Heritage Science

RESEARCH

Open Access

Paleoproteomic profiling for identification of animal skin species in ancient Egyptian archaeological leather using liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS)

Abdelrazek Elnaggar[1*], Aya Osama[2], Ali Mostafa Anwar[2], Shahd Ezzeldin[2], Salma Abou Elhassan[2], Hassan Ebeid[1], Marco Leona[3] and Sameh Magdeldin[2,4*]

Abstract

Ancient protein analysis provides clues to human life and diseases from ancient times. Paleoproteomics has the potential to give a better understanding of the modes of fabrication of ancient materials, their composition, and pathways of degradation, as well as the development of animal fibers through domestication and breeding. Thus, this study aimed at providing guidance for choosing proteomics workflows to analyze leather samples and their capacity to distinguish between unknown archeological species. Here, we performed shotgun proteomics of archeological animal skin for the first time. The raw output data were analyzed using three different software (Proteome Discoverer, Protein Pilot, and Peptide Shaker) with their impeded algorithms. The study found that the best species identification percentage was obtained using protein piolet with protein database. Particularly prevalent and relatively high collagen expression suggests its resistance to degradation, despite the samples' exposure to environmental and chemical alterations. The success of this case study indicates that further analyses could assist in reworking historical baseline data for putative identification of unknown archeological samples.

Keywords: Collagen, Peptide, Chromatography, Fragmentation, Databases, Algorithms

Introduction

Leatherworking is one of the significant industries in ancient Egypt where different animal skins (sheep, goat, cow, cattle, and gazelle) were tanned by creating chemical bonding between amino acids of the dermal network of the collagen protein and the vegetal or mineral molecules of tanning material [1–4]. Animal skins have similar structures, mainly consisting of the top grain layer, bundles of collagen fibers, and fibrils that form the middle layer (corium) and the flesh layer [4]. Collagen is the most abundant protein in the animal skin, consisting of a characteristic repeated amino acid sequence (mainly glycine, proline, and hydroxyproline) and comprises the triple helix formed of three polypeptide chains linked together covalently through peptide bonds. The protein information of the archaeological leather could provide significant insights into the leather technology, authentication, degradation, preservation needs, and environmental impact over the years [5].

Proteomics is an innovative analytical technique to study biological samples using chromatography coupled

*Correspondence: elnaggar@arch.asu.edu.eg; sameh.magdeldin@57357.org

[1] Archaeological Science and Excavations Department, Faculty of Archaeology, Ain Shams University, Cairo 11566, Egypt
[2] Proteomics and Metabolomics Research Program, Basic Research Department, Children's Cancer Hospital, Cairo, Egypt
Full list of author information is available at the end of the article

with mass spectrometry [6]. The latter is a technique most often used for proteomic profiling, where a given protein's sequence of amino acids could be identified by breaking it down into smaller peptides and analyzing their unique mass/ charge value [7]. Proteomics is an emerging area of research for studying cultural heritage materials [8, 9], mainly associated with a recent subdiscipline named paleoproteomics. Paleoproteomics represents the analysis of a set of proteins in resolving species identification and evolutionary relationships of extinct taxa [7, 10, 11], protein degradation, or authentication [12]. Proteomic analysis of ancient animal skins, used for different purposes in ancient Egypt, could provide valuable information on leather technology and degradation, considering the animal skin species' flexibility, durability, and possible defects [13].

Furthermore, proteomic evidence could also be used for studying ancient leather authentication, dating, provenance, and trade [14]. However, the patterns of age-induced degradation and other factors such as the temperature, pH, glutamine deamidation, tertiary structures of proteins, and the geographical variations in the animal's skin sourcing must be considered [12, 15]. Nevertheless, there is currently little research on leather studies concerning differences in archaeological animal species identification [16, 17].

One reason behind the research shortage is the difficulty of identifying the archaeological leather origin. The characterization of protein components in degraded leather could be affected by environmental aging and the presence of different degraded organic materials, including lubricants, oils, and tanning materials. Compared to other traditional methods relying on visual evaluation, mass spectrometry has been tested to identify animal skin species in archaeological leather in several research attempts [18–22].

Zooarchaeology by Mass Spectrometry (ZooMS) is another simple, minimally destructive low-cost proteomics method that uses diagnostic peptides of the dominant collagen protein as a fingerprint of species [6, 23]. It allows animal species identification while conserving the sample's integrity with minimal destruction. The extracted collagen is digested into peptides that are subsequently analyzed by soft-ionization mass spectrometry, usually Matrix Assisted Laser Desorption Ionization Time of Flight (MALDI-ToF) mass spectrometry. However, ZooMs is generally more successful with well-preserved collagen samples. However, the technique still faces some limitations, including the standardization and centralized repository of reference data.

The mass spectrometric techniques used in proteomic profiling are based on assembling the identified peptide. For instance, MALDI-TOF is one of the techniques used to analyze leather and animal skin identification by generating species-specific peptide patterns [24–26]. However, gas chromatography (GC-MS/MS) and liquid chromatography coupled to tandem mass spectrometry analysis (LC-MS/MS) can provide a more sensitive and definite identification of proteins in more complex samples by detecting the sequence of peptide mixture in a higher resolution [27]. However, utilizing these techniques for archaeological sample identification may suffer challenges due to the possible sample contamination during excavation, handling, and conservation treatments. Therefore, the quality of fragmentation and subsequent manual confirmation is critical [15].

A new complementary analytical strategy was utilized in this research for protein identification. Here, we investigated the best predictive model in terms of software, search engines, and databases to identify the animal skin species in archaeological leather samples along with fresh and aged model leather samples using LC-MS/MS. In brief, we compared three different software (Proteome Discoverer, Protein Pilot, and Peptide Shaker), which utilize unique search engines/ algorithms (SEQUEST, Paragon, and X! Tandem, respectively) to find the best algorithms. In addition, two different databases were constructed; one includes all the protein retrieved for the species, and another one consists of all unique peptide sequences only. Unlike previous methods relying on collagen for identification, we broadened the search to include any matched protein with high scores and tested different matching algorithms and databases for the best match (best algorithm scoring as a result of higher PSM matching).

