Greengenes, the ARB-compatible chimera-checked 16S rRNA gene database, introduces new online tools, including NAST (a flexible massive multiple sequence alignment tool).

DeSantis, T.Z. Hugenholtz, P. Keller, K. ***Brodie, E.L.*** Larsen, N., Piceno, Y.M. Phan, R. Andersen, G.L.

Lawrence Berkeley National Laboratory, Center for Environmental Biotechnology, Berkeley, CA, USA.
DOE Joint Genome Institute, Microbial Ecology Program, Walnut Creek, CA, USA.
Danish Genome Institute, Aarhus, Denmark.
Lawrence Berkeley National Laboratory, Virtual Institute for Microbial Stress and Survival, Berkeley, CA, USA.
University of California, Quantitative Biomedical Research, Berkeley, CA, USA

greengenes.lbl.gov

Abstract

Microbiologists conducting surveys of bacterial and archaeal diversity often require comparative alignments of thousands of 16S rRNA genes collected from a sample. The computational resources and bioinformatics expertise required to construct such an alignment has inhibited high-throughput analysis. It was hypothesized that an online tool could be developed to efficiently align thousands of 16S rRNA genes via the NAST (Nearest Alignment Space Termination) algorithm for creating multiple sequence alignments (MSA). The tool was implemented with a web-interface at http://greengenes.lbl.gov/cgi-bin/nph-NAST_align.cgi. Each user-submitted sequence is compared to Greengenes' "Core Set", comprising approximately 10,000 aligned non-chimeric sequences representative of the currently recognized diversity among bacteria and archaea. User sequences are oriented and paired with their closest match in the Core Set to serve as a template for inserting gap characters. Non-16S data (sequence from vector or surrounding genomic regions) is conveniently removed in the returned alignment. From the resulting MSA, distance matrices can be calculated for diversity estimates and organisms can be classified by taxonomy. The ability to align and categorize large sequence sets using a simple interface has enabled researchers with various experience levels to obtain prokaryotic community profiles.

Introduction

Multiple sequence alignments (MSA) are used for annotating conserved versus variable gene loci by observing heterogeneity along the columns. They are critical for a variety of analyses, and yet they create a bottle-neck in data analysis. Frequently, when adding a candidate sequence to a MSA profile, one or more internal insertions will be discovered that cannot be accommodated in the profile. This event requires a researcher to make one of two choices:

1) allow the column count to grow whenever an insertion is required, which requires each sequence to gain more characters or
2) allow a local misalignment within a sequence (row) so that the insertion does not disrupt the entire alignment format of the profile.

Until now, only the former was available. NAST (10) was created to allow the second choice. It is intended to facilitate comparison of thousands of user-supplied 16S rRNA gene sequences from bacteria and archaea.

One unique feature is that NAST can output the MSA in a standard, consistent format of 7,682 characters per sequence, so that similar loci are located at dependable positions from batch to batch (necessary for large, ongoing projects). An optional pre-processing of data based on chromatogram quality scores is allowed and post-processing options include distance matrix creation and taxonomic classification using five independent curators' nomenclature.

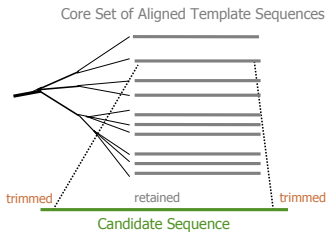


Figure 1. Locating a NAST alignment template for a user-supplied candidate sequence. Candidate sequence in green is matched to a near-neighbor aligned template in Greengenes' Core Set (grey) by tallying 7-mers in common. The alignment "template" is BLAST aligned to the candidate parameter g = -1 (favors long match). The candidate is then trimmed of flanking sequence data such as rRNA, intergenic spacer regions, vector sequence, 23S rDNA and sequence outside of the high-scoring pair (HSP) boundaries. If the HSP pairs opposite strands, then the candidate is reverse complemented.

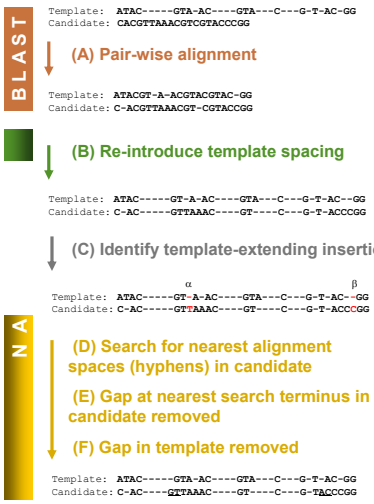


Figure 2. Example of **NAST (Nearest Alignment Space Termination)** compression of a BLAST pair-wise alignment using a 38 character aligned template. Template and candidate is extended to 40 characters after BLAST gap insertion (A) and retention of original template spacing (B). Nucleotide insertions in the candidate relative to the template which force additional characters to be added in the template are identified at positions α and β (C). A bi-directional search for the nearest alignment space (hyphen) relative to the insertion terminates at the positions indicated by the black arrows (D). The leftward search from the α position was shorter in distance compared to the rightward, thus the space left of 'GT' was removed. The search from the β position encountered the alignment edge on the right, thus the position to the left of 'AC' was removed (E). Lastly, the two template-extending spaces are deleted from the template (F). The NAST removal of two characters from both sequences allowed local misalignments (underlined) while preserving the 38 character format of the global multiple sequence alignment.

