Hunting the Monsters in the Database

A critical review of computational tools for detecting chimeric 16S PCR amplification

Koshlan Mayer-Blackwell | March 13, 2011 | Computational Molecular Biology

> "The Chimera had the head of a lion and the tail of a serpent, while her body was that of a goat, and she breathed forth flames of fire; but Bellerophon slew her, for he was guided by signs from heaven."

- Book IV, Homer's *Iliad*. circa 800 BCE

> "Unprecedented diversity in a range of samples has been reported using pyrosequencing, and has been interpreted as evidence of an important and pervasive rare biosphere." When applied, "rigorous chimera checking … reduced diversity estimates based on pyrosequencing by a factor of 10."

-Hass et al. *Chimeric 16S detection using chimera slayer.* 2011

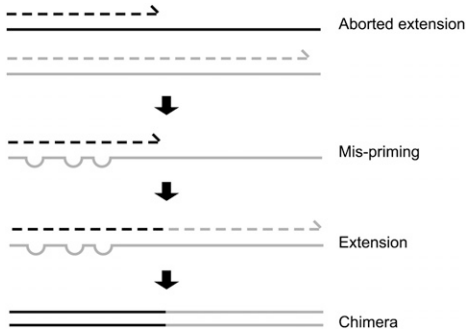
Table of Contents

1 Introduction: Chimeras the monsters in the database

As a molecular marker of microbial diversity and abundance, the ribosomal 16S small-subunit gene is the workhorse of both environmental and medical microbiology. Whether probing microbial diversity in the 70°C water of a Yellowstone hot spring (Pitulle et al. 1998) or the human foreskin (Price et al. 2010), 16S sequences can be amplified by polymerase chain reaction using broad specificity primers. Such culture-independent surveys have dramatically expanded knowledge of microbial diversity (Pace 1997). In the last decade, the arrival of next-generation nucleotide sequencers increased the number of 16S sequences that could be feasibly recovered from a given sample, permitting the discovery of low-abundance organisms – that make up the so-called "rare biosphere" -- previously imperceptible within conventional clone libraries (Sogin et al. 2006).

While next-generation sequencing instruments, can achieve staggering coverage depth (generating more than a million 500 base-pair reads in a single machine-run), they do not fully obviate the PCR amplification step required to enrich for the 16S gene. For instance, the best alternative method -- whole-genome shotgun sequencing of a total DNA sample -- yields less than 0.5% of the 16S reads achievable by a targeted approach (Hass et al. 2011; Shah et al. 2011), making PCR-enrichment the mainstay of studies examining diversity over a time-course or between multiple environments.

A major drawback in PCR-targeted surveys is the potential to form chimeric junctions due to incomplete PCR amplification. Formation of chimeric sequences can occur when DNA polymerase terminates prematurely or when a sheared template is incomplete. In the next round of denaturing and annealing, the portion of the incomplete strand near the breakpoint may "mis-prime" a similar sequence, acting as a primer for extension on a new template. In subsequent steps, the original primers can further amplify the newly formed chimeric sequence. (See informative figure below, taken directly from Haas et al. 2011).



Chimeric sequences join two or more parent strands and can lead to false estimates of diversity due to novel sequences from non-existent organisms

(Huggenholtz et al. 2003). The experimentally observed rates of PCR chimera formation from a mixed group of bacterial genomes is significant, with chimera formation reported as high as 30% of total sequences under low-stringency conditions (Wang and Wang 1997).

The recognition of this problem is not new (Shuldiner et al. 1989), but tools to detect chimeric PCR amplification have lagged behind the ability of researchers to sequence new environments and seed the public sequence databases. In 2005, well before the boom in high-throughput16S surveys, a study of the Ribosomal Database Project (RDP) with a new computational detection tool revealed that 5% of sequences were corrupt, with chimeras as the leading culprit (Ashelford et al. 2005).

