Identifying sources of tick blood meals using unidentified tandem mass spectral libraries

Özlem Önder[1], Wenguang Shao[2,] Brian Kemps [1], Henry Lam[2,3,*], Dustin Brisson[1,*]

[1]Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19014-6019, USA.

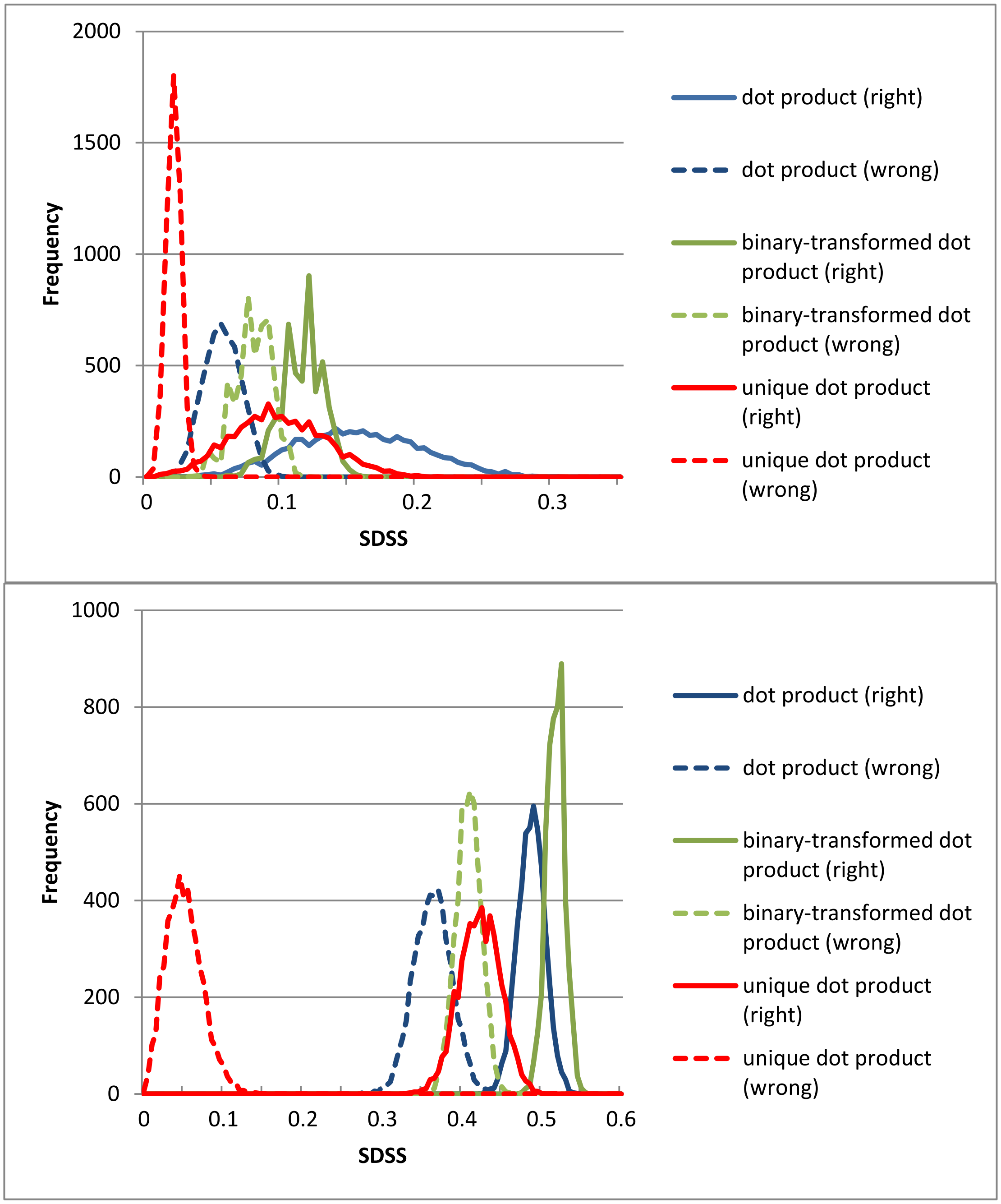
[2]Bioengineering Graduate Program, Division of Biomedical Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

[3]Department of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

* Co-corresponding authors

Supplementary Information

Supplementary Figures



Supplementary Figure S1. Score histograms of the top-scoring, correct library match (solid lines) and second-highest-scoring, incorrect library match (dashed lines), using the three SDSS functions as described, for the mouse-fed molted nymph dataset (top panel), and the tiger blood dataset (bottom panel). The distributions are obtained by bootstrapping the data 5000 times.