Methods

Study design and subjects

Ten ancient Egyptian leather samples were selected from the leather collections at the Metropolitan Museum of Art, New York, USA [16] (Additional file 1; archio. samples). To identify the archaeological animal skin species, 18 model reference leather samples made of common domesticated animal skin species used in ancient Egypt (goat, sheep, deer, bovine, camel, cobra, crocodile, and ostrich), brought from the Leather Conservation Centre, Northampton, UK were used for the mass spectrometric analysis and subsequent database search (Additional file 1; reference samples). For peptide sequencing analysis, a sample measuring approximately 2×2 mm was cut off from all archaeological and model samples.

Shotgun proteomics analysis

Sample processing and protein extraction

After blinding all the samples (both standards and archeological), samples were subjected to proteomics

workflow (Fig. 1). Leather pieces were homogenized by placing 1 ml lysis solution (8 M urea, 500 mM Tris HCl, pH 8.5) with a complete ultra-proteases inhibitors mixture (Roche, Mannheim) using an ultrasonic homogenizer. The protein extract was obtained after incubation at 37 °C for 1 h with an occasional vortex. The extract was then centrifuged at 12,000 rpm for 20 min and assayed using the BCA method (Pierce, Rockford IL) at Å562 nm before digestion [28].

In-solution digestion

Thirty µg of protein extract from each sample was subjected to in-solution digestion [29]. In brief, protein pellets were reduced with 200 mM 1,4-Dithiothreitol (DTT) for 30 min. Alkylation of Cysteine residues was performed using 1 M Iodoacetamide (IAA) for 30 min in a dark area. Before digestion with trypsin, samples were diluted to a final concentration of 2 M urea with 100 mM Tris–HCl, pH 8.5. For endopeptidase digestion, Pierce Trypsin protease MS-grade (Sigma, Germany) was added at 30:1 (protein: activated porcine trypsin) and incubated overnight in a thermo-shaker at 600 rpm at 37 °C. Digested peptide solution was acidified using 100% formic acid to a final pH of 2.0. The resultant peptide mixture was then cleaned using the stage tip [29]. Peptides w were assayed using the BCA method (Pierce, Rockford, IL) at Å562 nm before injection (1ug/10ul).

Liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS)

To identify the animal skin species' specific peptides, LC-MS/MS analysis was performed using TripleTOF 5600^{+}(AB Sciex, Canada) interfaced at the front end with Eksigent nanoLC 400 autosampler with Ekspert nanoLC 425 pump. In trap and elute mode, peptides were trapped on CHROMXP C18CL 5um (10 × 0.5 mm) (Sciex, Germany). MS and MS/MS ranges were 400–1250 m/z and 170–1500 m/z, respectively. A design of a 55-min linear gradient 3–40% solution (80% ACN, 0.2% formic acid). The 40 most intense ions were sequentially selected under data-dependent acquisition (DDA) mode with a charge state 2–5. For each cycle, survey full scan MS and MS/MS spectra were acquired at a resolution of 35.000 and 15.000, respectively. External calibration was scheduled and run during sample batches to ensure accuracy to correct possible TOF deviation. Samples were run twice to have a single high-stringency dataset of reproducibly identified proteins present at each time point.

Proteomics data analysis

Raw LC-MS/MS data were searched using Protein pilot software (version 5.0.1.0, 4895) with the paragon algorithm (version 5.0.1.0, 4874). Using porcine trypsin as a digestion factor for the peptides identified from MS/MS spectra, the Pro Group™ Algorithm assembles peptide identifications into a list of reliable protein identifications. Iodoacetamide was selected as the Cys Alkylation.

In the second software, Mascot generic format (mgf) files were generated from raw files using a script supplied

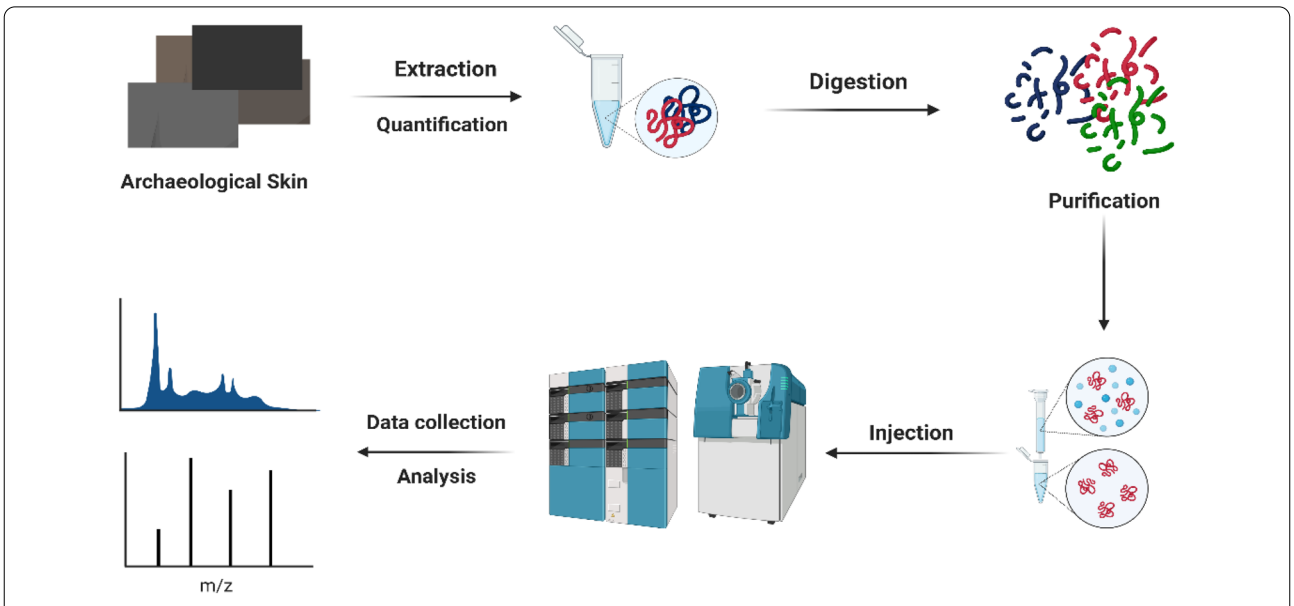

Fig. 1 General workflow used for the proteomic analyses of skin archeological samples

