

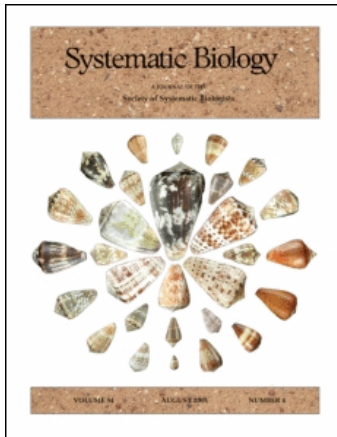
This article was downloaded by: [Wageningen UR]

On: 17 December 2008

Access details: Access Details: [subscription number 789193022]

Publisher Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Systematic Biology

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t713658732>

## Statistical Assignment of DNA Sequences Using Bayesian Phylogenetics

Kasper Munch <sup>a</sup>; Wouter Boomsma <sup>b</sup>; John P. Huelsenbeck <sup>a</sup>; Eske Willerslev <sup>c</sup>; Rasmus Nielsen <sup>d</sup>

<sup>a</sup> Department of Integrative Biology, University of California, Berkeley, California, USA <sup>b</sup> Bioinformatics Centre, University of Copenhagen, Denmark <sup>c</sup> Department of Biology and Centre for Ancient Genetics, University of Copenhagen, Denmark <sup>d</sup> Department of Biology, University of Copenhagen, University of Copenhagen, Denmark

First Published on: 01 October 2008

**To cite this Article** Munch, Kasper, Boomsma, Wouter, Huelsenbeck, John P., Willerslev, Eske and Nielsen, Rasmus(2008)'Statistical Assignment of DNA Sequences Using Bayesian Phylogenetics',*Systematic Biology*,57:5,750 — 757

**To link to this Article:** DOI: 10.1080/10635150802422316

**URL:** <http://dx.doi.org/10.1080/10635150802422316>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

## Statistical Assignment of DNA Sequences Using Bayesian Phylogenetics

KASPER MUNCH,<sup>1</sup> WOUTER BOOMSMA,<sup>2</sup> JOHN P. HUELSENBECK,<sup>1</sup> ESKE WILLERSLEV,<sup>3</sup> AND RASMUS NIELSEN<sup>4</sup>

<sup>1</sup>Department of Integrative Biology, University of California, Berkeley, California 94720-3140, USA

<sup>2</sup>Bioinformatics Centre, University of Copenhagen, Ole Maaloes Vej 5, 2200 København N, Denmark and <sup>3</sup>Department of Biology and Centre for Ancient Genetics, University of Copenhagen, Universitetsparken 15, 2100 København Ø, Denmark

<sup>4</sup>Department of Biology, University of Copenhagen, University of Copenhagen, Universitetsparken 15, 2100 København Ø, Denmark and Departments of Integrative Biology and Statistics, University of California, Berkeley, California 94720-3140, USA

**Abstract.**—We provide a new automated statistical method for DNA barcoding based on a Bayesian phylogenetic analysis. The method is based on automated database sequence retrieval, alignment, and phylogenetic analysis using a custom-built program for Bayesian phylogenetic analysis. We show on real data that the method outperforms Blast searches as a measure of confidence and can help eliminate 80% of all false assignment based on best Blast hit. However, the most important advance of the method is that it provides statistically meaningful measures of confidence. We apply the method to a re-analysis of previously published ancient DNA data and show that, with high statistical confidence, most of the published sequences are in fact of Neanderthal origin. However, there are several cases of chimeric sequences that are comprised of a combination of both Neanderthal and modern human DNA. [Assignment; barcoding; Bayesian; phylogenetics.]

The identification of organic material through comparisons of DNA sequences from a sample to DNA sequences from a database is an important research tool in a number of scientific disciplines. In the zoological and ecological literature, identification of unknown specimens based on cytochrome oxidase I (COI) has become known as DNA barcoding (Floyd et al., 2002; Hebert et al., 2003; Remigio and Hebert, 2003; Moritz and Cicero, 2004). DNA barcoding has found a wide range of applications, from identification of specimens in conservation biology and molecular ecology to identification of birds that have collided with aircraft. A similar methodology is applied in metagenomics (Tringe and Rubin, 2005; Breitbart et al., 2002; Venter et al., 2004; Rusch et al., 2007; Yooseph et al., 2007) where genomic sequences from environmental samples are obtained and compared to database sequences.

The topics of this article are the methodological issues relating to the assignment of DNA sequences to taxa represented in a sequence database. The classical procedure for such identification has been the use of Blast searches (Altschul et al., 1997). There are, however, at least three statistical problems associated with this: (1) Blast searches provide a score based on local alignments and not global alignments, leading to a loss of information; (2) Blast searches ignore the population genetic and phylogenetic issues associated with species identification; and (3) the measures of confidence associated with Blast searches represent significance of local sequence similarity and not significance of taxonomic assignment. Blast thus offers no information to help researchers choose among multiple close matches. Whereas the local alignment problem can be circumvented using global alignments, the remaining two problems cannot be addressed without a statistical evaluation of the phylogenetic associations among species.

Several new methods have been developed that attempt to address the problems associated with the use of Blast to identify sequences (Matz and Nielsen, 2005; Meyer and Paulay, 2005; Steinke et al., 2005; Nielsen and Matz, 2006; Abdo and Golding, 2007); most of these

methods focus on identifying species affiliation. This question is difficult to address as the evolutionary relationship among genetic markers may not truly reflect the evolutionary relationship among species. In cases where reciprocal monophyly cannot safely be assumed, an analysis quantifying within- and between-species genetic variation forms a more correct basis of assignment. Such analyses, however, require a comprehensive database coverage that is generally not available to the biologist. In this article we describe a purely phylogenetic solution to the DNA barcoding problem. We will not address the species problem but instead attempt to devise an automated method for the assignment of sample sequences to taxa based on the position of the sample sequence in the phylogeny of life. This method leads to improved accuracy and, importantly, it provides a measure of statistical confidence associated with the barcoding assignment.