The good news is that by 2011 multiple computational tools are available to detect 16S chimeras. In some cases, these tools can be applied retroactively to flag blatant chimera sequences from public databases. This paper reviews the development of computational tools aimed at detecting 16S chimeras. A recent article uses a "synthetic microbial community" generated from genomic DNA of fully sequenced organisms to compare the performance of two popular detection algorithms, Bellerophon and Pintail, against a new contender: Chimera Slayer. Readers interested in data-based comparison of these methods should see Haas et al. 2011. In this paper, I mention the result of this study. But the emphasis is placed on describing the differences between chimera detection algorithms, past and present, and their shortcomings. The paper compares five distinct computational tools, discussing each in the chronological order they were released. Suggestions for future concerted empirical and computational collaboration are suggested in the concluding section.

2 PCR-16S Chimera Detection

2.1 Overview of selected 16S Chimera detection methods

Computational Tool (year released)	**Fragments query**	**Database Dependent**	**Description**	**Metric**
Check_Chimera (1994)	Yes	Yes	Nearest-neighbor method.	Improvement Score(IS)
Bellerophon (2006)	No	No	Each fragment is subject to a MSA-distance method	Preference Score
Pintail (2005)	No	No	Combines 16S variability profile with single alignment	Deviation statistics (DE)
KmerGenus (2011)	Yes	No	Looks for two 50-mers specific to separate genus	Match between two different genus-specific k-mers.
Chimera Slayer (2011)	Yes	Yes	Dynamic programming to generate i*n silico* chimeras	Minimum Divergence

2.2 Check_Chimera and the Nearest-Neighbor Alignment Methods

The program **Check_Chimera** was included in the early release of the Ribosomal Database Project. In 1994 it was described in the electronic mail server commands (Maidek et al. 1994):

CHECK__CHIMERA	Analyze a user-supplied sequence for evidence of chimeric structure. Options allow the user to add their own sequences to the database used in the analysis and to ignore short matches with partial sequences.

It has since received a number of modifications, but the principle of the original class of nearest-neighbor method is as follows:

The candidate sequence is split into two parts. Both parts are searched against a database for the nearest neighbor, with the goal of detecting inconsistent "phylogenetic affiliation" between the fragments. The difference in similarity scores generated by the alignment of the two fragments with their respective nearest neighbors vs. the best possible single alignment of the original query yields an "improvement score" (Robinson-Cox et al. 1995). This general operation is repeated for multiple breakpoints along the length of sequence to search for highest possible improvement score.

A limitation of this method's metric, the improvement score (IS), is the lack of a statistical confidence measurement of whether a sequence is chimeric. Rather the magnitude of the improvement score is really a measure of the degree of sequence difference between two potential parental fragments. A high improvement score usually occurs when a chimera formed between distantly related parent fragments, the kind of recombination that is more obvious in hindsight. However, the magnitude of the improvement score is low, and much less informative, when the

generated from a chimera between two closely related parent sequences. It is also important to remember in the case of similar parent sequences joined by PCR artifact, the results of nearest neighbor method can be sensitive to the scoring matrix or penalties used to calculate alignment scores. (Komatsoulis & Waterman 1997).

Robinson-Cox et al. (1995) and Komatsoulis and Waterman (1997) both offered modifications to Check_Chimera, improving its sensitivity and discrimination slightly. These nearest-neighbor methods share a common reliance on single sequence alignment. That is, the each query fragment is compared against a single parent sequences. The advantage of this approach is that one need not enter a query of specific sequence length. However, a major liability of this approach can occur if a chimera makes it into the database. It can hinder the Check_Chiamera's detection of similar ones.

2.3 Bellerophon – A Treeing-Distance Method

In 2003, a new approach was introduced in the form of the tool Bellerophon, cleverly named after the Greek mythic chimera slaying hero in Homer's *Iliad*. Whereas, the nearest neighbor methods rely on single alignments scores, Bellerophon is based on comparing the branching patterns of a multiple sequence alignment. This partial-treeing method has the following key features (Huber 2004):

- The Query sequence is split into fragments on either side of an arbitrary breakpoint and each fragment is aligned with the best hits from either (i) a reference dataset of 16S sequences or (ii) together with the other sequences that make up a newly generated library.

