Short Communication

Development of a host blood meal database: *de novo* sequencing of hemoglobin from nine small mammals using mass spectrometry

Ünige A. Laskay¹, Jennifer Burg¹, Erin J. Kaleta¹,
Inger-Marie E. Vilcins²*, Sam R. Telford III³,
Alan G. Barbour² and Vicki H. Wysocki¹,*

¹The University of Arizona Department of Chemistry and Biochemistry, Tucson, AZ 85721, USA
²Departments of Microbiology and Molecular Genetics and Medicine, University of California Irvine, Irvine, CA 92697, USA
³Tufts University, Cummings School of Veterinary Medicine, Division of Infectious Diseases, North Grafton, MA 01536, USA

*Corresponding author
e-mail: vwysocki@email.arizona.edu

Abstract

We report the successful *de novo* sequencing of hemoglobin using a mass spectrometry-based approach combined with automatic data processing and manual validation for nine North American species with currently unsequenced genomes. The complete α and β chain of all nine mammalian hemoglobin samples used in this study were successfully sequenced. These sequences will be appended to the existing database containing all known hemoglobins to be used for identification of the mammalian host species that provided the last blood meal for the tick vector of Lyme disease, *Ixodes scapularis*.

Keywords: LC-MS/MS; PepNovo; species identification; tick-borne diseases.

Hemoglobin is a protein present in virtually all kingdoms of living organisms and is particularly highly abundant in the blood of vertebrates, constituting approximately 97% of the dry weight of the red blood cells (Weed et al., 1963). The increasingly sensitive separation and detection methods, such as polyacrylamide gel electrophoresis (PAGE), thin-layer isoelectric focusing (TL-IEF), capillary electrophoresis (CE), high-performance liquid chromatography (HPLC), and mass spectrometry (MS), allow unambiguous identification of small variations in the hemoglobin sequence such as single point mutations and deletional and non-deletional gene mutation products. As a result, severe hemoglobinopathies like sickle cell anemia or thalassemia can be diagnosed using a routine and cost efficient laboratory procedure (Boemer et al., 2008; Michlitsch et al., 2009).

Mutations in the hemoglobin sequence of small mammals, in particular that of the deer mouse (*Peromyscus maniculatus*), have been previously correlated with high-altitude habitats (Storz et al., 2009). Storz and coworkers have recently reported that hypoxic stress results in the increase of oxygen blood conductance by change in the oxygen affinity of hemoglobin. In this study, as many as five simultaneous adaptive modifications of the α chain were observed (Storz et al., 2007).

In the study of adaptive and non-adaptive evolution, it has been recognized that hemoglobin-like genes are present in many organisms besides animals. Plants (Hoy et al., 2007) as well as microorganisms such as fungi or bacteria (Vasudevan et al., 1991) all possess hemoglobin-like proteins. The expressed proteins of these genes may bind other gases besides O₂ and may have entirely different roles in each individual organism. Ultimately, phylogenetic trees may be constructed based on sequence homology studies (Hardison, 1998).

Lyme disease, the most common vector-borne infection in North America, is maintained in enzootic cycles between ticks of the *Ixodes persulcatus* species complex and diverse reservoir hosts, mainly rodents. The contribution of individual species of reservoir hosts may differ from site to site, and thus, specifically identifying the main source of infectious blood meals for ticks would help in locally adapting intervention methods. Classical methods for blood meal identification from mosquitoes and other rapidly feeding vectors (such as the use of specific antihost antibodies) cannot be used for ticks because they feed only once at each host-seeking life stage (larva, nymph, adult) and undergo dramatic developmental remodeling (molting), thereby leaving inadequate amounts of immunoglobulin. DNA amplification methods have been described (Humair et al., 2007; Alcaide et al., 2009; Allan et al., 2010) for detecting remnants of host nucleic acid but may fail as frequently as half of the time due to DNA degradation.

Because ticks slowly process blood using an intracellular mode of digestion and the gut itself is not remodeled during molting, it may be that sufficient protein remnants remain to be

---

¹Present address: Vector-Borne Disease Section, California Department of Public Health, Ontario, CA, USA
identified (Jasinski et al., 2000; Wickramasekara et al., 2008). A protein-based method for identification of the blood meal from the previous life stage may thus be useful. For this purpose, we are exploring the use of sequence variations of hemoglobin among small mammals as a means of identifying host species from the blood meal of *Ixodes scapularis* ticks. Accordingly, the mammals selected for the present work were targeted based on their potential role as hosts that might be involved in the transmission cycle of tick-borne diseases and because their amino acid sequences are not available.

