

Abstract

16S DNA sequencing uses ribosomal targets which are multi-copy genes. In about 5% of isolates submitted for identification, copies are not the same, requiring the use of software to correct for polymorphic positions caused by insertions or deletions. Manual methods analyze bacterial sequence data through a conventional assembly process and compare sample sequences against proprietary libraries, resulting in data interpretation with assigned confidence levels based on phylogenetic analysis. The fully automated commercial method handles all sequence files in the same manner regardless of the data quality and provides the top match, relying on the end user to determine the true identity of the organism. Commercial software weighs polymorphic positions (assigns Quality Values) differently than single copy sequences. Use of the manual process includes both auto-assembly and editing of base caller errors, demonstrating far better repeatability than the fully automated method. This poster compares both processes, utilizing an extensive data set that includes representatives of various sequence types. Results of using a proprietary library that contains a significantly higher number of encountered environmental species shows a 2X reduction in unidentified strains over the commercial library. Utilizing full consensus sequences, the manual method demonstrated 6X better precision, clearly making it the preferred tool for tracking and trending sequence data. Clusters of subspecies can be seen with the improved precision. In many cases, the ability to resolve sequence differences or similarities at the subspecies level can eliminate the need to use a strain typing method. Many factors that affect the accuracy and repeatability of sequence-based identifications are discussed.

Introduction

The publication of all new bacterial species required a 16S DNA sequence as of 1994. In 1998, the first commercial sequence-based identification system became available, required MAC software and was based on a semi-automated assembler and manual correction of base caller errors. In 1999, Accugenix started providing sequenced-based identifications to the pharmaceutical industry, while developing extensions to the original commercial libraries. In 2005, the second commercial version (MicroSeq® v 2.0) was released by Applied Biosystems for use on PC Windows, that utilized a completely different assembly method and algorithm to analyze sequences. In addition, bacterial and fungal libraries supported on this system were not primarily targeted to the pharmaceutical EM market. By 2007, Accugenix had created its own full coverage libraries focusing primarily on organisms relevant to environmental monitoring programs. The software used by Accugenix allows for a manual base-caller editing approach, which yields high accuracy, full length consensus sequences for report generation. Factors affecting accuracy of identifications include library coverage for intended use, the method of sequence comparisons, and interpretation guidelines. Precision is determined by how the software handles insertions and mixed bases, detection of base calling errors and length of the read sequence.

Methods

The data used in this study consisted of 442 sequences (forward and reverse), derived from unknown bacterial environmental isolates from the pharmaceutical and sterile manufacturing environments.

MicroSEQ® v 2.0 – Raw sequence data (Forward and Reverse .ab1 files) were analyzed with MicroSEQ® v 2.0 automated assembly software and searched against the newly released MicroSEQ® v 2.2 Bacterial Library. The same data set was also searched against the Accugenix Bacterial Library 11Oct10, which was uploaded according to the Applied Biosystems MicroSEQ® ID Analysis Software Version 2.0 manual. Samples that had a percent similarity to their closest match of 98% or greater were considered species level identifications. The Analysis Protocol for both analyses defined a “Minimum Clear Length” of 400 base pairs.

Accugenix Sequencing – Sequence data was assembled manually and searched against the validated Accugenix proprietary sequence library (effective 11Oct10). Species level was determined by qualified data review scientists at Accugenix.

Results and Discussion

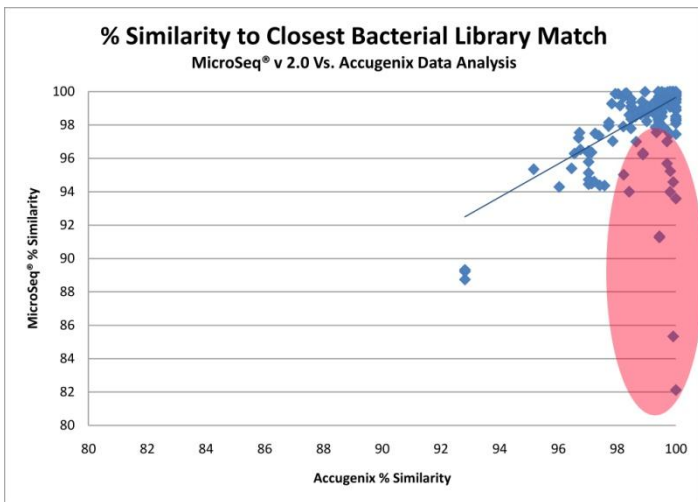


Figure 1. A comparison of % similarities to the closest match after 442 sequences were analyzed by both the Accugenix and MicroSeq® methods, utilizing current libraries, Accugenix Bacterial Library 11Oct10 and MicroSeq® v 2.2 Bacterial Library, respectively. The red ellipse contains data points for missing MicroSeq® v 2.2 library entries (approximately 19%).

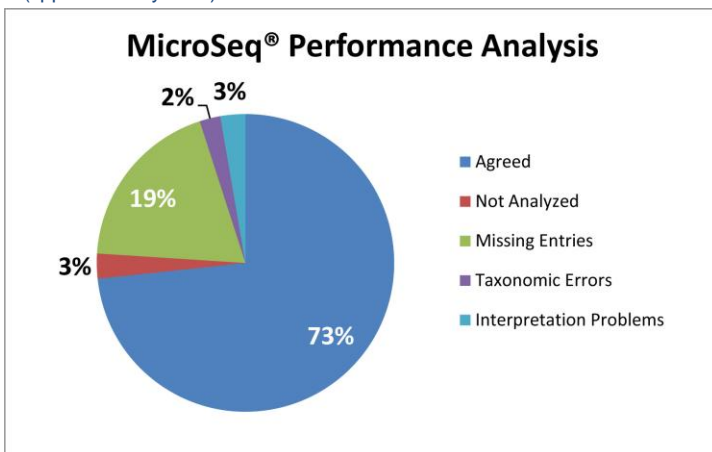


Figure 2. MicroSeq® identification results compared to Accugenix identification results.