Supplementary Tables

Species	Animal_ID. Tick_ID	Hemoglobin only (sequence-based)		
		Best sequence match	Bootstrap support	Success Rate
Lab mouse	MM1.l1	Lab Mouse	>99%	5/5
	MM1.l2	Lab Mouse	>99%	
	MM1.l3	Lab Mouse	>99%	
	MM1.l4	Lab Mouse	>99%	
	MM2.l1	Lab Mouse	>99%	
White-footed mouse	PL1.l1	White-footed mouse	>99%	4/4
	PL2.l1	White-footed mouse	>99%	
	PL3.l1	White-footed mouse	>99%	
	PL4.l1	White-footed Mouse	>99%	
Chipmunk	TS1.l1	Chipmunk	>99%	3/3
	TS2.l1	Chipmunk	>99%	
	TS3.l1	Chipmunk	>99%	
Squirrel	SC1.l1	Chipmunk [a]	93.4%	3/3[a]
	SC1.l2	Chipmunk [a]	88.4%	
	SC1.l3	Chipmunk [a]	>99%	
TOTAL				15/15

Supplementary Table S1. The species on which an engorged larvae (*I. scapularis*) previously parasitized can be confidently determined through detection of hemoglobins by sequence searching. ([a]) *S. carolinensis* squirrel was not included in the sequence database. The identification was regarded as successful if the most frequently identified hemoglobin sequence was from *T. striatus* chipmunks, evolutionarily the most closely related species.

Species	Tick Age	Animal_ID. Tick_ID	Hemoglobin only (sequence-based)		
			Best sequence match	Bootstrap support	Success Rate
Lab Mouse	1 Month	MM1.n1	Lab Mouse	>99%	6/6
		MM1.n2	Lab Mouse	>99%	
		MM1.n3	Lab Mouse	>99%	
		MM2.n1	Lab Mouse	>99%	
		MM2.n2	Lab Mouse	>99%	
		MM2.n3	Lab Mouse	>99%	
	3 Month	MM1.n4	Lab Mouse	>99%	6/6
		MM1.n5	Lab Mouse	>99%	
		MM1.n6	Lab Mouse	>99%	
		MM2.n4	Lab Mouse	>99%	
		MM2.n5	Lab Mouse	>99%	
		MM2.n6	Lab Mouse	>99%	
	6 Month	MM1.n7	Lab Mouse	>99%	4/6
		MM1.n8	Lab Mouse	>99%	
		MM1.n9	Lab Mouse	>99%	
		MM2.n7	Lab Mouse	>99%	
		MM2.n8	No Match [a]	X	
		MM2.n9	No Match [a]	X	
White-footed mouse	5 Month	PL5.n1	White-footed mouse	>99%	2/6
		PL6.n1	No Match [a]	X	
		PL6.n2	No Match [a]	X	
		PL7.n1	White-footed mouse	>99%	
		PL7.n2	No Match [a]	X	
		PL7.n3	No Match [a]	X	
Chipmunk	5 Month	CM4.n1	No Match [a]	X	0/2
		CM5.n1	No Match [a]	X	
TOTAL					18/26

Supplementary Table S2. The species on which molted nymphs (*I. scapularis*) parasitized during their larval stage is less effective using hemoglobin sequence searching than proteome profiling. ([a]) No match were found to any of the hemoglobin sequences.

