research articles Journal of proteome research

Identification of Proteins from Non-model Organisms Using Mass Spectrometry: Application to a Hibernating Mammal

Kevin P. Russeth,[†] LeeAnn Higgins,[‡] and Matthew T. Andrews*,[†]

Department of Biology, University of Minnesota Duluth, 1035 Kirby Drive, Duluth, Minnesota 55812, and Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, 140 Gortner Lab, 1479 Gortner Avenue, St. Paul, Minnesota 55108

Received September 13, 2005

A major challenge in the life sciences is the extraction of detailed molecular information from plants and animals that are not among the handful of exhaustively studied "model organisms." As a consequence, certain species with novel phenotypes are often ignored due to the lack of searchable databases, tractable genetics, stock centers, and more recently, a sequenced genome. Characterization of phenotype at the molecular level commonly relies on the identification of differentially expressed proteins by combining database searching with tandem mass spectrometry (MS) of peptides derived from protein fragmentation. However, the identification of short peptides from nonmodel organisms can be hampered by the lack of sufficient amino acid sequence homology with proteins in existing databases; therefore, a database search strategy that encompasses both identity and homology can provide stronger evidence than a single search alone. The use of multiple algorithms for database searches may also increase the probability of correct protein identification since it is unlikely that each program would produce false negative or positive hits for the same peptides. In this study, four software packages, Mascot, Pro ID, Sequest, and Pro BLAST, were compared in their ability to identify proteins from the thirteen-lined ground squirrel (*Spermophilus tridecemlineatus*), a hibernating mammal that lacks a completely sequenced genome. Our results show similarities as well as the degree of variability among different software packages when the identical protein database is searched. In the process of this study, we identified the up-regulation of succinyl CoA-transferase (SCOT) in the heart of hibernators. SCOT is the rate-limiting enzyme in the catabolism of ketone bodies, an important alternative fuel source during hibernation.

Keywords: hibernation • nonmodel • MS/MS • *Spermophilus tridecemlineatus* • algorithm • protein identification

Introduction

Proteomic research has moved dramatically forward with great improvements in mass spectrometers and the subsequent development of improved database searching algorithms. The ability to determine the identity of proteins by comparison with characterized sequences in available databases can now be performed with a high rate of success. Peptide matching with the most widely used search programs requires: (1) high homology or identity, (2) relevant amino acid modifications included in the searching parameters, or (3) a more general error tolerant search for amino acid mutations or post-translational modifications.[1,2] Protein identification is dependent upon the following: (1) the quality of the data, (2) the software tool used to extract searchable data from raw data, (3) the protein database content, and (4) the algorithm of the database searching program. The user selection of an algorithm is a variable that is not often examined. Protein identity searches using a single program may yield fewer statistically significant peptides hits per protein than a strategy that uses multiple programs. As a result, a protein identity search using a single method may not identify all peptides correctly and thus a valid protein identification could be overlooked. This challenge is even greater when the goal is to identify proteins from a nonmodel organism that lacks a sequenced genome or an extensive sequence database since the number of exact matches between experimentally and theoretically derived peptides may be small.

In this paper, we show the molecular characterization of an important nonmodel eukaryotic organism by a proteomics approach at the nanoscale level[3,4] using unfractionated peptide mixtures, mass spectrometry and database searching with four commonly used software programs: Mascot, Pro BLAST, Pro ID, and Sequest. We used either matrix assisted laser desorption ionization quadrupole-time-of-flight (MALDI-QqTOF) or electrospray ionization quadrupole-time-of-flight (ESI-QqTOF) mass spectrometry to identify proteins from active and hibernating thirteen-lined ground squirrels (*Spermophilus tridecemlinea-*

* To whom correspondence should be addressed. Tel: (218) 726-7271. Fax: (218) 726-8142. E-mail: mandrews@d.umn.edu.

[†] Department of Biology, University of Minnesota Duluth.

[‡] Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota.

10.1021/pr050306a CCC: $33.50

Published on Web 03/09/2006

tus). In all cases, protein identifications among all programs were in agreement, regardless of whether manual inspection revealed a valid match. However, the number of peptides found and the ranking of common peptides among programs varied. Our results show the degree of variability inherent in the results from a select group of software packages. In general, we found protein coverage was increased by the use of multiple programs when proteomic studies of a nonmodel organism is attempted. In addition, we offer a brief overview of the scoring mechanisms and the type of output generated by the four programs to assist in understanding the limitations and advantages of each.

Hibernating mammals display novel phenotypes but do not have well-characterized nucleic acid and protein sequences. These animals can survive up to six months of near-freezing body temperatures with little or no food by lowering their basic metabolic requirements. During deep hibernation, body temperatures are maintained at 4−8 °C, heart rates are 5−10 beats per minute (bpm) versus 200−300 bpm in an active animal, and oxygen consumption is 2% of the active state.[5] The identification of proteins shown to be differentially expressed during hibernation in ground squirrels is dependent on homology with nucleic acid and protein sequences from other mammals.[6−8] In this paper, we describe a strategy for protein identification using multiple software packages and manual inspection of all MS data. We found that femtomole amounts of protein (estimated) from two dimensional polyacrylamide gel electrophoresis (2D-PAGE) gel spots provide sufficient material for the successful identification of proteins from a nonmodel organism that lacks a completely sequenced genome.

Experimental Procedures

Animals. Animal care and use was in accordance with the Institutional Animal Care and Use Committee (IACUC) guidelines. Thirteen-lined ground squirrels (*Spermophilus tridecemlineatus*) were purchased from TLS Research (Hanover Park, Illinois), and delivered within 3−4 days after wild capture during the first week of August. Squirrels were raised on a diet of Purina Rodent Chow #5001 supplemented with sunflower seeds and water ad libitum under 12/12, light/dark conditions from late-March through October. Squirrels were observed daily and maintained at 23 °C from late-March through August, 17 °C in September and 11 °C in October. From November through mid-March the ambient temperature is 5 °C, food is absent and the animals are housed in total darkness. Lengths of individual hibernation bouts were confirmed by the sawdust technique.[9] Hibernating animals were at least in day 2 of a torpor bout when they were sacrificed. Active and hibernating animals were sacrificed by decapitation and organs were surgically removed, placed in cryovials, frozen, and stored in liquid nitrogen. Body temperatures (rectal) were measured at the time of sacrifice.

Protein Preparation and 2D-PAGE. Heart or skeletal muscle proteins were prepared as either membrane preparations or acetone powders.[10] The total solubilized protein concentration was determined using the BCA Protein Assay Reagent (Pierce, Rockford, IL).

All apparati for first-dimension isoelectric focusing (IEF) and vertical SDS-gel electrophoresis were purchased from Amersham Biosciences (Piscataway, NJ). Immobilized pH gradient (IPG) gel rehydration was performed using the Immobiline DryStrip Reswelling tray. IEF was performed using a Pharmacia Multiphor II with an EPS 3501 XL power supply. A Hoefer SE 600 Ruby with the EPS 601 power supply was used to run the vertical second dimension SDS-PAGE gels. Methods for protein separation are reviewed in Gorg et al.[11] The completed gels were fixed for 30 min in 10% methanol:7% acetic acid solution, stained with Sypro Ruby overnight (Bio-Rad, Hercules, CA) and destained for 4 × 30 min each in 10% methanol:7% acetic acid solution.[4] Digital images of the gels were analyzed with Phoretix 2D Expression (version 2004, Nonlinear Dynamics, Newcastle upon Tyne) to determine spot intensity changes between gels from active animals vs. hibernating animals. All protein spots were prepared for mass spectrometric analysis according to the trypsin in-gel protein digestion protocol as described previously.[12]

ESI Mass Spectrometry. Approximately half of the total peptide mixture (pH ≤ 3) was desalted in a glass purification capillary (Proxeon, Denmark) using 20 μm particle size Poros R2 reversed-phase resin (ABI, polystyrene divinylbenzene). Briefly, the peptide mixture was loaded onto the R2 resin in an uncoated purification needle, washed three times with ∼7 μL of 95:5 H_2O:acetonitrile (ACN), 0.5% formic acid, eluted with ∼1.5 μL of 30:70 H_2O:ACN, 0.5% formic acid into a coated nanoelectrospray needle and mounted on a nano-ESI source (Proxeon). Mass spectra were acquired using a QSTAR (ABI, Foster City, CA) QqTOF MS. The ESI voltage was 1000 V, the TOF region acceleration voltage was 4 kV and the injection pulse repetition rate was 6.0 kHz. The $[M + 3H]^{3+}$ monoisotopic peak at 586.9830 m/z and $[M + 2H]^{2+}$ monoisotopic peak at 879.9705 m/z from human renin substrate tetradecapeptide (Sigma-Aldrich, St. Louis, MO) were used for external calibration. Mass spectra were the average of approximately 300 scans collected in positive mode over a 5 min acquisition period.

