Poster #436

asm2014
114th General Meeting | American Society for Microbiology
May 17–20, 2014 | Boston, Massachusetts

Integration of Average Amino Acid Identity (AAI) and Percentage of Orthologous Genes in a Single Phylogenomic Metric, the Reciprocal Orthology Score Average (ROSA).

Andrew Gale[1], Jordan Krebs[1], Eileen Peluso[2] ; Jeff Newman[1];
Department of Biology[1] and Mathematical Sciences [2], Lycoming College, Williamsport, PA.

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 Reciprocal Orthology Score Average (ROSA) metric described here is calculated using Average Amino Acid Identity (AAI) and percent bi-directional best-hit (%BBH) genes at its core. We have developed a JavaScript-based tool (http://lycofs01.lycoming.edu/~newman/rosa/) that calculates AAI, %BBH, and ROSA using the output from the "Sequence-based comparison" tool on the Rapid Annotation with Subsystems Technology (RAST) service (rast.nmpdr.org). The ROSA metric has a range from <3 when comparing genomes from different domains to >98 when comparing closely related strains. Organisms at every level from subspecies to domain are clustered more accurately using ROSA thresholds than with any existing published metric because it takes into consideration both similarity of orthologs as well as percentage of the genome composed of orthologs.

Background/Review of Existing Phylogenomic Metrics

70% DNA-DNA Hybridization (DDH)-Official Species definition (Wayne *et al*. 1987)
- Physical integration of ortholog similarity and percentage of orthologs
- Examines entire genome
- Few labs have capability, error prone, applicable only at species level
- Requires physical reciprocal experiment for each organism, cannot database

70% estimated DDH using Genome-Genome Distance Calculator (GGDC
- Easily accessible web-based tool (Meier-Kolthoff *et al.* , 2013)
- Examines entire genome sequence, range similar to DDH, statistical support
- "Black box" calculation derived from unclear metric and complex statistics
- No widely accepted thresholds above species level

98.65% 16s rRNA similarity (Stackebrandt & Ebers, 2006; Kim *et al.*, 2014)
- Simple, very inexpensive, large databases.
- Easily accessible stand alone and web-based tools (Kim *et al.*, 2012)
- Examines only a single gene, many separate species have >99% similarity
- No widely accepted thresholds above species level

95-96% Average Nucleotide Identity (ANI) (Goris et al. , 2007)
- Accessible tools (Jspecies - Richter & Rossello´-Mora, 2009;
  Kostas Lab http://enve-omics.ce.gatech.edu/ani/;
  ChunLab http://www.ezbiocloud.net/ezgenome)
- Metric is moderately clear and intuitively understood at a basic level
- Value reflects average similarity of only highly similar (>60%) sequences
- Does not consider distant or non-orthologous sequences (min ~60%)
- No widely accepted thresholds above species level

95% Average Amino Acid Identity (AAI) (Konstantinidis & Tiedje, 2005)
- Higher amino acid sequence conservation detects more orthologs than ANI
- Metric is clear and intuitively understood, reflects unit of selection/conservation (proteins rather than randomly broken DNA seq)
- Applicable above species level, but no widely accepted thresholds (yet)
- Does not consider non-orthologous sequences
- No easily accessible tools

Goal: Create a genome-based metric to differentiate bacterial species.
- Incorporate orthology between distantly-related organisms (AAI)
- Incorporate core genome percentage due to HGT/reductive evolution (%BBH)
- Comparable to DDH using full range from 0 – 100
- Accessible via web using meaningful components, intuitive to understand

RAST Rapid Annotation using Subsystem Technology | The SEED Viewer

Rapid Annotation with Subsystem Technology (RAST)
http://rast.nmpdr.org/ (Overbeek *et al.*, 2014)
- Users upload multi-sequence fasta files
- RAST calculates GCmol%, genome size, identifies CDSs & RNAs, Integrates into subsystems.
- Sequence-Based Comparison tool identifies unidirectional and bidirectional best hits between a "Reference Genome" and up to ten "Comparison Genomes"
- Exportable color-coded table and graphic shows best matches.
- Screen "mouseover" or columns in exported table provide protein name and AA sequence identity of each.

Figure 1. Screenshot of Sequenced-Based Comparison tool results.

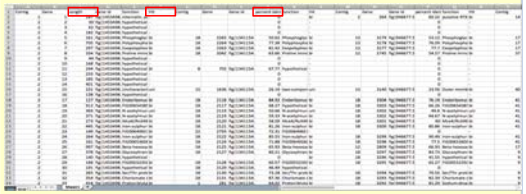

Figure 2. Exported table with Sequenced-Based Comparison tool results. Important columns are indicated with a red box.