by AB Sciex. MS/MS spectra were searched using X! Tandem in Peptide shaker (version 1.16.26). Searching all fully and semi-tryptic peptide candidates adjusted up to 2 missed cleavages with at least 6 amino acids. Precursor mass and fragment mass were identified with an initial mass tolerance of 20 ppm and 10 ppm, respectively. Carbamidomethylation of cysteine (+57.02146 amu) was considered as a static modification and oxidation at Methionine (+15.995), Acetylation of protein N- terminal and K (+42.01 amu), and pyrrolidone from carbamidomethylated C (− 17.03 amu) as variable modification.

Proteome discoverer 1.4 3 (version 2.4.0.305, Thermo Scientific) is the third software used as a raw data post-processing interface to select scan events for peptide/protein identification. Sequest HT was used as a search engine, and to avoid any bias, the variable amino acid modification and fixed modifications were set as peptide-shaker. Trypsin was selected as the enzyme, with two potential missed cleavage. Peptide and fragment ion tolerance was 20 ppm and 0.5 Da, respectively. The false discovery rate (FDR) was kept at 1% at the protein level to ensure high-quality results in the three software mentioned.

Database construction

Two databases were tested against the 3 software mentioned above; the protein database and the unique peptide database (Fig. 2). The protein database was done on a combined database of 11 species (*Sus scrofa* (Pig), *Ovis aries* (Sheep), *Bos Taurus* (Bovine), *Capra hircus* (Goat), *Camelus dromedaries* (Camel), *Struthio camelus* (Ostrich), *Crocodylus niloticus* (Nile crocodile), *Naja haje* (Egyptian cobra), and *Dama Mesopotamia, Axis axis, Cervus elaphus* (Deer)), downloaded from Uniprot database with a total 299,884 protein sequence.

The unique peptide database was constructed by *in-silico* digestion of the aforementioned combined database after deduplication with two maximum missed cleavages. This database holds 59,823,706 peptides. Duplicate peptides were refined through the following steps; (1) peptides with sequences of less than six amino acids and more than thirty-five amino acids were excluded from the database, and (2) all duplicated peptides with identical sequences were removed to keep unique characteristic sequences. As a result, a unique peptide database of 1,670 peptides was constructed.

Bioinformatics analysis

The analysis was done based on each software with its embedded algorithm. Firstly, data retrieved from the protein pilot using the Pro Group™ Algorithm was analyzed using peptide sequence identification. Proteins sorted based on the highest identified peptide number, and the protein identified with the highest peptide number (95% confidence) was selected. Species were then determined based on these chosen proteins. These criteria were applied to all data retrieved using the two different databases.

Secondly, X! Tandem in the peptide-shaker software platform was analyzed using the protein database based on Peptide Spectrum Matches (PSMs). PSMs scoring is a value representing the total number of identified peptide spectrum matches for a protein, including the redundant matches identified. Our analysis strategy selected the top

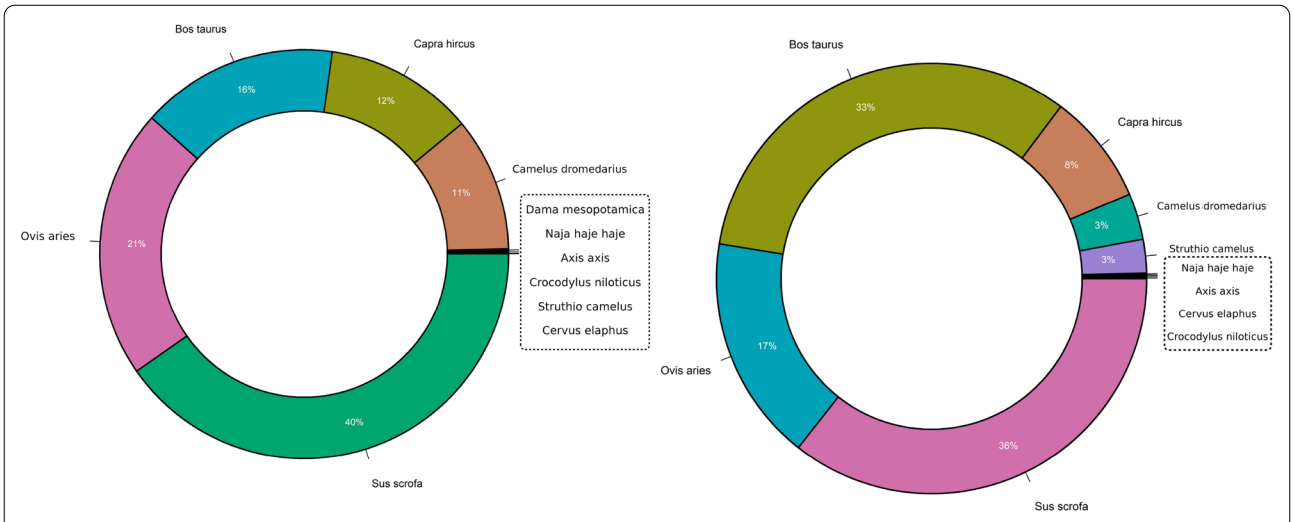

Fig. 2 Describes the proteins or/and peptides distributions in both databases used as background searches for protein/species identification. **A** Proteins distribution percentage assigned by the species name, used for protein search. **B** Peptides distribution percentage assigned by the species name, used for unique peptide search

highest and validated PSMs for each sample. Proteins linked to these peptides were inspected to determine their associated animal species. For the unique peptide database, X! Tandem was used for peptide identification. The species identification was based on the number of unique peptides, varying in at least 1 amino acid residue. The higher the number of unique peptides in a protein group, the more the sample is closely assigned to the species recognized from that protein group.