MALDI Mass Spectrometry. Prior to analysis, the sample was desalted with a Millipore C18 ZipTip using the protocol described by Millipore (Bedford, MA). Approximately 1 μL of desalted sample was spotted on the MALDI target with 1 μL dihydroxybenzoic acid (Agilent Technologies, Palo Alto, CA) as matrix and allowed to air-dry. Spectra were collected on an *o*MALDI-QSTAR Pulsar *i* QqTOF mass spectrometer. External calibration was performed using human angiotensin II (monoisotopic mass $[M+H]^{+1}$ 1046.5417, Sigma-Aldrich, St. Louis, MO) and adrenocorticotropin hormone (ACTH) fragment 18−39 (monoisotopic mass $[M+H]^{+1}$ 2465.1989; Sigma-Aldrich). Laser pulses were generated with a nitrogen laser at 337 nm, 33 μJoules of laser energy using a laser repetition rate of 20 Hz. Mass spectra were the average of approximately 100 laser shots collected in positive mode.

MS Data Analysis. Our method for protein identification using tandem MS data is outlined in the diagram shown in Figure 1.

Software. Three 'mass-based' software programs used for MS/MS interpretation were: Pro ID 1.1/BioAnalyst 1.1.5/Analyst QS 1.1 (ABI, Foster City, CA); TurboSequest (referred to as Sequest) v.27 rev 12/BioWorks 3.1 (Thermo Electron, Waltham, MA) and Mascot v 2.0 (in-house server) (Matrix Science Ltd., Boston, MA). The 'alignment-based' program used for MS/MS interpretation was Pro BLAST 1.1/BioAnalyst 1.1.5/Analyst QS 1.1 (ABI). The Peptide Mass Fingerprint (PMF) search was performed using the in-house Mascot software.

Protein Database. A subset protein database was extracted from NCBI's (http://www.ncbi.nlm.nih.gov) nr (nonredundant) protein database from April 20, 2005. The subset database contained 1,182,676 protein sequences from the sub-kingdom 'Metazoa.' ProID's pre-indexed Interrogator database was

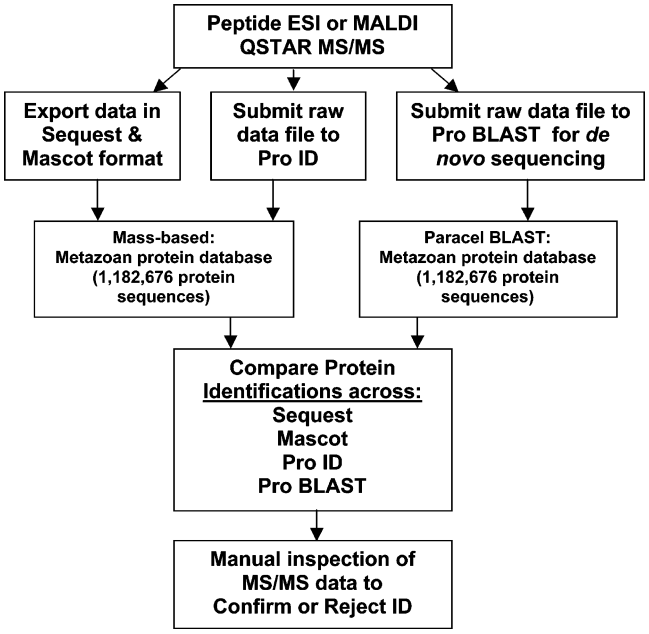

Figure 1. Scheme for the identification of thirteen-lined ground squirrel proteins using mass spectrometric analysis of peptides derived from 2D-PAGE gel spots.

Table 1. Parameters for MS/MS Searches and de Novo Sequencing for Four Software Programs[a]

program	peptide tolerance (Da)	fragment ion tolerance (Da)	missed cleave sites	refs
Pro ID 1.1	1.0	0.8	1	35
Mascot v2.0 (in-house)	1.0	0.8	1	39,40
TurboSequest v.27 rev 12	1.0	0.0	1	41,42
Pro BLAST 1.1	0.08	0.8	N/A	2,13−15

[a] The protein database used for all searches was extracted from NCBI's nonredundant database (April 20, 2005), and included 1 182 676 protein sequences in the "metazoan" taxonomy category. All programs recognized cam-Cys and ox-Met protein modifications.

formatted with the following parameters: custom amino acids were *J* = carbamidomethyl-cysteine (cam-Cys), maximum 3/peptide and O = oxidized methionine (ox-Met), maximum 3/peptide; 4 total modifications allowed per peptide, 1 trypsin missed cleave site. A nonindexed protein database was used for Sequest searches.

MS/MS Search Parameters. See Table 1. Peptide Tolerance: 1.0 Da for ProID, Mascot and Sequest; Fragment ion tolerances: 0.8 Da for ProID and Mascot; 0.0 Da for Sequest; ProID, Mascot, and Sequest: variable modifications were cam-Cys and ox-Met; 1 trypsin missed cleave site. *Note*: Mass tolerance values for Pro ID, which was written for and tested with high accuracy QqTOF data, are typically set between 0.15 and 0.35 Da for both peptide and fragment ion tolerances; however, our Pro ID tolerance parameters were chosen for consistency among programs. Searches with Pro ID peptide and fragment ion tolerances set to 0.35 Da showed all peptides in the report except for one as well as slightly lower confidence values for some peptides (data not shown) when compared to the Pro ID searches with the lower mass tolerances.

PMF Search Parameters. The Bayesian Peptide Reconstruct tool in BioAnalyst 1.1 software was used to create a peak list for a PMF search for sample Ht3. The mass tolerance for ions in a charge series was 0.2 Da and the S/N parameter was 3 for the reconstruction. A list of 91 values was generated from which 38 peaks were manually selected for the PMF based on signal intensity above the noise level (S/N approximately 3, as judged by visual inspection) and the presence of an isotope series containing at least 2 peaks. The trypsin autolysis peptide peak at 842.5 Da was omitted from the search query. Search parameters were as follows: 100 ppm peptide mass tolerance; 1 trypsin missed cleave site; variable modifications: ox-Met and deamidation of Gln and Asn; fixed modification: cam-Cys.

Homology-Based Searches. See Table 1. Pro BLAST first initiates the BioAnalyst de novo sequencing program and then invokes the local Paracel MS BLAST server. The peptide mass tolerance for de novo sequence candidates was 0.08 Da. Custom amino acids assigned prior to de novo sequences (in the BioAnalyst data dictionary) were *J* = cam-Cys and O = ox-Met. Any assignments to modified amino acids were automatically converted back to the unmodified amino acid code before alignment. Ten de novo-sequence candidate peptides per MS/MS scan were submitted to the Paracel BLAST server. The expected number of unique peptides was set to one.

Algorithm Overview and Scoring. Mass-Based MS/MS Search Programs. The three programs, Sequest, Mascot, and Pro ID, share the fundamental method by which database searching is performed, but they differ in the method for producing a score—a number which imparts a level of confidence in the peptide hit and protein match. Each program considers as candidates only those peptides whose theoretical and experimental masses fall within the specified mass tolerance set in the search parameters. For each candidate, the programs look for matches between theoretical and experimental peptide fragment ions. The scoring algorithms are unique but share some fundamental features. Scores are generated based on the number and type of fragment ion matches, fragment ion intensities and other parameters. Differences among programs in the assignment of a level of confidence in a peptide match reflect the differences in scoring algorithms. Other database searching programs are available but were not utilized for this project.

