Bioinformatics, 34(22), 2018, 3929–3930
doi: 10.1093/bioinformatics/bty448
Advance Access Publication Date: 1 June 2018
Applications Note

Sequence analysis

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

Simon Whelan[1,*], Iker Irisarri[2] and Fabien Burki[2,3,*]

[1]Department of Evolutionary Genetics, Program in Evolutionary Biology and [2]Department of Organismal Biology, Program in Systematic Biology and [3]Science for Life Laboratory, Uppsala University, Uppsala 75236, Sweden

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on December 20, 2017; revised on May 18, 2018; editorial decision on May 25, 2018; accepted on May 30, 2018

Abstract

Summary: Phylogenomic datasets invariably contain undetected stretches of non-homologous characters due to poor-quality sequences or erroneous gene models. The large-scale multi-gene nature of these datasets renders impractical or impossible detailed manual curation of sequences, but few tools exist that can automate this task. To address this issue, we developed a new method that takes as input a set of unaligned homologous sequences and uses an explicit probabilistic approach to identify and mask regions with non-homologous adjacent characters. These regions are defined as sharing no statistical support for homology with any other sequence in the set, which can result from e.g. sequencing errors or gene prediction errors creating frameshifts. Our methodology is implemented in the program PREQUAL, which is a fast and accurate tool for high-throughput filtering of sequences. The program is primarily aimed at amino acid sequences, although it can handle protein coding DNA sequences as well. It is fully customizable to allow fine-tuning of the filtering sensitivity.

Availability and implementation: The program PREQUAL is written in C/C++ and available through a GNU GPL v3.0 at https://github.com/simonwhelan/prequal.

Contact: simon.whelan@ebc.uu.se or fabien.burki@ebc.uu.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

The assembly of phylogenomic datasets heavily relies on automation, but automated tools for some important quality control steps that ensure the accuracy of datasets are still lacking. As a result these controls are either ignored or performed manually. One such step is the quality-proofing of sequences to identify and remove non-homologous (erroneous) regions before phylogenetic re-construction. This is different from the standard trimming of MSA to remove poorly aligned sites; here, we deal with non-homologous residues in individual sequences that are not necessarily associated with regions of poor alignment but should be excluded before phylogenetic inference. Such stretches of non-homologous characters are often present in phylogenomic datasets, resulting from frameshifts due for example to poor overall sequence quality, or genome annotation errors.

To allow automated and high-throughput detection of these non-homologous regions, we present PREQUAL, a new program for PRE-alignment QUALity filtering. PREQUAL uses an explicit probabilistic model to test evidence of homology between amino acid residues in pairs of unaligned sequences, and residues showing no statistical evidence of homology are filtered. The probability model defining the relationship between sequence pairs is a pair hidden Markov model (Supplementary Material). Given a parameterized pairHMM, PREQUAL calculates the posterior probability (PP) of a character being related to a character from another sequence using the forward-backward algorithm described in (Durbin *et al.*, 1998), and characters with insufficient evidence of shared ancestry are filtered.

Similar tools that can detect portions of non-homologous characters are rare. To our knowledge, HMMCleaner is the only related tool that has been applied to phylogenomic datasets, although it has not been formally published, its performance is untested and its design is fundamentally different since it uses homology from a given MSA to detect non-homologous regions (Amemiya *et al.*, 2013). We benchmarked PREQUAL and compared its performance relative to HMMCleaner v1.8 using simulated gene alignments representative of a recently published phylogenomic dataset (Whelan *et al.*, 2017).

Table 1. Performance of PREQUAL and HMMCleaner[a] on simulated data

			Misannotations			Frameshifts		
			No. characters	Accuracy	Errors captured	No. characters	Accuracy	Errors captured
Gappyness	Low	PREQUAL	2 339 354	95.47%	99.67%	2 250 400	95.60%	93.21%
		HMMCleaner	2 339 354	85.54%	99.67%	2 250 400	86.57%	97.72%
	Mid	PREQUAL	2 345 437	91.84%	99.68%	2 255 919	91.89%	94.60%
		HMMCleaner	2 345 437	83.05%	99.65%	2 255 919	83.90%	97.52%
	High	PREQUAL	2 362 243	73.10%	99.76%	2 272 380	73.14%	96.61%
		HMMCleaner	2 362 243	68.23%	99.59%	2 272 380	67.24%	97.50%
Number of errors	Low	PREQUAL	2 300 164	92.26%	99.81%	2 255 919	92.36%	94.24%
		HMMCleaner	2 300 164	83.41%	99.67%	2 255 919	84.21%	97.30%
	Mid	PREQUAL	2 390 366	91.47%	99.78%	2 255 919	91.49%	94.62%
		HMMCleaner	2 345 437	83.05%	99.65%	2 255 919	83.90%	97.52%
	High	PREQUAL	2 390 366	91.47%	99.78%	2 255 919	91.49%	94.62%
		HMMCleaner	2 390 366	82.82%	99.68%	2 255 919	83.61%	97.68%
Expected error length	10 AA	PREQUAL	2 296 403	91.77%	99.16%	2 255 919	91.90%	68.97%
		HMMCleaner	2 296 403	83.23%	99.26%	2 255 919	84.12%	94.70%
	20 AA	PREQUAL	2 390 366	91.47%	99.78%	2 255 919	91.49%	94.62%
		HMMCleaner	2 345 437	83.05%	99.65%	2 255 919	83.90%	97.52%
	30 AA	PREQUAL	2 389 967	91.99%	99.87%	2 255 919	91.98%	98.15%
		HMMCleaner	2 389 967	83.04%	99.75%	2 255 919	83.80%	97.91%

