Updating the 97% identity threshold for 16S ribosomal RNA OTUs

Robert C. Edgar

Supplementary Material

OTU quality metrics

Richness Ratio

Let S be the number of species and N be the number of OTUs. I defined the richness ratio (RR) as $RR = \min(S, N) / \max(S, N)$. RR thus has a maximum of 1 indicating that the number of OTUs is equal to the number of species while values <1 indicate that there are more or fewer OTUs than species.

Bijection

Define an OTU to be *split* if any species in that OTU also appears in another OTU. An OTU is *lumped* if it contains more than one species. If an OTU is neither lumped nor split, it is *bijective*, i.e. is in 1:1 correspondence with a species. Let K be the number of bijective OTUs, then the bijection metric (Bij) is defined as $Bij = K/N$. Bij ranges from a minimum of 0 when all OTUs are split and/or lumped to a maximum of 1 when all OTUs are bijective.

Normalized mutual information

Let X be a discrete random variable with one value per species and Y be a discrete random variable with one value per OTU. Let N be the total number of sequences, n_x be the number of sequences assigned to species x, k_y be the number of sequences assigned to OTU y and j_{xy} be the number of sequences for species x which are assigned to OTU y. Probabilities are calculated as observed frequencies: $p(x) = n_x/N$, $p(y) = k_y/N$ and $p(x, y) = j_{xy}/N$. The entropy H of random variable X is:

$$\mathrm{H}(X) = -\sum_{x \in X} p(x) \log p(x).$$

The mutual information of two random variables *X* and *Y* is:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right).$$

The normalized mutual information (*NMI*) of species and OTU assignments is then (Cover and Thomas, 1991):

$$NMI(X;Y) = \frac{I(X;Y)}{\sqrt{\mathrm{H}(X)\mathrm{H}(Y)}}\ .$$

The value of *NMI* ranges from 0 when *X* and *Y* are independent variables to 1 when knowledge of the value of *X* gives certain knowledge of the value of *Y* and vice versa, which can occur only if species and OTUs are in perfect 1:1 correspondence.

Matthews' correlation with species

Matthews' Correlation Coefficient (*MCC*) (Matthews, 1975; Baldi *et al.*, 2000) is a metric used to assess the accuracy of predictions by a binary classifier on a dataset annotated with known classifications. Let *TP* be the number of true positives, *FP* false positives, *TN* true negatives and *FN* false negatives. MCC is then calculated as

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}\ .$$

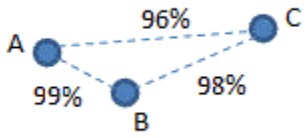
The value of MCC ranges from a maximum of 1, indicating perfect correlation (all pair-wise classifications are true), to a minimum of –1 indicating perfect anti-correlation (all pair-wise classifications are false). I defined the correlation of OTUs with species (MCC_{sp}) by considering an OTU assignment algorithm to be a pair-wise classifier which predicts whether a pair of sequences belongs to the same species. Pair-wise assignments are defined to be true or false as follows.

True positive: pair of sequences from the same species in the same OTU.
True negative: pair of sequences from different species in different OTUs.
False positive: pair of sequences from different species in the same OTU.
False negative: pair of sequences from the same species in different OTUs.

This differs from the definition proposed by (Schloss and Westcott, 2011) (MCC_{SW}) where the truth standard is based on pair-wise sequence identity as measured by mothur (e.g., a true positive is a pair of sequences with ≥97% identity assigned to the same OTU), while MCC_{sp} uses species assignments by taxonomists as the standard of truth. A drawback of both MCC_{SW} and MCC_{sp} is that with typical data, a large majority of pairs belong to different OTUs and the number of true negatives is then much larger than the number of true positives. For example, if 10 000 sequences are evenly distributed into 1 000 OTUs, then each OTU will contain ten sequences. Assuming perfect clustering, the number of true positives is (number of OTUs) × (number of pairs of sequences per OTU) = 1 000 × (10 × 9) / 2 = 45 000 while the number of true negatives is (number of pairs of sequences) – (number of true positives) = (10 000 × 9999)/2 – 45 000 = 49 950 000. Thus, in this example, TN is three orders of magnitude larger than TP, illustrating that the MCC metrics are strongly biased towards maximizing true negatives over true positives. Also, MCC metrics are not always mathematically well-defined as shown below.

Example where* $\mathbf{MCC_{SW}}$ *fails

Suppose there are three sequences A, B and C with identities AB=99%, BC=98% and AC=96% as shown below.



This is an adverse triplet as defined in the main text. There are five possible sets of OTUs of three sequences. MCC_{SW} values for the triple {ABC} are given in the table below.

