Document Title:	Identifying Individuals through Proteomic Analysis: A New Forensic Tool to Rapidly and Efficiently Identify Large Numbers of Fragmentary Human Remains
Author(s):	City of New York, Office of Chief Medical Examiner
Document Number:	254583
Date Received:	March 2020
Award Number:	2014-DN-BX-K014

Final Summary: NIJ Grant 2014-DN-BX-K014, *Identifying Individuals through Proteomic Analysis: A New Forensic Tool to Rapidly and Efficiently Identify Large Numbers of Fragmentary Human Remains*.

This summary follows the NIJ Post Award Reporting Requirements issued March 28, 2019 and is divided into the four prescribed sections: 1) Purpose, 2) Research Design & Methods, 3) Data Analysis & Findings, and 4) Implications for Criminal Justice Policy and Practice.

ABBREVIATIONS

ACN = Acetonitrile
ADD = accumulated degree days
BMI = body mass index
CV = cross validation
DDA = direct data analysis
Ev = electron volts
FAC = Forensic Anthropology Center
HPLC = high performance liquid chromatography
Q-TOF = TripleTOF mass spectrometer
MAF = minor allele frequencies
ms – millisecond
MS = mass spectrometry
mse = mean square error
MS/MS = tandem mass spectrometry
NYC OCME = New York City Office of Chief Medical Examiner
NYU = New York University
PMI = postmortem interval
SAAP = single amino acid polymorphism
SNP = single nucleotide polymorphism
STR = short tandem repeat
TBS = total body score
TOF = time-of-flight
UTK = University of Tennessee, Knoxville

1. Purpose – To determine if the genetic information present in proteins can be accurately and efficiently detected by mass spectrometry for use in human identification and, if so, the stability of protein markers to taphonomic decay. The use of proteomic genetic information for human identification would be valuable in cases where DNA is too degraded for analysis, or when there are substantial numbers of physically unidentifiable human remains, e.g. following a mass disaster, as they could be more quickly identified by mass spectrometry.

1.1 STATEMENT OF THE PROBLEM: Incidents such as airline crashes, infrastructure failures (e.g. bridges), industry explosions and natural catastrophes can result in large numbers of fragmentary human remains spread over large areas of land or water. These remains may be burned or contaminated with caustic substances (e.g. jet fuel and industrial chemicals) that may degrade

DNA or interfere with DNA analysis. In cases where comingling of taphonomically degraded biological material occurs (e.g. common graves), insufficient DNA may remain for analysis. In such cases, individual identification may not be possible. Additionally, in cases of mass disasters, where thousands of fragmentary remains may be present, the significant amount of time required for DNA testing could delay funeral arrangements by grieving families who want to wait for all remains to be identified and returned, as well as potentially impede investigators.

Because proteins, like DNA, possess genetic information and, because they are more stable than DNA and can be detected even when partially degraded, they offer practical alternative for individual identification. Additionally, current methods for protein identification by mass spectrometry are both confirmatory and relatively fast. Consequently, the use of proteomic mass spectrometry (**MS**) for human identification would be useful not only where DNA degradation or contamination preclude DNA analysis, but also for separating large numbers of fragmentary human remains into groups of single individuals, and consequently reducing DNA STR testing to one or a few sample fragments.

1.2 OBJECTIVES: The major goals of this project are: **i)** to identify informative single amino acid polymorphisms (**SAAPs**) in human muscle and bone proteins that can be used for individual identification, to confirm these variants by genetic analysis, and to determine the maximum population size that they can meaningfully discriminate, **ii)** to evaluate taphonomic effects on protein degradation over time in different seasons, and **iii)** to assay three different body areas in order to assess possible variations in protein expression.

2. Research Design & Methods: The two main objectives of this application were i) to identify protein polymorphisms and determine if they can be used for human identification, and ii) to evaluate the effects of taphonomic decay on protein markers. To achieve these objectives, i) a sufficient number of human tissue samples needed to be collected to survey for polymorphisms and determine their discriminative power, and ii) samples from the same tissue needed to be assayed over time as they decayed. Consequently, this work required the expertise of three distinct disciplines at three institutions: 1) the Forensic Anthropology Center (**FAC**) at the University of Tennessee, Knoxville (**UTK**) for tissue collection, 2) the New York City Office of Chief Medical Examiner (**NYC OCME**) for protein analysis by mass spectrometry, and 3) New York University (**NYU**) for bioinformatic analysis of the discriminatory power of the identified SAAPs. Experimental design for each of these sections are described below.

