necessary regulatory regimes to facilitate responsible biotech and establish nurturing environments for biotech initiatives to thrive. If they do so in an expedient and efficient manner, biotech could ultimately make a contribution to alleviating the plight of Africa's nutritiously deficient and catalyzing Africa's renaissance.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors would like to acknowledge Muffy Koch, Daniel Kamanga, Adrian Dubock and Andy Shoyer for helpful comments on an earlier draft of this paper. J.A.S. and A.S.D. are supported by a grant from the Bill and Melinda Gates Foundation through the Grand Challenges in Global Health Initiative.

Jerome Amir Singh[1,2] & Abdallah S Daar[3]

[1]CAPRISA, Private Bag X7, Congella, 4013, Durban, South Africa, and Howard College School of Law, University of KwaZulu-Natal, King George V Avenue, Durban, South Africa. [2]Department of Public Health Sciences and Joint Center for Bioethics, University of Toronto, 88 College Street, Toronto, Canada. [3]McLaughlin-Rotman Centre for Global Health, Program on Life Sciences, Ethics and Policy, University Health Network; McLaughlin Centre for Molecular Medicine at University of Toronto, MaRS Centre, South Tower, 101 College Street, Suite 406 Toronto, Ontario, Canada M5G 1L7.
e-mail: Singhj9@ukzn.ac.za

1. Koenig, R. *Science* **315**, 748 (2007).
2. http://www.iisd.ca/africa/brief/briefing0702e.html
3. http://www.africa-union.org/root/AU/Conferences/2007/November/HRST/AMCOST/docs/pdf/AU-EXP-ST-16_III_-ENG-Regional%20Workshop%20Modified%20Orga.pdf
4. http://www.tralac.org/scripts/content.php?id=2843
5. http://www.gcgh.org/Projects/ImproveNutrition/NutrientRichPlants/default.htm
6. European Commission. Biotech, European Communities—Measures Affecting the Approval and Marketing of Biotech Products, Reports of the Panel, WT/DS291/R, WT/DS292/R, WT/DS293R, Circulated 29 September 2006. (EC, Brussels, 2006).
7. http://www.tralac.org/scripts/content.php?id=4520
8. http://www.scidev.net/Editorials/index.cfm?fuseaction=readEditorials&itemid=220&language=1
9. http://www.nepadst.org/biopanel/index.shtml
10. http://www.biowatch.org.za/main.asp?include=about/faq.html
11. http://www.scidev.net/Editorials/index.cfm?fuseaction=readEditorials&itemid=220&language=1
12. http://www.scidev.net/Editorials/index.cfm?fuseaction=readEditorials&itemid=220&language=1
13. High-Level African Panel on Modern Biotechnology (Calestous Juma and Ismail Serageldin, co-chairs). *Freedom to Innovate: Biotechnology in Africa's Development* (African Union, Addis Ababa, Ethiopia, July 2006).http://www.nepadst.org/doclibrary/pdfs/biotech_africarep_2007.pdf

BLOSUM62 miscalculations improve search performance

To the editor:

The BLOSUM[1] family of substitution matrices, and particularly BLOSUM62, is the *de facto* standard in protein database searches and sequence alignments. In the course of analyzing the evolution of the Blocks database[2], we noticed errors in the software source code used to create the initial BLOSUM family of matrices (available online at ftp://ftp.ncbi.nih.gov/repository/blocks/unix/blosum/blosum.tar.Z). The result of these errors is that the BLOSUM matrices—BLOSUM62, BLOSUM50, etc.—are quite different from the matrices that should have been calculated using the algorithm described by Henikoff and Henikoff[1]. Obviously, minor errors in research, and particularly in software source code, are quite common. This case is noteworthy for three reasons: first, the BLOSUM matrices are ubiquitous in computational biology; second, these errors have gone unnoticed for 15 years; and third, the 'incorrect' matrices perform better than the 'intended' matrices.

