Revised ms: BIOINF-2003-0817

Bellerophon; a program to detect chimeric sequences in multiple sequence alignments.

Thomas Huber[1]*, Geoffrey Faulkner[1] and Philip Hugenholtz[2]

[1]ComBinE group, Advanced Computational Modelling Centre, The University of Queensland, Brisbane 4072, Australia and [2]Department of Environmental Science, Policy and Management, 151 Hilgard Hall, University of California Berkeley, Berkeley, CA 94720-3110, USA.

*To whom correspondence should be addressed

Tel: Int + 617-3365-7060

Email: huber@maths.uq.edu.au

Running head: Bellerophon: chimera detection program

Journal section: Application note

ABSTRACT

Summary: Bellerophon is a program for detecting chimeric sequences in multiple sequence datasets by an adaption of partial treeing analysis. Bellerophon was specifically developed to detect 16S rRNA gene chimeras in PCR-clone libraries of environmental samples but can be applied to other nucleotide sequence alignments.

Availability: Bellerophon is available as an interactive web server at http://foo.maths.uq.edu.au/~huber/bellerophon.pl

Contact: huber@maths.uq.edu.au

INTRODUCTION

A PCR-generated chimeric sequence is usually comprised of two phylogenetically distinct parent sequences and occurs when a prematurely terminated amplicon reanneals to a foreign DNA strand, and is copied to completion in the following PCR cycles. The point at which the chimeric sequence changes from one parent to the next is called the conversion, recombination or break point. Chimeras are problematic in culture-independent surveys of microbial communities because they suggest the presence of non-existent organisms (von Wintzingerode *et al.,* 1997). Several methods have been developed for detecting chimeric sequences (Cole *et al*., 2003; Komatsoulis and Waterman, 1997; Liesack *et al*., 1991; Robinson-Cox *et al*., 1995) that generally rely on direct comparison of individual sequences to one or two putative parent sequences at a time. Here we present an alternative approach based on how well sequences fit into their complete phylogenetic context.

METHOD

Bellerophon detects chimeras based on a partial treeing approach (Wang and Wang, 1997; Hugenholtz and Huber, 2003), that is, phylogenetic trees are inferred from independent regions (fragments) of a multiple sequence alignment and the branching patterns are compared for incongruencies that may be indicative of chimeric sequences. No trees are actually built during the procedure and the only calculations required are distance (sequence similarity) calculations. A full matrix of distances (*dm*) between all pairs of sequences are calculated for fragments left and right of an assumed break point. The total absolute deviation of the distance matrices (distance matrix error *dme*) of *n* sequences is then

$$dme = \sum_{i}^{n} \sum_{j}^{n} \left| \mathrm{dm}^{\mathrm{left}}[i][j] - \mathrm{dm}^{\mathrm{right}}[i][j] \right|$$

where $dm[i][j]$ denotes the distance between two sequences *i* and *j*.

The largest contribution to the *dme* is expected to arise from chimeras, since fragments from these sequences have distinctly different locations relative to all other sequences in the dataset, and therefore distinctly different distance matrices. To rank the sequences by their contribution to the *dme* value, we calculate the ratio of the *dme* value from all sequences over the *dme* value ($dme[i]$) of a reference dataset lacking the sequence *i* under consideration. This ratio is called the preference score of the sequence.

$$preference[i] = \frac{dme}{dme[i]} \qquad (1)$$

The ratio for chimeric sequences will have a preference score >1, whereas non-chimeric sequence scores are expected to be ~1. To detect all putative chimeras in a dataset, preference scores have to be calculated for all sequences. Naïvely, the calculation would

require a computationally expensive distance matrix comparison for each sequence in the dataset. This can, however, be implemented more efficiently by taking advantage of previously performed calculations. Because the calculation of the *dme* involves column sums in the form of

$$col[i] = \sum_{j}^{n} \left| \mathrm{dm}^{\mathrm{left}}[i][j] - \mathrm{dm}^{\mathrm{right}}[i][j] \right|$$

and the distances between identical sequences $dm[i][i]$ are by definition zero, equation (1) can be rewritten as:

$$preference[i] = \frac{dme}{(dme - 2 \cdot col[i])}$$

which only involves calculation of a single matrix and some intermediate storage of the column sums.

To determine the optimal break point for putative chimeras, all sequences are scanned along their length by dividing the alignment into fragment pairs at 10 character intervals. Distances are calculated from equally sized windows (200, 300 or 400 characters) of the fragments left and right of the break point to obtain similar signal-to-noise ratios for each fragment. The highest preference score calculated for each sequence in all fragment pairs indicates the optimal break point. Sequences are ranked according to their highest recorded preference score and reported as potentially chimeric if that score is >1. Absolute preference scores are dataset-dependent and should only be used for relative ranking of putative chimeras within a given dataset.

For manual confirmation of identified chimeras and phylogenetic placement of the chimeric halves, it is necessary to specify the most likely parent sequences in the dataset, giving rise to the chimera. Parent sequences are assigned to each putative

chimera by selecting the two sequences with the highest opposing paired distance contributions ($\Delta dm[i][j]$) to the *dme* at the optimal break point. The parent sequences of a chimera are most likely to be found in the same PCR-clone library and therefore as many sequences as possible from the one library should be included in the analysis. However, even if the exact parent sequences of a given chimera are not present in the dataset, Bellerophon will identify and report the closest phylogenetic neighbors of the parents. In addition, the output from Bellerophon includes the location of the optimal break point relative to an *E. coli* reference alignment (Brosius *et al.*, 1978) and the percentage identities of the parent sequences to the chimera either side of the break point. These features aid in verification of chimeras. Mutually incompatible chimeras are screened from the Bellerophon output. That is, once a sequence (A) has been identified as chimeric, subsequent putative chimeras with lower preference scores, that identify sequence A as one of the parents, are removed from the output list.