### METHODS

Sequences can be assigned to taxa using a number of different statistical frameworks. Here we pursue a Bayesian approach that allows us to estimate the probability that the sample sequence is part of a monophyletic group, identified with database sequences. We will thus not address the population genetic questions latent in species assignment but reduce the question to a purely taxonomic, or cladistic, question of assigning the sample sequence to a particular clade in an established phylogeny. The procedure is summarized graphically in Figure 1 and described in detail below.

In the Bayesian framework (e.g., Pawitan, 2001), the relevant probability of interest is the posterior probability that the query species belong to a particular taxonomic group:

$$P(X \in T_i | X, \mathbf{D}) = \frac{P(X, \mathbf{D} | X \in T_i)P(X \in T_i)}{\sum_{j=1}^k P(X, \mathbf{D} | X \in T_j)P(X \in T_j)}$$

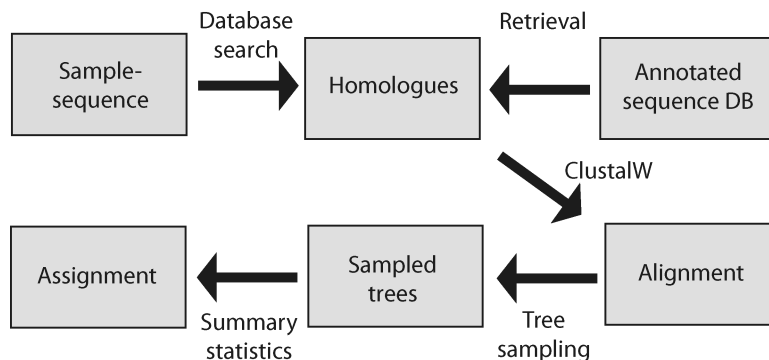


FIGURE 1. Flowchart of the assignment procedure. A set of homologues is compiled using information from Blast searches and annotation from NCBI's Taxonomy database. The relevant sequences are retrieved from GenBank and aligned using ClustalW. Based on the resulting multiple alignment a large number of phylogenetic trees are sampled and these are then used to calculate posterior probabilities of assignment.

where  $X$  is the sample-sequence,  $T_i$  is taxon  $i$ , and  $\mathbf{D}$  is the set of database sequences representing  $k$  disjoint groups. Because the denominator contains a sum over sequences represented in a database, the probability calculated using this approach is the probability of assignment to a taxonomic group given that the sequence has to be assigned to one of the groups represented in the database.

The posterior probability involves a summation over all possible phylogenetic trees and, for each tree, a multiple integral over all combinations of substitution parameters. Hence, the posterior probability cannot be evaluated analytically. However, Markov chain Monte Carlo (MCMC; e.g., Huelsenbeck and Ronquist, 2001) can be used to sample trees in proportion to their posterior probabilities. The fraction of the time the MCMC sampler visits trees that place the sample sequence within a specific monophyletic group ( $X \in T_i$ ) is a valid approximation of the posterior probability that the query sequence falls within that group.

Ideally, each sample sequence should be compared to the entire tree of life or as much of it as is represented in the available sequence database. For obvious reasons this is not possible, and a heuristic is required to extract a limited representation of the database. To this end we use sequence homology between the sample sequence and sequences obtained using remote Blast searches against GenBank. A taxonomic annotation for each homologue is retrieved from NCBI's taxonomy browser. Homologues with insufficient taxonomic annotation are disregarded.

The vast majority of taxa represented in the sequence database are not relevant to the analysis because the posterior probabilities of grouping monophyletically with these taxa are not appreciably large. The bulk of sequence homologues representing these taxa can be avoided by including only homologues with a Blast score of at least half that of the best matching homologue.

More often than not, however, this relative similarity cutoff does not reduce the number of sequence homologues to a set that can be handled computationally. To obtain the best possible taxonomic coverage in a limited set, only the best-matching sequence homologue for each species is included. If available, up to 30 different species homologues are included. If, at this point, the

relative cutoff described above has not been reached, up to 20 homologues providing further taxonomic diversity are added progressively including up to 10 genera, six families, five orders, three classes, and two phyla in the set. If the relative cutoff is reached before 50 homologues have been included in the set, additional sequences are added for the species already represented in the set by including homologues previously rejected as suboptimal representatives for the species.

The analysis is discontinued if the compiled set does not include at least five Blast hits with an E-value below 0.1. An alignment of the sample sequence and the set of homologues is produced using ClustalW in slow/accurate mode with default parameters.

Like any other comparable method, our approach can only assign sequences to taxonomic groups represented in the database. Hence, if only a single taxon represents the clade in which the sample sequence belongs, the sample sequence will be assigned to this taxon with probability one. We have in our approach made no attempt to model the structure and sampling representation of the databases to evaluate the probability that the sequence truly belongs to some other taxon not represented in the database.