Set	Clusters	TP	TN	FP	FN	MCC_{SW}
#1	{ABC}	2	0	1	0	*undefined*
#2	{A}, {B}, {C}	0	1	0	2	*undefined*
#3	{AB}, {C}	1	1	0	1	0.5
#4	{A}, {BC}	1	1	0	1	0.5
#5	{AC}, {B}	0	0	1	1	–1

With sets #1 and #2, MCC_{SW} is mathematically undefined because the denominator is zero. Sets #3 and #4 have the same score, but #3 is better because the identity of AB is higher than BC. This illustrates that MCC_{SW} considers all identities ≥97% to be equivalent, when in fact a pair of sequences is more likely to belong to the same biologically-defined group (e.g., genome, strain or species) if it has higher identity.

This example shows that MCC_{SW} will fail to identify the best OTUs for all adverse triplets present in the data. Such triplets are common in the HiQ databases and the soil100 dataset (see main text) and are therefore probably common in practice.

Supplementary References

Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**: 412–424.

Cover T, Thomas J. (1991). Elements of Information Theory. Wiley.

Matthews BW. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA - Protein Struct* **405**: 442–451.

Schloss PD, Westcott SL. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl Environ Microbiol* **77**: 3219–26.

Supplementary Table

Identity	FL	FL_1	V4	V4_1
100% ≥ *d* > 99.5	0.7617	0.5297	0.4994	0.2053
99.5% ≥ *d* > 99	0.7399	0.1957	0.0758	0.0257
90% ≥ *d* > 98.5	0.6617	0.1005	0.046	0.009
98.5% ≥ *d* > 98	0.1025	0.0269	0.022	0.0024
98% ≥ *d* > 97.5	0.0053	0.0098	0.0034	0.0008
97.5% ≥ *d* > 97	0.0014	0.0047	0.0059	0.0007
97% ≥ *d* > 96.5	0.0006	0.0013	0.0004	0.0005
96.5% ≥ *d* > 96.0	0.0003	0.0009	0.0001	0.0001

Table S1. Conspecific probabilities $P_{sc}(d)$ for the HiQ databases. FL is HiQ16S, FL_1 is HiQ16S_1, V4 is HiQV4, V4_1 is HiQV4_1. Identities are binned into intervals of 0.5%.

Supplementary Figure

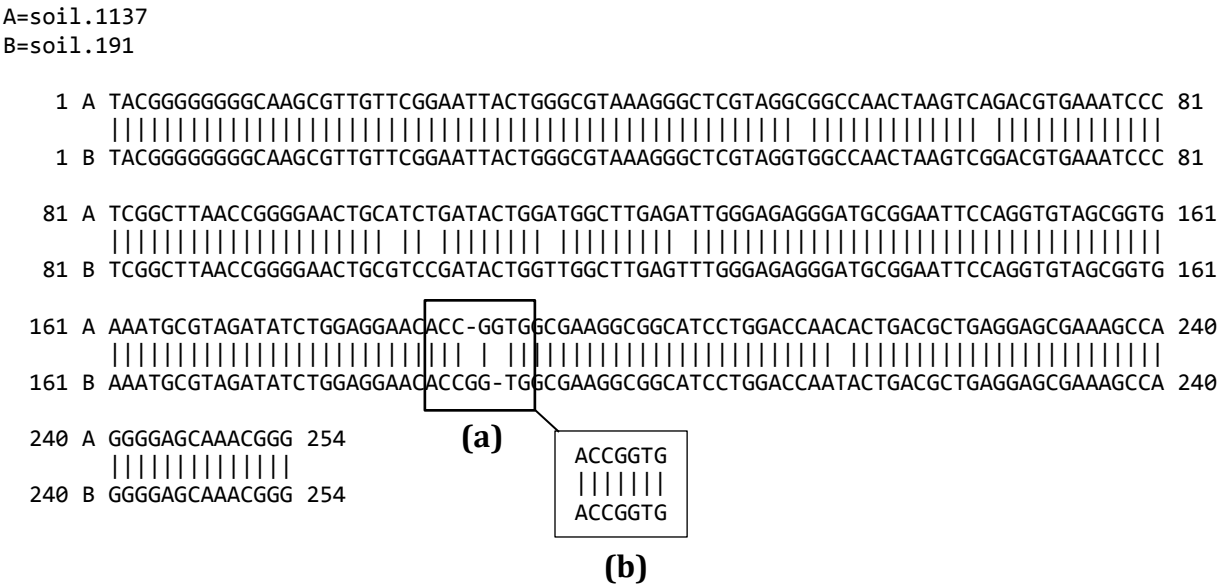


Fig. S1. Typical misalignment by mothur. The misalignment is highlighted (a) with the correct alignment given in the call-out (b). In this example, the identity of the mothur alignment is 245/254 = 96.5% while the identity of the correct alignment is 247/252 = 98%. This type of misalignment cannot occur with pair-wise dynamic programming algorithms, which are therefore superior for calculating identity of closely related sequences.