A web-based method to calculate average amino acid identity (AAI) between prokaryotic genomes.

Authors: Andrew N. Gale[1], Jordan E. Krebs[1], Thomas C. Sontag[1], Victoria K. Keyser[1], Eileen M. Peluso[2] & Jeffrey D. Newman[1]
[1] Biology Department, Lycoming College, 700 College Place, Williamsport, PA, 17701
[2] Department of Mathematical Sciences, Lycoming College, 700 College Place, Williamsport, PA, 17701

Abstract

With the decreasing cost of NextGen sequencing and the subsequent increase in the availability of microbial genome sequences, it has been suggested that the prokaryotic species definition should change from physical measurements of DNA-DNA hybridization (DDH) to computationally-determined genome-wide metrics. The method described here calculates one such metric, average amino acid identity (AAI), using easily accessible, web-based tools. The AAI calculation is based on the protein-length-weighted pairwise identity of orthologous proteins as determined by bidirectional best hits between a reference genome and up to ten comparison genomes. A JavaScript-based calculator (www.lycoming.edu/~newman/aai) analyzes the output from the "Sequence-based comparison" tool on the Rapid Annotation with Subsystems Technology (RAST) service (rast.nmpdr.org) and yields values similar to previously-published, but not web-accessible AAI calculation methods.

Background

- For over 40 years, the threshold for species clustering has been 70% DDH. (Fig 1)
- Next, species differentiation was based on DNA sequences such as 16srRNA and Multi-Locus Sequence Analysis (MLSA). (Stackebrandt et al., 2002). (Fig 2.)

Figure 1. DDH

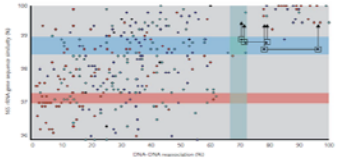

Figure 2. 16S rRNA % similarity vs. DDH (Stackebrandt & Ebers, 2005).

- NextGen sequencing has led to an increase in the availability of whole bacterial genome sequences. The genome-based metrics 95% Average Amino Acid Identity (AAI), and 95% Average Nucleotide Identity (ANI). have been suggested as replacement metrics. (Konstantinidis & Tiedje, 2005; Goris et al., 2007). AAI requires Perl or Python programming.

Figure 3. ANI calculator JSpecies uses BLAST. (Richter & Rosello-Mora 2009).

Figure 4. A web based Genome-Genome Distance Calculator (GGDC) estimates DDH values from genome sequences. (Meier-Kolthoff et al., 2013)

- GGDC/DDH is only valid at the species level.
- Due to low protein-coding gene (DNA) sequence conservation, ANI is only valid to the genus level.
- AAI is valid for more distantly-related phylogenetic groups due to greater conservation of amino acid sequences.
- Here we describe a **simple web-based tool to calculate AAI** from the output of the sequence-based comparison tool (Overbeek et al, 2005) on the Rapid Annotation with Subsystem Technology (RAST) server (http://rast.nmpdr.org/) (Aziz et al., 2008) .
- The RAST sequence-based comparison tool bidirectional best hits, and the "two way BLAST conserved genes" described by Konstantinidis and Tiedje both use the BLASTP algorithm to identify orthologs.

Methods

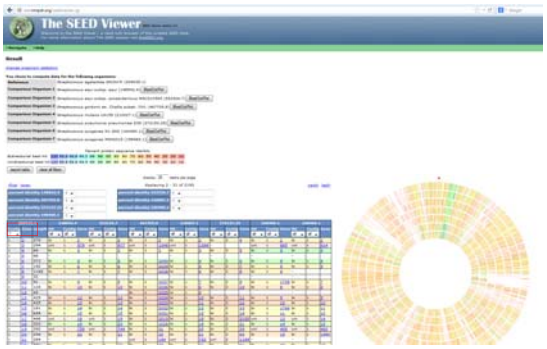

Figure 5. Result of sequence based comparison with *Streptococcus agalactiae*2603V/R as the reference organism.

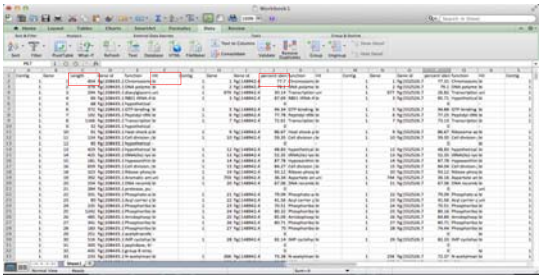

Figure 6. Results of the export table. Important columns are indicated.

