Supplementary Material

PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences

Simon Whelan, Iker Irisarri, Fabien Burki

Methodology and Implementation

The probabilistic methods used in PREQUAL are based on a pair hidden Markov model. These pairHMMs consist of three hidden states defining the relationship between sequences X and Y: match states, which define a shared ancestor between X and Y through a process of substitution; and insert and delete states, which describe gain and loss in sequence X, respectively. Given a parameterized pairHMM, one can calculate the posterior probability (PP) of a character from X being related to a character from Y using the forward-backward algorithm (see (Durbin *et al.*, 1998) for an overview of pairHMMs and their associated algorithms).

Our approach requires a slightly different PP to be calculated, which captures the PP of a character in sequence $\mathrm{X} = \{x_i\}$ sharing a common ancestor with any character in the set of n sequences being considered $\mathbf{S} = \{\mathrm{Y}^1, \ldots, \mathrm{Y}^{\mathrm{n}}\}$, where $\mathrm{X} \in \mathbf{S}$, the set of sequences without X is $\mathbf{S}' = \mathbf{S} - \mathrm{X}$, and a pairHMM can be run on the pair (X, Y) to obtain the posterior probability of $\Pr(x_i, y_j)$ using forward-backward. First, we can calculate the maximal PP of x_i sharing a common ancestor with any character in Y as $\Pr\,(x_i, \mathrm{Y}_*) = \max_{y_j \epsilon \mathrm{Y}}\{\Pr(x_i, y_j)\}$. Then we want to find the maximal PP of x_i sharing a common ancestor with any of the other sequences: $\Pr\,(x_i|\text{Ancestor}) = \max_{\mathrm{Y}\epsilon\mathbf{S}'}\{\Pr\,(x_i, \mathrm{Y}_*)\}$. The value of $\Pr\,(x_i|\text{Ancestor})$ can be computed for every character of every sequence in **S**, and an appropriate threshold τ can be used as a cutoff to discriminate between characters with adequate evidence of shared ancestry that should be carried through to the alignment phase and those that should be discarded.

PREQUAL calculates posterior probabilities using a bounded pairHMM with a substantially modified version of Zorro (Wu *et al.*, 2012). It uses a heuristic approach for calculating $\Pr\,(x_i|\text{Ancestor})$ by choosing a set of sequences from $\mathbf{S}'$ based on evolutionary divergence (Bogusz and S. Whelan, 2017) and sequence coverage to reduce the number of pairHMM calculations that need to be performed. This heuristic samples only a subset of sequences to find the max(p $(x_i|\text{Ancestor})$) ensuring that the closest sequences are included and they have adequate similarity over enough of the characters. We choose a default of 10 close sequences to examine, ensuring that 3 of these are equal to or greater than the median length of the all the sequences. We also avoid computing the whole dynamic programming matrix for the pair HMM where possible, by bounding the distance any path can take through that matrix from the diagonal.

Simulation

INDELible v.1.03 (Fletcher and Yang, 2009) was used to simulate 100 gene alignments under each of three levels of gaps (see below). The alignments were simulated along a recently published phylogenomic tree, which contains both long and short terminal branches (Supplementary Figure 17 in (N. V. Whelan *et al.*, 2017)), using the WAG model with 4 gamma categories. Two values for the gamma shape parameter alpha were tested (1.8 and 0.5) to take into account little and higher among-site rate heterogeneity. Since PREQUAL performed similarly under both conditions, only the simulation under one condition ($\alpha = 1.8$) is reported in Table 1. The gaps were inserted according to a Zipfian/Power law indels (max length of 20 and alpha=1.7). Simulated alignments were 500 amino acids long, formed by a core region (450 amino acids) flanked by more gappy regions (25 amino acids each) representing a typical exon alignment. The three gap levels were: low (0.01 gap rate for the core and 0.02 for flanking regions), medium (0.02, 0.05), and high (0.1, 0.2). Individual alignment files were then corrupted by inserting errors in random sequences (proportional to sequence length) and locations. Errors corresponded to random amino acids drawn proportionally from the residue frequencies of the WAG model. Errors were either inserted at random positions, mimicking misannotation errors such as wrong gene models, or replaced parts of the original sequences, mimicking frameshifts. Three error rates were tested: low (0.001 errors per amino acid, ~22 errors per file), medium (0.02: ~44 errors) and high (0.003; ~66 errors). Different lengths of individual errors were also tested, which were decided using a geometric distribution so that the final expected lengths of the erroneous stretches were 10, 20, and 30 amino acids. A total of 108 experimental conditions were simulated; gene files were then subjected to PREQUAL v.1.0 and HMMCleaner v.1.8 and the results are summarized in Table 1 of the main text.