Homology-Based Search Programs. Pro BLAST, which incorporates MS BLAST,[2] is a homology searching tool optimized for aligning short amino acids strings typically obtained from de novo interpretation of tandem mass spectral data to protein sequences in a database (public or in-house) using well-established amino acid sequence alignment rules.[13] De novo sequences are generated by first calculating a mass difference between two product ion peaks and correlating the difference to an amino acid mass, then continuing this process in order to find successive amino acids from a common fragment ion series, mainly b and y ions, from an MS/MS spectrum.[14,15] Pro BLAST was used to submit de novo sequences from BioAnalyst to the local Paracel BLAST server. The protein database used for sequence alignment was the same Metazoan database (see above) used for the mass-based searches. Other programs available for automatic de novo sequence interpretation include (but are not limited to) Lutefisk (http://www.hairyfatguy.com/Lutefisk/), PEAKS (http://www.bioinformaticssolutions.com/), Shrenga (SpectrumMill, Agilent Technologies) and proprietary programs from mass spectrometer vendors. Alignment algorithms optimized for short peptide sequences include FASTS,[16] CIDentify,[17] and MS−Shotgun.[18]

Manual Inspection of Peptide Matches. All peptide matches reported by each program were manually inspected. Manual

inspection of each MS/MS was performed by comparing theoretical fragments ions (singly and doubly charged b, y, a, internal fragments and fragments with neutral losses of ammonia and water) to experimental ions. The nomenclature for peptide fragment ions produced from collision induced dissociation in a mass spectrometer,[3,19] as well as the fundamental interpretation of MS/MS spectra, have been discussed thoroughly.[14,20] We categorized the peptide candidates into three categories: high (H), medium (M), and low (L) quality. High quality matches were considered valid peptide matches based on the presence of a continuous stretch of four or more y or b ions and at most one unassigned ion. Medium quality (M) matches were possible valid matches due to the presence of four or more continuous y or b ions and 2−5 unassigned intense fragment ion peaks. Low quality matches (L) were believed to be false positives based upon the lack of continuous y and/or b ion series and the presence of many (more than 5) unassigned peaks.

Results

The identification of ground squirrel proteins employed state-of-the-art mass spectrometric methods and database searching using one homology-based and three mass-based search programs. A total of seven protein spots were chosen and excised from select 2D-PAGE gels. Two spots from heart membrane preparations (Ht1 and Ht2) and two spots from heart acetone powders (Ht3 and Ht4) were chosen for further analysis. In addition, three 2D-PAGE-separated protein spots from hibernating ground squirrel skeletal muscle (Skm1, Skm2, and Skm3) were also excised and analyzed. After in-gel proteolytic digestion, femtomole amounts (estimated) of peptides were used for protein identification. We employed nanospray infusion on 6 spots (~5 nL/min infusion rate) and MALD ionization on 1 spot of desalted peptide mixtures in conjunction with QqTOF peptide MS/MS acquisition for protein ID.

MS/MS Search Results. Search results from Mascot, Sequest, Pro ID, and Pro Blast as well as a qualitative assessment of each peptide candidate (based on manual validation) are tabulated in Tables 2 and 3. A comparison of select peptide sequences, protein hits [listed by GenBank Identifier (gi) accession numbers] and corresponding scores from Mascot, Pro ID, and Sequest are reported in Table 2 from seven samples that were analyzed by tandem mass spectrometry followed by database searching. One dataset (Ht3 with MALDI MS/MS data) could not be analyzed by Pro ID (due to file incompatibility) therefore only Mascot and Sequest results are shown. In most cases, the software programs reported multiple protein hits for a single MS/MS spectrum due to database redundancy and protein homology; however, only one of the matches is reported for simplicity.

Overall, 49 peptides from 49 MS/MS scans were used for the IDs of the seven potential proteins. Seventeen MS/MS spectra could not be searched by all three programs for multiple reasons: (1) the precursor charge state for one peptide searched using Pro ID was determined incorrectly (and could not be corrected manually); (2) query files for individual MS/MS spectra were not generated by BioAnalyst's export scripts (12 occurrences with Sequest, indicated in Table 2 as "no dta" and once with Mascot, marked with "no query"); (3) three peptides (spot Ht3) were collected by MALDI-QqTOF manually where the file structures are not compatible with Pro ID. In some cases, the BioAnalyst data export scripts determined precursor charge states, and thus peptide precursor MW, incorrectly. Peptide MW values were manually corrected for 5 and 4 precursors after export to Sequest and Mascot, respectively.

Seven proteins were identified as candidates with as few as 1 or as many as 16 unique peptides. Manual inspection provided confidence for only five proteins. The proteins identified as β-Crystallin (spot Skm1) and superoxide dismutase (Skm3) could not be validated manually due to a poor quality MS/MS and the presence of only 1 peptide match, respectively. Of the 32 MS/MS datasets that were submitted to all three programs, 18 peptides (56%) were matched by all three programs and 14 peptides (44%) were matched by 1 or 2 of the programs. Of the 18 peptides matched by all 3 programs, 14 were judged to be valid hits after manual inspection (high quality and confidence) and four others were deemed invalid or inconclusive. Of the 17 out of 49 peptide matches that were made from scans submitted to only 1 or 2 software programs, 11 were considered to be valid high quality spectral matches after manual inspection. Altogether, 25 peptides out of 49 total MS/MS spectra were considered high quality matches after manual inspection.

Homology-Based Search Results. The proteins identified by Pro BLAST after automatic de novo sequencing using BioAnalyst were homologous to those identified by Mascot, Sequest, and Pro ID (Table 2) except for spots Skm2 and Skm3, for which no identifications were made. A previous version of Pro BLAST, which was interfaced with the EMBL MS Blast server (http://dove.embl-heidelberg.de/Blast2/msblast.html), produced a hit to creatine kinase (the protein identified by the mass-based search program) for spot Skm2, but the result could not be reproduced with the upgraded software. Eighteen peptide matches were unique to the Pro BLAST search results; of these, 10 were considered high quality matches after manual verification. Overall, 43 peptides were identified by Pro BLAST from the seven samples: 25 peptides with high confidence, 9 with medium confidence and 9 with low confidence from manual inspection of the spectra with the corresponding de novo sequences.

Identification of 2D-PAGE Spots. Ht1-ATP Synthase. Protein sequence coverage reported by all four programs for spot Ht1 is shown in Figure 2A. Pro ID found 13 peptides and Mascot found 14 peptides that matched to human ATP synthase. Sequest found 8 peptides matching ATP synthase, one of which was unique to Sequest but had low quality (peptide 1286.69). In five cases, the data exporting script in BioAnalyst QS (ABI), which format MS/MS data into Sequest search input files, failed to export searchable files for individual MS/MS spectra. One MS/MS spectrum query file for the Mascot search was also not exported by BioAnalyst.

Pro BLAST found 15 peptides that matched to human ATP synthase with a significant protein score of 875. Nine alignments match to peptides identified by the mass-based programs. Four peptide strings were unique to Pro BLAST and were not identified in the other three programs, two of which have high confidence as judged by manual inspection. The high quality MS/MS spectrum for unique peptide 1450.76 shows 10 y ions, 1 b ion and numerous other fragment ions that could be assigned to the de novo sequence LGTAEMSSLLEER which appears valid. Five regions in the protein (see peptides 1315.80, 1552.84, 1574.82, 2119.10, and 2324.23) were matched using all four software packages as shown by the alignment of peptide matches from all programs against human ATP synthase gi|4757810 (Figure 2A).

Table 2. Results from the 'Mass-Based' Search Programs for the Seven Ground Squirrel Protein Spots Excised from a 2D Gel, Digested with Trypsin and Analyzed by Tandem MS