The last analysis was applied to the data retrieved from the proteome discoverer using the protein database, where validated PSMs and identified peptide numbers were used for protein identification. Proteins with the top highest validated PSMs were selected. While in peptide number identification, the higher the number of unique peptides to a protein group, the more the sample is to be closely assigned to the species recognized from that protein group. Thus, proteins with the highest number of peptides were selected, and both methods were matched with the reference. Proteins identified using the peptide database relied on species identification through the number of unique peptides and Sequest score. The latter determines the peptide sequence yielding the best correlation between the experimentally observed and the theoretical MS/MS spectra of peptides present in the databases [30]. The proteins identified with the highest number of peptides and those with the highest Sequest score were selected and reported. Lastly, we implemented the three different methodologies to identify species of the unknown leather samples in terms of sensitivity and precision.

Results

Identification success of reference samples based on different search engines (algorithms) and databases

Our observation revealed that different search engines and databases yielded different results. The most successful approach for determining the leather samples' animal origin was using protein pilot software via the paragon algorithm. About 78% (14/18) of the samples were identified correctly compared to their corresponding pre-identified hits (Additional file 1: table S1). Correct hits were ranked based on the number of validated identified peptides. We assumed that the largest number of peptides most likely identifies the correct species by either a protein database or a unique peptide database. Detailed protein identification and peptide sequences could be accessed in Additional file 2. Additionally, a second copy could be accessed in an external database https://doi.org/10. 5281/ zenodo. 7143692. The successfully predicted species based on the constructed protein database allowed the identification of 67% (12/18) of the taxa, which was twice as determined through the peptide database. In contrast, the percentage of correct species identification decreased to 28% (5/18) when the peptide database was used (Additional file 1: table S1a). Yet, the number of peptides through the peptide database was the only method that correctly predicted the ostrich standard sample. Overall, the species identified using the paragon search engine were as follows: 7 skin samples belonging to bovine, three identified as sheep, two as goat, one as pig, and one as ostrich (Additional file 1: table S1a).

On the other hand, the least correctly predictive model was the Peptide Shaker software-based tool through the X! Tandem search. It correctly predicted about half of the samples, given only a 56% (10/18) species identification rate (Additional file 1: table S2). Herein, the validated PSM was matched to animal species protein databases to identify expressed proteins in the samples. At the same time, the number of peptides was adopted as a search parameter matched to the peptide database (Fig. 3). Among the 18 samples, 44% (8/18) of the identified taxa were detected through the validated PSM rather than the number of peptides approach that identified only 6% (1/18) of the samples, showing better results compared to the peptide database (Additional file 1: table S2a). The only sample identified through the number of peptide searches, but not through the PSM method, was one out of the 2 sheep standard samples. Using this X! Tandem model either through the PSM or the Number of peptides, none of the goat samples were correctly identified, nor was the ostrich sample; however, it was amenable to detect all sheep samples and the pig sample. Thus, the 10 recognized species were 6 bovine, 3 sheep, and one pig (Additional file 1: table S2a).

Our results showed successful identification regarding proteome discoverer software and the SEQUEST search engine. Since 67% (12/18) of the samples were correctly identified (Additional file 1: table S3). Based on the protein database constructed, we had two search methods for species identification: the validated PSM and the number of peptides. On the other hand, based on the peptide database, the unique peptides and the SEQUEST score were used as the search parameters (Fig. 3). The number of peptides method showed that 67% (12/18) of peptides were assigned to the correct species, while 61% (11/18) of spectra were correctly assigned via the validated PSM method (Additional file 1: table S3a). On the other hand, the species identification rate decreased to 33% (6/18) when the unique peptides were searched based on the peptide database. It dropped to 11% (2/18), where only two samples were correctly identified through the SEQUEST score (Additional file 1: table S3). Adopting this predictive model through SEQUEST engine in proteome discoverer, the 12 identified species were: 6 bovines, 3 sheep, 2 goat, and one pig.