2.1 SAMPLE SELECTION & COLLECTION - UTK, FAC - Muscle was chosen as an ideal tissue for this work for several reasons: i) it would to be commonly found in body fragments in the aftermath of a mass disaster, ii) it possesses large numbers of different proteins and therefore increases the likelihood of identifying SAAPs, iii) it contains large amounts of blood (and blood proteins) which would be found in many other biological samples, iv) it is possible to evaluate different types of muscle (arm, legs and chest) to determine if all SAAPs are expressed universally, and v) it is an abundant tissue that can be easily sampled in a longitudinal taphonomic study. Bone was also obtained for a smaller trial study aimed only at identifying bone protein SAAPs.

- *MUSCLE:* Muscle samples were obtained from the 14 cadaver donors over the course of four seasonal trials. Three muscle samples from different anatomical areas were collected from each individual in order to evaluate possible variations in protein expression in different area of the body. Initial, day 0 samples, were used to i) identify protein polymorphisms, ii) extract DNA to confirm candidate protein polymorphisms and iii) represent a “total protein” baseline for comparison in subsequent taphonomic studies. For three of the trials (winter, spring and summer), muscle samples were then taken at four additional time points: days 15, 30, 45, 60. During the third (fall) trial samples were taken every 2 days for 60 days for two of three donors, and every 10 days for 60 days for a third donor. This was done because of early rapid decay during the late summer at the beginning of this trial. Samples were stored at -80^{o}C and shipped to the NYC OCME on dry ice.

• *BONE:* Bone samples (hallux (large toe) phalanges) were taken from four of the 14 individuals at the beginning of the first trial (day 0) to identify bone polymorphisms. Again, samples were stored at -80^{o}C and shipped to the NYC OCME on dry ice.

2.2 PROTEIN EXTRACTION & ANALYSIS – NYC OCME - Samples arrived on a rolling basis after the completion of each seasonal trial at UTK FAC and were stored at -80^{o}C. Sample processing consisted of protein extraction, quantitation, and qualitative analysis, followed by protein digestion, peptides separation and mass spectrometry. A summary of methods follows.

• PROTEIN EXTRACTION – Muscle samples were homogenized on ice in 20 volumes (w/w) extraction buffer (7 M urea, 2 M thiourea, 50 mM Tris-HCl pH 7.5, 20 mM DTT and 1 mM EDTA) for 1-3 min using motorized glass-to-glass homogenizers except for 60-day samples which were solubilized in 10 volumes extraction buffer (w/w). After solubilizing, samples were transferred to 1.5 ml tubes and spun at 18,000 x g for 30 min at 4^{o}C to pellet debris. Supernatant protein concentration was measured by the Bradford protein assay with BSA as standard.

Bones were milled and 100 mg demineralized/extracted in 10 volumes (w/w) of 1.2 M HCl incubated at 4°C overnight. After incubation, samples were each centrifuged for 30 min at 4°C at 18,000 x g. Supernatants were concentrated and buffer exchanged in 10 kDa MWCO spin columns to 50 mM ammonium bicarbonate. Protein was measured using the bicinchoninic acid assay (**BCA**) protein assay.

• PROTEIN DIGESTION – Twenty micrograms of supernatant protein were reduced and alkylated prior to digestion with 1 µg of trypsin at 37 °C overnight. Following digestion samples were acidified and dried down. Sample peptides were resuspended in 2% acetonitrile (**ACN**) and 0.1 % trifluoroacetic acid (**TFA**).

• PEPTIDE SEPARATION & MASS SPECTROMETRY ANALYSIS - Peptides were separated by reverse phase (C18) high performance liquid chromatography (**HPLC**) typically using a 3-hour linear gradient of 2.0% to 40% ACN. Eluted peptides will be analyzed by a Q-TOF (TripleTOF 6600

System, Sciex, Framingham, MA) mass spectrometer operated in data dependent analysis (**DDA**) top 20 mode. MS and tandem MS (**MS/MS**) scans were set for 500 and 150 milliseconds (**ms**), respectively. A 20 sec elution window avoided repeating data collection on the same peptide. Rolling collision energy with a collision energy spread of 15 ev was used for fragmentation.

