Reviewer's report

Title: Automated analysis of ARISA data using ADAPT system

Version: 1 **Date:** 16 September 2009

Reviewer number: 1

Reviewer's report:

0. (overall) This manuscript describes software and a database for the interpretation of ARISA data. The basic idea is a very good one and the authors are technically adept. It is great that the program automatically looks for matches to the database and handles metadata well. However I have several serious concerns, mostly regarding limitations of the database and limitations of data interpretation. A terrible problem is that this implementation is almost certain to report on organisms not really present, due to accidental (or incidental) matches. As currently implemented, it is very likely to report on the presence of nasty pathogens even when they are very unlikely to be present (explained below). I believe that this particular implementation would really be useful only for analysis of mixtures of microorganisms from culture collections as opposed to “wild” environments. On the bright side, I think there are potential fixes (not difficult) that would make it much more useful. The main fix would be to allow restrictions of database searches to a more “relevant” subset, and to have the database go beyond cultured characterized organisms, ideally including user-provided lists of identifications relating to particular ARISA lengths.

The following call for “Major Compulsory Revisions”

1. Database misses critical organisms. Most organisms BY FAR in natural environments are not known, and only a minuscule proportion are represented by annotated gene sequences with characterized taxonomy. But this database only looks for matches with organisms that have such sequences in NCBI or SEED. Unfortunately, NCBI does not allow users to annotate the taxonomy of environmental clones, so for uncultured sequences that include 16S and ITS and 23S rRNA there is not a taxonomy in the database (even though one could get decent taxonomy from the sequences themselves). Hence only cultured organisms are included. And unfortunately you cannot expect that a “similar” ARISA length tells you that an organism is related to one you know. So with environmental samples, the overall exercise of using this software is akin to the cliché of looking for the keys you lost in a large parking lot by crawling around under a very small streetlamp. The database needs to include more realistic environmental clones. But that is not the worst problem.

2. There are too many matches to irrelevant organisms. Problem 1 above might not be so bad if users were only interested in the characterized sequenced organisms and if there were many thousands or millions of unique ARISA identifiers. But as the manuscript shows, ARISA, and the required binning, places all possible sequences into a limited number of bins, on the order of a few

hundred bins (depending on parameters) – so all the potentially millions of species of bacteria are put into a few hundred bins. Therefore if there are only 1 million species of bacteria and 300 bins, there is an average of 3000 species per bin. Therefore, searching blindly for matches in a UNIVERSAL bacterial database like NCBI or SEED will undoubtedly lead to many false matches, and the odds say this probably happens MUCH more often than not, as I suspect was probably the case with some of the pathogen matches from the island results (especially given how the database is strongly biased towards pathogens). I ran the program with some ARISA data from relatively pristine offshore marine samples, and the output list reported organisms causing anthrax, Q fever, tularemia, bubonic plague, cholera (the nasty El Tor strain), gangrene, tuberculosis and more (all sorts of well known bugs)!! I worry that if this software is applied as currently implemented by unsophisticated users (e.g. a college microbiology class), the wrong impressions may lead raids by Homeland Security and detention of foreign students! On the other hand, the program missed several reasonable matches to many of these ARISA lengths available in Genbank from uncultivated marine organisms likely to represent ordinary harmless marine bacteria, because the annotations did not meet the authors' criteria - many environmental sequences of 16S and adjacent ITS and 23S regions have been published but do not have a formal taxonomy associated with them.

It is very useful that the program can require that for organisms with multiple operons, all appropriate ARISA lengths must match . This is a great feature – however it should require they all have similar heights rather than just being present, as some ARISA outputs have many small peaks. However, that only reduces the odds of a problem, it doesn't eliminate it (especially given the very large number of organisms with single rrn operons or multiple ones with identical ARISA lengths). I think the problem could be reduced by filtering the results by habitat, location, etc., in other words just looking for known marine organisms when studying marine samples, or tissue pathogens and commensals when examining a tissue sample – which would admittedly remove the possibility of finding an organism other then where it "belongs." Unfortunately such filtering has not been implemented in this software, so there is not an option. But I think that even such filtering would leave the likelihood of many false matches. Ideally the database would allow for users to upload their own set of matching lengths and identifiers that they may generate with clone libraries from similar samples and locales, as reported by Brown et al. (2005). Users should be able to add these manually, if needed, as well as their own table of ARISA lengths with matching taxonomic identifiers. That would be a great addition that would improve the utility of the program.

The upshot is that ARISA is useful for identifying organisms only within a restricted set of circumstances, where that particular environment has been previously characterized well. Trying to find a "one database fits all" approach is just not suitable. The program just provides long lists of possible matches, the large majority of which are completely inappropriate for the sample. At some point a user must select which organisms (if any) are relevant, and the rationale

for that choice needs to be developed..

Other General Comments requiring Major Compulsory Revisions

3. I would need more details on the transformation of electropherogram data into fragment lengths and peak heights. It is not a simple problem with real, as opposed to idealized, data. How does the program handle peaks that have been split or merged, especially at the larger end, or moving baselines, choppy fluorescence signals, etc.? There are many sophisticated programs (justifiably expensive) fully dedicated to correct interpretation of peaks in chemical chromatograms, that correctly find small peaks on the tails of other larger peaks, moving baselines, etc. It is not clear how sophisticated this current one is in that regard, but it is important to do this part well. To approve this as a comprehensive software package, I would want to see how it handles typical real data, not just the best and easiest outputs (e.g. showing where it selects peaks out of noisy data).

4. The analysis uses peak height to determine relative contribution. What is the justification vs using area since longer fragments tend to be wider and shorter? Most chemical analysis uses area when there is such a range of peak shapes, and I suggest this is better.

5. Larger fragments (> 1000bp) may have apparent ARISA lengths considerably different from the actual length (by 15-20 bp), due to mobility differences compared to standards with different G+C content. This can lead to mismatches.

6. Classification of autotrophy or heterotrophy based on phylum level designations are problematic and oversimplified – likely often wrong. Wouldn't it make more sense to designate trophy based on the species/genus level? Since the database is mainly known organisms with genomes sequenced, this should be easy.

7. Figures: I think a figure of the sample output that shows the peak length with possible identities would be more useful than the current Figure 3. Otherwise the user has no introduction to the major result/purpose of the database within the paper.

8. In the input data section under data analysis, how is the text input data parsed, or is it not? This needs explanation

9. Is it reasonable to include fragments at 3000bp range? Is there an ITS that's 3000bp?

10. I would like the ability for 0.5bp bins using the x ± bin set value, to get bins of 2 instead 3 being the lower bound, and the ability for even number-sized bins.

11. I like the idea of comparative metatables, e.g. for the trophy status, etc. What about one for general phylogenetic classification at the kingdom or class level?

12. p 14, "bacterial DNA was extracted from the filtrate" filtrate of what. Granted the paper is referenced, but it might be better to include a few more details and rewrite the sentence so that it

provides enough information.

Not compulsory revisions:

13. Is it possible to allow input of metadata for part of the comparison? Or provide tables that can be used for other comparisons beyond the pie graphs?

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.