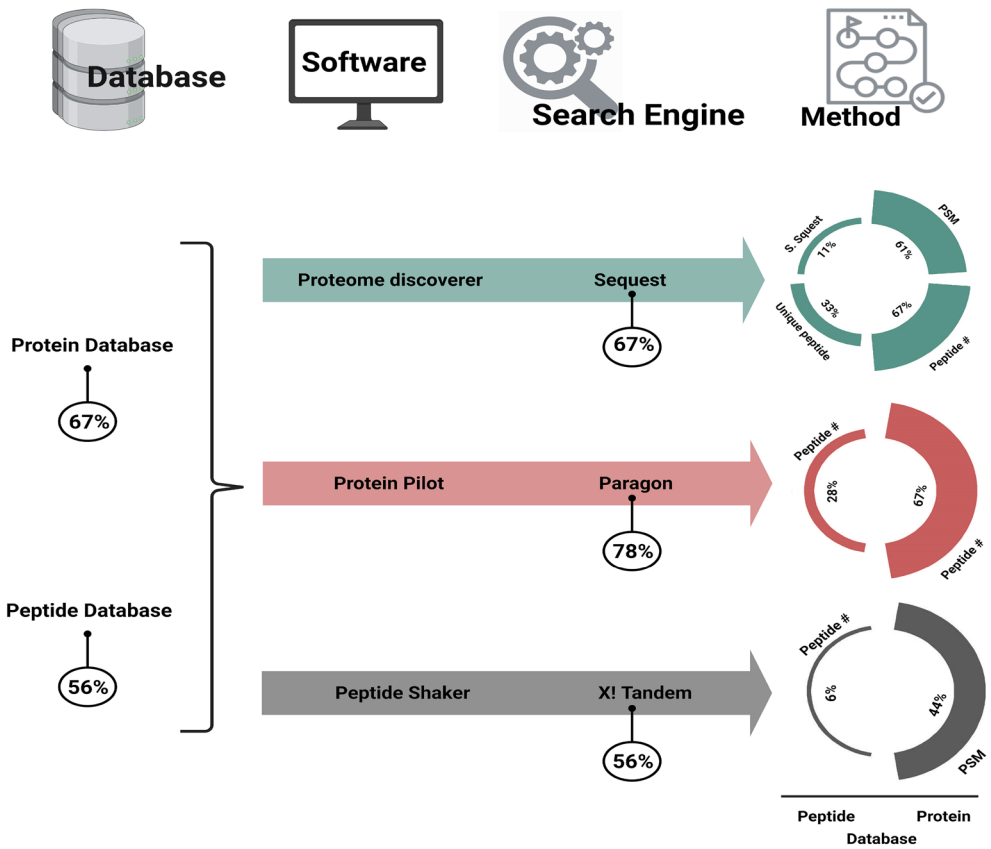

Fig. 3 Describes the used databases, softwares's algorithims and the metods used for analysis with the percentage of correct identification when matched with the reference

To this point, we compared all methods used in this study to determine which search method/parameter and database were most accurate in species identification. We found that the whole peptides search approach based on the protein database, used in Protein Pilot and Proteome discoverer, achieved the highest results with a 67% identification rate and detected all pig with 100% accuracy. In contrast, its accuracy has decreased to 85.7% for the bovine identification as it detected 6 out of the seven bovine samples and 67% (2/3) of goat samples. Coming second, the validated PSM approach, used in Proteome Discoverer and Peptide Shaker, with an overall accuracy of 61% (11/18) in identifying the proper species that matches the known standard. Furthermore, its accuracy in determining the pig and bovine samples remained 100%.

Nonetheless, compared to the number of peptide approaches, the overall accuracy of the PSM decreased to 33% (1/3) for goat. This goat sample was identified through validated PSM in Proteome Discoverer; however, the PSMs in peptide shaker did predict none of the goat samples. Despite the high accuracy of the number of peptides based on the protein database in identifying the standard samples, the number of peptides that relied on the peptide database was less precise. One explanation is the lower chance to match a unique sequence with 55.46% for the complete identification. However, we noticed that determining the animal taxa through the number of peptides relying on the peptide database was poor for some species, particularly goat. None of the goat samples were identified using this method.

Moreover, its accuracy in recognizing the bovine samples declined to 71.4%, making it more fallible than its counterpart, which relies on the protein database. Finally, the SEQUEST score has proven to be the least informative parameter to identify animal taxa since it barely managed to detect 11.11% (2/18) of the standard samples, harboring it unreliable. These 2 samples belonged to the pig sample and one of the bovine samples. All in all, the most predictive search parameter was the number of peptides through the protein database, followed by the validated PSM, whereas the poorest one was the SEQUEST score.

To validate the appropriate database that aids for accurate identification, either protein or unique peptide database, both searches were compared. Findings have shown that the protein database search identified 67% (12/18) of the standard samples. Using the protein database, the successfully predicted species were bovine, sheep, pig, and goat. On the other hand, only 56% (10/18) of the samples were correctly predicted using the unique peptide database, which included bovine, sheep, pig, and ostrich species. The protein database was more efficient in identifying goat samples. However, the peptide database correctly predicted the ostrich sample, even when it did not recognize any goat samples. Regardless of the database or the search method used, although many species were successfully defined via the different approaches adopted, the introduced methods could not identify some species due to the paucity of collagen sequences in the database. We have observed that they are 4 samples that include some species such as Egyptian cobra, deer, Nile crocodile, and a goat in the sample (Sample #3).

Species identification of the unknown archaeological samples

Having the results mentioned earlier on the reference samples, we subsequently attempted to identify the 10 unknown archaeological samples by the three different models. Still, we expected the protein pilot model to be the most accurate. First, using the proteome discoverer software with implemented SEQUEST search engine algorithm, we identified the unknown archeological objects to be as follows: (5261), (19.3.10), (25.3.223), and (31.3.73) were identified as goat, while samples (24.8f) and (31.3.98) as goat or bovine.

On the other hand, using the paragon search engine through the protein pilot software, we found that samples (5261), (19.3.10), (24.8f), and (31.3.98) were detected to be goat, sheep, or bovine, and sample (31.3.73) was recognized as goat or sheep, or pig. These results overlapped findings with those identified through the SEQUEST model, showing more identification confirmation. At Last, the X! Tandem of Peptide Shaker model recognized samples (19.3.10), (24.8f), and (31.3.98) to be identified as sheep, while sample (31.3.73) was highly matched with bovine, and sample (25.3.223) was close to either pig or sheep.

Identified hallmark proteins and unique peptides in reference samples

In this study, two critical variables influenced accurate species identification, the number of proteins identified in a sample and the thoroughness of a species' representation in the protein/ unique peptide database. Most of the proteins identified were collagens, including Collagen type I, type II, type III, and type VI. Most importantly, collagen marker peptides were common with different isoforms and patterns of the presence or absence of marker peptides between Sheep, Goat, Bovine, and Pig. All examined samples have produced collagen fingerprints, and the in-depth analysis revealed that each contains a combination of the collagen type I protein. Collagen type I alpha 1 (COL1A1) had the largest number of peptides in all samples, followed by collagen type I alpha 2 (COL1A2). Interestingly, the ten unknown samples showed an extensive list of protein signatures generated from the collagen type I alpha 1 as the most abundant protein in leather, with different isoforms in each species.