protein spot, ID	ExpPrec MW[a]	MI[b]	peptide	Pro ID			Mascot		Sequest					
				acc no.	conf[c]	score[d]	acc no.	Ions score[e]	acc no.	Xcorr[f]	ions[g]	int[h]	b	y
Ht1	814.45	M	ELIIGDR	gi|4757810	1	29	no query [i]		NF [j]			246	1	4
ATP	875.49	M	QOSLLLR	NF			gi|4757810	39	no dta [k]			125	0	6
synthase	891.56	H	LELAQYR	wr ch [l]			gi|4757810	35	NF[m]			110	2	6
	999.65	H	VLSIGDGIAR	NF			gi|4757810	39	no dta			215	1	5
	1025.66	H	AVDSLVPIGR	gi|4757810	98	22	gi|4757810	36	no dta			54	3	6
	1170.69	H	VVDALGNAIDGK	gi|4757810	99	25	gi|4757810	88	no dta			100	2	8
	1286.69	L	HALIIYDDLSK	NF			NF		gi|4757810	1.56	9/20	17	2	3
	1315.80	H	**TSIAIDTIINQK**	gi|4757810	99	23	gi|4757810	74	gi|4757810	3.07	14/22	36	3	8
	1437.95	L	**GIRPAINVGLSVSR**	gi|4757810	5	9	gi|4757810[k]	32	gi|4757810[m]	1.38	15/52	19	4	3
	1552.84	H	**EAYPGDVFYLHSR**	gi|4757810	99	24	gi|4757810[k]	40	gi|4757810	2.97	20/48	120	1	7
	1574.82	H	**ILGADTSVDLEETGR**	gi|4757810	99	29	gi|4757810	108	gi|4757810	3.55	16/28	42	1	11
	1623.92	H	TGAIVDVPVGEELLGR	gi|4757810	99	27	gi|4757810	69	no dta			109	4	9
	1682.78	L	NVQAEEOVEFSSGLK	gi|4757810	99	19	gi|4757810	44	NF[m]			26	2	5
	2119.11	H	**GOSLNLEPDNVGVVVFGNDK**	gi|4757810	99	20	gi|4757810	43	gi|4757810	2.79	24/76	40	3	9
	2324.23	H	**QGQYSPOAIEEQVAVIYAGVR**	gi|4757810	99	33	gi|4757810	41	gi|4757810	4.71	33/80	94	3	11
	2337.21	L	**EVAAFAQFGSDLDAATQQLLSR**	gi|4757810	79	14	gi|4757810	47	gi|4757810	2.28	17/84	8	1	4
Ht2	814.45	H	ELIIGDR	NF			NF		gi|114402	1.71	9/12	172	1	5
ATP synthase	875.49	H	QOSLLLR	gi|9256947	5	12	gi|4757810[m]	30	no dta			88	0	5
	999.59	H	VLSIGDGIAR	gi|9256947	60	15	gi|4757810	24	no dta			134	2	5
	1025.61	H	**AVDSLVPIGR**	gi|9256947	81	17	gi|4757810	33	gi|114402	1.78	9/18	284	2	6
	1170.64	H	VVDALGNAIDGK	gi|9256947	99	24	gi|4757810	56	no dta			92	3	10
	1286.72	H	**HALIIYDDLSK**	gi|9256947	59	15	gi|4757810	15	gi|114402	1.53	10/20	44	4	8
	1315.70	H	TSIAIDTIINQK	NF			NF		gi|114402	1.99	11/22	88	3	8
	1552.77	M	**EAYPGDVFYLHSR**	gi|9256947	58	15	gi|4757810[m]	25	gi|114402	2.32	20/48	130	1	5
	1623.90	H	**TGAIVDVPVGEELLGR**	gi|9256947	99	22	gi|4757810	46	gi|114402	2.77	15/30	33	4	8
	2119.04	M	GOSLNLEPDNVGVVVFGNDK	gi|9256947	30	10	gi|4757810	17	NF			42	0	5
	2324.08	H	**QGQYSPOAIEEQVAVIYAGVR**	gi|9256947	56	15	gi|4757810	42	gi|114402	3.12	28/80	54	4	9
	2337.18	H	**EVAAFAQFGSDLDAATQQLLSR**	gi|9256947	99	30	gi|4757810	95	gi|114402	3.82	26/84	74	7	12
Ht3	1043.48	H	EAFOLFDR	NA[n]			NF		gi|127149	1.50	9/14	290	4	5
myosin	1395.75	H	ALGQNPTQAEVLR	NA			gi|226007	14	NF			144	6	8
	1500.68	H	DTGTYEDFVEGLR	NA			gi|226007	43	gi|226007	4.30	18/24	239	7	10
Ht4	1254.57	H	**GOGGAODLVSSAK**	gi|4557817	98	16	gi|1519052	38	gi|10280560	1.37	10/24	66	1	8
succ coA tr	1632.74	H	**OVSSYVGENAEFER**	gi|4557817	99	22	gi|1519052	63	gi|10280560	2.94	12/26	26	2	11
Skm1	920.50	M	FSVNLDVK	NF			gi|57580	16	no dta			653	1	6
β-Crystallin	985.53	M	HFSPEELK	gi|57580	93	20	gi|57580	22	no dta			487	5	5
	1087.50	L	QDEHGFISR	NF			gi|57580	22	gi|57580[m]	1.05	9/16	363	1	6
	1164.60	M	**VLGDVIEVHGK**	gi|9716999	32	13	gi|57580	31	gi|57580	1.58	11/20	1085	2	7
	1477.69	L	APSWIDTGLSEOR	NF			gi|57580	4	NF			100	4	1
Skm2	947.53	M	VLTPDIYK	NF			gi|125307	27	NF			174	2	7
creatine	1092.62	H	ITQGQFDER	NF			gi|61882049	46	no dta			229	1	8
kinase	1106.66	L	VPPPLPQFGR	gi|125313	32	14	gi|57537	14	NF			144	1	4
	1230.65	H	**DLFDPIIQDR**	gi|125303	5	10	gi|125307	41	gi|50437	1.50	9/18	267	3	7
	1317.60	H	GQSIDDOIPAQK	NF			gi|125307	46	gi|103901	2.16	11/22	224	1	7
	1379.63	H	LSEOTEQDQQR	NF			gi|61882049	33	NF			72	1	7
	1388.70	M	LFPPSADYPDLR	gi|125313	32	14	NF		no dta			218	0	5
	1506.80	M	GGDDLDPNYVLSSR	gi|125303	32	14	NF		gi|103901	1.95	11/26	68	4	8
	"	L	LSVEALNSLTFEFK	NF			gi|125307	4	NF			68	2	2
	2108.00	L	GTGGVDTAAVADVYDISNIDR	NF			gi|61882049	67	gi|61882049[m]	1.61	9/40	63	4	6
Skm3 superox dism	1409.81	H	**GDVTAQVALQPALK**	gi|51949931	81	19	gi|108408	49	gi|108408	2.61	13/26	56	4	7

ESI−MS/MS data was acquired for samples Ht1, Ht2, Ht4, Skm1, Skm3, Skm3, and MALDI MS/MS data was acquired for sample Ht3. In most cases, the reported experimental peptides were shared by multiple, homologous proteins in the database but only one accession number per search program was reported here for simplicity. The ranges of scores for good quality peptide candidates (as judged by manual inspection) were: ProID: Scores of 10−33 (avg = 21 and average of 80% Confidence); Mascot: 15−104 Ions score (avg = 46); Sequest: 1.4−4.7 Xcorr (avg = 2.6). The peptides in bold font represent identical proteins found by all three programs; 'O' represents oxidized methionine. The b and y ion values represent the number of each of the fragment ion found upon manual inspection of MS/MS scan. [a] Experimental Precursor MW. [b] Manual Inspection 'MI' determined the confidence of the peptide match: H = high, M = medium, L = low. [c] ProID percent Confidence. [d] ProID MS/MS score. [e] Mascot Ions scores. [f] Sequest cross correlation score. [g] Sequest ratio of number of experimental to theoretical fragment ions. [h] Base peak intensity. [i] No data for peptide included in query generated by BioAnalyst. [j] Not Found = Peptide in list not found (typically unrelated peptides with different sequences were found but not reported here). [k] "no dta" = failure of BioAnalyst software to export a Sequest MS/MS search input file. [l] Charge state was determined incorrectly (cannot be corrected in ProID query). [m] Charge state for precursor was manually corrected prior to search, since BioAnalyst software determined charge state incorrectly during data exportation step. [n] Not Applicable: ProID does not search non-IDA MALDI datafiles.

Ht2-ATP Synthase. Spot Ht2 was located immediately adjacent to Ht1 on 2D gels and was also identified as ATP synthase. Protein sequence coverage reported by all four programs is shown in Figure 2B. Pro ID and Mascot matched the same 10 spectra to peptides from ATP synthase. Six Sequest hits matched results from Pro ID and/or Mascot, plus two additional peptides (814.45 and 1315.70) were assigned, each with high confidence after manual inspection. Overall, 6 peptides were found as identical peptide matches across all three mass-based programs.