A computer program, written in C++ by J.P.H., performs the MCMC analysis. This program takes as input the sequence alignment and a file describing any constraints on the topology of the tree. The constraints are of the form of a backbone constraint. In other words, the constraint tree may include only a subset of the sequences included in the alignment. Here, all sequences except the sample sequence are included in a constraint tree specified by the taxonomic annotation. The program assumes that nucleotide substitutions occur according to the general time reversible model (Tavare, 1986) and assumes that the rate of substitution at a site is a random variable drawn from a mean-one gamma distribution (Yang, 1993, 1994). The Markov chain explores the space of all of the parameters of the model, including the substitution rates, nucleotide frequencies, gamma-shape parameter, and topology/branch lengths of the tree subject to the specified constraints. The proposal mechanisms for all of the non-tree parameters have been described

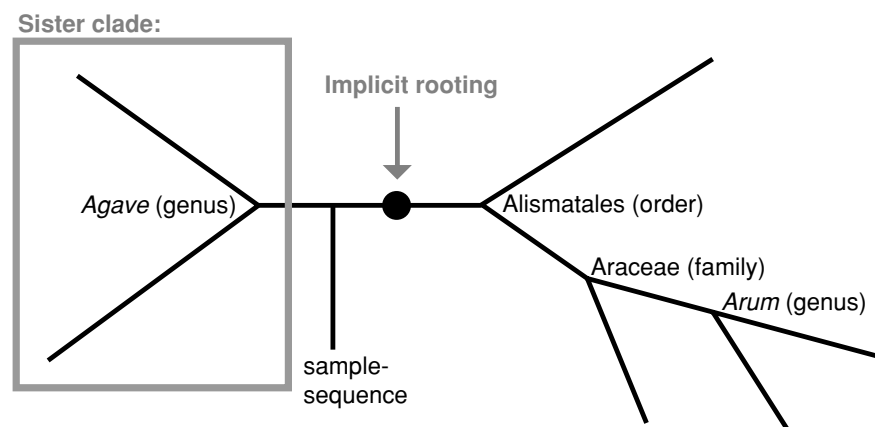


FIGURE 2. Assignment of the sample sequence in each sampled tree is done by assuming the root implied by the taxonomic annotation of homologues and then recording the consensus taxonomy for all members of the sister clade from the highest taxonomic level to the most specific level shared by all clade members.

elsewhere (e.g., Huelsenbeck et al., 2004). We propose new topologies using a stochastic variant of the SPR (subtree pruning and regrafting) tree perturbation often used to find optimal trees in a parsimony or maximum likelihood framework. Ten thousand unrooted trees sampled from the MCMC analysis are analyzed to obtain posterior probabilities of assignment to all taxa represented in the compiled set of homologues.

The retrieved taxonomic annotation is mapped onto each sampled tree by associating each clade in the tree with the taxon with lowest taxonomic rank that includes all sequences in the clade (see Fig. 2). By assuming the rooting implicit from the taxonomic annotation the sister clade to the sample sequence is identified. For some trees the position of the root relative to the sample sequence cannot be deduced from the taxonomic annotation. In these cases the taxonomic assignment of all sequences in the tree is recorded. The posterior probability of forming a monophyletic group with a given taxon is then calculated as the fraction of sampled trees where the sister clade to the sample sequence is a member of that taxon.

The posterior probability serves as a confidence measure associated with each assignment and has a straightforward statistical interpretation as the posterior probability that the assignment is correct given the available sequence information and a uniform prior on tree topology. Posterior probabilities are produced for all levels of taxonomic annotation. This allows the sample sequence to be assigned to a higher ranking taxon, such as genus or family, in cases where homology information is too ambiguous to allow a reliable assignment at the species level. The implementation of our approach, SAP (Statistical Assignment Package), generates scalable vector graphics summarizing assignment results. An example of this is shown in Figure 3.

The computational time to compile a homologue set relies heavily on a number of external factors such as the current response time of the online Blast server and bandwidth of the Internet connection for retrieval of sequences and annotation. On a 2-GHz Intel processor,

the alignment of fifty 1000-bp sequences in ClustalW takes about 2 minutes. The sampling of trees amounts to about an hour and represents the bulk of the computational time for the full analysis. The post-processing of the MCMC output may take up to 10 minutes.

The software can be accessed at <http://fisher.berkeley.edu/cteg/software/munch>.

## RESULTS

### Benchmarking

A benchmark analysis was carried out by assigning a data set of cytochrome oxidase I (COI) and tRNA-Leu (trnL) sequences to taxa. All COI entries for the class *Insecta* (true insects), and all trnL entries for the class *Liliopsida* (monocots) are downloaded from GenBank. Taxa represented by only one sequence in GenBank as well as database entries not explicitly targeting the relevant genes are not retrieved. The correct taxonomic annotation associated with each entry was downloaded from NCBI's Taxonomy database. From the 10,804 *Insecta* and 640 *Liliopsida* sequences, 500 are randomly chosen from each set to serve as test sample sequences. Taxonomic assignment of each sample sequence was performed as described, with the exception that the sample sequence itself was disregarded when identified as a homologue in GenBank.

The distribution of posterior probabilities associated with correct and wrong assignments are shown in Figure 4. At the levels of species, genus, and family, 90%, 99%, and 99% of assignments of *Insecta* sequences are correct and 51%, 90%, and 100% of assignments of *Liliopsida* sequences are correct. The false assignments generally have low probabilities and 86% of correct assignments of *Insecta* sequences and 60% of correct *Liliopsida* assignments have posterior probabilities above 0.95. The few false assignments primarily arise when lineage sorting disrupts the true phylogenetic relationship between taxa. False assignments may also arise when the correct taxon and one or more wrong taxa all obtain equally high