To separate from other blood components those red blood cells that did not lyse during freezing, 50 µl of each of the samples was centrifuged at 1000 g for 5 min, and the supernatant was discarded. The blood cells were washed with 1.2% NaCl and lysed with 1 ml of ultrapure water (Millipore, Billerica, MA, USA). Fifty microliters of the lysis solution was reduced with 2 µl 100 mM DTT (Bio-Rad Laboratories, Hercules, CA, USA) for 15 min at room temperature, alkylated for 15 min at room temperature in the dark using 3 µl 100 mM iodoacetamide (Sigma-Aldrich St. Louis, MO, USA) and digested with proteomics-grade trypsin (Sigma-Aldrich) at a 1:20 enzyme:protein ratio. To confirm the peptide sequences and obtain information about missing regions, 50 µl solution was removed from each lysis sample for digestion with GluC (Protea Biosciences, Morgantown, WV, USA).

After 12 h of digestion at 37°C, the peptides were desalted and cleaned using a C18 solid phase extraction cartridge (Varian, Harbor City, CA, USA) and eluted with 600 µl 90% acetonitrile 0.1% formic acid (EMD Chemicals, Gibbstown, NJ, USA). The samples were dried to 10 µl using a speed-vac and acidified with 40 µl 0.1% formic acid.

For the mass spectrometric analyses, a 5-µl sample was used for each liquid chromatography (LC)-MS/MS experiment; separation of peptides was performed on a 120-µm O.D. C18 capillary column (packing material Zorbax Eclipse XDB-C18, Agilent Technologies, Santa Clara, CA, USA) and eluted with 600 µl 90% acetonitrile 0.1% formic acid (EMD Chemicals, Gibbstown, NJ, USA). The samples were dried to 10 µl using a speed-vac and acidified with 40 µl 0.1% formic acid.

The XCalibur .raw data files generated by each LC-MS/MS experiment were subjected to Sequest search algorithm with a database compiled from all existing hemoglobin sequences in the National Center for Biotechnology Information (NCBI) database. For sequencing of the less conserved regions the .dta files were subjected to PepNovo; spectra were assigned to hemoglobin based on sequence homology to other hemoglobins and were manually sequenced for confirmation.

Protein sequencing is performed routinely using automated sequencing apparatus after tedious separation and cleanup procedures including 1D or 2D SDS-PAGE separations, immunoprecipitation, column chromatography, etc. (Maita et al., 1981). In the present work, we used MS to detect and sequence peptides derived from the proteins of interest, and we have used this method without any strenuous sample preparation methods. This is possible for several reasons. First, hemoglobin is one of the few proteins that are present in large quantities in a blood sample; therefore, the likelihood of matrix effects and contamination is minimal. Second, due to the large sequence homology of hemoglobins originating from different mammals, the identified peptides have similar sequences to known hemoglobins, therefore allowing identification in a sample that contains many other proteins while also allowing sequencing of variable regions.

Molecules present in the eluent from the LC column are ionized via electrospray ionization (ESI) forming singly- or multiply-protonated peptides. In a routine data-dependent MS experiment, two types of scans are used. First, the instrument takes a scan of all ionized species eluting from the LC column at a particular moment (MS scan). In a series of subsequent events, the software automatically selects the most intense peaks (user defined number of peaks are used, up to an instrument specific maximum) and performs an MS/MS scan on each of these individual peaks. In this scan, the selected peptide peak (i.e., precursor ion) is fragmented via collision-induced dissociation (CID), a fragmentation method regularly used in mass spectrometry of peptides. The resulting N- and C-terminal fragments are used to sequence the peptide.