- Two quantities are calculated: (i) an aggregate distance matrix error "**dme"** and (ii) the aggregate distance matrix error "**dme[i]"** if a single sequence is excluded from the matrix**.**

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

where d[i][j] = the distance between two sequences i and j

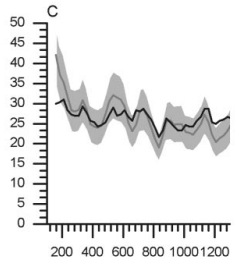
- The fraction $\frac{dme}{dme[i]}$ yields a preference score. Where a chimeric sequence accounts for significant portion of the total distance matrix error, the preference score > 1.
- The algorithm repeats, scanning the sequence at 10 nucleotide intervals. The maximum preference score is predictive of a chimeric junction.
- The algorithm requires equal sized window on either side of the putative break points.

An attractive feature of Bellerophon treeing is the ability to use a one's own newly generated library against its self. This avoids potential problems of hidden chimeras in the database and allows for the processing of a whole library in single test. If sequences are from a particularly rare environment, the alignments in a large library may be more informative than one formed with a database. A web implementation of Bellerophon can be accessed as a stand-alone (http://comp-bio.anu.edu.au/bellerophon/bellerophon.pl) or together with a filter provided by GreenGenes at the Lawrence Berkeley National Lab (http://greengenes.lbl.gov/cgi-bin/JD_Tutorial/nph-ChimeraSteps1.cgi). The LBNL implementation removes sequences that are highly similar to known non-chimeras to reduce the computation time required to process a large library.

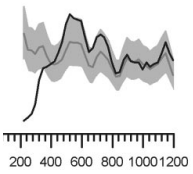
The chief limitation of Bellerophon is its requirement for large input sequences to permit equal sized windows (no sequence shorter than twice the window size can be analyzed). With a minimum recommended window size of 200 nucleotides, the method does poorly at detecting of chimeric junctions at the beginning or end of the 16S. Ballerophon is less well suited for work with short <500bp reads from a 454 pyrosequencer. It should be noted that, a reimplementation of Ballerophon for use with 454 length reads by Haas and colleagues (2011) showed low sensitivity and high false positive for the most challenging to detect chimeras (parental divergence < 15%).

2.4 Pintail – A Combined Profile + Alignment Method

The fundamental assumption of the Pintail method is that two non-chimeric "rRNA sequences of known overall evolutionary distance will vary by roughly the same amount over the length of the gene" (Ashelford et al. 2005). Before jumping into a discussion of the mechanics of the method, it is worth setting the stage by looking at the visual output of a legitimate vs. a blatant chimeric sequence. (Graphic images selected from Ashelford et al. 2005)



Non Chimeric Sequence



Chimeric Sequence

The ***x-axis*** is the base position of each nucleotide aligned to an *E.coli* reference 16S sequence. The ***y-axis*** is the percentage of nucleotides within in sliding window that differ from a user-selected reference. The black line represents our query sequence and the gray lines are a measure of expected variation combining information about the known regions of variability along the 16S gene and the overall divergence of the query and the reference. Notice how in the chimeric case, only a portion of the black line tracks with prediction.

Now let's discuss, the backbone of the model: a multiple sequence alignment derived from the 16S genes of all sequenced isolates in the RDP database, with the the known E.coli 16S as a reference. (16S sequences from cultured isolates are presumed to be much less likely to be chimeras than those from environmental samples). The *consensus* sequence represents the most likely nucleotide in each position. The rate of agreement with the consensus is not uniform over the length of the 16S. To the contrary, certain regions are known to be hyper-variable, and nucleotides there have a higher chance of being *divergent.* To its credits, pintail incorporates this biological fact wisely. From the RDP alignment, the method generates a "probability profile" reflecting the likelihood that each position will diverge form the consensus is stored in array **Q.**

$$\mathbf{Q} = \{q_j: q_1, q_2, q_3 \ldots . q_{1542}\}$$

This information can be smoothed into a shorter array of <u>average</u> expected divergence from the consensus within *m* windows of specified size. For *m* windows, *Q is transformed to* $\mathbf{Q_{average}}$.

