LETTER

In defense of statistical methods for detecting positive selection

In a highly publicized article, Nozawa et al. (1) claimed that the branch-site model (BSM) (2, 3) was unreliable because it produced excessive false positives in their simulation experiment. BSM uses a likelihood ratio test to detect positive selection that affects particular branches and codons in protein-coding genes, indicated by accelerated nonsynonymous substitution rates. The authors' conclusion, if true, would be important. But it is contradicted by their simulation results.

The study generated 14,000 datasets under a null model that postulated no positive selection and found that BSM falsely detected positive selection in 32 cases. Nozawa et al. (1) claimed that those false positives were "not supposed to be obtained theoretically" and indicated "abnormal behaviors" of the likelihood ratio test. Those claims are false: the false-positive rate is only 0.23% (32 of 14,000), much lower than the nominal significance level (5%). Contrary to Nozawa et al.'s claims, the test is thus conservative. Nozawa et al. preferred a parsimony-based approach, which averages rates over the whole protein and achieved 0% false-positive rate in their simulation. The authors did not examine the power of the tests. In previous simulations (4), such parsimony-based methods were found to have little power, even when the likelihood ratio tests detected positive selection with ≈100% power.

We suggest that sensible use of statistical methods for detecting positive selection such as BSM (5) is valuable in comparative analysis of genomic data. They can generate biological hypotheses for experimental verification, narrowing down possibilities for test in the laboratory. Nozawa et al.'s results, interpreted correctly, support this view, as do many studies in which the statistical predictions were validated in the laboratory.

***Ziheng Yang*[a,1], *Rasmus Nielsen*[b], *and Nick Goldman*[c]**

[a]*Department of Biology, Galton Laboratory, University College London, London NW1 2HE, United Kingdom;* [b]*Departments of Integrative Biology and Statistics, University of California at Berkeley, Berkeley, CA 94720-3140; and* [c]*EMBL–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom*

1. Nozawa M, Suzuki Y, Nei M (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci USA* 106:6700–6705.
2. Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118.
3. Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22:2472–2479.
4. Wong WSW, et al. (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051.
5. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.

Author contributions: Z.Y., R.N., and N.G. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: z.yang@ucl.ac.uk.