UCHIME in practice

Single-region sequencing

UCHIME is designed for experiments that perform community sequencing of a single region such as the 16S rRNA gene or fungal ITS region. While UCHIME may prove useful in other contexts, at the present time UCHIME has been validated only on 16S rRNA. Changes to the algorithm or parameters may give better results on other regions.

Reference database mode

The reference database mode of UCHIME assumes that the database contains high-quality sequences close to the true biological sequences in the sample. The most common problems with the reference database approach are: (i) the lack of a suitable reference database, (ii) inadequate phylogenetic coverage of the community being studied in available databases, and (iii) poor-quality sequences in available databases.

In practice, reference databases will usually be incomplete, and false negatives should be expected due to missing parents. Unknown species will of course be absent. Even if a given species has a high-quality reference sequence, it may have additional copies of the sequenced gene due to duplications (paralogs, pseudo-genes or segmental duplications) that are absent from the database. Phylogenetic coverage should therefore be understood in terms of all sequences in the community that are homologous to the gene and match the chosen primers, rather in terms of species.

Both false positives and false negatives can be caused by bad sequences. A false negative will occur if the query sequence is a chimera and the database contains a sufficiently similar chimera. Noisy sequences can cause both false negatives and false positives. Noise can reduce the score of a chimeric model below the *h* threshold (note that a 'no' vote is weighted much more highly than a 'yes' vote with default parameters, and noise may increase the number of 'no' votes as well as reduce the number of 'yes' votes). To see how noisy sequences can produce false positives, let X be a correct biological sequences, X_L be a prefix of X, X_R be a suffix of X and X' be a "noisy" copy of X, i.e. a copy of X with spurious substitutions and/or indels. Suppose there are two noisy copies of X^1 and X^2 in the database with asymmetric noise, such that X^1 has more noise on the left and X^2 has more noise on the right, i.e. $X^1 = X'_L X_R$, $X^2 = X_L X'_R$. Then a good copy of X may appear to be a chimera $X = X^2_R X^1_L$ formed from parents X^1 and X^2. If X' and a chimera $C = X_L Y'_R$ are present in the reference database, but not Y, this can cause a false positive identification of Y, which may appear to be a chimera formed as $Y = X'_L C_R$.

Correct sequences in the reference database may give rise to false positives if evolutionary rates in different regions of the gene vary in different lineages. Suppose the gene contains two regions r_1 and r_2, and there are three lineages A, B and C where r_1 evolves faster in A than in B or C, and r_2 evolves faster in B than in A or C. Now suppose the database contains A and B but not C, then C may appear to be a chimera formed from A and B.

These considerations present conflicting goals in the design of a reference database: high phylogenetic coverage and high-quality sequences. Increased phylogenetic coverage generally requires incorporating sequences from unfinished genomes and/or from environmental sequencing studies, both of which tend to have higher error rates than finished genomes. This can be mitigated by using the reference database mode of UCHIME to check a candidate reference database against itself using the *--self* option, which excludes the query sequence as a possible parent (otherwise all sequences would trivially be annotated as non-chimeric due to self-matches). Hits reported using *--self* are 3-way alignments in which either one or two of the sequences are putative chimeras. It should not be assumed that the query sequence is the chimera in this case. Further evidence is required to determine which, if any, of the sequences in the 3-way alignment are PCR artifacts. For example, if two of the sequences are derived from high-quality, finished genomes and the third is from an environmental sequencing study, then the third is most likely to be an artifact and should be discarded from the database. Any remaining sequences found in 3-way alignments can be annotated as unresolved. Hits to experimental data that have an unresolved parent can be treated differently. Whether they should be included or discarded depends on the goals of the study, which will determine whether sensitivity or specificity of chimera detection is more important. Discarding questionable hits will tend to improve specificity at the expense of sensitivity; including them will tend to improve sensitivity at the expense of specificity.

It is often the case that a reference database contains full-length sequences while a shorter region is sequenced. Here it may be advantageous to trim the database to the shorter region. This can improve computational efficiency because the time required to make a dynamic programming alignment scales with the square of the sequence length (Durbin et al, 1998). This may also reduce the number of false negatives due to failures to identify the correct parent which may be caused by the *k*-mer count heuristic filter (Edgar, 2010) that is used to improve search speed.

De novo *mode*

The *de novo* mode of UCHIME assumes (i) sequences correspond to unique sequences in the amplified sample, (ii) the abundances of those sequences have been estimated with sufficient accuracy, (iii) errors due to amplification and sequencing can be neglected, i.e. are adequately suppressed preprocessing of the sequences and/or by the UCHIME scoring function, and (iv) chimeras have abundance less than their parents, as specified by the abundance skew parameter. At the present time, it is not known how well these assumptions hold in practice, except for the mock communities described in the main text. It is an open research problem to determine how predictive these mock communities are of experiments on natural communities.

An advantage of the *de novo* approach is that we expect most or all parent sequences to be present in the reads, which may enable higher sensitivity to be achieved compared with a pre-existing reference database, which will generally be incomplete. A disadvantage of *de novo* mode is that "raw" reads are required, which may not be readily available, and robust estimates of false positive and false negative rates are not yet available.