The error that had the most impact was an incorrect normalization during a weighting procedure; this procedure, the error and its impact are discussed in greater detail in **Supplementary Note** online. Recalculated matrices are also available in the **Supplementary Note**, and differences from the original matrices are highlighted. These two matrices differ in 15% of their positions. Both the corrected and the original source code are also available through a link in the **Supplementary Note**. It is worth noting that the relevant comparison for BLOSUM62 is not with the revised BLOSUM62 (which we call RBLOSUM62) because matrices can only be 'fairly' compared if they have the same relative entropy[3]. We found that this relative entropy (when calculated from raw matrix values), which is a measure of the information content in a substitution matrix, was inflated in the BLOSUM matrices due to the errors. Thus, BLOSUM62 is best 'fairly' compared with RBLOSUM64 based on raw matrix value entropies. (Comparisons based on rounded matrix values show largely similar results and are presented in **Supplementary Note**.)

To investigate the effects of these differences, we used the pairwise sequence comparison evaluation methods and software developed by Price *et al.*[4]. We compared matrices' performance using two alignment algorithms: the exhaustive Smith-Waterman[5] approach (as implemented in ssearch[6,7]), and the heuristic BLAST[8] approach. These searches were used to determine each matrix's effectiveness at locating distant homologs from within the ASTRAL database[9], a set of hand-curated structure-based protein homologs derived from the SCOP database[10]. We used mostly default parameters for both search methods, with notable exceptions of gap penalties (which were varied to find the optimal values for each matrix) and statistical parameters for BLAST (which were calculated using routines provided by Stephen Altschul (personal communication)). These search methodologies, comprising over four billion individual sequence alignments, are described in detail in the **Supplementary Note**.

Surprisingly, 'fixing' the matrices does not improve performance (see **Fig. 1**); the RBLOSUM64 matrix performs consistently worse than BLOSUM62 across a wide range of errors per query cutoffs using both Smith-Waterman and BLAST search tools. (An errors-per-query cutoff is approximately equivalent to the E-value cutoff that one would use in a BLAST search, but is calculated by averaging the results of numerous searches.) Although the performance difference is statistically significant, it is, however, relatively small in magnitude. More detailed analyses about the statistically significant performance differences caused by the errors, as well as the potential origins of these performance differences, are provided in the **Supplementary Note**.

We find it interesting that the BLOSUM62 matrix is used every day (and more interesting still that its derivation is a common topic in computational biology classes), and yet we can find no previously published mention of any of the errors discussed here. We did find that some of the errors were fixed in later tangential work by the original authors[11], but the 'correct' matrices have never been published or adopted. We also note that the existence of statistically significant improvements due to (essentially random) software errors supports the notion that there is significant room for improvement in our understanding of protein evolution. Of course, software errors are quite common and nothing