$$\mathbf{Q_{av}} = \{a_j: a_1, a_2, a_3 \ldots . a_m\}$$

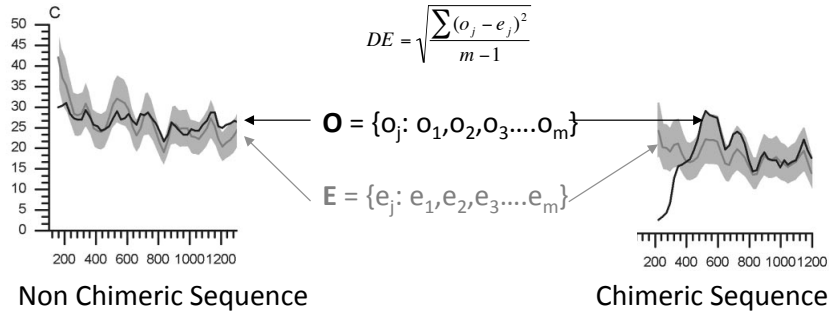
Up to this point we have only considered the architecture of Pintail and said nothing about the user input. The user inputs two sequences: a query sequence $\boldsymbol{S_q}$ that she wishes to test for anomalies and a second user-selected subject sequence $\boldsymbol{S_{s.}}$, which should be somewhat similar to the query and confirmed as legitimate (i.e. non-chimeric). The subject sequence - $\boldsymbol{S_s}$ can but need not be the nearest neighbor of the query sequence $\boldsymbol{S_q}$**.**

To match the dimensions of $\mathbf{Q_{av}}$, an array **O** is generated with information from the alignment of S_q and $S_{s.}$ For each window 1 to m, $\mathbf{O} = \{o_j: o_1, o_2, o_3 \ldots . o_m\}$ indicates the percentage of matching positions between S_q and S_s.

Once **O** is generated, it is possible to generate a proxy (Σo_j /m) for the total overall diversity between S_q and $S_{s.}$ A scaling factor (α): is multiplied against each entry of $\mathbf{Q_{av}} = \{a_j: a_1, a_2, a_3, \ldots a_m\}$ to generate an expected percentage of divergence within each window that combines information about overall diversity between S_q and S_s and position specific variability:

$$\alpha = \frac{\frac{\sum o_j}{m}}{\frac{\sum a_j}{m}} \; ; \; \mathbf{E} = \{a_1\alpha, \; a_2\alpha \; , a_3\alpha, \; .. a_m\alpha\} = \{e_1, e_2 \; , e_3 \ldots . e_m\}$$

Further discussion is much helped by re-examining the visual output for a chimeric and non-chimeric sequence produced by the Pintail interface. (Images selected directly from Ashelford et al. 2005, *annotation mine*):



The ***x-axis*** is the base position of each nucleotide aligned to an *E.coli* reference sequence. The ***y-axis*** is the percentage of nucleotides that differ from the consensus at all positions within a window. The dark ***black line*** is a plot of **O** = {oj: o1,o2,o3…om} against base position. The ***gray line*** is the expected variation -- **E** = {e_1, e_2 ,e_3….e_m} ploted against base position.

Note that the peaks in the expected sequence coincide with known hyper-variable regions. The grey area around the expected line represent +/– 5% expected difference. The standard deviation **DE** between each element of **O** and **E** is the final metric reported to the user (see figure above).

The major benefit of this method over previous methods, beyond accounting for hyper-variable regions, is an attribution of a statistical confidence measure to its output metric DE. The observed DE value can be compared against a previously calculated set of comparisons between sequences of similar overall diversity. The degree to which the DE of the query and reference exceeds the maximum legitimate DE of the reference and another legitimate comparison allows for the estimate of a p value.