PHYLOGENETIC CONSIDERATIONS

Alignment

As with all phylogeny-based methods, alignment ensuring positional homology is critical. If sequences are poorly aligned, Bellerophon will produce meaningless results. By default, sequences are assumed to be aligned. However, the align sequence option (set to yes) will automatically align up to 300 unaligned sequences using ClustalW (Thompson *et al*., 1994) in preparation for running on Bellerophon.

Filtering

Regions of 16S rRNA with variable secondary structure between different organisms are

routinely removed from sequence alignments prior to phylogenetic inference (e.g. Lane, 1991) because they often cannot be unambiguously aligned and therefore introduce noise into the analysis. We have implemented an automatic consensus filter to remove variable positions from the alignment prior to calculating pairwise distances. Sequence positions (columns) that have nucleotides in less than 50% of all sequences are removed, and from the resulting consensus only sequences longer than two times the window size are used in the analysis. However, we advise removing sequences shorter than 2*window size + 100 prior to submission if the dataset contains many partial sequences, since they will cause over-filtering of the alignment. The length of the filtered alignment also dictates how close to either end of the alignment break points can be detected. For instance if the largest window size of 400 is selected, break points can only be detected ≥400 characters from either end of the alignment.

Distance correction

In phylogenetic analysis, the distance (nucleotide divergence) between two sequences is commonly corrected for unseen nucleotide substitutions to more accurately reflect their evolution. For the detection of chimeric sequences, evolutionary distance corrections may not be ideal since they accentuate long distance relationships (e.g. Jukes and Cantor, 1969). Chimeras are more likely to result from recombination of closely related parent sequences, which led us to develop an empirical distance function (Huber-Hugenholtz correction):

$$d = \sqrt{1 - id}$$

where d is the corrected distance between two sequences with a fraction id of identically aligned bases. This function strongly differentiates between highly similar sequences and

puts less emphasis on differences between remote homologs. It should be noted that this is not a phylogenetic correction as it does not attempt to correct for unseen nucleotide substitutions.

Evolutionary distance calculations are sensitive to partial length sequences in an alignment, and in the worst case, non-overlapping partial sequences result in infinite pairwise distances. This problem is alleviated in the Huber-Hugenholtz correction where non-overlapping sequences have a bound value of 1. Even though this allows the analysis of datasets containing sequences of different lengths, non-overlapping partial sequences can produce artificial skews in distance matrices of fragments of the alignment resulting in potentially unreliable chimera calls. We strongly recommend using alignments of similar length overlapping sequences.

CONCLUSIONS

Bellerophon detects all putative chimeras in a multiple sequence alignment in a single analysis. This means that sequences from a PCR-clone library can be analyzed and results interpreted in parallel instead of serially as is the case with other commonly used programs such as CHIMERA_CHECK (Cole *et al.*, 2003). This also has the advantage that individual sequences do not become "invisible" to the program since they are not compared in isolation to a reference dataset (Hugenholtz and Huber, 2003). Bellerophon has been available to the scientific community since January 2003 and has been trialed by over 100 different users (ca. 70 submissions per month). As with all chimera-detection programs, putatively identified chimeras should be manually verified, for example by inspecting the sequences for signature shifts (Wang and Wang, 1997).

REFERENCES

Brosius,J., Palmer,M.L., Kennedy,P.J. and Noller,H.F. (1978) Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, **75**, 4801-4805.

Cole,J.R., Chai,B., Marsh,T.L., Farris,R.J., Wang,Q., Kulam,S.A., Chandra,S., McGarrell,D.M., Schmidt,T.M., Garrity,G.M. and Tiedje,J.M. (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res*., **31**, 442-443.

Hugenholtz,P. and Huber,T. (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int. J. Syst. Evol. Microbiol.,* **53**, 289-293.

Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In Munro, H.N. (ed), *Mammalian Protein Metabolism, vol. 3*. Academic Press, New York, pp. 21-132.

Komatsoulis,G. and Waterman,M. (1997) A new computational method for detection of chimeric 16S rRNA artifacts generated by PCR amplification from mixed bacterial populations. *Appl. Environ. Microbiol.,* **63**, 2338-2346.

Lane,D.J. (1991) 16S/23S rRNA sequencing. In Stackebrandt,E. and Goodfellow,M (eds), *Nucleic Acid Techniques in Bacterial Systematics*, John Wiley and Sons, New York,.pp. 115-175.

Liesack,W., Weyland,H. and Stackebrandt,E. (1991) Potential risks of gene amplification by PCR as determined by 16S rDNA analysis of a mixed-culture of strict barophilic bacteria. *Microb. Ecol.,* **21**, 191-198.

Robinson-Cox,J.F., Bateson,M.M. and Ward,D.M. (1995) Evaluation of nearest-neighbor methods for detection of chimeric small-subunit rRNA sequences. *Appl. Environ. Microbiol.,* **61**, 1240-1245.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673-4680.

von Wintzingerode,F., Göbel,U.B. and Stackebrandt,E. (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol. Rev.,* **21**, 213-229.

Wang,G.C.Y. and Wang,Y. (1997) Frequency of formation of chimeric molecules is a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl. Environ. Microbiol.*, **63**, 4645-4650.