

The following resources related to this article are available online at www.sciencemag.org (this information is current as of November 16, 2009):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/321/5892/1040c>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/cgi/content/full/321/5892/1040c#related-content>

This article **cites 9 articles**, 4 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/321/5892/1040c#otherarticles>

This article has been **cited by** 1 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/321/5892/1040c#otherarticles>

This article appears in the following **subject collections**:

Paleontology

<http://www.sciencemag.org/cgi/collection/paleo>

Technical Comments

http://www.sciencemag.org/cgi/collection/tech_comment

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>

Response to Comment on “Protein Sequences from Mastodon and *Tyrannosaurus rex* Revealed by Mass Spectrometry”

John M. Asara,^{1,2*} Mary H. Schweitzer,³ Lewis C. Cantley,^{1,2} John S. Cottrell⁴

Endogenous peptide sequences extracted from a 68-million-year-old *Tyrannosaurus rex* fossil bone and obtained by mass spectrometry have been shown to be statistically significant based on protein database searches using two different search engines and similarity comparisons to authentic tandem mass spectrometry spectra. Specifically, we have validated the sequence GVVGLP(OH)GQR.

Asara *et al.* previously reported collagen peptide sequence fragments from endogenous protein from 68-million-year-old *T. rex* fossil bone (1–3). The tandem mass spectrometry (MS/MS) spectra were rigorously validated by multiple methods, including target-decoy database searches, conservative Sequest score cut-offs, manual inspection of individual spectra (FuzzyIons and GraphMod in Proteomics Browser Software, Thermo Fisher Scientific) and, where possible, comparison of fragment ion relative intensities and chromatographic elution times with synthetically derived peptides or peptides from modern samples. Additionally, the mass spectrometry results were supported by various analytical and biochemical techniques, including *in situ* immunohistochemistry, electron microscopy, and atomic force microscopy (4).

Some database search tools, such as Mascot (5), report an expectation value based on internal computation of the distribution of scores that would be obtained for matches to random sequences. Other search tools, such as Sequest (6), report a score based on a more arbitrary measure, such as the correlation coefficient, from which an expectation value can be computed by empirical fitting of the observed distribution of scores (7). The expectation value is directly analogous to the E-value reported for a BLAST search, and is the number of times one would expect to get a match with a score at least as high by chance.

Whatever the source or perceived reliability of the expectation value, it has become common practice to make an experimental estimate of the false discovery rate (FDR) by performing a target-decoy search (8). In this procedure, the data are searched against both the target protein database and a second database, the decoy, which is of equal size and amino acid composition, but in which no matches are expected. Typically, the sequences in the decoy database are either random sequences or reversed target database entries. The number of matches found

in the decoy database is taken to be a good estimate of the number of false positives among the matches reported for the target database.

All MS/MS spectra from all data acquired after the sample extraction and purification (RP-SCX-RP) procedure had been optimized (1) were combined. This gave a collection of 48,216 spectra from seven liquid chromatography MS/MS experiments, which were searched against Swiss-Prot (release 55.1; 359,942 sequences, 129,199,355 residues) using Mascot 2.2.04. Target-decoy (reversed) searches were used to establish the FDR. Search conditions were directly equivalent to those used in the original Sequest searches. The number of high-scoring matches to collagen was low, as expected given the degraded ancient sample. The target-decoy search showed that using a Mascot expectation value of 0.24 as a cut-off corresponded to a peptide identification FDR of 5%. All six reported *T. rex* peptides (2, 3) are the top-ranked peptide matches in both the Mascot and Sequest searches, but two sequences have scores below the 5% threshold. One (Mascot score 32.0) was qualified in (1) using a MS/MS spectral comparison with a synthetic version of the peptide sequence. A second (Mascot score 22.1) was computationally qualified (Table 1). A third sequence (Mascot score 42.4) is close to the 5% FDR threshold, and we are confident in the accuracy of this match because six spectra from four separate samples give top-ranking matches to the same sequence, independently.

It is important to note that the majority of high-scoring protein hits were to contaminants not endogenous to bone, mainly human keratin, a common proteomics contaminant introduced by handling protein samples. Keratin is derived from hair and skin and is commonly found in laboratory dust. It seems more than coincidence that the only mass spectrometric fragmentation patterns from *T. rex* that statistically matched anything in vertebrate protein databases (other than human keratin contamination) was collagen and that this was also true for the 160- to 600-thousand-year-old mastodon fossil bone,

¹Beth Israel Deaconess Medical Center, Boston, MA 02115, USA. ²Harvard Medical School, Boston, MA 02115, USA. ³North Carolina State University, Raleigh, NC 27695, USA. ⁴Matrix Science Ltd., London, UK.