Table 3. Homology-Based Search Results. Pro BLAST de novo Search Results Show High Scoring Pairs (HSP Score), Regions of Sequence Similarity Between de novo Sequences and Protein Database Entries for Five Proteins[a]

Protein Spot, ID	Acc #	Total Score	HSP Score	Exp PrecMW	*De novo* sequence Subject Peptide	% Pos	Int, cnts	MI	b	y
Ht1 ATP synthase	gi: 30583257	875	32	562.29	**HPPGR** RRPPGR	83	34	L	0	2
			45	814.46	ELLLGDR RELIIGDR	100	246	M	0	3
			41	875.57	ZMSLLLR RQMSLLLR	100	125	M	0	4
			80	1170.69	VVDALGNALDGK RVVDALGNAIDGK	100	100	H	2	8
			33	1286.69	HAAAZMNLL(303.3) AAKMN	100	17	L	2	4
			62	1315.80	TSLALDTLLNGAK KTSIAIDTIIN	100	36	H	4	8
			81	1450.76	**LGTAEMSSLLEER** GTAEMSSILEER	100	42	H	1	10
			38	1499.48	(912.1)GDGLAR GDGIAR	100	70	L	0	3
			50	1533.78	(678.2)DSLVPLGR DSLVPIGR	100	32	H	0	5
			71	1552.84	TVYPGDVMYLHSR YPGDVFYLHSR	90	121	H	2	7
			99	1574.82	LLGADTSVDLEETGR RILGADTSVDLEETGR	100	42	H	2	11
			43	1623.92	LAGTVDVPVGMGEVHK VDVPVG	100	109	H	5	9
			72	1682.9	NVZVETMVEMSSGLK RNVQAEEMVEFSSGLK	81	26	L	2	5
			38	2119.1	(1086.6)VGVVVMGTZK VGVVVFGNDK	70	40	M	0	6
			90	2324.23	TPGLMEYALEEZVAVLYAGVR AIEEQVAVIYAGVR	100	94	H	6	11
Ht2 ATP synthase	gi: 4757810	627	45	814.46	ELLLGDR RELIIGDR	100	172	H	1	5
			41	875.60	ZMSLLLR RQMSLLLR	100	80	H	1	5
			45	891.51	**LELAZYR** KLELAQYR	100	88	H	1	6
			41	1025.61	AVDSLVLPGR KAVDSLV	100	284	H	3	7
			80	1170.60	VVDALGNALDGK RVVDALGNAIDGK	100	92	H	3	10
			42	1286.71	HALLLYNAASSK KHALIIYD	100	44	M	3	6
			57	1326.80	**YLCLYVALGZK** CIYVAIGQK	100	48	L	0	3
			36	1437.86	**LGPVZLNVGLDASR** INVGLSVSR	77	129	M	3	5
			81	1450.71	**LGTAEMSSLLEER** GTAEMSSILEER	100	37	H	0	10
			64	1544.83	**LVNGSZVAVLYAGVR** QVAVIYAGVR	100	109	H	0	7
			40	1623.9	LGATVDVPGZLGCVAPK GAIVDVP	85	33	H	2	6
			55	2359.16	**LGLVTGWWPEDTSVDLEEADK** DTSVDLEE	100	39	H	3	11
Ht3 myosin	gi: 738460	183	86	1500.68	DTGTYEDMVEGLR KDTGTYEDFVEGLR	92	239	H	6	10
			58	1639.89	**AAAAPAPAPAPELLGVPK** AAAPAAAPEPE	81	668	H	9	8
			39	1639.89	**AAAAPAPAPAPELDGPPK** AAPKAAPAP	77	668	H	9	8
Ht4 succ coA transf	gi: 27574274	339	33	1030.62	**SSNAGVVGVNK** SNNAGV	83	107	L	2	3
			52	1196.46	**DTDCDMAVSPK** TGCDFAVSPK	80	79	H	1	6
			49	1196.46	**SEDCDMAVSPK** KSTGCDFAVSPK	75	79	H	1	6
			66	1254.57	GCTGAMDLVSSAK GAMDLVSSAK	100	66	H	0	7
			59	1254.57	MGNAMDLVACAK MGGAMDLVSSAK	75	43	M	1	6
			46	1632.74	**MVSSYVWNAELYR** RMISSYV	100	26	H	3	10
			34	1632.74	**VMSSYVWNAEMER** NAEFER	83	26	H	3	10
Skm1 β-crystallin	gi: 265053	259	62	952.56	**GDVLEVHGK** GDVIEVHGK	100	300	M	2	5
			49	985.53	HMSPEELK KHFSPEELK	88	487	M	2	5
			34	998.56	**BLTAVDAAPNK** AVTAAPKK	75	421	L	1	3
			49	1164.68	BVLGDVVTVHVK KVLGDVIEVHGK	83	1085	M	3	3
			33	1509.72	**BMTSPGLDASLTDMR** IDTGLSEMR	66	84	L	5	9
			32	1887.58	**RPPSYNGHVLCGGNMDR** RPPSF	100	48	L	2	3

[a] The scores for each HSP are reported as well as the total score, which is the sum of the individual scores for each HSP. De novo sequences for all spectra were generated in automatic mode after the newest version of ABI's blast script was installed, "Pro BLAST." The table includes the protein name and accession number from the highest scoring protein (excluding trypsin), the experimental precursor molecular weight (Exp PrecMW), as determined by ABI software, subject peptide, % positive (% pos), intensity of the tandem mass spec in counts (Int, cnts), and the numbers of experimental b and y ions that match subject peptide fragments. The BLAST program aligns the subject peptide, which is the peptide from the protein in the database, to a region of the full de novo sequence derived from ABI. All Pro BLAST hits were manually inspected and labeled "MI" in the table based upon the quality and the presence of candidate b and y ions from the full de novo sequence and the outcome of the search. Manual Inspection 'MI' determined the confidence of the peptide match: H = high, M = medium, L = low. Sequences in bold font are unique to Pro BLAST and were not found in the mass-based searches. Matches indicated by the letter 'Z' represent either amino acids Q or K.

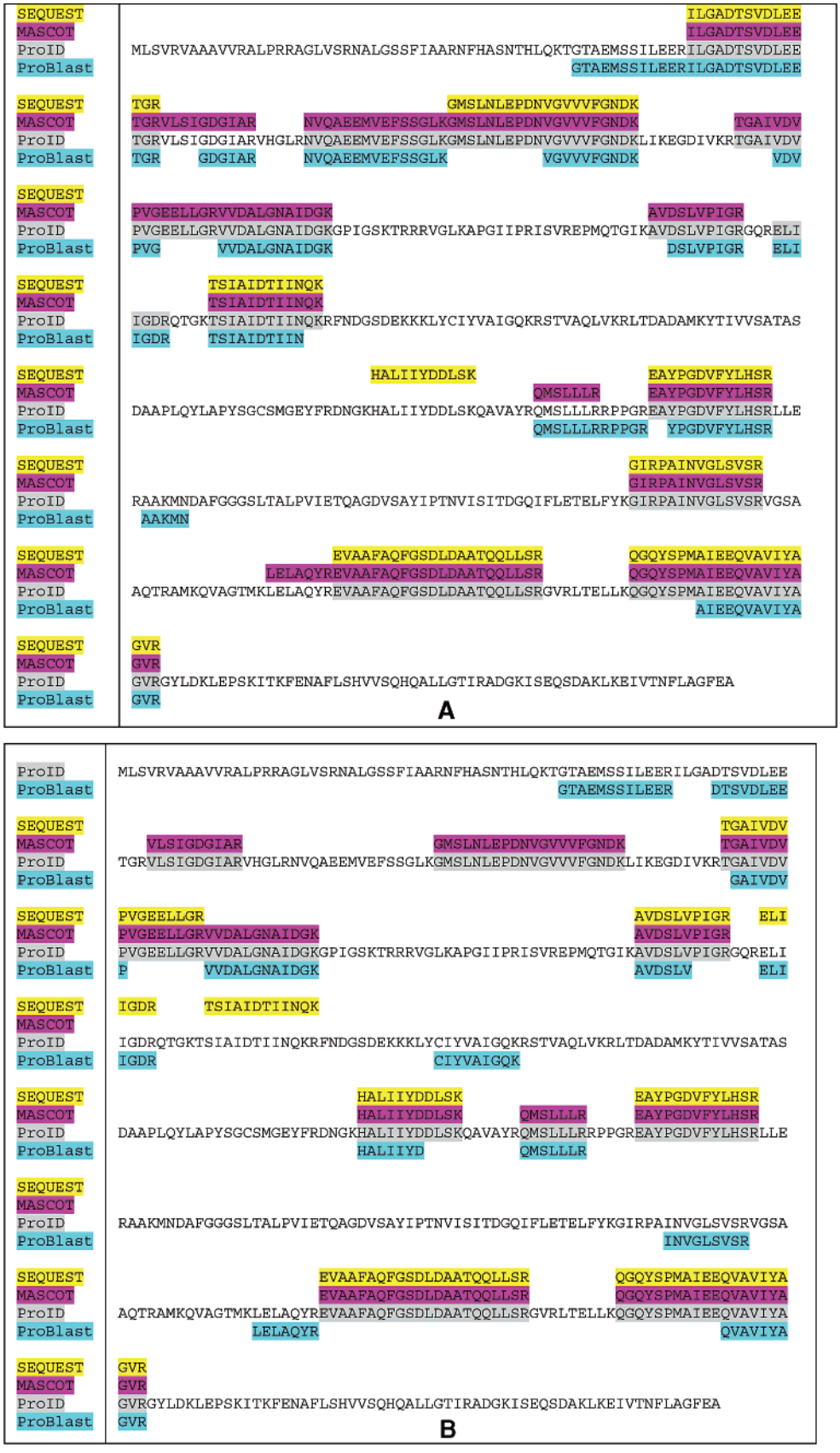

Figure 2. Protein sequence coverage reported by four software programs: Sequest, Mascot, ProID and Pro BLAST for two distinct 2D-PAGE gel spots. (*A*) Peptides matched to *Homo sapiens* ATP synthase, H^+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1 [gi:4757810] from ground squirrel heart protein Ht1. (*B*) Peptides matched to *Homo sapiens* ATP synthase, H^+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1 [gi:4757810] from ground squirrel heart protein Ht2.

