protein or genomic sequences. Parsimonious lists of identified proteins, protein families, peptides, and associated species as well as comparisons across samples were generated with MassSieve software. All other analysis was done in Excel.

Preliminary Data

A trend toward the use of seeds, nuts, and legumes in food has furthered the need to explore metaproteomic methods to identify plants in food. Pecans and walnuts share highly homologous proteins. Prior to the recent release of a pecan genome, pecan data searched against a database of 57,250 available walnut and 465 pecan protein sequences identified 295 walnut proteins and 5 pecan proteins. After release of the pecan proteome (31,194 sequences), 234 proteins with unique peptides in pecan and 119 proteins shared between pecan and another species were identified. Still, 61 proteins were identified with peptides that identified as unique to the walnut proteome despite being derived from pecans, likely because plant genome assembly is complex and often not complete. Now multiple pecan genomes are available, bringing the number of available pecan protein sequences to 295,000. This significantly increases the search space but may provide pecan sequences to account for the 61 walnut proteins. Data will be searched against a pecan multi-cultivar database and compared to a single reference proteome. Results will inform selection of protein sequences for plant metaproteomic databases. For most plants, the hurdle for identification is too few known protein sequences. In 2019 fortified cereal

distributed as food aid caused four deaths and 300 illnesses in Uganda. Non-targeted small molecule analysis identified the toxin. Proteomics showed that the toxin came from a plant, identified the part of the plant, and narrowed the plant to a phylogenetic family. Genome skimming was used to identify the specific plant. Data will outline the proteomics informatics workflow used to identify the presence of a contaminant plant, how the identity was narrowed to a plant Family, and clues that this was an unsequenced plant in a complex mixture of other plants. Current efforts to harness genomic technology will be discussed.

Novel Aspect

Identification of unknown plants in complex mixtures by proteomics.

Conflict of Interest Disclosure

The authors declare no competing financial interest.