*To whom correspondence should be addressed. E-mail: jasara@bidmc.harvard.edu

Table 1. The list of peptide sequences extracted from a 68-million-year-old *T. rex* fossil bone with Mascot search engine scores, molecular weight information, and the method for validation. m/z(obsd) represents the experimental mass/charge ratio acquired from the LTQ ion trap mass spectrometer. Mr(Calc)

represents the theoretical molecular mass based on the peptide sequence. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

Spectrum no.	m/z (obsd)	Mr(Calc)	Mass error	Rank	Mascot score	Expectation value	Validation	Peptide
13032	450.04	897.50	0.56	1	22.1	49	ostrich peptide	GVVGLP(OH)GQR
19231	581.59	1161.59	-0.42	1	32.0	5.5	synthetic peptide	GVQGGP(OH)GPQGPR
28920	787.10	1571.77	0.42	1	42.4	0.44	multiple matches	GATGAP(OH)GIAGAP(OH)GFP(OH)GAR
29032	790.06	1577.82	0.28	1	46.8	0.16	search stats.	GLPGESGAVGPAGPIGSR
26727	730.53	1458.69	0.36	1	62.3	0.0048	search stats.	GSAGPP(OH)GATGFP(OH)GAAGR
22865	645.76	1289.64	-0.12	1	71.1	0.00067	search stats.	GAPGPQGPSGAP(OH)GPK

where many collagen fragments were found and where no challenge to the identification has been raised. This result, combined with the anti-collagen antibody binding results, makes a very strong argument that the peptides are authentic. Clearly, collagen survives better than other proteins in bone.

Database search tools are intended for rapid screening of large proteomics data sets against large protein databases. Peptide matching is based on limited information, primarily fragment ion mass values. Neither Mascot nor Sequest makes use of the variation in fragment ion intensities within a spectrum, which is characteristic of the sequence. In a case such as this, where confidence in an individual match is unusually important, the relative intensity information provides an independent basis for confirming or rejecting a putative match.

We previously validated the second-lowest-scoring GVVGLP(OH)GPQGPR sequence using a synthetic peptide (1). Other peptides were validated by high database matching scores against the target protein database. To validate the *T. rex* collagen database match for the peptide sequence GVVGLP(OH)GQR in question by Pevzner *et al.* (9), we compared the spectrum of a peptide of the same sequence from ostrich (strong spectral intensity) with that from *T. rex* (weak intensity) to quantify the degree of spectral similarity. The spectral library search tool was MS Search 2.0, from the National Institute of Standards (Gaithersburg, MD). The ostrich spectrum was added to the existing libraries containing ~200,000 peptide spectra, and the *T. rex* spectrum was searched against these libraries using default settings. The dot product score for the match to the ostrich spectrum was 934, and the score for the second-best match was 89. According to the documentation (10), a perfect match results in a value of 999, and spectra with no peaks in common result in a value of 0. As a

general guide, 900 or greater is an excellent match; 800 to 900, a good match; and 700 to 800, a fair match. Less than 600 is a very poor match. The results are shown in Fig. 1. We are not surprised that the *T. rex* spectrum shows increased noise peaks because the spectrum is much weaker than the ostrich spectrum and was derived from a sample with high levels of unidentified brown-colored contaminants.

In the analogy used by Pevzner *et al.* (9), the monkey and the typewriter represent the mass spectrometer and the boy and his dictionary represent the database search. The authors suggest that, because 7 out of 10,000 six-letter words are likely to be found in a dictionary by chance, the same is true for mass spectra. Their analogy is inaccurate because the mass spectrometer is not producing spectra of uniform quality, containing six letters of information each time. In any large-scale proteomics experiment, most of the collected spectra are of low quality or nonpeptidic and are thus incapable of giving any meaningful matches. For this reason, the false discovery rate adopted by the proteomics community is the ratio of the number of matches found in a decoy database to the total number of matches found in the combined target and decoy databases (7). The ratio of matches to total spectra is simply not useful.