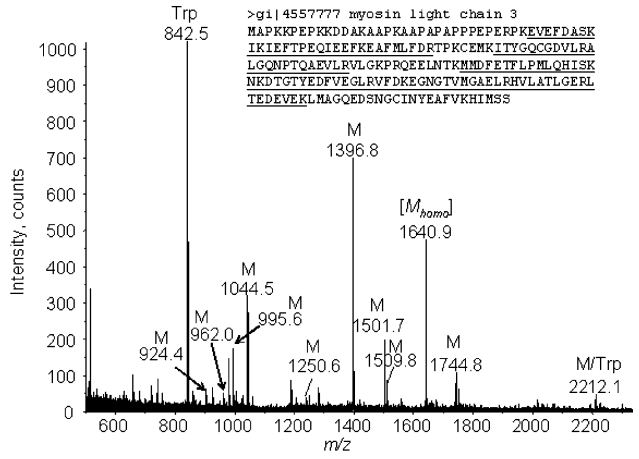

Figure 3. Full Scan MALDI QqTOF MS of ground squirrel protein Ht3. A Mascot PMF search identified the protein as ventricular myosin light chain 3 based on homology to the human protein (gi:4557777). Abbreviations: M, myosin peptide; M_{homo}, peptide with homology to a myosin peptide, as shown from Pro BLAST search (see Table 3); M/Trp, myosin or trypsin peptide (indistinguishable based on *m*/*z* only); Trp, trypsin autolytic peptide. Fourteen mass values (11 are shown) matched to myosin which represented 60% sequence coverage and the mass error falls within the range of −9.9−8.6 ppm. Underlined amino acids show the combined sequence coverage from ground squirrel myosin light chain 3.

Pro BLAST results show matches to 12 ATP synthase sequence strings, 6 of which are unique to Pro BLAST. Of the 6 unique peptides, 4 have very high confidence after manual inspection. Four peptides in the protein were found using all four software packages. Discrete spots Ht1 and Ht2 were both identified as ATP synthase, H^+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle. Despite the lack of an extensive sequence database for ground squirrels, we were able to identify these proteins with high specificity using multiple software packages due to highly conserved regions in the protein. Migration of the proteins to two different isoelectric points could be explained by differences in post-translational modifications. From our gel analysis, Ht1 has an apparent p*I* of 8.9 and MW of 60.6 kD, whereas Ht2 has an apparent p*I* of 8.7 and a MW of 60.6 kD.

Ht3-Myosin. MALDI QqTOF MS data was acquired for spot Ht3. Figure 3 shows a peptide mass fingerprint (PMF) search for spot Ht3 using Mascot that matched 14 ground squirrel peptides (11 peptides are shown) to human ventricular myosin light chain 3 (gi|4557777), out of a total of 38 *m*/*z* values submitted in the query list. The Mascot match to myosin has a score of 125, which falls above the 95% probability level according to Mascot's Probability-based MOWSE (MOlecular Weight SEarch) scoring scheme. With external calibration the accuracy of the 14 myosin peptides ranged from −9.9 to 8.6 ppm. Peptide 2210.08 matches the mass value of a trypsin autolysis peptide as well as a myosin peptide to within 9 ppm, and we are unable to distinguish the true peptide identity without tandem MS data. Twelve myosin peptides fall within the list of the top 19 most intense peaks in the TOF spectrum. The amino acid sequence coverage is 60% of the total sequence when *m*/*z* 2210.08 is counted as a myosin peptide (see Figure 3).

MALDI-QqTOF MS/MS scans were acquired on the four most intense myosin peaks, excluding the trypsin $[M+H]^{+1}_{mono}$ autolysis peak at 842.5 *m*/*z*. Since our version of Pro ID cannot process MALDI MS/MS data due to file incompatibility, Table 2 does not contain results from Pro ID for sample Ht3. Mascot reported two peptides, both of which are common to human myosin light chain 1 and 3, which share 97% homology. Our PMF or MS/MS data does not allow us to distinguish between myosin light chains 1 and 3. Sequest identified two peptides, one in common with Mascot results. The peptide hit unique to the Sequest search (peptide 1043.48) is included in the PMF search results. Neither Mascot nor Sequest produced a match for peptide 1639.89, a spectrum of exceptionally high data quality.

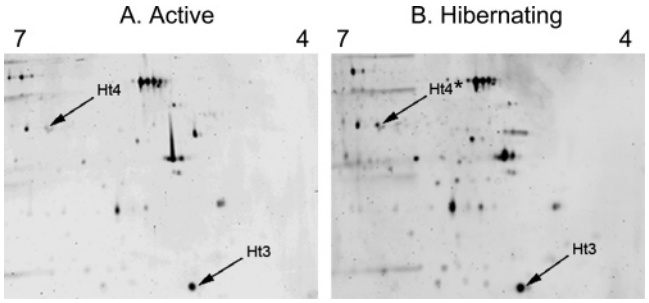

Figure 4. Two-dimensional polyacrylamide gel electrophoresis of thirteen-lined ground squirrel heart protein from (A) active T_b = 37 °C and (B) hibernating T_b = 5 °C animals. Ht3 is ventricular myosin light chain (nonsignificant change, active, n = 5, hibernating, n = 4) and Ht4 is succinyl CoA transferase (6-fold increase; ∗, $p < 0.01$, active, n = 5; hibernating, n = 4). The numbers 7 and 4 denote p*I* range in the first dimension.

Pro BLAST matched three de novo sequence strings to human myosin (Table 3). The alignment of peptide 1639.89 to a myosin peptide was unique to Pro BLAST. Two high-scoring pairs (HSP's) were reported from de novo sequences from peptide 1639.89. These two subject peptides from the HSP's for peptide 1639.89 are discrete, nonoverlapping sequence strings from myosin that share homology but are not repeats. The assignment of >1 HSP to a single MS/MS spectrum falsely elevated the total score.

Other Ground Squirrel Proteins. Other thirteen-lined ground squirrel proteins (see Tables 2 and 3) include succinyl Coenzyme A (CoA) transferase (Ht4) in the heart; β-cystallin (Skm1), creatine kinase (Skm2), and superoxide dismutase (Skm3) in skeletal muscle. Matches to β-cystallin and superoxide dismutase (SOD) could not be manually validated after visual inspection due to medium and low quality spectra (β-cystallin) and the presence of only 1 peptide match (SOD). Succinyl CoA transferase (SCOT), a valid protein identification based on two high quality MS/MS, showed significantly higher protein levels in the heart of winter hibernators compared to summer active animals. Figure 4 shows two-dimensional electrophoretic gels containing protein from the heart of both active and hibernating ground squirrels. Spot analysis of multiple two-dimensional gels using Phoretix 2D Expression (version 2004) showed that ventricular myosin light chain 3 (spot Ht3) had a nonsignificant increase of 42% from active to hibernating animals. However, SCOT showed a significant increase of 6-fold in hibernators over active animals. SCOT is an extremely important enzyme because it catalyzes the rate-limiting step in the metabolism of ketone bodies, an important fuel source during hibernation resulting from fatty acid breakdown.

During hibernation, fatty acids from white adipose tissue (WAT) become the dominant fuel source. Ketone bodies

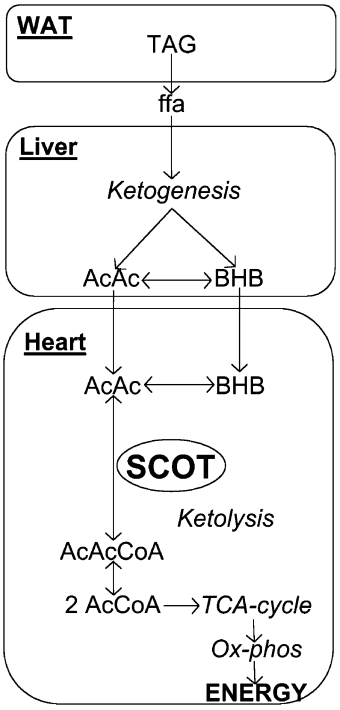

Figure 5. Model showing the role of the up-regulated heart protein succinyl-CoA transferase (SCOT) in the metabolism of ketone bodies beta-hydroxybutyrate (BHB) and acetoacetate (AcAc). Abbreviations: CoA, coenzyme A; ffa, free fatty acid; TCA cycle, tricarboxylic acid cycle; TAG, triacylglycerol; Ox-phos, oxidative phosphorylation.