Separate comparisons must be conducted using **each organism as a reference** and the others (up to 10) as comparison organisms.
- Can be used to construct Venn Diagrams

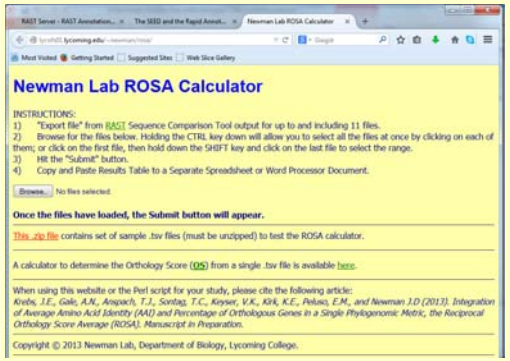

Figure 3. Screenshot of ROSA Calculator.

Logic of Phylogenomic Metric Design

- AAI → Due to variation in percent identity of CDS, (see fig. 2) length included as a factor.

$$AAI = \frac{\Sigma(\%identity * length)_{bbh}}{\Sigma length_{bbh}}$$

- Because Horizontal Gene Transfer (HGT) and reductive evolution would decrease DDH values between two organisms, we calculated this as %BBH.

$$\%BBH = \frac{\Sigma length_{bbh\,(comparison)}}{\Sigma length_{all\,(reference)}}$$

- Because 95% AAI corresponds to 70% DDH (the species barrier), we concluded that decreasing AAI would decrease DDH faster than a decrease in %BBH

Orthology Score (OS) = AAI² * %BBH

- Differences in reference genome size create different %BBH values in reciprocal comparisons

Reciprocal Orthology Score Average (ROSA)

$$ROSA = (OS_{AB} + OS_{BA})/2$$

Example ROSA Clusters

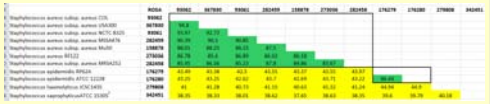

Table 1. Same species >65; Same genus, different species 35-65

Table 2. Same family, different genus 15-35; different family <15

Table 3. Same domain, different phylum 3-6; different domain <3

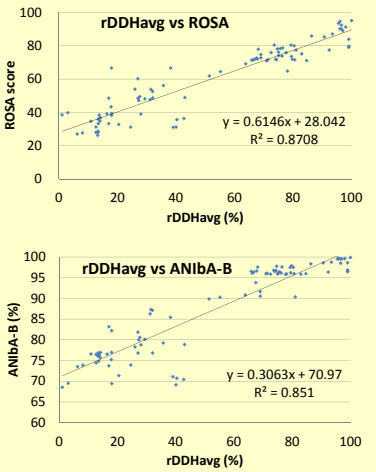

Figure 4. Relationship between reciprocal DDH and ROSA or ANI

Level	Same	Different	expected range	min	max	mean	n=	below range	above range
8	species	strain	>65	49	99.59	85.98	312	3	N/A
7	genus	species	35-65	6.85	95.8	36.63	521	294	38
6	family	genus	15-35	5.8	54.99	18.86	524	125	19
5	order	family	10-15	4.75	23.19	11.58	314	79	22
4	class	order	8-10	4.15	14.99	8.22	286	121	58
3	phylum	class	6-8	4.3	11.4	6.62	123	42	14
2	domain	phylum	3-6	1.85	7.66	4.45	211	10	6
1		domain	<3	1.36	4.44	2.42	47	0	9

Table 4. ROSA Values at Different Phylogenetic levels.

Conclusions:

- ROSA metric correlates well with DDH because it incorporates level of similarity and percentage of orthologs in a genome.
- If Microbial Systematics is to become based on phylogenomics, much taxonomic revision will be necessary.
- Comparison of shared and unique orthologs can allow prediction of differentiating traits between groups

References
- Auch, AF., Klenk HP, and Göker M (2010). Standard Operating Procedure for Calculating Genome-to-genome Distances Based on High-scoring Segment Pairs. *StGS* 2.1 (2010): 142–48.
- 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. *IJSEM* 57:81–91.
- Kim, Mincheol, Hyun-Seok Oh, Sang-Cheol Park, and Jongsik Chun. Towards a Taxonomic Coherence between Average Nucleotide Identity and 16S RNA Gene Sequence Similarity for Species Demarcation of Prokaryotes. *IJSEM* 64 (2014): 346-51.
- 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, JP., Auch AF, Klenk H-P, and Göker M. "Genome Sequence-based Species Delimitation with Confidence Intervals and Improved Distance Functions." *BMC Bioinformatics* 14.1 (2013): 60.
- Overbeek *et al.*, (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucl Acids Res.* 42:D206-D214.
- Richter, M. and R. Rosselló-Móra. 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.
- Wayne, L. G., D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, W. E. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, and H. G. Truper. "Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics." *IJSB* 37.4 (1987): 463-64.