Discussion

Paleoproteomics has been proven to be better at global protein identification when taxonomic resolution is required [31]. This is because LC-MS/MS has become more accurate and sensitive in the last few years as it can handle and separate complex peptide and protein mixtures more than before [32, 33]. LC-MS/MS, in particular, has been adopted to identify the animal source of various archaeological leather artifacts, including skin, shoes, clothing, etc. [15, 22, 34–36]. However, some barriers restrain protein identification including the database quality and protein evolutionary conservation among taxa. Therefore, this paper aimed to assess the best approach to correctly identify leather skin archaeological samples using three predictive models: SEQUEST through Proteome Discoverer, Paragon through Protein Pilot, and X! Tandem via Peptide Shaker. Inside each model, we approached other search methods, including high PSM scoring selection, the number of peptides, and/or SEQUEST score that best match to a protein or a peptide database. We sought to address the limitations harboring each method and the improvements that could be further applied. Increasing the sensitivity of species-specific identification is instrumental for multiple applications that infer the past human ecology, providing a better understanding of human–environment interactions, the development of agriculture, and distinctions between prehistoric civilizations.

Our analysis found collagen I followed by collagen II, as the most abundant protein in all samples due to its resistance to degradation, despite the samples' exposure to environmental and chemical alterations during excavation and tanning processes. Type 1 Collagen consists of two alpha I chains and one alpha II chain that are twisted on top of each other to form a solid and stable triple helix [37, 38]. Notwithstanding the robustness of collagen structure, it is known for its slow evolutionary rate, over ~ 450 million Years (Myr). This slow rate

of evolution gives rise to a few amino acid substitutions that could be derived from conservative missense mutations. As a result, exon and intron homologies are highly preserved among their isoforms and across vertebrate species [39] which subsequently provides homologous protein sequences that not only have similar masses but are accordingly hard to differentiate by the mass spectrometry [40, 41].

Accordingly, collagens have a tight potential to discern between closely related species. This complication even worsens when a species has limited representation in the database since another species will inescapably be recognized with multiple spectra [42]. For example, in the high-scoring PSMs, samples were from sus scrofa as an animal source of origin, and most of the bovine samples were correctly identified, indicating that pig and bovine were the most thorough and represented species in the database.

Using the number of unique peptides matched to a protein database was the most effective method for correct species identification. This method uses significantly more data and increases accurate species identification. It sums the species of all peptides found in a sample- both unique and shared- and assumes that the most frequently identified species would likely be the correct species. On the other hand, the validated PSM method showed the second-highest identification rate of the correctly assigned proteins. Thus, this could be due to different reasons. First, PSMs consider counting the redundant spectra, resulting in more data points for a detailed analysis. Second, PSMs reduce the effect of amino acid substitutions that could be generated from a single spectrum, which, in return, increases the accuracy of species identification. Furthermore, this method may be sufficient when plenty of proteins are available. However, when few highly conserved proteins are present for analysis, this method may suffer limitations in light of high-quality spectra identifying erroneous peptides.

A plethora of methods for taxa identification using leather/skin samples have been proposed [31, 43, 44]. Multiple methods rely on detecting species-specific peptide(s) to make these determinations. However, as proposed here, the frequency of identifying high-quality, unique peptides from species other than the sample species limits the usefulness of these methods. Compared to the other predictive methods in this study, the number of peptides matched to a peptide database showed poor taxa identification. This could be due to the fact that some peptides could show some discrepancies across samples even when they are derived from the same protein due to variations in post-translational modifications [45]. Hence, the higher the PTMs considered during the search, the less the identified peptides, and the less it is matched to the peptide database. Indeed, variable PTMs lead to a steep increase in the search space size, which makes search engines struggle with comparative analysis of PTMs and increases false positives because of incorrect PSMs. Another reason is that some species will get more peptide hits than others, and, in the same manner, some species will have few unique peptides assigned to them due to the overlapping of their proteome with other species.

Nonetheless, this method was able to recognize the ostrich Sample, unlike any other method, inferring that this method could help in finding species-specific peptides to a limited range since it gets also influenced by the animal representation in the database as well as the paucity of peptide markers among most animals due to the slow evolutionary rate of collagen proteins. Finally, using the SEQUEST score was the least successful identification method. Again, the homology and high similarity between collagen-identified sequences among species play a major role in the probability-based and cross-correlation analysis between the mass spectra and peptide sequences. As the number of amino acid substitutions in peptides decreases, the uniqueness and the number of the identifiable mass spectrum will diminish, giving lower SEQUEST scores, which inevitably reduces the accuracy of taxa identification.

Our study casts light on many practicable predictive models to identify the animal source of archaeological samples. However, possible limitations of the study would be the sample size, which could have been more significant to give more informative percentages for each identification methodology. Yet, it is impossible to obtain several artifacts and study them using destructive methods, such as mass spectrometry. Additionally, considering the PTMs in the sample could contribute to poor protein identification, as previously explained. However, to tackle this in the future, we support the previous suggestions by Chen et al., 2020, to use the iterative search for identifying PTMs (ISPTM) approach or a hybrid database of de novo and database search (i.e., InsPecT algorithm), which could help not only in controlling the search space, but also in increasing the spectral identification rate by improving the accuracy of isoform identification in non-canonical proteomes [46].