(acetoacetate, AcAc; β-hydroxybutyrate, BHB; and acetone) are produced in the liver during the oxidation of fatty acids and exported to peripheral tissues as an alternative fuel source during starvation, diabetes and hibernation.[21] It has been previously shown that there is a 15-fold increase in circulating ketones in hibernating animals.[22] Ketones, unlike larger fatty acids from which they are derived, are short 4-carbon carboxylic acids that easily pass in an unbound form from hepatocytes to the circulation and can even be transported across the blood brain barrier. SCOT is a mitochondrial enzyme responsible for transferring CoA from succinyl CoA to acetoacetate forming acetoacetyl-CoA and releasing succinate.[23] Acetoacetyl-CoA can be further metabolized into 2 molecules of acetyl CoA that can enter the tricarboxylic acid (TCA) cycle for energy production. We have proposed a model based on our results showing ketone body utilization in the heart during hibernation (Figure 5). The differential expression of several genes regulating fuel selection has been recently shown by digital transcriptome analysis of heart tissue from active and hibernating animals.[24]

Discussion

This paper presents a mass spectrometric-based proteomics approach for the identification of proteins from eukaryotic organisms lacking nucleic acid and/or protein sequence databases. Identification of individual proteins isolated by 2D-PAGE can be accurately analyzed using mass spectrometry and database searching because MS can provide amino acid sequence data from peptides at the nanogram level.[25,26] In the case of model organisms, the likelihood for success is relatively good since the sequences of proteins under investigation are usually present in the databases used for peptide MS/MS interpretation. However, many of the important nonmodel species that show novel phenotypes have little or no sequence information,[2] thus confounding database searching and results interpretation. Some nonmodel organisms share enough protein homology with sequenced organisms that small peptide sequences may be identical and the 'mass-based' software programs can report a valid ID.

To achieve the goal of identifying gel separated proteins using tandem MS, we used multiple software programs available in our core facility for database searching of tandem MS data. Since the genome of our species of interest is not completely sequenced, and since all software programs have discrete, unique features, we expected the use of multiple programs could improve our ability to identify proteins and help to improve our confidence level for protein matches. Upon viewing the results, we noted interesting differences and similarities among each software package and we reported differences in peptide hits and scores. The use of a nonmodel organism shows that despite the lack of protein sequences in the public databases for *Spermophilus tridecemlineatus,* and despite the differences in software programs, protein identification was successful.

No universal method exists for imparting confidence on a protein hit found upon database searching with MS data; however, guidelines are being proposed for a standardized method of reporting peptide matches to precursor ions and subsequent protein identifications, and for reporting the parameters used during data preprocessing and database searching.[27] Even before formal guidelines are established, one must be aware of the level of accuracy associated with peptide matches and protein IDs from database search results. Researchers with little or no MS or peptide tandem MS software background that are dealing with results interpretation must realize the discretion one should use in reporting protein coverage and homology. An understanding of the parameters that affect the processing and outcome of a database search, as well as the degree of variation in output as a function of software program, are crucial for researchers using MS for protein identification. We have provided references to the programs used for our searches and have listed some pros and cons for these programs, based upon our experience with each software program. Our results show the type of variability inherent in search results from multiple programs, and that the process of exporting raw data to multiple programs can differ.

A comparison of results derived from the search programs shows various differences. The rank order of scores is not consistent among the four programs, although the lowest scoring peptides for Ht1 and Ht2 are common to all three mass-based search programs. The inherent differences in scoring algorithms among programs most likely account for many of the differences in rank order of scores and the absence of peptides in the results lists from various programs. In multiple cases, one program produced a high score for a particular peptide that was missing from other search results (see peptide 1092.62 from Skm2, for example). Some of the peptides that were absent from one or more of the reports were categorized as accurate matches based on manual inspection. Some peptide matches and protein identifications would be missed altogether if simple cutoff rules based upon software manufacturer's recommendations were applied.

We noted a striking consistency among all programs with one particular dataset consisting of multiple MS/MS spectra.

Spot Skm1 was identified as β-Crystallin by all four programs, with a range of two to five peptide hits for the mass-based programs and six peptide hits for the homology-based program. However, manual inspection of all MS/MS spectrum shows low and medium quality MS/MS, and we therefore do not have confidence in the protein ID.

Overall, the scores for the peptide matches from each of the mass-based programs are not in the range typically designated as "high scoring," which guides the researcher to matches that are likely to be valid. Only 4 out of 31 individual peptide Scores for Pro ID are >28, which is typical for a significant match, although 14 out of 31 have % Confidence levels of 98−99. For Mascot, 8 out of 42 total matches have Ions Scores >49, which typically indicates extensive homology. For Sequest, 15 of 27 total peptides show Xcorr scores >2.0, a 'general' cutoff for a good-quality match, although significant scores have been reported to increase with increasing charge state[28] as follows: ≥1.9 for +1, ≥2.2 for +2 and ≥3.75 for +3 precursors. The fact that numerous scores are low among all programs is most likely explained by the fact that many MS/MS spectra (about half) have low S/N, which was apparent upon visual inspection of the data. After manual validation of the MS/MS data, we categorized numerous matches as valid hits despite the fact that scores from some of the programs were low.

The peptide hits we validated after manual inspection were almost all assigned with success by each program. We showed that of 18 spectra that were searchable by all three mass-based programs, and judged as valid (high quality spectra and excellent match to pertinent peptide fragment ions) after manual inspection, contained only 3 peptides that were not found (NF) by one or more programs. Matches judged as "low" or "medium" quality showed 10/14 results that produced a peptide hit by only one program.

The use of multiple programs for MS/MS interpretation along with manual validation can provide identification of a greater number of peptides per protein as compared to the use of a single program. This is important when the goal of a project is to maximize protein coverage because protein ID relies on sequences from other species due to the lack of a sufficient database for the organism of interest. Homology-based searches have the potential to find unexpected modified peptides, peptides generated from protease nonspecificity and genetic variants for both sequenced and non- or partially sequenced genomes. The use of both mass-based and homology-based software may provide expanded sequence coverage when compared to the use of only one method. For our project, the percentage of coverage based on amino acid sequence for most proteins identified by MS/MS was increased as a result of the use of multiple programs. The commercially available program Scaffold (http://www.proteomesoftware.com/) can be used to combine results from Mascot and Sequest and statistically analyze results further with proven algorithms.[29,30] Scaffold also shows that identifications resulting from more than one search increases protein coverage.

Large volumes of MS/MS data typical of 1D or 2D LC−MS/MS analyses can produce hundreds to thousands of spectra, which makes the task of manual inspection or validation challenging and time-consuming. The need for automated, computational validation methods has driven the creation of programs such as Qscore,[31] Peptide Prophet,[32] and others[33] for a second-tiered level of data evaluation and methods such as PROT_PROBE[34] and Protein Prophet[29] for validation and grouping of protein hits. The programs incorporate probability-based statistics into their algorithms in order to systematically eliminate or reduce false positive hits and produce a nonredundant list of proteins, or groups of related proteins, and provide a way to compare large datasets.

Pros and Cons of Software Programs. ProID. Pros: (1) Modification tolerant database search algorithm is fast due to extensive and unique method of indexing both peptide and fragment ion masses;[35] (2) numerous differential modifications of arbitrary or known mass can be chosen without compromising score; (3) high accuracy for peptide and fragment ion masses for QqTOF data adds discrimination; and (4) convenient interface between search results and raw data, which allows for detailed and thorough manual inspection of spectra. **Cons:** (1) Protein database is pre-indexed with trypsin as the only option for proteolytic enzyme; (2) new protein sequences cannot be added to existing indexed databases; and (3) only datafiles from ABI instruments can be searched.

Sequest. Pros: (1) FASTA-formatted protein database can be used for search, to which sequences can be added easily; (2) custom proteolytic enzymes and 'no enzyme' features can be used; (3) cluster version is available, which substantially decreases search time; (4) various add-on programs and software upgrades have been written to expand upon data organization and to provide additional peptide validation tools;[31,36,37] and (5) input files are simple text-based files. **Cons:** (1) Scoring algorithm was developed using low resolution data; therefore, high-accuracy fragment ion m/z values do not add discriminating power with the current version; (2) cross correlation (Xcorr) scores can be falsely low for short peptides and falsely elevated for long peptides; and (3) searches against large databases (e.g., NCBI nr) can produce false negative and positive results.