Pevzner *et al.* further argue that many other sequences would fit equally well to the GVVGLP(OH)GQR spectrum. This is not unreasonable. However, no one has suggested that this is the best match out of all possible sequences. Practically every database search result reported in the literature would fail if this were the criterion. Chicken collagen is a very reasonable assignment for this peptide on the basis that (i) collagen is known to be present from statistically significant matches to other collagen peptides; (ii) most likely, contaminant proteins (keratins, caseins, bacteria, and the like)

are present in the Swiss-Prot database, and none of these give a better match to this spectrum; (iii) multiple search engines find the same match, independently; and (iv) the spectrum shows extremely strong and statistically significant similarity to a spectrum of the authentic peptide from ostrich. The presence of hydroxylated prolines in the majority of peptides recovered from dinosaur bone lends strong support for the peptides being derived from collagen, because this posttranslational modification is most frequently found in collagen. The other sequences that Pevzner *et al.* list as possibilities for this spectrum are not consistent with all of the observed major fragment ions.

Pevzner *et al.* state that there are thousands of peptides that match the fifth *T. rex* spectrum reported in (1) even better than the alleged *T. rex* peptide GVVGLP(OH)GQR (FPR = 1.3×10^{-6}) and that “[t]his implies that if one tries to match this spectrum against a small database of 10^6 amino acids, there is a good chance of matching this spectrum simply by chance.” We are not sure where the 1.3×10^{-6} value comes from, but the suggestion that you can get an equally good match by chance in a database of 1 million residues is clearly incorrect. Swiss-Prot contains 130 million residues representing nearly 360,000 proteins, and GVVGLP(OH)GQR is the highest-scoring match. If Pevzner *et al.* were correct, this match should have been lost in 130 equally good or better matches.

Overall, the comment by Pevzner *et al.* does not change any of our original conclusions that collagen fragments from a 68-million-year-old *T. rex* fossil bone were extracted and sequenced and that they match better as a group to chicken collagen than to any other protein from any other organism. It is obvious that the other peptides are strong hits, and we find it interesting that the authors only target the lowest-scoring peptide match in their critique. Ultimately, the hypothesis that our data support, preservation of original biomolecules in dinosaur fossil material, will be more robustly supported with additional data on other specimens currently under study. The data gathered to date from this and other dinosaur material continue to support the hypothesis of preservation.

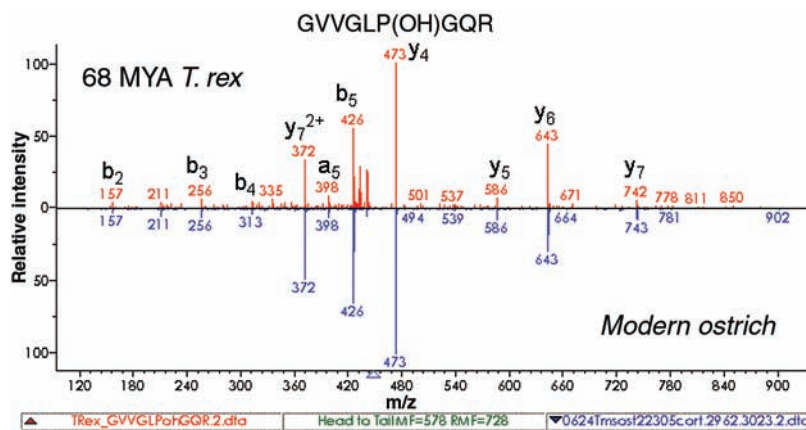


Fig. 1. The MS/MS spectral comparison of the 68-million-year-old *T. rex* peptide sequence GVVGLP(OH)GQR compared to a modern ostrich version of the same sequence using the computational algorithm MS Search 2.0. The spectral similarity score of 934 is very high and validates the *T. rex* sequence. The sequence was a top match for collagen alpha 1 type I against the Swiss-Prot protein database.

References

1. J. M. Asara, M. H. Schweitzer, L. M. Freemark, M. Phillips, L. C. Cantley, *Science* **316**, 280 (2007).
2. J. M. Asara *et al.*, *Science* **317**, 1324 (2007).
3. J. M. Asara, M. H. Schweitzer, *Science* **319**, 33d (2008).
4. M. H. Schweitzer *et al.*, *Science* **316**, 277 (2007).
5. D. N. Perkins *et al.*, *Electrophoresis* **20**, 3551 (1999).
6. J. Eng *et al.*, *J. Am. Soc. Mass Spectrom.* **5**, 976 (1994).
7. I. Nesvizhskii *et al.*, *Nat. Methods* **4**, 787 (2007).
8. J. E. Elias, S. P. Gygi, *Nat. Methods* **4**, 207 (2007).
9. P. A. Pevzner, S. Kim, J. Ng, *Science* **321**, 1040 (2008); www.sciencemag.org/cgi/content/full/321/5892/1040b.
10. H. Lam *et al.*, *Proteomics* **7**, 655 (2007).

18 April 2008; accepted 28 July 2008
10.1126/science.1157829