The authors suggest that a major benefit of Pintail compared to other tools is the lack of dependence on user-prepared database (Asherford et al. 2005). This is true if the user is only checking a small number of sequences, but is not really an advantage when the user analyzes a clone library, the most common application in need of chimera detection. Second, the tool is highly specific to investigation of chimeric 16S. Other methods could be more easily adapted to studying chimera in a PCR-amplified functional genes.

2.5 NAST Alignments for high-throughput

In the past decade, new sequencing instruments, particularly the 454-Roche pyrosequencer, have emerged as popular tools for targeted 16S amplification libraries. The 454 instrument yields a far lower cost per read compared to that associated with Sanger sequencing, however individual read length is reduced by more than half. Consequently, it is now common to only amplify and sequence a segment of the 16S covering one or more of the hyper-variable regions. Not all computational chimera detection tools are well suited for the shift to shorter sequences. As mentioned previously, the original Bellerophon algorithm was meant to detect chimeric junctions on sequences 2x its minimum 200bp window size. What is more, simply converting thousands of reads of differing length and quality

into a multiple sequence alignment (MSA) of a fixed length needed to run the Bellerophon algorithm creates its own difficulties.

To facilitate high-throughput 16S studies, DeSantis and colleagues (2006) developed Nearest Alignment Space Termination NAST tool capable of aligning a high number of 16S genes within a fixed number of columns. The details of the NAST algorithm are beyond the scope of this paper and are discussed in the original paper by DeSantis et al. (20056). Suffice it to say, that without NAST the automated alignment of numerous sequences with insertions (either real or due to sequencing error) would cause misalignments or expansion of columns to the point where the MSA no longer resembled the spacing of intact 16S sequence.

2.6 Explicitly high throughput chimera detection tools

The NAST algorithm has enabled the newest set of chimera detection tools released in 2011. Haas et al. (2011) have developed **Chimera Slayer** and **KmerGenus** as well as re-implementing Bellerophon and Pintail to be NAST compatible.

KmerGenus is relatively simple. It mines a database of 16S sequences with "validated taxonomic predictions" for all overlapping 50-mer sequences and assigns them to a genus. Queries that match unique 50-mers from two distinct genus groups are flagged. The main advantage of KmerGenus is that is can analyze a query of any reasonable length and requires no multiple sequence alignment. However, its authors show that it is much less sensitive at detecting chimeric sequences from similar parents than other available methods (Haas et al. 2011).

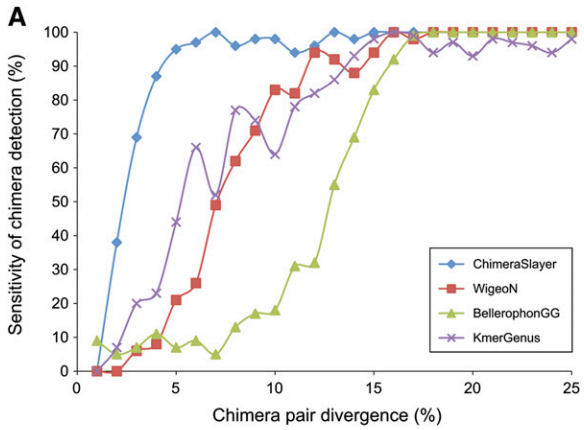
Ironically the approach taken in **Chimera Slayer** most closely resembles the Check_Chimera nearest-neighbor methods modified by Waterman in 1997. The main features of Chimera Slayer method are as follows (Haas et al. 2011):

- The terminal ends of the original search query are split into two fragment.
- Relying on a database 16S sequences putatively free of Chimeric sequences, the 15 nearest neighbors to each fragment are retrieved and NAST formatted. A the nearest NAST-formatted neighbor sequences are used as hypothetical parents fragments to make in silico chimeras of the query.
- By combing hypothetical parent fragments at all possible break points, the highest scoring alignment with the query is identified using dynamic programming. Like the dynamic programming approach utilized for global sequence alignment by Needleman and Wunsch (1970) the discovery of the highest-scoring alignment is guaranteed for a given the gap and substitution penalties.
- The percent identity between (i) the original query sequence and best *in silico* chimera is divided by (ii) the percent identity of the query with either the single parent alone. A "minimum divergence" ratio of 1.007 or above was used to flag possible PCR anomalies.
- For those flagged sequences, further computations can be done to validate location of the breakpoint (see Haas et al 2011).