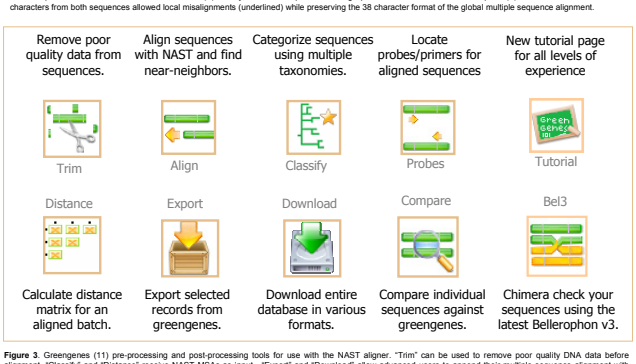
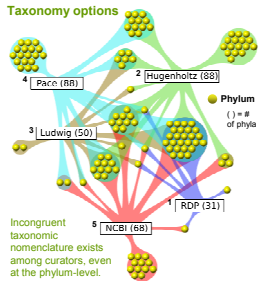
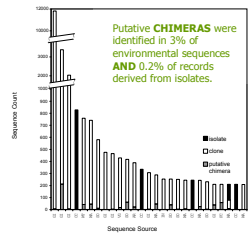


Figure 3. Greengenes (11) pre-processing and post-processing tools for use with the NAST aligner. "Trim" can be used to remove poor quality DNA data before alignment. "Classify" and "Distance" receive NAST MSAs as input. "Export" and "Download" allow advanced users to append their multiple sequence alignment with select sequences from the public repositories. Bel3 allows chimera checking against either your own library or a core set of non-chimeric sequences. The new tutorial page has been developed to cater for all levels of greengenes user experience including high school students.





Additional Features in Greengenes

ARB compatibility

- import ←→ export methods
- simplified personal ARB maintenance

Taxonomy tracked from multiple curated databases

- Hugenholtz, Pace, Ludwig, NCBI, RDP, Andersen Phylochip

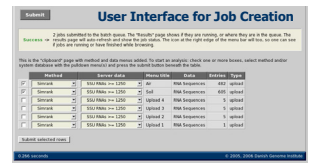
Chimera evaluation of each 16S rRNA gene record

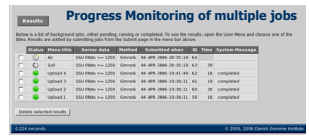
- by alignment to core set
- by Bellerophon v3

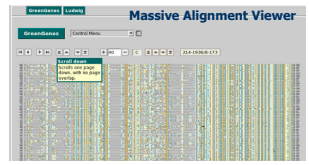
Community curation

- suggest a group name
- improve an alignment
- contest a chimera call
- add to Core Set
- correct sequence descriptions

In Development







Summary

A "Core Set" of 10,578 non-chimeric (6) sequences was assembled. Each sequence is representative of a 96% identity cluster with all sequences being over 1250 nt in length. The Core Set is considered the profile MSA and consists of the template sequences aligned into 7,682 columns.

Local misalignments, spanning from the insertion base to the deleted alignment space, are permitted to preserve the global multiple sequence alignment format.

Comparing the submitted length to the post-NAST length can alert the user of unexpected sequence truncation.

- Minor truncation of one to five bases occurs when terminal bases cannot be accurately aligned.
- Large truncations indicate that either non-gene data was in the record or that BLAST found matches distributed to multiple Core Set sequences, possibly suggesting chimeric content.
- Long insertions are reported for identification of sequences divergent from the Core Set.

Examples

An 1,800 sequence set from uranium contaminated soil, deep sub-surface water, and urban aerosols was NAST aligned, allowing analysis of sample diversity as well as evaluation of parallel 16S rRNA microarray results (7, 8).

From sea water along the Atlantic coast of the U.S., 8,690 sequences were aligned, 7,244 of which were full-length genes (9).

NAST's utility is not limited to 16S rRNA data. Sizeable MSAs of other genes or proteins, such as those encoding 18S rRNA, *rpo*B, or *rec*A can be built and maintained.

References

(1) Cole, J. R., B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje. 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res 33:D294-6.

(2) Hugenholtz, P. 2002. Exploring prokaryotic diversity in the genomic era. Genome Biol 3:1-8.

(3) Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. H. Schleifer. 2004. ARB: a software environment for sequence data. Nucleic Acids Res 32:1363-71.

(4) Pace, N. R. 1997. A molecular view of microbial diversity and the biosphere. Science 276:734-40.

(5) National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/.

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

(7) DeSantis, T.Z., Brodie, E.L., Moberg, J.P., Zubieta, I.X., Piceno, Y.M. and G.L. Andersen. High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microb. Ecol. (in press)

(8) Brodie, E.L., DeSantis, T.Z., Joyner, D.C., Baek, S., Larsen, J.T., Andersen, G.L., Hazen, T.C., Herman, D.J., Tokunaga, T.K., Wan, J.M. and Firestone, M.K. Application of a High-Density Oligonucleotide Microarray Approach To Study Bacterial Population Dynamics during Uranium Reduction and Reoxidation. Appl. Environ. Microbiol. (in press).

(9) A. Shaw, unpublished.

(10) DeSantis, T.Z., Hugenholtz, P, Keller, K., Brodie, E.L., Larsen, N., Piceno, Y.M., Phan, R., Andersen, G.L. 2006. NAST: A multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res 34:W394-9.

(11) DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L. 2006. greengenes: Chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Envir. Microbiol. 2006 72: 5069-5072

Acknowledgements

The computational infrastructure was provided in part by the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL Program and the Natural and Accelerated Bioremediation Research Program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy. Web application development was funded in part by the Department of Homeland Security under grant number HSSCHC04X00037. A huge thanks to Jonathan Davies for creating the latest tutorial.

Contacts

Todd DeSantis – TDeSantis@lbl.gov Eoin Brodie – ELBrodie@lbl.gov Gary Andersen – GLAndersen@lbl.gov