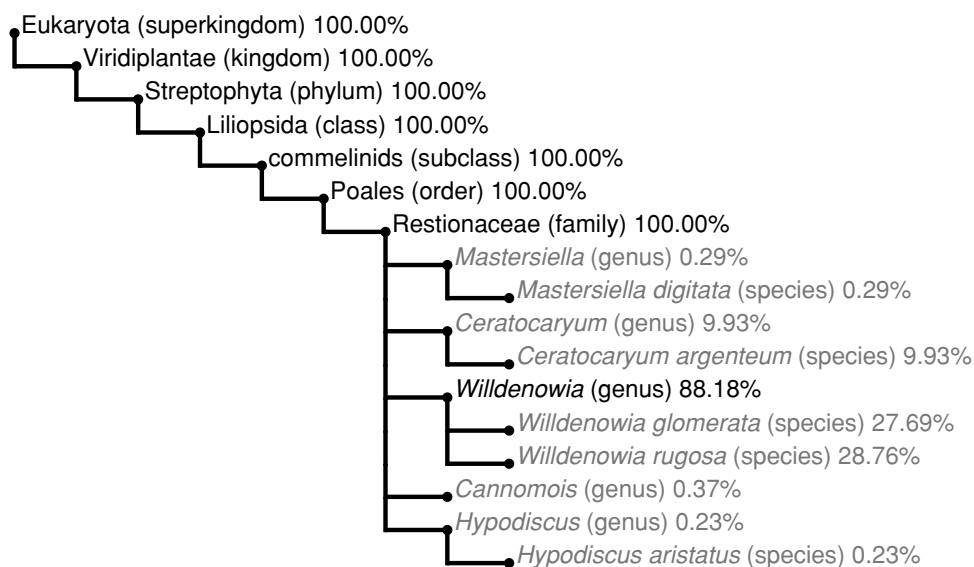


FIGURE 3. Graphic representation of assignment. The taxonomic tree shows all taxa obtaining positive probabilities of assignment. For clarity, assignment probabilities below 50% are shaded. In the example shown, sequence evidence is substantial but too ambiguous to allow a reliable assignment at the species and genus level. The evidence at family level, however, is decisive.

assignment probabilities. In these cases, the small error in the estimation of assignment probabilities may cause that of a wrong taxon to be marginally greater than that of the correct one, resulting in an incorrect assignment. This problem, however, only affects assignments with probabilities below 0.5. A global alignment may not always constitute an optimal alignment of all homologues to the sample sequence so that the relative distances to

the sample sequence are all represented correctly. However, only the part of each homologue corresponding to the sample sequence is submitted to the multiple alignment leaving little room for incorrect alignment. In addition, the clustering algorithm used by ClustalW assures that faulty alignment is least likely to occur between the most similar sequences in the multiple alignment. This minor source of error is therefore expected to mainly

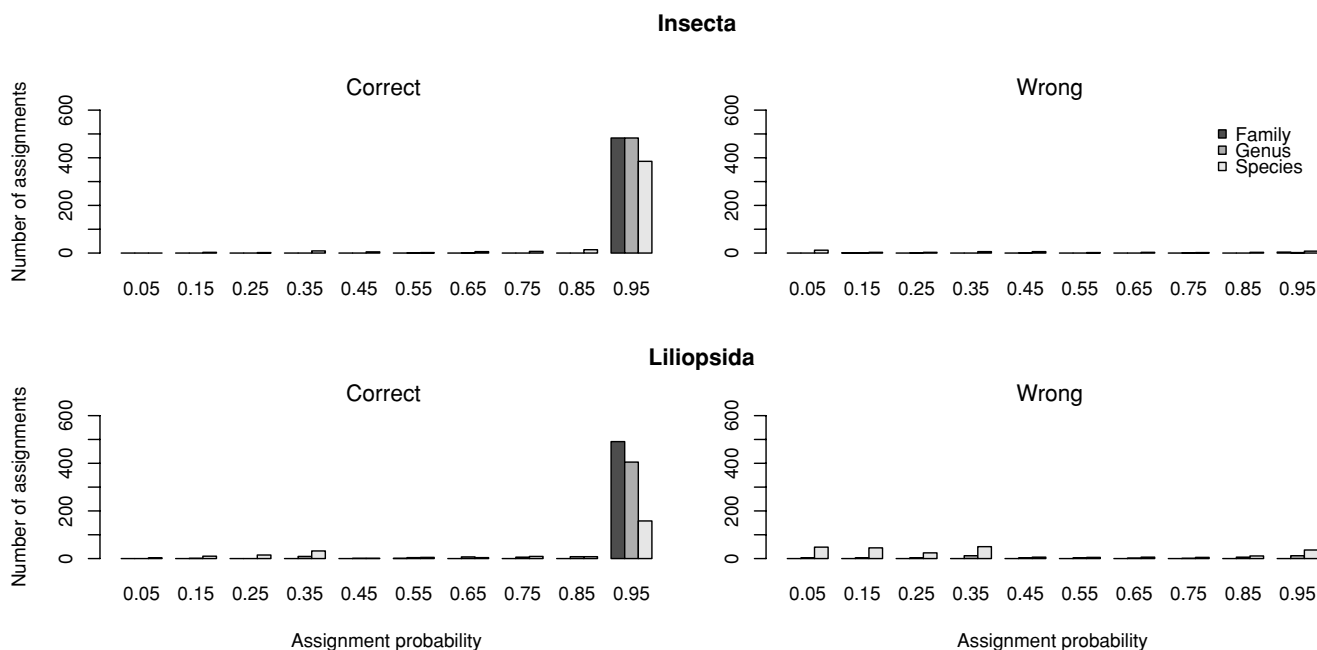


FIGURE 4. Distributions of assignment probabilities for correct and wrong assignments. At the levels of species, genus, and family, 90%, 99%, and 99% of assignments of *Insecta* sequences are correct and 51%, 90%, and 100% of assignments of *Liliopsida* sequences are correct. Wrong assignments are generally associated with low probabilities, whereas most correct assignments achieve probabilities above 95%.