De novo sequencing using MS is an attractive feature because of the inherent sensitivity of state-of-the-art instruments, fast data collection, and compatibility with LC. The main difficulty – and ultimately, the bottleneck of sequencing via mass spectrometry – lies in the interpretation of the several thousand spectra that are generated during a routine LC-MS/MS run. Most approaches rely on identification of sequences that exist in databases (NCBI, SwissProt, etc.), while others focus on de novo sequencing. For this, several software packages have been designed, such as PEAKS (Ma et al., 2003), NovoHMM (Fischer et al., 2005), and Lutefisk (Taylor and Johnson, 1997) and implemented with success in real-life applications (Pevtsov et al., 2006; Pitzer et al., 2007). In this work, we used PepNovo (by the Pevzner group from University of California San Diego, CA, USA) and validated the results manually, as described in detail below. The sequences inferred from MS data are presented here.
De novo sequencing of hemoglobin using mass spectrometry

In this example, the forward (i.e., N- to C-terminus) amino acid sequence was found by identifying the b_n ion series in the spectrum between residues 4 and 14. The C-terminal amino acid is arginine (R), as found by subtracting the precursor mass and the last identified peak. The reverse (i.e., C- to N-terminal) sequence was similarly found by using the y_n ion series in the spectrum (shown in Figure 1). The N-terminal residues LGG were identified from the doubly charged y_n series. The fourth amino acid was determined to be histidine (H) from the mass difference between y_132+ and y_142+.

In our approach, we compiled a list of five hemoglobin sequences from human as well as mouse and other small mammals. When aligning these proteins, it was readily visible that there are regions where the sequence is highly conserved across species, and there are also regions where a single or multiple amino acids are different. For example,

Table 1  Example of a PepNovo algorithm output for a single MS/MS scan.

<table>
<thead>
<tr>
<th>#Index</th>
<th>RnkScr</th>
<th>PnvScr</th>
<th>N-Gap</th>
<th>C-Gap</th>
<th>[M+H] Charge</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.777</td>
<td>177.363</td>
<td>0</td>
<td>0</td>
<td>1529.736</td>
<td>LGGHAGEYGAEALER</td>
</tr>
<tr>
<td>1</td>
<td>2.456</td>
<td>151.359</td>
<td>227.076</td>
<td>0</td>
<td>1529.736</td>
<td>HAGEYGAEALER</td>
</tr>
<tr>
<td>2</td>
<td>2.445</td>
<td>153.978</td>
<td>113.166</td>
<td>0</td>
<td>1529.736</td>
<td>NHAGEYGAEALER</td>
</tr>
<tr>
<td>3</td>
<td>2.187</td>
<td>156.659</td>
<td>170.406</td>
<td>0</td>
<td>1529.736</td>
<td>GHAGEYGAEALER</td>
</tr>
<tr>
<td>4</td>
<td>1.937</td>
<td>154.814</td>
<td>156.446</td>
<td>0</td>
<td>1529.736</td>
<td>AHAGEYGAEALER</td>
</tr>
<tr>
<td>5</td>
<td>1.895</td>
<td>160.414</td>
<td>113.166</td>
<td>0</td>
<td>1529.736</td>
<td>GHAGEYGAEALER</td>
</tr>
<tr>
<td>6</td>
<td>1.815</td>
<td>148.683</td>
<td>113.166</td>
<td>0</td>
<td>1529.736</td>
<td>NHAGEYGAEALER</td>
</tr>
<tr>
<td>7</td>
<td>1.732</td>
<td>172.068</td>
<td>113.166</td>
<td>0</td>
<td>1529.736</td>
<td>LGGHAGEYGAEALER</td>
</tr>
<tr>
<td>8</td>
<td>1.697</td>
<td>150.758</td>
<td>364.268</td>
<td>0</td>
<td>1529.736</td>
<td>AEGYGAELER</td>
</tr>
<tr>
<td>9</td>
<td>1.645</td>
<td>171.767</td>
<td>0</td>
<td>0</td>
<td>1529.736</td>
<td>LNHAEGYGAELER</td>
</tr>
</tbody>
</table>

The header in the table contains the .raw file name (UL_Hb_Animal1) as well as the scan number of the MS/MS spectrum in the file (in this case, 3766). The last digit (2) is the assumed charge state for the precursor ion. The “RnkScr” and “PnvScr” are score values assigned by the algorithm based on fragmentation probabilities and represent the likelihood that the peaks present in the spectrum follow known peptide fragmentation chemistry (Frank and Pevzner, 2005). The N-Gap value is the mass of the non-sequenced peptide portion on the N-terminus, and C-Gap is the C-terminal non-sequenced mass. In this example, the top peptide (Index 0) is the true match that was confirmed by manual sequencing of the corresponding spectrum (shown in Figure 1).

