Comment on "Protein Sequences from Mastodon and *Tyrannosaurus rex* Revealed by Mass Spectrometry"

Pavel A. Pevzner,* Sangtae Kim, Julio Ng

Asara *et al.* (Reports, 13 April 2007, p. 280) reported sequencing of *Tyrannosaurus rex* proteins and used them to establish the evolutionary relationships between birds and dinosaurs. We argue that the reported *T. rex* peptides may represent statistical artifacts and call for complete data release to enable experimental and computational verification of their findings.

Imagine a monkey typing random keys on a typewriter and let us assume that the monkey is given 100,000 attempts to generate six-letter words. One would be surprised if the monkey typed a six-letter word from Webster's dictionary on the first attempt; indeed, the probability of this is rather low. However, nobody would be surprised if some of the 100,000 words turned out to be correctly spelled English words.

Now imagine a boy who watches the monkey and discovers that 7 out of 100,000 words are actually spelled correctly. The boy is so surprised that he writes a paper called "My monkey can spell!" and publishes it in a scientific journal. Some scientists are not convinced, and they request the list of all words the monkey generated in addition to the seven correctly spelled words. The boy does not understand the reason for such requests; indeed, if all other words are just junk, what is the point of asking for them?

One often feels like a monkey (and a boy) when trying to interpret peptide mass spectra. Indeed, a randomly chosen spectrum can easily match a word in Webster (if English letters are interpreted as amino acids) or in any protein database. Scientists fail to interpret the lion's share of mass spectra generated worldwide, resulting in billions of uninterpreted or "junk" spectra. If we matched these junk spectra against Webster we would surely find that some of them spell English words. Unfortunately, we would not be able to publish a paper called "Mass spectrometers can spell!" because false protein identifications are unavoidable in the field of proteomics. Scientists learned how to cope with them by establishing the Proteomics Publication Guidelines that require authors to provide the error rates of their identifications.

Asara *et al.* (*1*) reported the sequencing of proteins from 68-million-year-old *T. rex* fossils and established similarities between dinosaur and chicken genomes. The authors generated seven *T. rex* peptides by matching mass spectra against collagen proteins. They did not reveal all generated spectra and never specified exactly how many spectra were generated. Because there are false identifications in every mass spectrometry experiment, without addressing the statistical significance problem, the results of (*1*) are no more convincing than the first sensational report of dinosaur DNA published in *Science* more than a decade ago (*2*).

In the spring of 2007, we notified Asara and *Science* of concerns about the statistical significance of some of the peptides. In a subsequent clarification letter (*3*), Asara *et al.* acknowledged some of the problems with their analysis in (*1*). In particular, they stated, "We have determined that one of the reported *T. rex* spectra for the peptide GLVGAPGLRGLPGK is statistically insignificant when searched against large protein databases...." (*3*, *4*). By admitting this point, Asara *et al.* implicitly (and probably unknowingly) acknowledged a much bigger problem with their original study (*1*). Indeed, the statistical significance (e.g., false positive rate or FPR) is a number that needs to be computed, but Asara *et al.* (*3*) never described how they computed statistical significance, and it is not clear whether they tried. If they computed the statistical significance, they would discern that other *T. rex* peptides do not fare much better. For example, it turns out that there are thousands of peptides that match the fifth *T. rex* spectrum reported in (*1*, *3*) even better than the alleged *T. rex* peptide GVVGLP*GQR [FPR or spectral probability equal to 1.3×10^{-6} (*5*)]. This 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. Or, equivalently, if one tries to match 1000 arbitrary spectra of similar quality against an arbitrary database of 1000 amino acids, there is a good chance to find an interpretation that is even better than the alleged *T. rex* peptide GVVGLP*GQR.

Asara *et al.* (*3*) must have generated at least hundreds of thousands of spectra, and their database is much larger than 1000 amino acids. This immediately characterizes the peptide GVVGLP*GQR as a statistical artifact, in addition to GLVGAPGLRGLPGK, which the authors acknowledge in (*3*). If Asara *et al.* (*1*) stand by the statistical significance argument given in (*3*), they should question all of the *T. rex* peptides identified in (*1*). Only one of these peptides was supported by chemical synthesis with a spectral correlation coefficient of 0.71, which although borderline significant, may also represent homeometric (*6*), but not identical, peptides. We argue that most of the peptides with GVVGLP*GQR-like spectra (e.g., 10,919 peptides with better InsPecT scores or 10,294 peptides with better X!Tandem scores than GVVGLP*GQR) would have produced spectra that are somewhat similar to the spectrum of GVVGLP*GQR, thus calling for more extensive synthesis-based verification of the results in (*1*). For example, one could potentially synthesize GVVGLP*GQR and discover that the resulting spectrum "looks like" one of the *T. rex* spectra, thus "proving" that GVVGLP*GQR is indeed a *T. rex* peptide. In this case, it is puzzling how Asara *et al.* selected the "correct" statistically insignificant peptide among hundreds of other statistically insignificant peptides. For example, peptides RVGLRAAR, RVGLPTKK, RVGP*PTKK, and thousands of others represent better InsPecT and X!Tandem spectral interpretations than GVVGLP*GQR (table S1) (supporting online material). If one is willing to argue that GVVGLP*GQR is a valid identification based on peptide synthesis, the peptides RVGLRAAR, RVGLPTKK, and RVGP*PTKK should also be synthesized and compared to the *T. rex* spectrum. Extraordinary science requires extraordinary proofs.

Since the publication of their report (*1*), Asara *et al.* have reinterpreted (*3*) four out of seven of the *T. rex* peptides originally reported. The most likely outcome of further criticism is that Asara and colleagues will continue changing their original interpretations until the critics give up. So far, five out of six of the remaining significant *T. rex* peptides have already emerged as identical to chicken peptides. Maybe *T. rex* was a chicken after all?

Recently, a group of 27 mass spectrometrists, bioinformaticians, and dinosaur experts published an insightful criticism of the *T. rex* protein analysis (*7*). Still, Asara and Schweitzer (*8*), refused to acknowledge the problems with their analysis. It is now the turn of the mass spectrometry community to question whether the monkey can actually spell. It is very easy to check; just ask the boy how many words (e.g., spectra) the monkey has generated and what tests of statistical significance were used to compute FPR. With this information in hand, the scientists can finally match all dinosaur proteins against Webster's dictionary to see whether mass spectrometers can spell and whether *T. rex* was a chicken.

Department of Computer Science and Engineering, University of California, San Diego, 9500 Gilman Drive, Mail Code 0404, La Jolla, CA 92093–0404, USA.

*To whom correspondence should be addressed. E-mail: ppevzner@cs.ucsd.edu

References and Notes
1. J. M. Asara, M. H. Schweitzer, L. M. Freimark, M. Phillips, L. C. Cantley, *Science* **316**, 280 (2007).

2. S. Woodward, N. Weyand, M. Bunnell, *Science* **266**, 1229 (1994).
3. J. M. Asara *et al.*, *Science* **317**, 1324 (2007).
4. 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.
5. S. Kim, N. Gupta, P. A. Pevzner, *J. Proteome Res.* **7**, 3354 (2008).
6. A. M. Frank, M. M. Savitski, M. L. Nielsen, R. A. Zubarev, P. A. Pevzner, *J. Proteome Res.* **6**, 114 (2007).
7. M. Buckley *et al.*, *Science* **319**, 33 (2008); www.sciencemag.org/cgi/content/full/319/5859/33c.
8. J. M. Asara, M. H. Schweitzer, *Science* **319**, 33 (2008); www.sciencemag.org/cgi/content/full/319/5859/33d.