DECIPHER: A Search-Based Approach to Chimera Identification for 16S rRNA Sequences

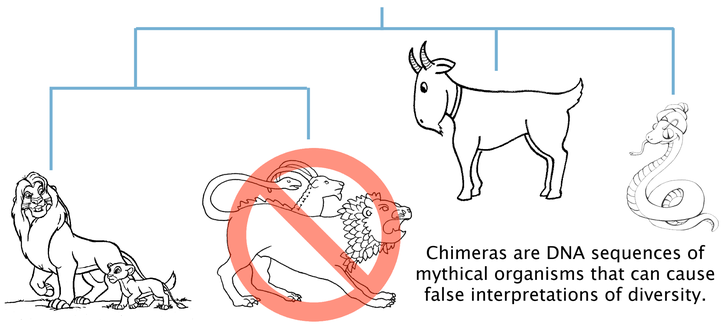
Erik S. Wright (DECIPHER@cae.wisc.edu), L. Safak Yilmaz, Daniel R. Noguera

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

DEPARTMENT OF
Civil and
Environmental Engineering
UNIVERSITY OF WISCONSIN-MADISON

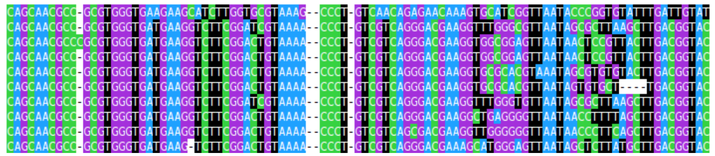
Introduction

Why are chimeras bad?

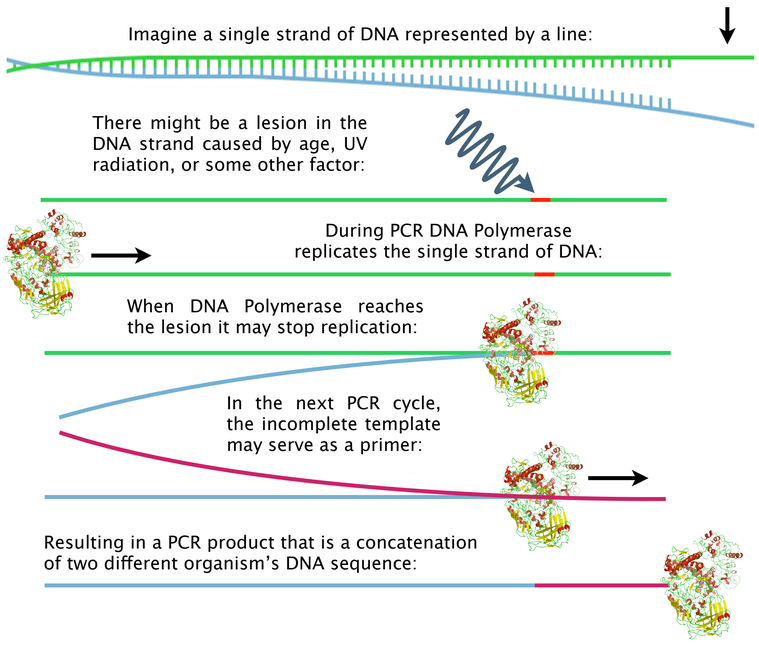


Chimeras are DNA sequences of mythical organisms that can cause false interpretations of diversity.

Which one of these sequences is a chimera?



Where do chimeras come from?



Methods

How do you find chimeras with DECIPHER?

This method for chimera finding was derived by modeling a set of several hundred real chimeric sequences present in public sequence repositories and thousands of artificial chimeras generated from the rRNA sequence of isolated organisms:

1. Break classified sequence into overlapping 30 nucleotide fragments:

2. Categorize fragments by prevalence (search hits):
   - High frequency in classified group:

   - Low frequency in-group and high frequency out-of-group (suspect fragments):

3. Apply Rules to chimeric region:
   - At least 6 suspect fragments for full-length sequences (fs):

     At least 2 suspect fragments for short-length sequences (ss):

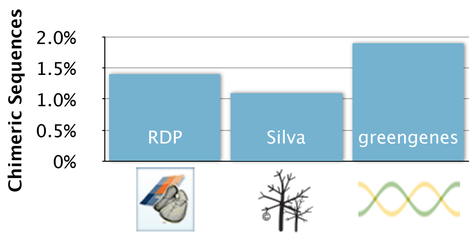
   - At least 70 nucleotides long for full-length sequences (fs):

     At least 40 nucleotides long for short-length sequences (ss):

   - At least 1 nucleotide overlapping first or last 200 nucleotides:

   - Greater than 60% coverage with suspect fragments:

How was this applied to a real problem?