- PROTEIN DATABASE SEARCHING: To identify proteins, data were searched against the non-redundant human protein database using ProteinPilot 5 (Sciex).
- DNA EXTRACTION & ANALYSIS – DNA extraction, primer design, amplification and sequencing of suspected SAAPs was performed by Genewiz (South Plainfield, NJ).
- INFORMATICS – Custom POLYMORPHISM DATABASE CONSTRUCTION – In order to identify muscle and bone SAAPs a custom variant database was made using the human Go Exome database constructed with over 6,500 individuals with ~200-fold coverage and a total of 700,337 missense mutations. Minor allele frequencies (MAF) are known and MS peptide data with a >95% confidence interval were used to search for polymorphisms.

3. Data Analysis & Findings: This section is divided into seven parts: 1) Cohort Demographics 2) Total Muscle Protein, 3) Muscle SAAP Identification, 4) Human Identification from Muscle SAAPs, 5) Muscle Taphonomy, 6) Bone SAAPs, and 7) Conclusions.

3.1 COHORT DEMOGRAPHICS – While age, sex, ethnicity and health would not interfere with identifying SAAPs, age and health might affect the taphonomic processes. The age of the 14 decedents ranged from 42 to 90; only three (21%; 2 males, one female) were under 60-years of age. Average age was ~70.2 years, the median ~73.5. All decedents were Caucasian (self-reported), four (~29%) were female. All but one died of natural causes (male, 42, blunt force trauma). Eight (~57%) died of cancer (4 M/4 F); two from heart disease (14%). Mean and median BMIs were 26, range 19-35; two were underweight, four overweight and four obese. Accumulation in adipose tissue has been associated with loss in quality and quantity of muscle mass.

3.2 Total Muscle Protein & Differences in Anatomical Expression (Day 0) – Each cadaver was sampled in three places and supernatant protein content measured. Typically sampling sites were upper arm (average ~21 μg protein /μl), chest (average ~24 μg protein/μl) and thigh (average ~25 μg protein/μl). Differences in protein expression of different areas were not statistically significant. The five proteins found most commonly in all samples were filamin, myosin, nebulin, titin and troponin.

3.3 Muscle SAAP Identification – To date, 13 SAAPs have been identified and confirmed by DNA sequencing with population frequencies ranging between 0.15 and 0.49 (see table below). Cohort frequencies vary from population frequencies as the sample size was small.

Muscle SAAPs Cohort and General Population Frequencies

Protein*	CA3-2	COL6A2-2	ENO3-3	ENO3-4	MYOM1-25	MYOM3-7	MYOM3-8	NEB	PGM1-4	SYNPO2-4	TTN-1	TTN-2	TTN-3
Cohort	0.50	0.86	0.86	0.71	1.00	0.86	0.86	0.93	0.36	1.00	0.79	0.43	0.43
Population	0.49	0.41	0.33	0.45	0.25	0.48	0.40	0.37	0.22	0.15	0.23	0.22	0.32

* Numbers following dash refer to exons with polymorphisms for except titin. TTN numbers are different SAAPs without regard to exon.

CA3 = carbonic anhydrase 3
COL6A2 = collagen 6 alpha 2
ENO3 = enolase 3
MYOM1 = myomesin 1
MYOM3 = myomesin 3 (1)
MYOM3 = myomesin 3 (2)
NEB = nebulin
PGM1 = phosphoglucomutase 1
SYNPO2 = synaptopodin 2
TTN1 = titin SAAP 1
TTN2 = titin SAAP 2
TTN3 = titin SAAP 3

3. 4 Human Identification from Muscle SAAPs – Individual identification, whether by single nucleotide polymorphisms (**SNPs**) or single amino acids polymorphisms (SAAPs), is dependent on the number of polymorphisms identified, their population frequency, and whether or not zygosity is known. It is important to note that zygosity could not always be determined using current methods. Of the 88 heterozygotes determined by DNA analysis, only 31 were observed by mass spectrometry. Homozygous wildtype (+/+, 50) and homozygous polymorphic (-/-, 44) similarly could not be determined by the current method. Zygosity, however, can be taken into consideration when determining likelihood of a correct identification. For example, in a set of 1000 simulated individual profiles using the 13 SAAPs identified above, without knowing zygosity the average profile probability is 0.05, meaning 1 in 20 people share a profile. When full zygosity is known, the average profile probability decreases to 0.00001, meaning 1 in 10,000. A reasonable approach to significantly improve the discriminatory power of a proteomic SAAP assay would be to use SWATH data acquisition, which would interrogate mass spectrometry data for both wildtype and polymorphic amino acids and would also allow significantly more SAAPs to be identified.