Figure 1  MS/MS spectrum of LGGHAGEYGAEALER peptide sequenced by PepNovo and validated manually. Note that the X-axis of the spectrum is the mass divided by the number of charges the ions carry.
the peptide LRVDPVNFK (hemoglobin α, residues 92–100) is a region where the sequence is conserved for a great variety of mammals, while the N-terminus of the protein (the beginning of the sequence) is extremely variant with a general sequence of (M)VLSXXDKXXXK, where X represents non-conserved residues. Therefore, when starting to de novo sequence, we first searched the PepNovo output file for scans where the sequence is conserved and then for sequence tags of three amino acids or longer to fill the missing positions. As all hemoglobin α and β chains have the same number of amino acid residues (141 for α and 146 for β chain), once a peptide is sequenced, its position in the protein sequence can be unambiguously determined.

In an instrument with unit mass resolution, such as the instrument we have used in our study, the MS/MS spectrum of a peptide is the most informative when all b and y ions are present and carry a single or two charges, at most. For this, the ideal length of the peptide should be between five and 10 residues. The sequence of peptides shorter than five residues has little or no information regarding position of the peptide in the protein sequence, while longer peptides can have multiple basic residues, and therefore more charges. The use of GluC, another proteolytic enzyme that selectively cleaves after glutamic acid (E) residues, was necessary for sequencing protein regions where two tryptic cleavage sites were yielding peptides that were either too long or too short.

When de novo sequencing is performed on mass spectra obtained using CID in an ion trap, one inherent disadvantage is that isoleucine and leucine cannot be distinguished due to their identical nominal mass. In the genetic code, leucine is encoded by six different codons, while isoleucine is encoded by three; therefore, there is a 2-fold statistical probability for the presence of leucine vs. isoleucine. As a consequence, in the sequences reported herein, we assume that leucine is present in all cases. For the purpose of our future studies (i.e., determine the host species based on their hemoglobin sequence), we will use similar mass spectrometry-based approaches; therefore, distinguishing between leucine and isoleucine is not critical for our experiments. If distinction of leucine and isoleucine should be desired in the future, keV CID, e.g., in a TOF TOF instrument, could be performed. A similar problem arises in the case of lysine and glutamine; however, these two amino acids can be readily distinguished when using trypsin as a proteolytic enzyme. Because trypsin cleaves after lysine and arginine, we can assume that all amino acids with nominal mass 128 at the C-terminus are lysine, while those in the beginning or middle of a sequence are more likely to be glutamine. An additional inconvenience is the fact that certain amino acid combinations have the same nominal mass as a single amino acid (such as GA or AG have the same mass as K or Q, GE or EG, VS or SV has the same mass as W). In these cases, we can rely on the fact that all hemoglobins gave the same number of amino acid residues in their sequence; therefore, the PepNovo results where the correct number of amino acids is present is accepted, even if it has a lower rank score than the first calculated sequence. The results of the approach described above are presented in Table 2, showing the sequence of the α and β hemoglobin chains determined from each small mammal.

Performing a basic local alignment search tool (BLAST) search of the studied hemoglobin sequences confirmed that there is a large sequence homology between mammals of different species but part of the same family. The sequest database search of the acquired LC-MS/MS data for the mammals with previously unknown hemoglobin sequence against all known hemoglobin sequences similarly returned species that belong to the same family.

A maximum likelihood phylogram was produced with PhyML (Guindon and Gascuel, 2003) as implemented by the SeaView version 4 suite of algorithms (Guoy et al., 2010). For example, Figure 2 shows that the sequences obtained for the α chains fit with expected phylogeny of the species.

The α chain sequences were aligned with known protein sequences from the database for these species: Peromyscus leucopus, Peromyscus maniculatus, Ondatra zebethicus (musk rat), Marmota marmota (marmot), Ctenodactylus gundi (northern gundi), Sorex araneus (European shrew), and Suncus murinus (house shrew). The numbers for each node show the percent bootstrap support for 100 iterations. The marker indicates the distance for branch lengths. The two Peromyscus species sequenced herein had sequences that clustered with the two known Peromyscus species; similarly, the shrew sequenced (Blarina brevicauda) clustered with two other insectivores.