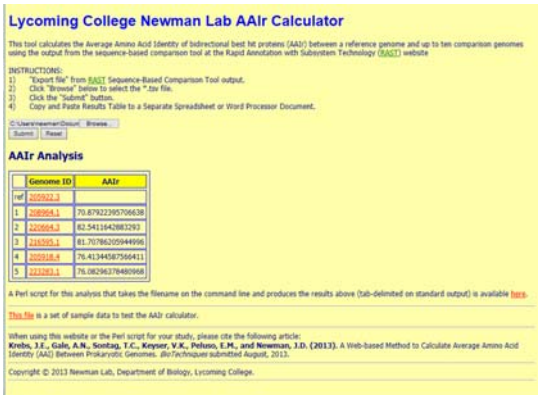

Figure 7. Calculator with NCBI Taxonomy database hyperlink as well as the respective AAIr values.

$$AAIr = \frac{\sum(percent\ identity_{bbh} * length_{bbh})}{\sum length_{bbh}}$$

Figure 8. AAIr calculator formula.

Results

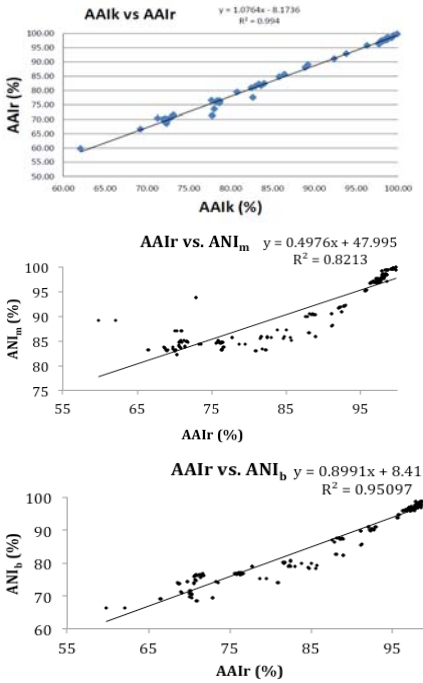

Figure 9. Comparison of AAIr to other genomic based metrics. **(a).** AAIr vs. AAIk. AAIk values were provided by Kostas Konstantinidis. **(b).** AAIr vs. ANIm. **(c)** AAIr vs. ANIb. Analyzed genomes, ANIm and ANIb values were chosen from Richter and Rossello-Mora, 2009.

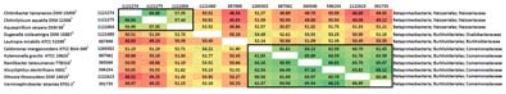

Figure10. AAIr effectively clusters organisms at the family level. Organisms within the same family generally have AAI values >55%. those in different families have AAIr values <55%

Conclusions

- AAIr values are **nearly identical** to AAIk values.
- The AAI metric **correlates well with ANI** at high values (>75)
- AAI is an effective genome-based classification tool at the **family** level.
- Because the **AAIr calculator relies on all web-based tools**, this will allow microbiologists with limited bioinformatics experience to utilize genomic based methods to differentiate bacterial species and can help facilitate the widespread use of this metric

Future Directions

- Incorporate AAI calculator into RAST's sequence based comparison tool as an output option.
- Investigate AAI ranges among type strains at higher taxonomic groups.

References

- **Aziz, R.K., D. Bartels, A.A. Best, M. DeJongh, T. Disz, R.A. Edwards, K. Formsma, S. Gerdes, *et al.*** 2008. The RAST server: Rapid annotations using subsystems technology. BMC Genomics *9*:75.
- **Goris, J., K.T. Konstantinidis, J.A. Klappenbach, T. Coenye, P. Vandamme, and J.M. Tiedje.** 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. *57*:81–91.
- **Konstantinidis, K.T., and J.M. Tiedje.** 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. USA. *102*:2567-2572.
- **Konstantinidis, K.T. and J.M. Tiedje.** 2005. Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. *187*:6258-6264.
- **Meier-Kolthoff, J.P., A.F. Auch, H.P. Klenk, and M. Goker.** 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics *14*:60.
- **Overbeek, R., T. Begley, R.M. Butler, J.V. Choudhuri, H.Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, et al.** 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. *33*:5691-5702
- **Richter, M. and R. Rossello-Mora.** 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. USA. *106*:19126-19131.
- **Stackebrandt, E., and J. Ebers.** 2006. Taxonomic parameters revisited: tarnished gold standards. Microbiol. Today *33*:152–155.
- **Stackebrandt, E., and B.M. Goebel.** 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Evol. Microbiol. *44*:846-849.
- **Stackebrandt, E., W. Frederiksen, G.M. Garrity, P.A. Grimont, P. Kampfer, M.C. Maiden, X. Nesme, R. Rosselló-Móra, *et al.*** 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. *52*:1043-1047.