How best to estimate amplicon sequences and their abundances depends on the sequencing technology. In the case of 454 flowgrams, PyroNoise (Quince et al., 2009) or AmpliconNoise (Quince et al., 2011) could be used. Clustering using a method such as UCLUST (Edgar, 2010) can be used for any technology. This clustering should be done using dereplicated reads (i.e., reads that are exact substrings of other reads have been discarded). Dereplication can also be done using UCLUST. It is important to sum the number of reads in each dereplicated set when calculating the cluster size. It is also important to include only reads from a single experiment (strictly, a single amplification stage), otherwise abundances will not be directly comparable. The size of a cluster can be used as an estimate of the relative abundance of the corresponding amplicon, and the representative sequence of the cluster can be used as an estimate of the amplicon sequence. This approach requires a minimum identity that determines which reads are assigned to a given cluster. The appropriate threshold depends on the sequencer error rate and the goals of the analysis. For example, if the sequencer error rate is believed to be ~1% and the subsequence step will be identification of OTUs at 97% identity, then a reasonable threshold for estimating amplicons could be 98% or 99% identity. In the case of UCLUST, sequences should be sorted in order of decreasing length for replication, and then sorted in order of decreasing dereplicated set size prior to the second clustering step used to determine amplicons and abundances. This is because longer reads tend to have more errors, especially towards the end of the read, so longer sequences are less suitable as cluster representatives for amplicon estimation.

Consistency check

Where possible, we recommend that the reference database mode and *de novo* modes be used to check each other. Hits found by both methods are more reliable than hits reported only by one method. A hit found by the reference database mode but not by *de novo* mode can be investigated by searching the estimated amplicons for the putative parent sequences. If these are present in the reads, then this is probably a false negative by the *de novo* mode, which could be due to poor estimates of amplicon sequences or abundances, a preceding false positive that incorrectly identified a parent as a chimera, or a violation of the assumption that the parents have higher abundance. If the parents are not found in the reads, then this could be a false positive by the reference database mode (see previous discussion of causes of false positives in this mode). A hit found by *de novo* mode but not by reference database mode may be explained by a missing parent sequence in the reference database, which can be verified by searching the reference database for the parents predicted by *de novo* mode.

Computational efficiency

Community sequencing experiments often produce very large numbers of reads that can be computationally expensive to process. It is generally recommended that the number of sequences be reduced before running UCHIME in order to save computational resources. Preprocessing steps can include dereplication (removing identical sequences), denoising (attempting to correct sequencing error) and data reduction (clustering at, say, 98% identity to reduce experimentally irrelevant variation in the sequences). Computational cost can also be substantially reduced by using the USEARCH (Edgar, 2010) implementation of the UCHIME algorithm. We strongly recommend using version 4.1.93 or later; earlier versions of UCHIME used a different scoring function (not published), and have significantly worse performance on the 16S benchmark tests used in the present work.

Parameter tuning

The default parameters of UCHIME were tuned to give lower error rates and higher sensitivity than ChimeraSlayer on the SIM2 benchmark. This strategy was chosen in order to demonstrate that UCHIME has better performance than ChimeraSlayer on a published benchmark (Haas et al., 2011) on which ChimeraSlayer was shown to be superior to previous methods and thereby establish that UCHIME is superior to all previously published methods. We believe that while these parameters probably represent reasonable default settings, different parameters may be optimal in some applications. It should be noted that the ChimeraSlayer validation emphasized sensitivity to closely related parents: the divergence measure used by Haas *et al*. is the distance between the parents A and B ($D = 100\% - id(A,B)$), while in this work we use the identity between the chimera Q and the closest parent ($T = 100\% - \max\{ id(Q,A), id(Q, B) \}$ in the case of bimeras). Generally we expect that $T \leq D/2$ since at last half of the chimera will

be identical to the closer parent. In many experiments, it is *T* rather than *D* that indicates whether the chimera is experimentally relevant. For example, if the goal is to identify OTUs by clustering at 97%, and a parent is successfully identified as the representative sequence for a cluster, then a chimera with $T \leq 3\%$ should be assigned to the parent cluster and will not create a spurious OTU. Such a chimera could have $D \geq 6\%$. By default, the minimum *T* divergence, set by the *--mindiv* option of UCHIME, is set to 0.5% to allow detection of chimeras with small *D*, which is required to achieve good performance on SIM2. Chimeras with divergence $T \gtrsim 0.5\%$ may have very small numbers of diffs and hence be difficult to discriminate from false positives, requiring a higher *h* threshold to suppress errors. These considerations suggest that in a typical OTU clustering experiment, higher sensitivity to experimentally relevant chimeras could be achieved with acceptable false positive rates by increasing *--mindiv* and reducing *h* (*--minh* option) and/or β (*--xn* option). In addition, SIM2 has no multimeras and adds simulated noise that is designed to indicate the general impact of sequencing error and natural variation on performance rather than to accurately model errors due to a given sequencing technologies or to model natural biological variation that can cause a reference sequence to differ from the true parent sequence. Ideally, parameters would be re-tuned on a benchmark that is tailored to the details of a particular experiment, including simulated errors based on estimates of error rates of the chosen sequencing technology. Designing and implementing such a benchmark would be challenging. Further work is needed to determine whether and how parameters should be varied according to the details of a particular experiment.

References

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) Biological Sequence Analysis. Cambridge University Press.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics* **26**(19), 2460-1.

Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E., Methe, B., Desantis, T.Z., Petrosino, J.F., Knight, R. and Birren, B.W. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons, *Genome Res*, **21**, 494-504.

Quince, C., Lanzen, A., Curtis,T.P., Davenport,R.J., Hall,N., Head,I.M., Read.,L.F. and Sloan,W. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data, *Nature Methods*, **6**(9), 639-641.

Quince, C., Lanzen, A., Davenport, R.J. and Turnbaugh, P.J. (2011) Removing noise from pyrosequenced amplicons, *BMC Bioinformatics*, **12**, 38.