We have therefore presented the results for de novo sequencing of hemoglobin from nine small mammals using tandem MS. This method can be readily used for sequencing of proteins when there is a large degree of sequence homology between the unknown and a previously known protein sequence, such as a related species or a protein that is highly conserved across species. It can also be applied when searching for single-point mutations in a known protein. Current work is in progress for the sequencing of hemoglobin from 33 small bird species known to be reservoirs for tick-borne diseases. The protein library containing these hemoglobin sequences will be used to identify the host species on which I. scapularis and other species of ticks fed. This data, along with existing sequence data (for example, for P. leucopus, pre-sumptively the main Lyme reservoir species in the Northeast), serves as the basis for analyzing host-seeking vector ticks to determine the identity of their previous blood meal and thus quantitating the relative roles of different species in the transmission of tick-borne diseases.

In a recently completed time-course study, we have found that mouse (Mus musculus) hemoglobin was present in detectable levels in laboratory-fed I. scapularis nymphs for as long as 1 year postfeeding. Although several other mouse proteins (such as actin, keratin, tubulin, histones) were also detected, we have targeted hemoglobin as the main protein of interest for host species identification for several reasons. For one, hemoglobin consists of two, relatively short protein chains, which facilitates the de novo sequencing when compared to other, longer proteins. Second, as illustrated in the present work, the amino acid sequence is preserved between various species to a certain extent; therefore, we can readily
Table 2  Hemoglobin α and β chain sequences of nine small mammals obtained by LC-MS/MS.

<table>
<thead>
<tr>
<th>Species</th>
<th>Hemoglobin α Chain Sequence</th>
<th>Hemoglobin β Chain Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamias merriami</td>
<td>A C  B 3 E W C 7;</td>
<td>A C  B 3 E W C 7;</td>
</tr>
<tr>
<td></td>
<td>OS Tamias merriami (Merriam’s chipmunk)</td>
<td>OS Tamias merriami (Merriam’s chipmunk)</td>
</tr>
<tr>
<td></td>
<td>VLSADKTNKAVSKWKLGGHAGAYGAEALRFLSFPTTKTYYPFHDFLSDHGAQVQGKVKVADALANAAGHLDDLPGALASLDDLHAKLRVDVPVNFKLSSICLLVTLAAHHPAEFTPAVHASLDKFLATVSTVLSKYS</td>
<td>VHLTDAEKLTVGLGWKVKPEDELGEALGLRLVYYPWTQRFDFDSDGDLSSASAMGNPKVKAHGKKVLDSDFSLKLHLDNLKGTFASLHELHDCKLHVDPNFKLLGNLVMVLALHLLGKDFTPAAQAAYQKVAGVANALAHKYH</td>
</tr>
<tr>
<td>Spermophilus beecheyi</td>
<td>A C  B 3 E W D 0;</td>
<td>A C  B 3 E W D 0;</td>
</tr>
<tr>
<td></td>
<td>OS Spermophilus beecheyi (Beechy ground squirrel)</td>
<td>OS Spermophilus beecheyi (Beechy ground squirrel)</td>
</tr>
<tr>
<td></td>
<td>VLSADKTNKASWWEKLGGHAGAYGAEALRFLSFPTTKTYYPFHDFLSDHGAQVQGKVKVADALANAAGHLDDLPGALASLDDLHAKLRVDVPVNFKLSSICLLVTLAAHHPAEFTPAVHASLDKFLATVSTVLSKYS</td>
<td>VHLTDAEKLTVGLGWKVKPEDELGEALGLRLVYYPWTQRFDFDSDGDLSSASAMGNPKVKAHGKKVLDSDFSLKLHLDNLKGTFASLHELHDCKLHVDPNFKLLGNLVMVLALHLLGKDFTPAAQAAYQKVAGVANALAHKYH</td>
</tr>
<tr>
<td>Sciurus carolinensis</td>
<td>A C  B 3 E W D 1;</td>
<td>A C  B 3 E W D 1;</td>
</tr>
<tr>
<td></td>
<td>OS Sciurus carolinensis (gray squirrel)</td>
<td>OS Sciurus carolinensis (gray squirrel)</td>
</tr>
<tr>
<td></td>
<td>VLSADKTNKASWWEKLGGHAGAYGAEALRFLSFPTTKTYYPFHDFLSDHGAQVQGKVKVADALANAAGHLDDLPGALASLDDLHAKLRVDVPVNFKLSSICLLVTLAAHHPAEFTPAVHASLDKFLATVSTVLSKYS</td>
<td>VHLTDAEKLTVGLGWKVKPEDELGEALGLRLVYYPWTQRFDFDSDGDLSSASAMGNPKVKAHGKKVLDSDFSLKLHLDNLKGTFASLHELHDCKLHVDPNFKLLGNLVMVLALHLLGKDFTPAAQAAYQKVAGVANALAHKYH</td>
</tr>
<tr>
<td>Peromyscus crinitus</td>
<td>A C  B 3 E W D 3;</td>
<td>A C  B 3 E W D 3;</td>
</tr>
<tr>
<td></td>
<td>OS Peromyscus crinitus (canyon mouse)</td>
<td>OS Peromyscus crinitus (canyon mouse)</td>
</tr>
<tr>
<td></td>
<td>VLSADKTNKAVSKWKLGGHAGAYGAEALRFLSFPTTKTYYPFHDFLSDHGAQVQGKVKVADALANAAGHLDDLPGALASLDDLHAKLRVDVPVNFKLSSICLLVTLAAHHPAEFTPAVHASLDKFLATVSTVLSKYS</td>
<td>VHLTDAEKLTVGLGWKVKPEDELGEALGLRLVYYPWTQRFDFDSDGDLSSASAMGNPKVKAHGKKVLDSDFSLKLHLDNLKGTFASLHELHDCKLHVDPNFKLLGNLVMVLALHLLGKDFTPAAQAAYQKVAGVANALAHKYH</td>
</tr>
<tr>
<td>Peromyscus californicus</td>
<td>A C  B 3 E W D 5;</td>
<td>A C  B 3 E W D 5;</td>
</tr>
<tr>
<td></td>
<td>OS Peromyscus californicus (California mouse)</td>
<td>OS Peromyscus californicus (California mouse)</td>
</tr>
<tr>
<td></td>
<td>VLSADKTNKAVSKWKLGGHAGAYGAEALRFLSFPTTKTYYPFHDFLSDHGAQVQGKVKVADALANAAGHLDDLPGALASLDDLHAKLRVDVPVNFKLSSICLLVTLAAHHPAEFTPAVHASLDKFLATVSTVLSKYS</td>
<td>VHLTDAEKLTVGLGWKVKPEDELGEALGLRLVYYPWTQRFDFDSDGDLSSASAMGNPKVKAHGKKVLDSDFSLKLHLDNLKGTFASLHELHDCKLHVDPNFKLLGNLVMVLALHLLGKDFTPAAQAAYQKVAGVANALAHKYH</td>
</tr>
</tbody>
</table>