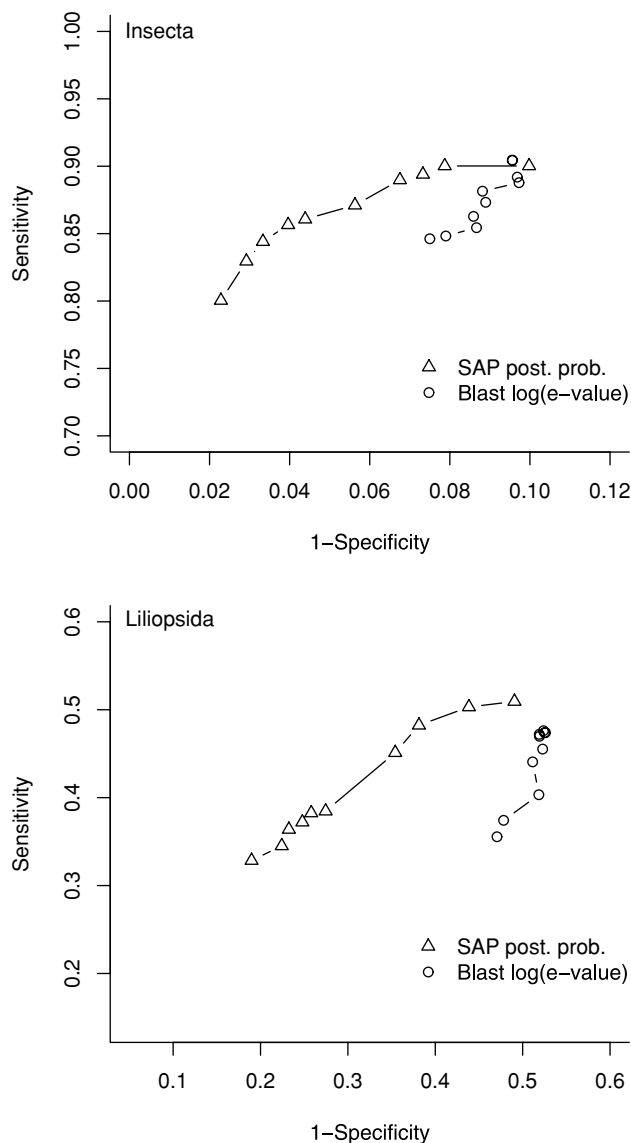


FIGURE 5. ROC (receiver operating characteristic) curves summarizing the tradeoff between sensitivity and specificity in the range of most to least stringent assignment criteria used. Sensitivity is the fraction of all sequences that are correctly assigned, specificity is the fraction of assignments that are correct. The performance of SAP exceeds that of Blast for any sensitivity-specificity combination except when blindly accepting all assignments.

affect assignment in cases where the homology evidence is ambiguous and will thus rarely if ever affect unambiguous assignments based on probabilities over 90%. As a safeguard, the alignment is presented to the user together with the assignment results and should be inspected whenever possible.

To compare the performance of our approach to that of simple Blast searches, all sample sequences are assigned using new Blast searches. To our knowledge there is no canonical way to use Blast for taxonomic assignment. Here we use the taxonomic annotation associated with the best Blast hit to GenBank, disregarding matches to the sample sequence itself. Blast results were retrieved

using remote Blast. In cases of equally high-scoring hits to multiple species, one of these was chosen at random to form the basis of assignment.

Figure 5 compares the two approaches by plotting the tradeoff between sensitivity and specificity in the range of most to least stringent assignment criteria used. Sensitivity is the fraction of sample sequences that are correctly assigned, whereas specificity is the fraction of accepted assignments that are correct. The posterior probability of assignment provided by SAP allows rejection of assignments that do not exceed a minimum assignment probability cutoff. Increasing the stringency of this assignment criterion imposes a more conservative sensitivity-specificity tradeoff. For Blast, the assignment criterion used was a maximum  $\log(E\text{-value})$  cutoff. The so called ROC plots in Figure 5 show how specificity of SAP can be raised at the expense of sensitivity by changing the assignment probability cutoff from zero to the maximal probability obtained in the analysis. For the *Insecta* set, sensitivity of Blast was almost identical to that of SAP when all assignments were accepted. For all other sensitivity-specificity combinations, however, the performance of SAP exceeded that of Blast. At the most permissive assignment criteria, the overlap in correct assignments of *Insecta* sequences was almost complete, with only 3% specific to SAP and 4% to Blast. For the *Liliopsida* set, the overlap was smaller, with 20% of correct assignments specific to SAP and 14% to Blast. The proportion of wrong Blast assignments avoided as a function of posterior probability cutoff (Fig. 6) shows that a large proportion of wrong Blast assignments would be rejected using a stringent assignment criterion in our approach.

Even though the curves for E-value cutoffs correlate with specificity in the two different sets of sample-sequences, the E-value does not constitute a reliable measure of confidence. The E-value only reports the probability due to chance that there is another database

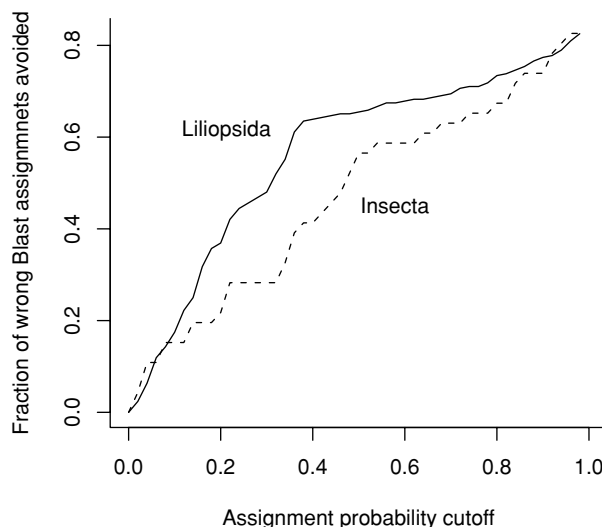


FIGURE 6. The proportion of wrong Blast assignments otherwise made that are avoided using different posterior probability cutoffs.

hit with a sequence similarity score greater than the one obtained and offers no information on the relative confidence in taxonomic assignment to one of multiple hits with similar high E-value. Alternatively, a length-normalized bit-score could be used as assignment criterion. This would reflect the sequence identity of Blast hits. However, such a measure would be associated with the same problems.