[a] HMMCleaner masks erroneous AA with the character X, but sometimes also introduces additional Xs in all sequences; these additional Xs were accounted for to generate the statistics.

These alignments were corrupted by inserting errors under different experimental conditions (Supplementary Material); the error-riddled sequences were then fed to both programs with default parameters (Table 1).

PREQUAL performed well under all conditions but was particularly efficient at detecting errors inserted at random positions with >98% captured. These types of errors mimic misannotation such as wrong gene models. Accuracy was in most cases >90%, although it went as low as 73% when a high level of gaps was considered. For errors replacing parts of the original sequences, i.e. mimicking frameshifts, PREQUAL generally detected >93% of erroneous residues, except for errors of ~10 residues long (69% of errors captured). The accuracy of frameshift detection followed the same trend as for misannotations. The comparison with HMMCleaner revealed that under these conditions both tools captured most (>99%) of the misannotation errors, but HMMCleaner additionally removed more correct residues leading to lower overall accuracy. For frameshifts, HMMCleaner captured more errors than PREQUAL (mean = 97% versus 91%, respectively), but again at the cost of removing a much higher proportion of correct residues. On error-free sequences, the proportion of correct residues removed by PREQUAL was in most cases ≤7%, but between 17% and 25% for HMMCleaner. The general performance of both methods was further examined by ROC curves, which suggested that while they perform similarly for misannotation, PREQUAL provides a more efficient classification of frameshifts than HMMCleaner across the range of possible thresholds (Supplementary Fig. S1).

The ROC curves were also used to derive the default PP threshold, chosen so that ≥95% of correct amino acids were retained while removing >90% of frameshift and misannotation errors. This threshold was further validated by using PREQUAL on empirical datasets characterized by very different levels of sequence divergences (ranging from 20 MYA to 1 BYA; Burki *et al.*, 2016; Irisarri and Meyer, 2016; MacLeod *et al.*, 2016). Supplementary Figure S2 shows an example of MSA from unfiltered and filtered sequences. The default PP threshold can be adjusted by the user to find a trade-off between true positives and false positives to match their needs. Additional functionalities in PREQUAL allow it to ignore known fast evolving sequences (e.g. from parasites) where it can be particularly difficult to tease apart genuine characters from non-homologous stretches. In addition, PREQUAL can input protein-coding nucleotide sequences, which are automatically translated, masked at the amino acid level and back-translated.

PREQUAL can handle typical datasets on a common laptop computer. For example, the analysis of 91 sequences with a maximal length of 718 amino acids took 63 s on a computer equipped with a 2 Ghz i7 processor, whereas a larger set of 272 sequences with a maximal length of 1149 amino acids took 293 s.

In closing, we believe that PREQUAL fills an important gap in phylogenomic pipelines and will help improving the reproducibility and construction of more accurate datasets. All options and functionalities to fine-tune the filtering are described in the '–h all' command line argument, or with further details in the manual available online.

Funding

S.W. is supported by Carl Tryggers Stiftelse and Uppsala University. This work was supported by a fellowship from Science for Life Laboratory available to F.B.

Conflict of Interest: none declared.

References

Amemiya,C.T. *et al.* (2013) The African coelacanth genome provides insights into tetrapod evolution. *Nature*, **496**, 311–316.

Burki,F. *et al.* (2016) Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc. R. Soc. B*, **283**, 20152802.

Durbin,R. *et al.* (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Irisarri,I. and Meyer,A. (2016) The identification of the closest living relative(s) of tetrapods: phylogenomic lessons for resolving short ancient internodes. *Syst. Biol.*, **65**, 1057–1075.

MacLeod,A. *et al.* (2016) The complete mitochondrial genomes of the Galápagos iguanas, *Amblyrhynchus cristatus* and Conolophus subcristatus. *Mitochondrial DNA*, **27**, 3699–3700.

Whelan,N.V. *et al.* (2017) Ctenophore relationships and their placement as the sister group to all other animals. *Nat. Ecol. Evol.*, **1**, 1737–1746.