References

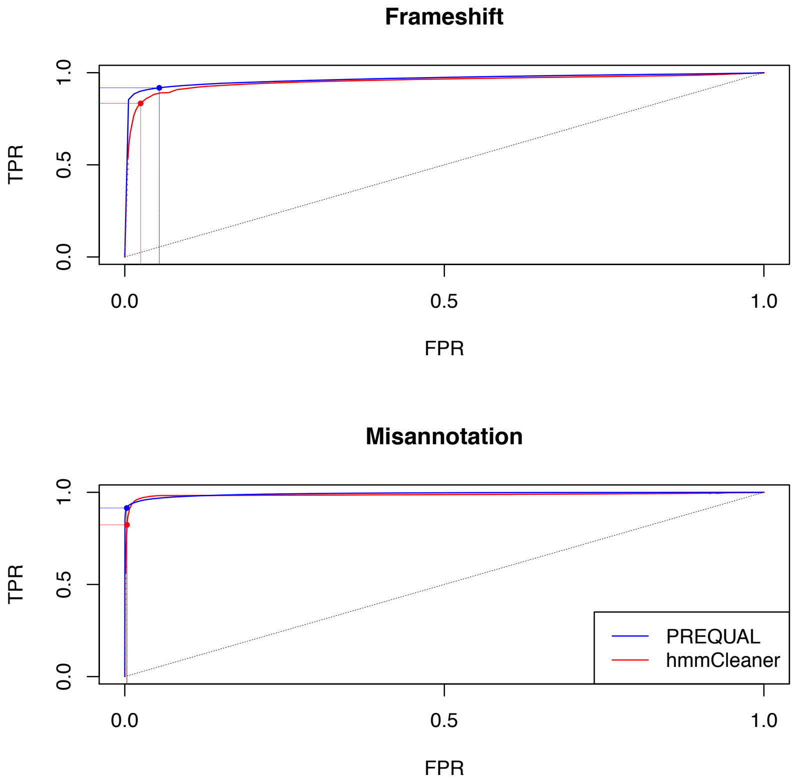
Bogusz,M. and Whelan,S. (2017) Phylogenetic Tree Estimation With and Without Alignment: New Distance Methods and Benchmarking. *Syst Biol*, **66**, 218–231.

Durbin,R. *et al.* (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.