The main source of wrong assignments using best hit from Blast was that Blast only evaluates local similarity to individual database homologues. Longer imperfect matches may obtain higher scores than shorter perfect matches leaving the homologue representing the correct taxon far down or off the list of Blast hits. This phenomenon also explains the striking difference in SAP's sensitivity between the *Insecta* and *Liliopsida* sets. For 31% of all *Liliopsida* assignments, this problem prevents the correct homologue from being included in the compiled set of homologues. To eliminate this source of error, the immense number sequences to which Blast produce hits would have to be downloaded and these hits would then have to be re-ranked based on individual global alignments to the sample sequence. Unfortunately, information on sequence homology in GenBank can only be accessed through Blast searches greatly hampering any database-driven assignment approach.

#### *Reanalysis of Neanderthal Sequences*

The fact that sequence similarity does not generally map well to assignment specificity prompts for a reanalysis of datasets where similarity has formed the basis of taxonomic inference such as in genetic barcoding and metagenomics. A less obvious but equally important application relates to the reconstruction of ancient DNA by tiling shorter sequence segments obtained using PCR or 454 pyro-sequencing. Traditionally, longer ancient DNA sequences are constructed by concatenating many short reads. Authenticity of the ancient DNA is then evaluated based on the concatenated sequence and not separately on the individual contributing fragments. Our objective here is to re-examine how much confidence can be assigned to each of the inferred sequences.

We have evaluated the probability of assignment of each of the PCR segments originally used to infer seven known mitochondrial Neanderthal sequences from hyper-variable region I: ElSidron (Lalueza-Fox et al., 2006), Feldhofer1 (Krings et al., 1997), Feldhofer2 (Schmitz et al., 2002), Mezmaiskaya (Ovchinnikov et al., 2000), MontiLessini (Caramelli et al., 2006), Sclandina (Orlando et al., 2006), and Vindija75 (Krings et al., 2000) as well as the two sequences from hyper-variable region II: Feldhofer1 (Krings et al., 1999) and Vindija75 (Krings et al., 2000). For the two Vindija75 sequences, however, the information on PCR sequences was not accessible. The five published sequences not in GenBank: Rochers-DeVilleneuve (Beauval et al., 2005), LaChapelleAux-Saints, Engis2, Vindija77, and Vindija80 (Serre et al., 2004), are very short and not produced using tiling of shorter segments.

To ensure a maximal coverage of Neanderthal diversity in the compiled set, it was assured that all unique Neanderthal sequences matching the query were included. The Neanderthal database sequence corresponding to the sample sequence was not included in the compiled alignments. The same number of unique homologues was allowed for all other species/subspecies represented. To accommodate the short length of sample sequences, no low complexity filtering was used in the Blast searches, the number of Blast hits considered each time was raised from 200 to 500, and the gap-open penalty in ClustalW alignments was raised to 25. These settings correspond to standard user options of SAP. Results from the analysis are summarized in Figure 7.

Whereas it is clear that most fragments with high probability are of Neanderthal origin, it is also clear that the confidence in some of them is low. More worrying, many of the fragments have close to zero percent posterior probability of being Neanderthal, suggesting that these fragments are in fact modern human contaminants. It is clear that the protocol used in ancient DNA studies of constructing larger sequences by concatenating many smaller sequences may easily lead to the artefactual production of chimeric sequences that are part Neanderthal and part modern human.

#### DISCUSSION

To reliably assign DNA sequences to taxonomic groups, a measure of confidence in the assignment is required. The traditional use of Blast for identification does not yield any such information. In contrast, the posterior probability of assignment supplied by SAP has a straightforward statistical interpretation as a measure of confidence that allows the researcher to judge whether a reliable assignment can be made. The value of being able to reject unreliable assignments is emphasized by the fact that ~80% of wrong Blast assignments otherwise made for the *Insecta* and *Liliopsida* sets at species level are rejected using a 0.95 assignment probability cut-off. A further advantage of our approach is that all taxonomic levels are associated with individual measures of confidence. This makes it possible to make a reliable assignment to higher taxonomic levels where sequence information is not sufficient for a reliable species-level assignment.

The Bayesian approach to tree sampling required to obtain a statistically meaningful confidence in assignment is computationally demanding. The motivation for this work was reliability in assignment rather than speed. Nevertheless, to be able to use this approach on very large datasets, such as environmental samples, alternative tree sampling approaches must be explored. A possible alternative could be a modified neighbor-joining algorithm allowing topological constraints. However, although bootstrap scores for assignment most likely correlates with probabilities of assignment, they do not have the same probabilistic interpretation as posterior probabilities. One advantage of the Bayesian approach is that it allows for the possibility of using decision theory