Mascot. Pros: (1) FASTA-formatted protein database can be used for search, to which sequences can be added easily; (2) taxonomy can be specified during database search without the need for individual database formatting (for some databases); (3) wide protease specificity option (customizable with in-house version); and (4) scores from different probability-based searches, such as Mascot's Sequence Tag, Peptide Mass Fingerprint, MS/MS ions, and from different databases can be compared to one another rationally. **Cons:** (1) Multiple (greater than 3) differential modifications applied to a search greatly reduce the discriminating power of the program; (2) large proteins and proteins with extended repeats may yield false positive scores and small proteins may yield false negative scores due to probability-based scoring scheme; and (3) nonindependent results, such as duplicate datasets for a single peptide (i.e., tandem MS twice or tandem MS on a modified plus a nonmodified peptide) violate the probability-based scoring scheme, which is based on the assumption that all measurements are independent.

Pro BLAST. Pros: (1) False positives are uncommon;[38] (2) local Paracel MS BLAST server provides means to search proprietary and in-house databases while the interface with EMBL provides means for searching public databases;[13] (3) user-specified modifications can be considered during the de novo sequencing step; 4) partial inaccuracies in de novo sequences are well tolerated. **Cons:** (1) Falsely high scores are generated for proteins with multiple repeats; (2) the program errs on the side of producing false negatives; (3) de novo sequencing is optimized for peptides derived from trypsin only; and (4) Paracel (local) MS BLAST can only be used with de novo

sequences generated from BioAnalyst as opposed to sequences determined by manual de novo methods, for instance.

Conclusion. One of the great challenges in biology is the ability to determine the molecular basis of phenotypic change. As the final product of gene expression, the level and activity of proteins are major determinants of phenotype. Therefore, the identification of differentially expressed proteins based on MS-derived sequence of short peptides is a powerful tool. In this paper, we have gone one step further by analyzing peptides as small as 6−7 amino acids to identify proteins from the thirteen-lined ground squirrel; a hibernating mammal with a recently initiated genome sequence project (www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=genomeprj). The power of the approach described in this paper is its widespread applicability in identifying differentially expressed proteins in previously understudied organisms that display novel phenotypes. Our results show the importance of utilizing multiple software packages because the peptide sequence string and the occurrence of a peptide hit varied among programs. In general, there is a high level of confidence in most of the protein hits found from all programs according to conclusions made after close manual inspection of all MS/MS spectra. Our results give the reader a sense of confidence relative to score and program, which is helpful to researchers unaccustomed to protein MS data interpretation and unaware of the variability in software algorithms.

Acknowledgment. The authors wish to thank C. Walker for his work with the 2D-PAGE experiments. Digital analysis of 2D-PAGE data was performed with the help of the Visualization and Digital Imaging Lab at UM-Duluth. Mass spectrometric analysis of tryptic-digested proteins was performed at UM Center for Mass Spectrometry and Proteomics at UM-Twin Cities. Bioinformatics support was provided by Zheng Jin Tu from the Minnesota Supercomputing Center. The paper was improved by the comments of L. Anderson, T. Griffin, B. Witthuhn, C. Hunter, D. Tabb, C. Reilly and members of the Andrews laboratory. This work was supported by a Grant-in-Aid from the University of Minnesota and NIH Grant HL-081100 to M.T.A.

References

(1) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−4399.
(2) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem* **2001**, *73*, 1917−1926.
(3) Biemann, K. *Annu. Rev. Biochem.* **1992**, *61*, 977−1010.
(4) Berggren, K.; Chernokalskaya, E.; Steinberg, T. H.; Kemper, C.; Lopez, M. F.; Diwu, Z.; Haugland, R. P.; Patton, W. F. *Electrophoresis* **2000**, *21*, 2509−21.
(5) Carey, H. V.; Andrews, M. T.; Martin, S. L. *Physiol. Rev.* **2003**, *83*, 1153−1181.
(6) Buck, M. J.; Squire, T. L.; Andrews, M. T. *Physiol. Genomics* **2002**, *8*, 5−13.
(7) Squire, T. L.; Lowe, M. E.; Bauer, V. W.; Andrews, M. T. *Physiol. Genomics* **2003**, *16*, 131−140.
(8) Epperson, L. E.; Dahl, T. A.; Martin, S. L. *Mol. Cell Proteomics* **2004**, *3*, 920−933.
(9) Pengelley, E. T.; Fisher, K. C. *Can. J. Zool.* **1961**, *39*, 105−120.
(10) Garfinkel, A. S.; Schotz, M. C. *J. Lipid Res.* **1972**, *13*, 63−68.
(11) Gorg, A.; Obermaier, C.; Boguth, G.; Harder, A.; Scheibe, B.; Wildgruber, R.; Weiss, W. *Electrophoresis* **2000**, *21*, 1037−1053.
(12) Jensen, O. N.; Wilm, M.; Shevchenko, A.; Mann, M. *Methods Mol. Biol.* **1999**, *112*, 513−530.
(13) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(14) Hunt, D. F.; Yates, J. R., 3rd; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233−6237.
(15) Shevchenko, A.; Chernushevich, I.; Ens, W.; Standing, K. G.; Thomson, B.; Wilm, M.; Mann, M. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1015−1024.
(16) Mackey, A. J.; Haystead, T. A.; Pearson, W. R. *Mol. Cell Proteomics* **2002**, *1*, 139−147.
(17) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067−1075.
(18) Huang, L.; Jacob, R. J.; Pegg, S. C.; Baldwin, M. A.; Wang, C. C.; Burlingame, A. L.; Babbitt, P. C. *J. Biol. Chem.* **2001**, *276*, 28327−28339.
(19) Roepstorff, P.; Fohlman, J. *Biomed. Mass Spectrom.* **1984**, *11*, 601.
(20) Papayannopoulos, I. *Mass Spectrometry Rev.* **1995**, *14*, 49−73.
(21) Laffel, L. *Diabetes Metab. Res. Rev.* **1999**, *15*, 412−426.
(22) Krilowicz, B. L. *Am. J. Physiol.* **1985**, *249*, R462−470.
(23) Bateman, K. S.; Brownie, E. R.; Wolodko, W. T.; Fraser, M. E. *Biochemistry* **2002**, *41*, 14455−14462.
(24) Brauch, K. M.; Dhruv, N. D.; Hanse, E. A.; Andrews, M. T. *Physiol. Genomics* **2005**, *23*, 227−234.
(25) Patterson, S. D.; Aebersold, R. *Electrophoresis* **1995**, *16*, 1791−1814.
(26) Yates, J. R., 3rd *J. Mass Spectrom.* **1998**, *33*, 1−19.
(27) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. *Mol. Cell Proteomics* **2004**, *3*, 531−533.
(28) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd *Nat. Biotechnol.* **1999**, *17*, 676−682.
(29) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646−4658.
(30) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466−1467.
(31) Moore, R. E.; Young, M. K.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 378−386.
(32) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.
(33) Fenyo, D.; Beavis, R. C. *Anal. Chem*. **2003**, *75*, 768−774.
(34) Sadygov, R. G.; Liu, H.; Yates, J. R. *Anal. Chem.* **2004**, *76*, 1664−1671.
(35) Tang, W. H.; Halpern, B. R.; I. V., S.; Seymour, S. L.; Keating, S. P.; Loboda, A.; Alpesh, A. A.; Schaeffer, D. A.; Nuwaysir, L. M. *Anal. Chem.* **2005**, ASAP.
(36) Sun, W.; Li, F.; Wang, J.; Zheng, D.; Gao, Y. *Mol. Cell Proteomics* **2004**, *3*, 1194−1199.
(37) Tabb, D. L.; McDonald, W. H.; Yates, J. R., 3rd *J. Proteome Res.* **2002**, *1*, 21−26.
(38) Habermann, B.; Oegema, J.; Sunyaev, S.; Shevchenko, A. *Mol. Cell Proteomics* **2004**, *3*, 238−249.
(39) Pappin, D. J.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993**, *3*, 327−332.
(40) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.
(41) Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426−1436.
(42) Eng, J. K.; McCormack, A. L.; Yates, J. R., 3rd *J. Am. Soc. Mass Spectom.* **1994**, *5*, 976−989.

PR050306A