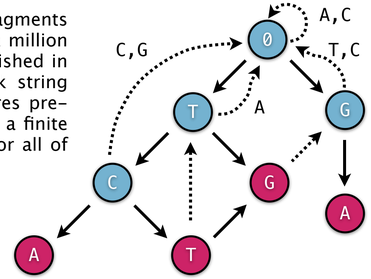
Major public DNA sequences repositories contain millions of 16S Ribosomal RNA sequences.

The curators of these databases screen for sequence abnormalities such as chimeras, yet additional chimeras are still present in these public sequence repositories.

How was the solution implemented?

Hundreds of thousands of suspect fragments can be queried against a set of over a million 16S rRNA sequences. This is accomplished in linear time by using an Aho-Corasick string search algorithm. This method requires pre-processing the 30-mer fragments into a finite state machine that enables querying for all of the sequence patterns simultaneously.

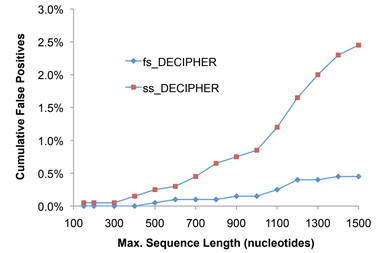
The example shows the state machine for the pattern set: {TG, TCA, TCT, GA}.

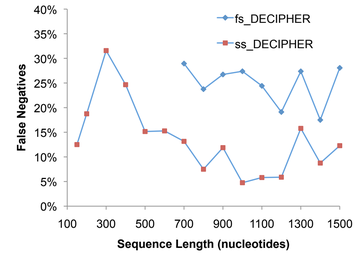


Results

What is the accuracy of this method?

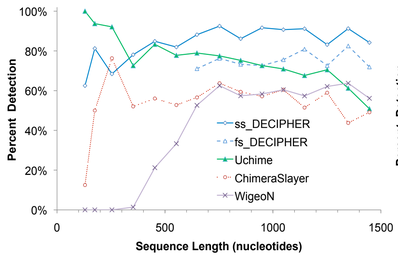
The method was validated by testing thousands of artificially generated chimeras developed by joining the rRNA sequences of randomly selected isolate organisms. Using this method resulted in many chimeras that were frequently difficult to differentiate from their parent sequences. Therefore, this sequence set served as an excellent benchmark for comparison between different chimera finding programs.

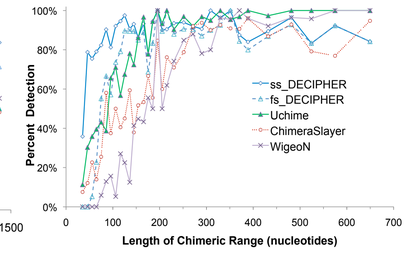




How did this method compare to others?

Qualitative rating for each characteristic	DECIPHER[1]		Uchime[2]	Chimera Slayer[3]	Pintail[4] (WigeoN)
	fs	ss			
Detection in short sequences	+	+++	++++	+	+
Detection in mid-range sequences	+	++++	+++	++	+
Detection in long sequences	+++	++++	+++	++	++
Detection of short chimeric regions	++	++++	++	++	+
Detection of complex chimeras	++++	++++	++++	+	+++
Detection of chimeras from low divergence parents	+	+	++++	+++	+
Independence from reference dataset	+++	+++	++	++	++
Low false positives	++++	++	++	+++	++





Where is there more information?

Find chimeras online at:
DECIPHER.cee.wisc.edu
Contact Erik at:
DECIPHER@cae.wisc.edu

1. **Wright, E. S., Yilmaz L. S., and Noguera D. R.** 2012. "DECIPHER: A Search-Based Approach to Chimera Identification for 16S rRNA Sequences", Appl. Environ. Microbiol., doi:10.1128/AEM.06516-11.
2. **Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince, and R. Knight.** 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics:Epub ahead of print, doi: 10.1093/bioinformatics/btr1381.
3. **Quince, C., A. Lanzen, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read, and W. T. Sloan.** 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. Nat. Methods **6**:639-U627.
4. **Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman.** 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl. Environ. Microbiol. **71**:7724-7736.