	Self			Core[b]	
	Service	**Cost[a]**	**Time**	**Service**	**Cost**
Sample prep	*Digestion*	$0.18	~18h[c]	*Digestion*	$40.00
	Sample clean up	$2.22	~30 min	*Sample clean up*	$35.00
LC-MS^{2}		$0.57[d]	90min		$350.00

Supplementary Table S3. Per sample cost of data acquisition, on a self-run LC-MS platform or in a proteomics core facility. **(a)** Per sample costs associated with reagents (trypsin, reducing/alkylating agents, buffers and columns) are estimated assuming approximately 15 samples per day for 250 days per year. (**b)** Core facility costs are derived from those advertised at the Children's Hospital of Philadelphia. Core facility costs vary considerably based on volume of samples and the affiliation of the researcher submitting the sample. (**c)** Overnight enzymatic digestion is recommended. Over 20 samples can be prepared at a time by a single investigator**.** **(d)** Service agreements are the primary cost associated with running LC-MS^{2} equipment. Factoring a service contract of $15,000/yr adds ~$4 per sample assuming 3750 samples per year resulting in a cost of ~$4.57 per sample when using our in-house LC-MS^{2}

Supplementary Discussion

Peptide identifications of spectra unique to Mus musculus

The library spectra unique to *M. musculus* blood, the only species for which a complete protein database is available, were identified to peptides by conventional sequence searching to reveal the distribution of the underlying molecules of the library spectra. It is likely that blood proteomes from other vertebrate species will follow similar protein distributions. As expected, the majority of identifiable spectra from *M. musculus* blood were derived from subunits α and β of hemoglobin, serum albumin, β-globin, and fibrinogen in addition to a few other proteins (Supplementary Fig. S1). Interestingly, over 80% of the spectra could not be identified to the *M. musculus* proteome despite the completeness of the protein sequence database. It is important to note that a 20% identification rate is comparable with that of similar experiments using this instrument, and that many high-quality spectra were not identified due to various limitations in the sequence searching process[61].

Comparison of scoring functions for the spectral dataset similarity score

We evaluated the discriminating power of three different scoring functions for the spectral dataset similarity score (SDSS). The first one is the dataset dot product: the number of query spectra matched to that library entry times the number of replicate spectra from a given source used to build that library entry -- summed over all library entries:

$$SDSS_{dot}(q,s) = \frac{\sum_i \left[m_q(i) \times r_s(i)\right]}{\sqrt{\sum_i \left[m_q(i)\right]^2 [r_s(i)]^2}}$$

where $m_q(i)$ is the number of spectra in the query dataset q that matches the i-th library spectrum, and $r_s(i)$ is the number of replicates from sample source s contributing to the i-th library spectrum. The normalization factor in the denominator scales all possible SDSS values from 0 (orthogonal) to 1 (identical profiles). Note that this *dataset* dot product measures the similarity of two datasets, and is similar in formula but different in concept from the *spectral* dot product, which measures the similarity of two spectra. This dot product SDSS accounts for the quantitative information in matching fingerprints. Namely, spectra that are more frequently found in the dataset used to build the library are also expected to be more frequently found in the query dataset. This score will therefore reward similar quantitative profiles of the library and query dataset, and penalized dissimilar profiles. Moreover, spectra that are matched more frequently will be more heavily weighted.

The second scoring function we evaluated was the “binary-transformed dot product”:

$$SDSS_{binary}(q,s) = \frac{\sum_i \left[\mathrm{sgn}(m_q(i)) \times \mathrm{sgn}(r_s(i))\right]}{\sqrt{\sum_i \left[\mathrm{sgn}(m_q(i))\right]^2 [\mathrm{sgn}(r_s(i))]^2}}$$

where sgn(.) denotes the signum function: sgn(x) = 1 if x > 0 and and sgn(x) = 0 if x = 0. In other words, this scoring function only counts the presence and absence of matches to a particular library entry, and does not take into account quantitative information.

The third scoring function, called the “unique dot,” uses the same formula as the dot product SDSS but discards all consensus spectra that originate from multiple species. In other words, all library entries owing to merging of replicates from two or more species are ignored altogether, with the implicit assumption that these spectra are due to noise or other non-discriminating molecules.

We tested the three scoring functions on two datasets: one of those from flat tick fed on mouse blood (MM1.n9 in Table 2 in main text), and a separate dataset of the tiger blood sample not used to build the library. Each dataset was bootstrapped 5,000 times, and the score histograms of the top scoring species (the correct answer) and the second highest-scoring species were plotted (Supplementary Fig. S1).