**Note:** -peptide also present in the sample.
6. *Tamiasciurus hudsonicus*

α:
- AC B3EW67:
  - DE RecName: Full=Hemoglobin subunit alpha;
  - OS *Tamiasciurus hudsonicus* (American red squirrel)
  - VLSADKTNVSKAWDCLKLGHGHAEGYAEALGRMFLSFTPTTKTYPFHFQDLHSHGAQPQGHGKVQAELATAAGHLDDLPGAL
  - SALSDLHAKHLRVDVPVNFKLLLHCLLVTLLAHMPAEFTPAVHASLDKFLASVSTVLTSKYR

β:
- AC B3EW68:
  - DE RecName: Full=Hemoglobin subunit beta;
  - OS *Tamiasciurus hudsonicus* (American red squirrel)
  - VHLSGEEKTALATLWGNVPAEVEQGGEALGRLLVYYPWTQRFQDFSGLDSLSSASAMNSAKVKAHGKKVLDSFSEGKLHLD
  - LGKFSSSLHCDDKHLVDPMFRRLLGNMLVLMHHLGKDFTPAAQAAYQKVAGVANALAHKYH

7. *Tamias striatus*

α:
- AC B3EW69:
  - DE RecName: Full=Hemoglobin subunit alpha;
  - OS *Tamias striatus* (Eastern chipmunk)
  - VLSAPDKTNLKAAWHKLGGHGEYAEALERMATFAPTPTTGTKYPFHFQDLHSHGAQPQGHGEKVADALLHAVGNLDDLPGAL
  - SALSDLHAKHLRVDVPVNFKLLLHCLLVTLLAHHPAEFTPAVHASLDKFLATVSTVLTSKYR

β:
- AC B3EWE0:
  - DE RecName: Full=Hemoglobin subunit beta;
  - OS *Tamias striatus* (Eastern chipmunk)
  - VHLTADEKVSLSSLWGGVLIDLEGALGRLLVYYPWTQRFQDFSGLDSLSSAVMNGNAVKVKAHGKKVLDSFSDGLKHD
  - LGKFSASSLHCDDKHLVDPMFRRLLGNVLVLMHHLGKEFTPQAQTFQKVAGVANALAHKYH
  - *VLSAPDKTNV* - peptide also present in the sample, where V replaces L after N.