Other suggestions and improvements could also be taken into consideration. For example, herein, we used untargeted proteomics analysis to search against a unique database comprising different species pertaining to the Ancient Egypt civilization. Nonetheless, working on targeting particular peptide markers in a Multiple Reaction Monitoring (MRM) mode could lead to more sensitive, specific, and rapid protein detection and species-specific identification because of its high

resolution in distinguishing the few amino acid substitutions in collagen among taxa [36]. Furthermore, we recommend utilizing this novel approach in addition to the emerging Data-Independent Acquisition (DIA) analysis. In contrast to the traditional DDA analysis used here, the DIA strategy enables higher reproducible and sensitive detection of peptides since all precursor ions on the survey scan (MS1) are objectively selected for fragmentation in MS2, giving more reliable and detailed data for peptide identification [47].

Regardless of the future work that suggests improving the taxonomic identification rate, protein identification for archaeological samples will remain problematic as long as the number of available proteins in samples, the accuracy of sequencing, and the database are insufficient and unthorough. Consequently, these problems hamper the proteins' informative value regarding speciation and make MS accuracy even harder to quash. However, the hope continues with the promising evolutionary trends in high-throughput genomic sequencing, resulting in more genes being translated into proteins and thus allowing protein completeness of various taxa in the database.

Abbreviations

LC-MS/MS: Liquid Chromatography Coupled with Tandem Mass Spectrometry; GC-TOF/MS: Gas chromatography time-of-flight/mass spectrometry; MALDI-TOF-MS: Matrix Assisted Laser Desorption Ionization-Time Of Flight Mass Spectrometry; PSMs: Peptide Spectrum Matches; Met: Metropolitan Museum of Art, New York, USA; DTT: Dithiothreitol; IAA: Iodoacetamide.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40494-022-00816-0.

Additional file 1. Species- samples identification using different search engines and databases.

Additional file 2. Detailed protein and peptide information using different search engines and databases. https://doi.org/10.5281/zenodo.7143692.

Acknowledgements

We would like to acknowledge Yvette A Fletcher & Anne Lama from the Leather Conservation Centre (UK) for providing the reference leather samples. The permission for the archaeological leather sampling was secured by the first author during his research fellowship under the supervision of Ann Heywood at the Metropolitan Museum of Art (New York, USA, 2012).

Author contributions

AE & SM provided oversight and leadership responsibility for the research activity planning, execution and design of methodology. AO, SE and SE conducted the wet lab proteomics experiment. AMA performed the bioinformatic analysis including statistical analysis and data visualization. HE & ML conducted a research and investigation process for leather analysis. AE and ML collected the archeological samples and facilitated permissions. All authors read and approved the final manuscript.

Funding

Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB). Part of this research was supported by the Andrew W. Mellon Foundation during a research fellowship of the first author at the Metropolitan Museum of Art (New York, USA, 2012). Additionally, part of this work was supported by the Egyptian Cancer Network, USA (ECN) and the Children's Cancer Hospital Egypt 57357.

Availability of data and materials

The experimental proteomics data supporting this study's findings were deposited in Proteomics identification database PRIDE (https://www.ebi.ac.uk/pride/), project ID: PXD033367.

Declarations

Competing interests

The authors declare no competing interests.

Author details

[1]Archaeological Science and Excavations Department, Faculty of Archaeology, Ain Shams University, Cairo 11566, Egypt. [2]Proteomics and Metabolomics Research Program, Basic Research Department, Children's Cancer Hospital, Cairo, Egypt. [3]The Scientific Research Department, the Metropolitan Museum of Art, New York, NY 10028, USA. [4]Department of Physiology, Faculty of Veterinary Medicine, Suez Canal University, Ismailia 41522, Egypt.

Received: 4 August 2022 Accepted: 24 October 2022
Published online: 09 November 2022

References

1. Wills B. The tanning of sheep and goat skins in North Africa. Newslett (Museum Ethnographers Group). 1987;20:84–90.
2. Khanbabaee K, Van Ree T. Tannins: classification and definition. Nat Prod Rep. 2001;18(6):641–9.
3. van Driel-Murray C. Leatherwork and skin products. In: Selin H, editor. Ancient Egyptian materials and technology. Dordrecht: Springer; 2000. p. 299–319.
4. Covington AD. Tanning chemistry: the science of leather. Cambridge: Royal Society of Chemistry; 2009.
5. Welker F, et al. Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. Nature. 2015;522(7554):81–4.
6. Warinner C, Korzow Richter K, Collins MJ. Paleoproteomics. Chem Rev. 2022;122(16):13401–46.
7. Kumazawa Y, et al. A novel LC–MS method using collagen marker peptides for species identification of glue applicable to samples with multiple animal origins. Herit Sci. 2018;6(1):1–9.
8. Leo G, et al. Proteomic strategies for the identification of proteinaceous binders in paintings. Anal Bioanal Chem. 2009;395(7):2269–80.
9. Vinciguerra R, et al. Proteomic strategies for cultural heritage: from bones to paintings. Microchem J. 2016;126:341–8.
10. Buckley M. Paleoproteomics: an introduction to the analysis of ancient proteins by soft ionisation mass spectrometry. In: Paleogenomics. Dordrecht: Springer; 2018. p. 31–52.
11. Coutu AN, et al. Palaeoproteomics confirm earliest domesticated sheep in southern Africa ca. 2000 BP. Sci Rep. 2021;11(1):6631.
12. Hendy J. Ancient protein analysis in archaeology. Sci Adv. 2021;7(3):9314.
13. Maidment C, et al. Comparative analysis of the proteomic profile of cattle hides that produce loose and tight leather using in-gel tryptic digestion followed by LC-MS/MS. J Am Leather Chem Assoc. 2020;115(11):399–408.
14. Fiddyment S, et al. Animal origin of 13th-century uterine vellum revealed using noninvasive peptide fingerprinting. Proc Natl Acad Sci. 2015;112(49):15066–71.
15. Brandt LØ, et al. Species identification of archaeological skin objects from Danish bogs: comparison between mass spectrometry-based peptide sequencing and microscopy-based methods. PLoS ONE. 2014;9(9):e106875.