From Supplementary Fig. S1, all three scoring functions demonstrate sufficient ability of discriminating the right and wrong answer. The "unique dot" SDSS appears to offer the best discrimination, and the "binary" score the worst, though the difference is quite small. This suggests that the quantitative information afforded by spectral counts appears to augment the discriminating power, although for these datasets, the quantitative information was not absolutely necessary, as the binary-transformed SDSS was also sufficiently discriminating. This is understandable, because distinguishing between blood samples of different organisms probably relies more on sequence differences of detected peptides and thus qualitatively different spectra, than on the quantities of identical peptides in the sample. However, the quantitative information may be more useful in other biological sample fingerprinting applications, such as distinguishing different samples from the same organism (e.g. different tissues, different biological states).

The removal of multiple-organism spectra in the library also increases the discrimination power, as the "unique-dot" SDSS appears to outperform the "dot" SDSS. Removing multiple-organism spectra should expectedly elevate the importance of the matches to library spectra that can unambiguously identify the sample source. However, this scoring function may not be as appropriate when the library contains several phylogenetically close animals. In this hypothetical case, many library spectra will be shared among multiple species, and removing all of them may not be advisable. For instance, in the present study, a large fraction of spectra found in the lion can also be found in the tiger. While one can still distinguish between these two cats using the "unique-dot" score, there may be cases where discarding all shared spectra removes too much information, to the point that even the ability to distinguish a cat and a non-cat is lost. The "dot" score, on the other hand, retains these shared spectra, and therefore will not be affected by the composition of the library. Based on these results, the "unique dot" score is chosen as the SDSS function used for blood meal identification as described in the main text.

Identification of blood meal by sequence *searching against hemoglobin sequences*

As a benchmark for our spectral matching-based proteome profiling method, the spectral datasets from engorged larvae and molted nymphal ticks were searched against a database of hemoglobins of *M. musculus*, *P. leucopus* and *T. striatus*, plus expected contaminants by conventional sequence searching (Supplementary Methods above). The results show that identification of the source of the blood meal from engorged larval ticks was effective using methods based on detecting specific blood proteins by sequence searching algorithms (Supplementary Table S1**)**. Similar to the results of the proteomic profiling method, larval ticks that had fed upon *S. carolinensis* could also be identified as ticks that had fed upon *T. striatus* if *S. carolinensis* hemoglobin sequences are not included in the database. However, identification of the source of the blood meal from molted nymphal ticks was less effective when peptide identification relied upon *de novo* sequenced hemoglobin subunits (Supplementary Table S2**)**, failing to identify the correct source of the blood meal from two ticks that had fed on a lab mouse, both ticks that had fed on a chipmunk, and 4 of the 6 ticks that had fed on white-footed mice. For these tick samples, which had been held at room temperature for 5 months or more post-molt, no confident identification to any hemoglobin sequences were found.

Accessibility and estimated cost

The spectral matching methodology proposed here is accessible to researchers with and without direct access to proteomic technologies. Sample preparation consists of standard laboratory protocols, which are also routinely performed by proteomics core facilities for an additional fee. Further, the per sample costs for blood samples and test samples are inexpensive relative to many other technologies and can be accomplished with either an in-house LC-MS^2 or through core facilities (Supplementary Table S3).

Supplementary Methods

Unidentified spectral library building

The software program SpectraST, available as part of the Trans-Proteomic Pipeline suite, was adapted to build spectral libraries from unidentified spectra. SpectraST obtains the spectra directly from open XML formats such as mzXML, mzData and mzML. The library building procedure consists of the following steps. First, spectrum filtering is performed to remove unwanted spectra that are likely due to noise. Only spectra meeting all the following criteria are kept: precursor *m/z* being greater than 350 Th, the range of fragment *m/z* (*i.e. m/z* difference between heaviest and lightest fragment) being greater than 350 Th, having at least 35 peaks, and having at least 5% of the total intensity at above the precursor *m/z*. The last criterion is intended to remove all singly-charged precursors, which are much more likely to be from non-peptide molecules and therefore not discriminating for the purpose of this application.

Second, spectra that pass the filters are clustered by similarity using the following algorithm. All spectra are first sorted by a measure of signal-to-noise ratio, and clustering proceeds from the highest quality to the lowest. A spectrum (the "root") is chosen from the list in that order, and compared to all spectra with precursor *m/z* within 2.5 Th of that of the root. The dot product, a measure of similarity, is calculated for all pairs of spectra compared, in the same way as defined in SpectraST[49]. All spectra with dot product greater than 0.7 are considered similar to the root and are included in a cluster together with the root. Then, the search for other similar spectra proceeds recursively, with all newly added spectra acting as the root in turn. Subsequent rounds of recursion use progressively tighter precursor *m/z* tolerance and dot product threshold. For more implementation details of this algorithm, the reader is referred to the source code, in particular the file SpectraSTPeakList.cpp.