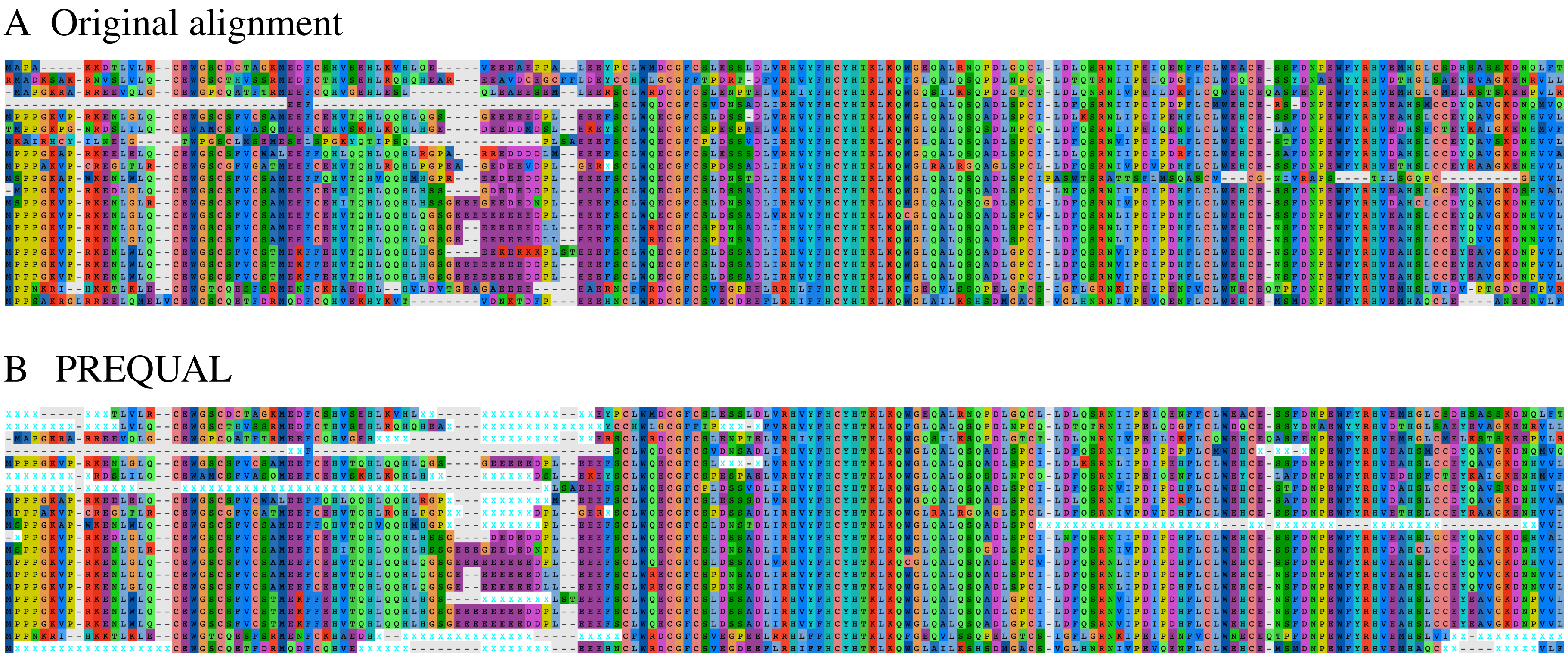
Fletcher,W. and Yang,Z. (2009) INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*, **26**, 1879–1888.

Whelan,N.V. *et al.* (2017) Ctenophore relationships and their placement as the sister group to all other animals. *Nat Ecol Evol*, **1**, 1–10.

Wu,M. *et al.* (2012) Accounting for alignment uncertainty in phylogenomics. *PLoS ONE*, **7**, e30288.



Supplementary Figure 1. ROC curves for PREQUAL (blue) and HMMcleaner (red) for our frameshift and misannotation simulation schemes. These plots show the trade-off between true positive rate (TPR) and false positive rate (FPR), with the best performing method being that which reaches closest to the top left-hand corner. The dots on the curves show the performance for the methods under the recommend thresholds. AUC values calculated using the trapezoidal approximation for PREQUAL are 0.965 (frameshifts) and 0.992 (misannotations), and for HMMCleaner 0.951 (frameshifts) and 0.985 (misannotations). PREQUAL offers an increase to classifier performance in terms of AUC for both frameshifts and misannotations of HMMCleaner.


A Original alignment

B PREQUAL

Supplementary Figure 2. Visualization of PREQUAL performance on a set of vertebrate HHIP orthologs (obtained from http://www.orthodb.org/). (A) Alignment of unfiltered sequences. (B) PREQUAL filters non-homologous residues by masking them with X (light blue). Note that PREQUAL works on unaligned sequences; here aligned for easy identification of masked residues. PREQUAL is effective at identifying and filtering non-homologous residues (true positives), but in common with all statistical methods there is a risk of filtering some genuinely homologous residues (false positives).