3.5 Muscle Taphonomy - To evaluate muscle decomposition over time, protein content (µg protein per sample mass) was determined for each of the four seasonal studies (concentrations are the averages of three different muscle samples taken from each cadaver on each day). As expected, average protein concentration decreased from day 0 to day 60 with day 60 having the smallest amount of protein for most individuals. Winter trial individuals as a whole showed the least amount of protein decomposition over the 60-day period and had the most consistent protein concentrations within and between individuals. The spring trial showed a slightly faster decline in protein content. The remaining two trials (summer and fall) showed a rapid and steady decrease in sample protein content and greater variation within and between individuals. These variations likely resulted from the areas sampled, decedents age, sex, BMI and general health. Differences in the rates of muscle protein decomposition correlated with temperature.

Using a set of 75 samples from 11 individuals at day 0, 30, and 60 time points, we analyzed the decomposition of over time of 7,863 peptides identified by MS/MS data searched against the UniProt human protein database. A gradient boosting decision tree machine learning algorithm was used to assess the ability of MS/MS data to predict accumulated degree days (**ADD**, a measurement of heat units calculated from ambient temperatures). Prediction accuracy was tested using k-fold cross validation with each validation fold composed of all time point samples from each individual (11 folds). Results showed an average cross-validation (**CV**) mean square error (**mse**) of 0.187, not quite as good as the 0.114 mean CV mse obtained when predicting ADD from total body score (**TBS**) using the same algorithm. However, when the combined MS/MS peptide data and TBS data was used to predict ADD, the mean CV mse showed a slight improvement to 0.105.

3.6 Bone SAAPs – A total of 40 proteins were identified from the four bone samples. As was done with muscle, these proteins were used to build a custom variant database from the human Go Exome database and used to search for polymorphisms (see methods). However, the bulk of extracted bone peptides (~60%) belonged to only two proteins (collagen alpha 1(1) and collagen alpha 2(1)), which are highly conserved, and no polymorphisms were detected in any of the four bone samples. The third most abundant protein, alpha-2-HS-glycoprotein, constituted ~9% of all identified peptides. Here too, no SAAPs were identified in the four samples. Of the remaining 37 proteins (not all were identified in each sample) seven proteins had between 4-8 peptides shared between the four samples, and 26 proteins were represented by fewer than three peptides, and consequently none were found in all samples. Because of the conserved nature of collagen, the limited number of non-collagenous peptides identified, and the small population tested (four), no SAAPs were identified in these samples. As described above with use of SWATH MS data acquisition, with a larger sample size, would likely improve results.

3.7 Conclusions – The major goals of this work were i) to determine if single amino acid polymorphisms can aid in individual human identifications when DNA analysis is not possible due to degradation or chemical contamination or as a way to expedite identifications when large numbers of samples require processing, and ii) if proteins can be used to help determine postmortem interval. Our data demonstrate that individual identifications are possible through detection of SAAPs in muscle. Critical to this process is the establishment of custom variant databases. Our research also shows that improvements in the ability to identify SAAPs and consensus sequences through the use of SWATH data acquisition would likely improve SAAP and wildtype identifications and consequently increase the discriminatory power of the assay.

Our data also demonstrates that the use of protein quantitation in conjunction with the total body score method may increase the accuracy of determining the postmortem interval.

4. *Implications for Criminal Justice Policy & Practice* – The ability to confidently identify individuals provides valuable evidence to criminal justice system as well as for use in criminal investigations. This information is also vital following a mass disaster or the discovery of a commingled grave where large numbers of otherwise unidentifiable remains may be present. DNA analysis is a long used and accurate method for individual human identification. However, DNA can degrade rapidly in soft tissue due to normal taphonomic processes, and often cannot be detected in samples that are accidently or intentionally contaminated with caustic chemicals or are burned. Proteins offer an attractive alternative for human identification when DNA is not available or when large numbers of sample must be rapidly identified. Proteins carry individual genetic information similar to DNA, and they are both more stable and abundant than DNA. In both mass disaster and commingled graves where there are large numbers of samples that need to be quickly identified, or where either DNA is not present or is contaminated by chemicals that inhibit DNA analysis, proteins may be used. As proteomic forensic methods expand, become more accurate and simpler to use - with respect to instrumentation, software analysis and access to specific and expanded databases - protein identification standards, best practices and stringent SOPs will need to be established. This is an area where policy makers can help organize the scientific expertise necessary for the practical implementation of proteomic identification of individuals into routine forensic casework.