Third, at the end of the process, all spectra belonging to a cluster is merged into a consensus spectrum using the same consensus algorithm for identified spectra, described in[39]. Spectra belonging to single-member clusters are included in the library but reduced to the most intense 150 peaks. The sample source(s) of its originating replicate spectra, together with corresponding quantitative information (spectral counts or precursor intensities) are recorded with each consensus spectrum.

Identification of library spectra from Mus musculus

The library spectra unique to *Mus musculus* blood, the only species for which a complete protein database is available, were identified to peptides by conventional sequence searching to reveal the distribution of the underlying molecules of the library spectra. Two sequence search engines, OMSSA (version 2.1.8)[62] and X!Tandem with K-score plugin (version 2009.10.01)[63] were used with the following search parameters: trypsin specificity on both termini, at most two missed internal tryptic cleavages, carbamidomethylation on cysteines as a fixed modification, oxidation on methionine as a variable modification, and mass tolerances of +/- 3 Da for precursors and +/- 1 Da for product ions. The database was constructed by combining the *M. musculus* sequences in SwissProt from UniProt[64], and contaminant sequences from *I. scapularis* and porcine trypsin, then appended with an equal-size decoy database generated by random shuffling of amino acids between tryptic sites[65]. The results were processed by the Trans Proteomic Pipeline (v4.5.2)[54] using non-parametric model for PeptideProphet[66] and combined using iProphet[67]. The identifications were filtered at an FDR of 1%.

To validate the clustering algorithm, a larger dataset of 40 runs of a SDS-PAGE fractionated protein digest sample of whole *M. musculus* blood was used. The sequence search engine SEQUEST[68] was used to search against a database of *M. musculus* sequences from SwissProt. The search parameters were: +/- 2 Da parent monoisotopic mass window, +/- 1 Da fragment ion mass window, tryptic cleavage sites, variable methionine oxidation (+16.0 Da) and variable C-term carboxymethylation (+57.0 Da). PeptideProphet was used to filter the search results at an FDR of 1%.

Sequence searching against hemoglobin sequences

The spectral datasets from engorged larvae and molted nymphal ticks were searched against a database combining *de novo* sequenced hemoglobin α and β subunits of *T. striatus* [27], and *P. leucopus* and *M. musculus* hemoglobins (from SwissProt), total proteins from *I. scapularis* (NCBI), and porcine trypsin, then appended with an equal-size decoy database generated by random shuffling of amino acids between tryptic sites. The same search engines and search parameters were used as for identification of library spectra as described in the above paragraph. The search results were similarly processed by TPP, and filtered at FDR 1%. The numbers of spectra identified to the hemoglobin sequences unique to each species were counted. The species with the greatest number of identifications is taken as the source of the blood meal. This procedure was repeated on 1000 bootstrap samples of the same spectral dataset, and the fraction of times each species is identified as the source of the blood meal is reported as the bootstrap confidence.

Supplementary References

61. Ning, K., D. Fermin & A. I. Nesvizhskii. Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. *Proteomics,* 10**,** 2712-2718 (2010).

62. Geer, L. Y., S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Y. Yang, W. Y. Shi & S. H. Bryant. Open mass spectrometry search algorithm. *Journal of Proteome Research,* 3**,** 958-964 (2004).

63. Craig, R. & R. C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics,* 20**,** 1466-7 (2004).

64. The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research,* 40, D71-D75 (2012).

65. Elias, J. E. & S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods,* 4**,** 207-14 (2007).

66. Keller, A., A. I. Nesvizhskii, E. Kolker & R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry,* 74**,** 5383-5392 (2002).

67. Shteynberg, D., E. W. Deutsch, H. Lam, J. K. Eng, Z. Sun, N. Tasman, L. Mendoza, R. L. Moritz, R. Aebersold & A. I. Nesvizhskii. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics,* 10**,** M111 007690 (2011).

68. Eng, J. K., A. L. Mccormack & J. R. Yates. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry,* 5**,** 976-989 (1994).