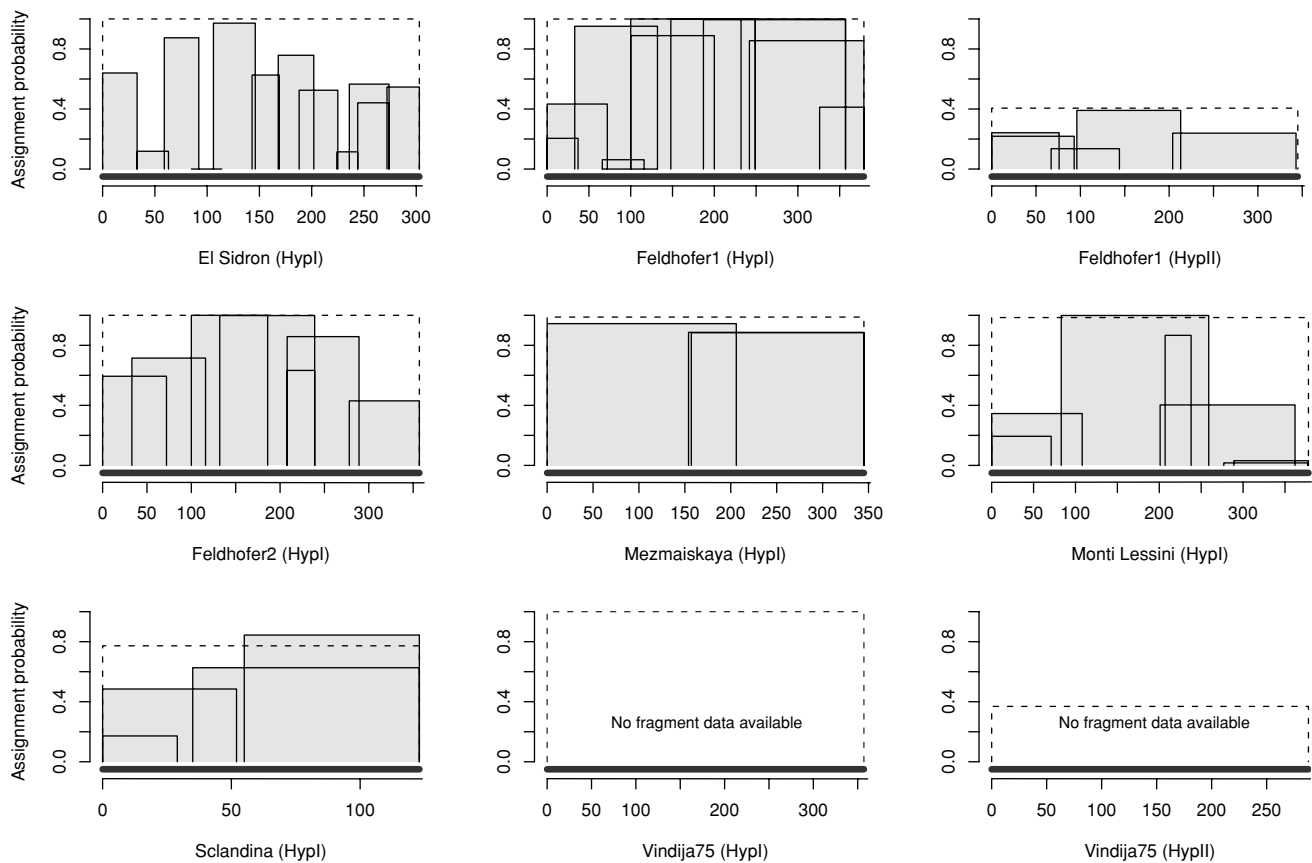


FIGURE 7. Summary of confidence analysis for published Neanderthal sequences. In each sub-figure, a bold bar represents the Neanderthal sequence analyzed. The overlapping boxes above it each represent the assignment probability of the sequence fragment spanned by the box. The dashed box represents the full inferred sequence, whereas shaded boxes represent individual contributing PCR fragments. For the Vindija75 sequences, no information on PCR fragments is available. The five short sequences not in GenBank obtain the following assignment probabilities: Engis2 (HypI): 0.88; LaChapelleAuxSaints (HypI): 0.88; RochersDeVilleneuve (HypI): 0.63; Vindija77 (HypI): 0.87; Vindija80 (HypI): 0.89.

to devise criteria for assignment (Abdo and Golding, 2007).

Even for genes used for genetic barcoding such as COI and *trnL*, each species is typically represented in the databases by only very few sequences. In addition, the sample sequence and the database sequences may stem from subpopulations with little ongoing gene flow. A rigorous population genetic analysis would involve assumptions about population structure and demography that may not be valid in general. With the limited amount of information available, a simpler phylogenetic approach, such as the one suggested here, may be an appropriate alternative. In this framework, within-species variability is not modeled and an assignment to a species-level taxon thus makes no implicit statement about how well the corresponding group of database sequences fits a species concept. However, as for other methods that are not based on explicit modeling the population genetics of the species in question, our method may also be misled by incomplete lineage sorting.

Common to all approaches for taxon assignment is the problem that no available reference database fully represents the tree of life. For many taxa the database coverage is still poor and this will lead to false assignments

in cases where the correct taxon is not represented in the database. In the analyses presented here, GenBank was used as reference database but the approach can be used with any database of taxonomically annotated sequences. Access to a local version of GenBank or other exhaustive database would greatly increase the speed of the analysis.

In the example analysis provided here, we demonstrated that several published Neanderthal sequences are likely composed of both Neanderthal and modern human DNA. However, we emphasize that most of the published sequences seem to contain 100% Neanderthal DNA, and none of the sequences have more than a few small segments that are likely to be of modern human origin. Nonetheless, it would be desirable in future ancient DNA studies to provide confidence measures for each position in the sequence. Such measures could be obtained using the method described here.

#### REFERENCES

- Abdo, Z., and G. B. Golding. 2007. A step toward barcoding life: A model-based, decision-theoretic method to assign genes to preexisting species groups. *Syst. Biol.* 56:44–56.

- Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped Blast and psi-Blast: A new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.
- Beauval, C., B. Maureille, F. Lacrampe-Cuyaubère, D. Serre, D. Pessinotto, J.-G. Bordes, D. Cochard, I. Couchoud, D. Dubrasquet, V. Laroulandie, A. Lenoble, J.-B. Mallye, S. Pasty, J. Primault, N. Rohland, S. Pääbo, and E. Trinkaus. 2005. A late Neandertal femur from Les Rochers-de-villeneuve, France. *Proc. Natl. Acad. Sci. USA.* 102:7085–7090.
- Breitbart, M., P. Salamon, B. Andresen, J. Mahaffy, A. Segall, D. Mead, F. Azam, and F. Rohwer. 2002. Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. USA.* 99:14250–14255.
- Caramelli, D., C. Lalueza-Fox, S. Condemi, L. Longo, L. Milani, A. Manfredini, M. de Saint Pierre, F. Adoni, M. Lari, P. Giunti, S. Ricci, A. Casoli, F. Calafell, F. Mallegni, J. Bertranpetit, R. Stanyon, G. Bertorelle, and G. Barbujani. 2006. A highly divergent mtDNA sequence in a Neandertal individual from Italy. *Curr. Biol.* 16:R630–R632.
- Floyd, R., E. Abebe, A. Papert, and M. Blaxter. 2002. Molecular barcodes for soil nematode identification. *Mol. Ecol.* 11:839–850.
- Hebert, P., A. Cywinska, S. Ball, and J. Dewaard. 2003. Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270:313–321.
- Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol. Biol. Evol.* 21:1123–1133.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
- Krings, M., C. Capelli, F. Tschentscher, H. Geisert, S. Meyer, A. von Haeseler, K. Grossschmidt, G. Possnert, M. Paunovic, and S. Pääbo. 2000. A view of Neandertal genetic diversity. *Nat. Genet.* 26:144–146.
- Krings, M., H. Geisert, R. W. Schmitz, H. Krainitzki, and S. Pääbo. 1999. DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen. *Proc. Natl. Acad. Sci. USA.* 96:5581–5588.
- Krings, M., A. Stone, R. W. Schmitz, H. Krainitzki, M. Stoneking, and S. Pääbo. 1997. Neandertal DNA sequences and the origin of modern humans. *Cell* 90:19–30.
- Lalueza-Fox, C., J. Krause, D. Caramelli, G. Catalano, L. Milani, M. L. Sampietro, F. Calafell, C. Martínez-Maza, M. Bastir, A. García-Taberner, M. de la Rasilla, J. Fortea, S. Pääbo, J. Bertranpetit, and A. Rosas. 2006. Mitochondrial DNA of an Iberian Neandertal suggests a population affinity with other European Neandertals. *Curr. Biol.* 16:R629–R630.
- Matz, M., and R. Nielsen. 2005. A likelihood ratio test for species membership based on DNA sequence data. *Philos Trans R Soc Lond B Biol Sci* 360:1969–1974.
- Meyer, C., and G. Paulay. 2005. DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biol.* 3:e422.
- Moritz, C., and C. Cicero. 2004. DNA barcoding: Promise and pitfalls. *PLoS Biol.* 2:e354.
- Nielsen, R., and M. Matz. 2006. Statistical approaches for DNA barcoding. *Syst. Biol.* 55:162–169.
- Orlando, L., P. Darlu, M. Toussaint, D. Bonjean, M. Otte, and C. Hänni. 2006. Revisiting Neandertal diversity with a 100,000 year old mtDNA sequence. *Curr. Biol.* 16:R400–R402.
- Ovchinnikov, I. V., A. Götherström, G. P. Romanova, V. M. Kharitonov, K. Lidén, and W. Goodwin. 2000. Molecular analysis of Neandertal DNA from the Northern Caucasus. *Nature* 404:490–493.
- Pawitan, Y. 2001. In all likelihood: Statistical modelling and inference using likelihood. Oxford University Press, Oxford, UK.
- Remigio, E., and P. Hebert. 2003. Testing the utility of partial Coi sequences for phylogenetic estimates of gastropod relationships. *Mol. Phylogenet. Evol.* 29:641–647.
- Rusch, D., A. Halpern, G. Sutton, K. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. Eisen, J. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. Venter, K. Li, S. Kravitz, J. Heidelberg, T. Utterback, Y.-H. Rogers, L. Falcón, V. Souza, G. Bonilla-Rosso, L. Eguiarte, D. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. Ferrari, R. Strausberg, K. Neilson, R. Friedman, M. Frazier, and C. Venter. 2007. The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.* 5:e83.
- Schmitz, R. W., D. Serre, G. Bonani, S. Feine, F. Hillgruber, H. Krainitzki, S. Pääbo, and F. H. Smith. 2002. The Neandertal type site revisited: Interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. *Proc. Natl. Acad. Sci. USA.* 99:13342–13347.
- Serre, D., A. Langanay, M. Chech, M. Teschler-Nicola, M. Paunovic, P. Menecier, M. Hofreiter, G. Possnert, and S. Pääbo. 2004. No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biol.* 2:E57.
- Steinke, D., M. Vences, W. Salzburger, and A. Meyer. 2005. Taxi: A software tool for DNA barcoding using distance methods. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360:1975–1980.
- Tavare, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect. Math. Life Sci.* 17:57–86.
- Tringe, S., and E. Rubin. 2005. Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* 6:805–814.
- Venter, C., K. Remington, J. Heidelberg, A. Halpern, D. Rusch, J. Eisen, D. Wu, I. Paulsen, K. Nelson, W. Nelson, D. Fouts, S. Levy, A. Knap, M. Lomas, K. Neilson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Tillson, C. Pfannkoch, Y. Rogers, and H. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74.
- Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10:1396–1401.
- Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* 39:306–314.
- Yooseph, S., G. Sutton, D. Rusch, A. Halpern, S. Williamson, K. Remington, J. Eisen, K. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. Miller, H. Li, S. Mashiyama, M. Joachimiak, C. van Belle, J.-M. Chandonia, D. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. Raphael, V. Bafna, R. Friedman, S. Brenner, A. Godzik, D. Eisenberg, J. Dixon, S. Taylor, R. Strausberg, M. Frazier, and Craig. 2007. The sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biol.* 5:e83.

First submitted 1 December 2007; reviews returned 11 February 2008;

final acceptance 4 June 2008

Associate Editor: Paul Lewis