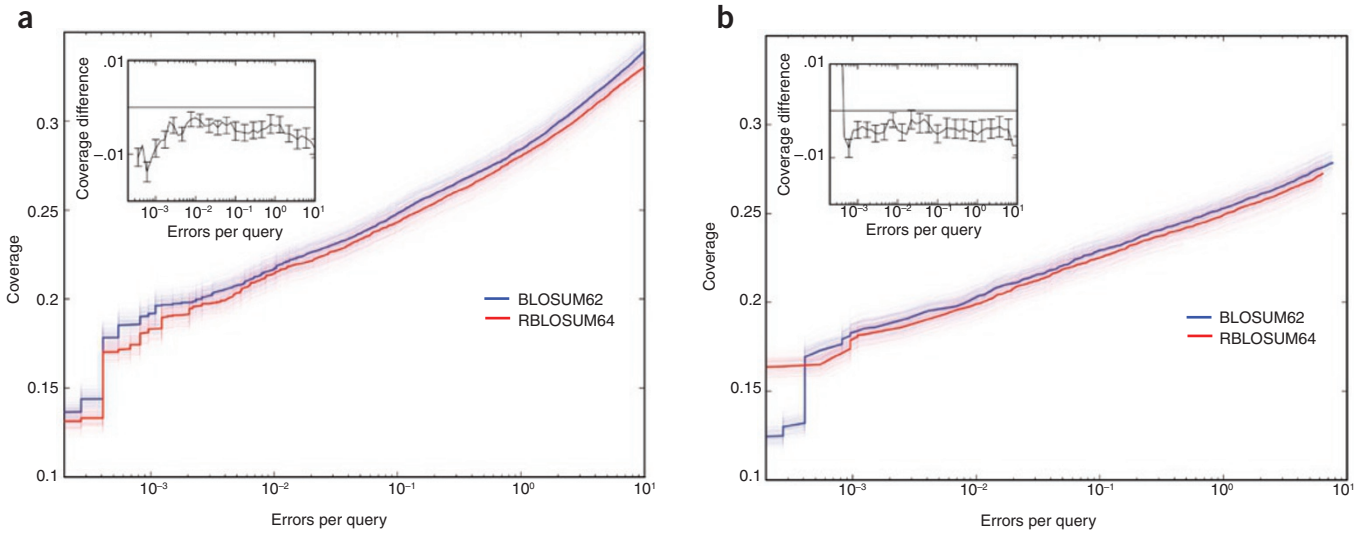

Figure 1 Performance difference between BLOSUM62 and RBLOSUM64 as assessed by pairwise sequence comparison evaluation (PSCE) tools[4]. (**a**) Performance when using ssearch. (**b**) Performance when using BLAST. Coverage is the fraction of true homologs that are identified at a given errors-per-query threshold. Notice that the blue line is above the red line over a wide range of errors-per-query values; roughly, this means that the BLOSUM62 matrix finds more true homologs when searching a protein database than RBLOSUM64 finds. Thick lines represent the original data, whereas thinner lines represent individual bootstrap replicates that are used to calculate the error bars in the inset. Concerted Bayesian bootstrapping is used in the PSCE software[4] to determine the statistical significance of the difference in matrices' effectiveness by evaluating whether slightly different reference databases would have yielded different performance. The insets plot the mean difference in performance between the two matrices. Error bars are 95% confidence intervals, such that error bars that do not cross the origin indicate statistically significant differences between the two matrices' performance.

special; however, it is at least a curiosity that these errors stayed buried for so long and have been improving BLAST searches (ever so marginally) for the past 15 years.

Note: Supplementary information is available on the Nature Biotechnology website.

***Mark P Styczynski*[1,5,6], *Kyle L Jensen*[2,3,6], *Isidore Rigoutsos*[4] *& Gregory Stephanopoulos*[1]**

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Room 56-469c, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [2]Public Intellectual Property Resource for Agriculture, University of California, One Shields Avenue, Department of Plant Sciences, Plant Reproductive Biology Building–Mail Stop 5, Davis, California 95616, USA. [3]Harvard-MIT Health Sciences and Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [4]IBM Research Division, Thomas J. Watson Research Center, 1101 Kitchawan Road, PO Box 218, Yorktown Heights, New York 10598, USA. [5]Current address: Broad Institute, Room 6175X, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [6]These authors contributed equally to this work.

1. Henikoff, S. & Henikoff, J.G. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919 (1992).
2. Henikoff, J.G., Greene, E.A., Pietrokovski, S. & Henikoff, S. *Nucleic Acids Res.* **28**, 228–230 (2000).
3 Altschul, S.F. *J. Mol. Biol.* **219**, 555–565 (1991).
4. Price, G.A., Crooks, G.E., Green, R.E. & Brenner, S.E. *Bioinformatics* **21**, 3824–3831 (2005).
5. Smith, T.F. & Waterman, M.S. *J. Mol. Biol.* **147**, 195–197 (1981).
6. Pearson, W.R. *Methods Enzymol.* **183**, 63–98 (1990).
7. Pearson, W.R. *Genomics* **11**, 635–650 (1991).
8. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. *J. Mol. Biol.* **215**, 403–410 (1990).
9. Brenner, S.E., Koehl, P. & Levitt, M. *Nucleic Acids Res.* **28**, 254–256 (2000).
10. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. *J. Mol. Biol.* **247**, 536–540 (1995).
11. Henikoff, S., Henikoff, J.G., Alford, W.J. & Pietrokovski, S. *Gene* **162**, GC17–GC26 (1995).