8. *Blarina brevicauda*

α:
- AC B3EWE1:
  - DE RecName: Full=Hemoglobin subunit alpha;
  - OS *Blarina brevicauda* (short-tailed shrew)
  - VLSASDKTNLKAWWKLGHGQAANYGAEALERTFASPTTGTKYPFHFQDLPSGASQVGHHGKVADALTAVGSLDDLPGAL
  - SALSDLHAKHLRVDVPVNFKLLLHCLLVTLLASHHPADFTPAPVHASLDKFLATVSTVLTSKYR

β:
- AC B3EWE2:
  - DE RecName: Full=Hemoglobin subunit beta;
  - OS *Blarina brevicauda* (short-tailed shrew)
  - VHLTAEEKSLTYGLGWKVNYVEAGGLEAGRLLLVYYPWTQRFQDFSGLDSLSSASAVMNGNKVKAHGKKVLQSMGDGLANLD
  - NLKGTFAKSLHCDDKHLVDPMFRRLLGNVLVVLARHFGKEFTPQVAAFQKVAGVATALAHKYK

9. *Microtus pennsylvanicus*

α:
- AC B3EWE3:
  - DE RecName: Full=Hemoglobin subunit alpha;
  - OS *Microtus pennsylvanicus* (meadow vole)
  - VLSGDDKSNLKTAWGKLGHHAEGYAEALERMFVAYPTTGTKYPFHFQDLHSHGAQPQGHGKVADALTAVGHLDDLPGAL
  - SALSDLHAKHLRVDVPVNFKLLLHCLLVTLLANHLPDFTPAPVHASLDKFLASVSTVLTSKYR

β:
- AC B3EWE4:
  - DE RecName: Full=Hemoglobin subunit beta;
  - OS *Microtus pennsylvanicus* (meadow vole)
  - VHLTDaedaAALSGLWGKANADAVGAELGRLLVYYPWTQRFQFHEFDGLDSLSSAVMNGNKVKAHGKKVLHAFADGLKHD
  - NLKGTFAKSLHCDDKHLVDPMFRRLLGNMLVLSHDLGDFTPAQAAPFQKVAGVASALAHKYH

Whole blood from nine mammals (*T. merriami* – Merriam’s chipmunk, *S. beecheyi* – California ground squirrel, *S. carolinensis* – Eastern gray squirrel, *P. crinitus* – canyon mouse, *P. californicus* – California mouse, *T. hudsonicus* – red squirrel, *T. striatus* – Eastern chipmunk, *B. brevicauda* – Northern short-tailed shrew, *M. pennsylvanicus* – meadow vole) was obtained from specimens captured in Lyme-endemic sites in Massachusetts and California and placed in vials with EDTA anticoagulant; the samples were stored at -20°C until analysis. All animals from Massachusetts were obtained and processed under approved Institutional Animal Care and Use Committee (IACUC) protocols (Tufts University G520-04, G895-07). California samples were collected following Centers for Disease Control (CDC) guidelines for rodent sampling by the California Department of Public Health Vector-Borne Disease Section. The protein sequence data reported here will appear in the UniProt Knowledge base under the indicated accession numbers. The AC line indicates protein accession number, DE shows the description while the OS line contains the organism species.
distinguish peptides originating from hemoglobin vs. other proteins from a complex matrix, such as an entire tick organism. More importantly, the amino acid substitutions allow us to discriminate not only birds from mammals, but also distinguish between mammals on the species level. The findings of our time-course study will be subject of a future publication.

Acknowledgments

Funding for this project was provided by the National Institutes for Health (NIH) grant AI-65359. We would like to thank Seung Whan Oh for sample preparation and Drs. Linda Breci and Arpad Somogyi for the helpful discussions regarding experiment design and database search. Unigue Laskay would like to thank Dr. Ari Frank from University of California San Diego for the kind assistance with PepNovo.

References


Received August 12, 2011; accepted December 24, 2011

Figure 2  Phylogenetic tree based on the α chain of the hemoglobin of the sequenced species (red). The numbers represent the percent bootstrap support for 100 iterations.