16. Elnaggar A, et al. The characterization of vegetable tannins and colouring agents in ancient Egyptian leather from the collection of the metropolitan museum of art. Archaeometry. 2017;59(1):133–47.
17. Abdel-Maksoud G. Analytical techniques used for the evaluation of a 19th century quranic manuscript conditions. Measurement. 2011;44(9):1606–17.
18. Hollemeyer K, et al. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry combined with multidimensional scaling, binary hierarchical cluster tree and selected diagnostic masses improves species identification of Neolithic keratin sequences from furs of the Tyrolean Iceman Oetzi. Rapid Commun Mass Spectrom. 2012;26(16):1735–45.
19. Izuchi Y, Takashima T, Hatano N. Rapid and accurate identification of animal species in natural leather goods by liquid chromatography/mass spectrometry. Mass Spectrometry. 2016;5(1):A0046–A0046.
20. Buckley M, et al. Distinguishing between archaeological sheep and goat bones using a single collagen peptide. J Archaeol Sci. 2010;37(1):13–20.
21. Hasegawa N, et al. Calcineurin binds to a unique C-terminal region of NBCe1-C, the brain isoform of NBCe1 and enhances its surface expression. BPB Reports. 2019;2(1):7–18.
22. Ebsen JA, et al. Identifying archaeological leather–discussing the potential of grain pattern analysis and zooarchaeology by mass spectrometry (ZooMS) through a case study involving medieval shoe parts from Denmark. J Cult Herit. 2019;39:21–31.
23. Buckley M, et al. Species identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. Rapid Commun Mass Spectrom. 2009;23(23):3843–54.
24. Pizzi A. Covalent and ionic bonding between tannin and collagen in leather making and shrinking: a MALDI-ToF study. J Renew Mater. 2021;9(8):1345–64.
25. Fiddyment S, et al. Girding the loins? Direct evidence of the use of a medieval English parchment birthing girdle from biomolecular analysis. R Soc Open Sci. 2021;8(3): 202055.
26. Hollemeyer K, et al. Species identification of Oetzi's clothing with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry based on peptide pattern similarities of hair digests. Rapid Commun Mass Spectrom. 2008;22(18):2751–67.
27. Hendy J. Ancient protein analysis in archaeology. Sci Adv. 2021. https://doi.org/10.1126/sciadv.abb9314.
28. Saadeldin IM, et al. Thermotolerance and plasticity of camel somatic cells exposed to acute and chronic heat stress. J Adv Res. 2020;22:105–18.
29. Magdeldin S, et al. Off-line multidimensional liquid chromatography and auto sampling result in sample loss in LC/LC–MS/MS. J Proteome Res. 2014;13(8):3826–36.
30. MacCoss MJ, Wu CC, Yates JR. Probability-based validation of protein identifications using a modified SEQUEST algorithm. Anal Chem. 2002;74(21):5593–9.
31. Buckley M. Species identification of bovine, ovine and porcine type 1 collagen; comparing peptide mass fingerprinting and LC-based proteomics methods. Int J Mol Sci. 2016;17(4):445.
32. McDonald WH, Yates JR 3rd. Shotgun proteomics: integrating technologies to answer biological questions. Curr Opin Mol Ther. 2003;5(3):302–9.
33. Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC— MS/MS. J Proteome Res. 2011;10(4):1785–93.
34. Brandt LØ, Haase K, Collins MJ. Species identification using ZooMS, with reference to the exploitation of animal resources in the medieval town of Odense. Danish J Archaeol. 2018;7(2):139–53.
35. Brandt LØ, Mannering U. Taxonomic identification of Danish Viking Age shoes and skin objects by ZooMS (Zooarchaeology by mass spectrometry). J Proteomics. 2021;231: 104038.
36. Kumazawa Y, et al. A rapid and simple LC-MS method using collagen marker peptides for identification of the animal source of leather. J Agric Food Chem. 2016;64(30):6051–7.
37. Shoulders MD, Raines RT. Collagen structure and stability. Annu Rev Biochem. 2009;78:929–58.
38. Engel J, Bächinger HP. Structure, stability and folding of the collagen triple helix. In: Brinckmann J, Notbohm H, Müller PK, editors. Collagen. Berlin: Springer; 2005. p. 7–33.
39. Welker F. Palaeoproteomics for human evolution studies. Quatern Sci Rev. 2018;190:137–47.
40. Dutta S, et al. Chemical evidence of preserved collagen in 54-million-year-old fish vertebrae. Palaeontology. 2020;63(2):195–202.
41. Azemard C, et al. Animal fibre use in the Keriya valley (Xinjiang, China) during the Bronze and Iron Ages: a proteomic approach. J Archaeol Sci. 2019;110: 104996.
42. Bogdanow B, Zauber H, Selbach M. Systematic errors in peptide and protein identification and quantification by modified peptides. Mol Cell Proteomics. 2016;15(8):2791–801.
43. Gu M, Buckley M. Semi-supervised machine learning for automated species identification by collagen peptide mass fingerprinting. BMC Bioinformatics. 2018;19(1):1–9.
44. Kirby DP, et al. Identification of collagen-based materials in cultural heritage. Analyst. 2013;138(17):4849–58.
45. Bugyi F, et al. Influence of post-translational modifications on protein identification in database searches. ACS Omega. 2021;6(11):7469–77.
46. Hu C, et al. The solute carrier transporters and the brain: Physiological and pharmacological implications. Asian J Pharm Sci. 2020;15(2):131–44.
47. Hu A, Noble WS, Wolf-Yadlin A. Technical advances in proteomics: new developments in data-independent acquisition. F1000Research. 2016;5:419.

Publisher's Note