3 Conclusion

Virtually all papers written on the topic of 16S chimera detection allude to the elephant in the PCR tube, the inability of computational tools to detect chimeras between two closely related parent fragments. Thus it is no surprise that the recent paper introducing Chimera Slayer (CS) highlights its high sensitivity. A third-party data-driven review of all modern algorithms is certainly required as CS's authors comparison is based on their own implementations of their competitors' software. Nonetheless CS's high sensitivity will certainly catch a number of chiameric sequences that would otherwise take up residence in the public databases. Pictured below is comparative data from Haas et al. showing > 90% detection of chimeras from closely-related parent sequences.

Haas et al.



Given the broad spectrum of bacterial genomes used in the study, this heightened sensitivity is good news indeed. However noticeably absent in Haas et al. 2011 paper was a discussion of a much harder measure to quantify: selectivity. That is, among closely related sequences how often is a truly novel sequence rejected as a chimera? While false sequences in the database are a problem, the high-throughput type II error is also problematic.

As chimera detection reaches a new level of sophistication, it may be worth investigating the secondary structure that foster premature polymerase termination. In the past, the "exact breakpoints [were] difficult to determine because the parent sequences are usually identical around the recombination site" (Hugenholtz 2003). However, the idea that chimerization is completely random has been rejected. The same chimeric sequences are often found in independent PCR

reactions (Haas 2011), raising the possibility that, with more empirical and computational collaboration, it may possible to predict chimeric junctions *ab initio*.

4 Work Cited

Acinas, S.G. et al., 2005. PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. Applied and Environmental Microbiology, 71(12), p.8966-8969.

Ashelford, K.E. et al., 2005. At Least 1 in 20 16S rRNA Sequence Records Currently Held in Public Repositories Is Estimated To Contain Substantial Anomalies. Applied and Environmental Microbiology, 71(12), p.7724-7736.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., et al., 2006a. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Applied and Environmental Microbiology, 72(7), p.5069.

DeSantis, T.Z., Hugenholtz, P., Keller, K., et al., 2006b. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Research, 34(Web Server), p.W394-W399.

Haas, B.J. et al., 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome research, 21(3), p.494-504.

Huber, T., Faulkner, G. & Hugenholtz, P., 2004. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics, 20(14), p.2317-2319.

Hugenholtz, P., 2003. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. International Journal of Systematic ….

Komatsoulis, G.A. & Waterman, M.S., 1997. A new computational method for detection of chimeric 16S rRNA artifacts generated by PCR amplification from mixed bacterial populations. Applied and Environmental Microbiology, 63(6), p.2338-2346.

Maidak, B.L. et al., 1994. The Ribosomal Database project. Nucleic Acids Research, 22(17), p.3485-3487.

Pace, N., 1997. A molecular view of microbial diversity and the biosphere. Science.

Price, L. et al., 2010. PLoS ONE: The Effects of Circumcision on the Penis Microbiome. PloS one.

Robison-Cox, J., Bateson, M. & Ward, D., 1995. Evaluation of nearest-neighbor

methods for detection of chimeric small- subunit rRNA sequences. Applied and Environmental Microbiology, 61(4), p.1240.

Shah, N. et al., 2011. COMPARING BACTERIAL COMMUNITIES INFERRED FROM 16S rRNA GENE SEQUENCING AND SHOTGUN METAGENOMICS. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, p.165-176.

Sogin, M., Morrison, H. & Huber, J., 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." In Proceedings of the …. Proceedings of the ….

Wang, G. & Wang, Y., 1997. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. Applied and Environmental Microbiology, 63(12), p.4645.