Figure 3 shows side-by-side Neighbor Joining Trees to demonstrate the importance of accurate consensus sequences and precise distance measurements when evaluating a system for its ability to track and trend. There are 5 groups (possible subspecies) whose members contain identical sequences (no horizontal distance) in the Accugenix tree. We would expect that, with an equivalent system, the same samples in the MicroSeq® tree would cluster together. In fact, none of the sequences are identical, suggesting that the variation in length or uncorrected base caller errors of the consensus sequences generated by MicroSeq® negates the ability to properly track and trend species. The Accugenix method demonstrates at least 6X better precision.

Removing the library variable, due to the missing library entries in MicroSeq® v 2.2, allowed for a more accurate comparison of data analysis methods. Figure 4 shows the resulting data when all 442 sequences were analyzed with MicroSeq® using the Accugenix 11Oct10 Bacterial Library. Again, 12 sequences were not analyzed. The second variable when comparing the systems is the distance measurement calculation. Where Accugenix treats all bases similarly, as though part of a genome sequence, MicroSeq® relies on the “Specimen Score,” or the “average consensus quality value (QV)” to calculate the distance between the consensus sequence and its closest match.

Figure 1 shows a comparison of the percent similarity to the closest match for 430 of the unknown sequences. If both methods of analysis and interpretation were equivalent, all data points would fall on a 45° trend line (Figure 1). The red ellipse indicates MicroSeq® analyzed samples that did not agree with the Accugenix identification. Further analysis of the identifications made by both methods (Figure 2) demonstrated that there were 118 discrepancies (27%). Twelve of the 442 sequences were not analyzed with the MicroSeq®, due to short sequences caused by insertions or deletions. Of the remaining 106 analyzed discrepancies, 84 were due to missing entries in the MicroSeq® v 2.2 Bacterial Library. Ten of the organisms had a species level match to MicroSeq® library entries that require taxonomic reclassification (invalid entries). For example, the MicroSeq® entry for *Acinetobacter genomospecies 10* is not a valid species and was reclassified as *Acinetobacter bereziniae* in 2010¹. The remaining 12 had a top match between 85.33% and 97.49% similarity, but a confidence level was not made, as there are no guidelines for report interpretation in the MicroSeq® manual. The remaining 73% agreed at the species level. This clearly demonstrates at least a 2X reduction in unidentified strains when using the Accugenix library.

Tracking and trending utilizing sequences can be a powerful tool for identifying possible excursions and clusters of organisms with nearly identical 500 base pair regions. Accugenix sequence data provides a useful method for tracking and trending sequences. Of the 442 sequences, 32 were identified by MicroSeq® as *Micrococcus luteus*, perhaps the most frequently identified organism in any manufacturing facility.

¹ Alexandri Nemeč et al. (2010), *Acinetobacter bereziniae* sp. nov. and *Acinetobacter guillouiae* sp. nov., to accommodate *Acinetobacter* genomic species 10 and 11, respectively. *Int J Syst Evol Microbiol* 60, 896-903

Accugenix Method full length (all 504 bp)

MicroSEQ® Method variable length (410 – 469 bp)

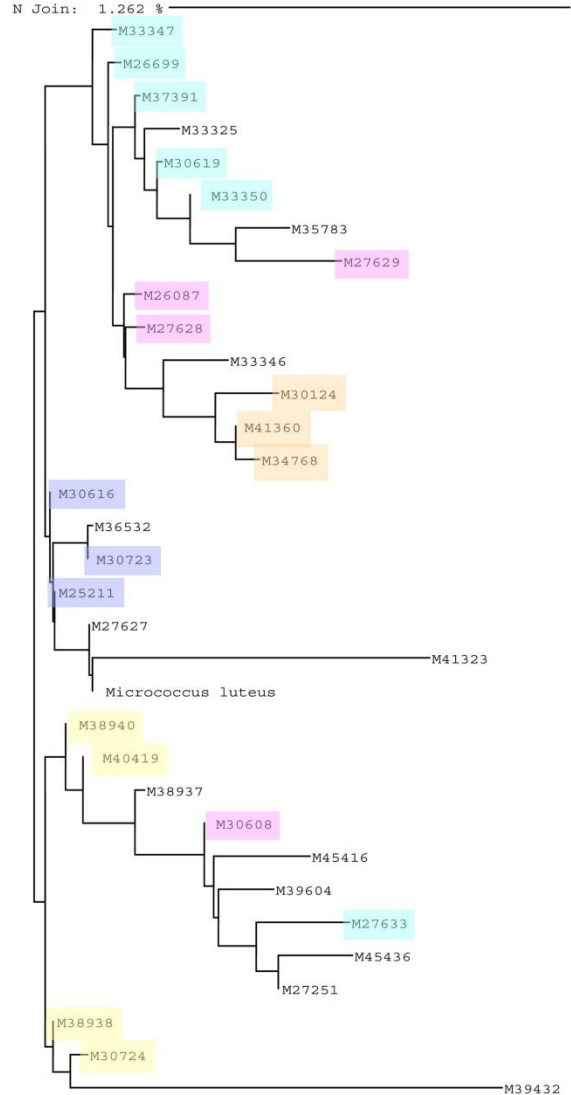
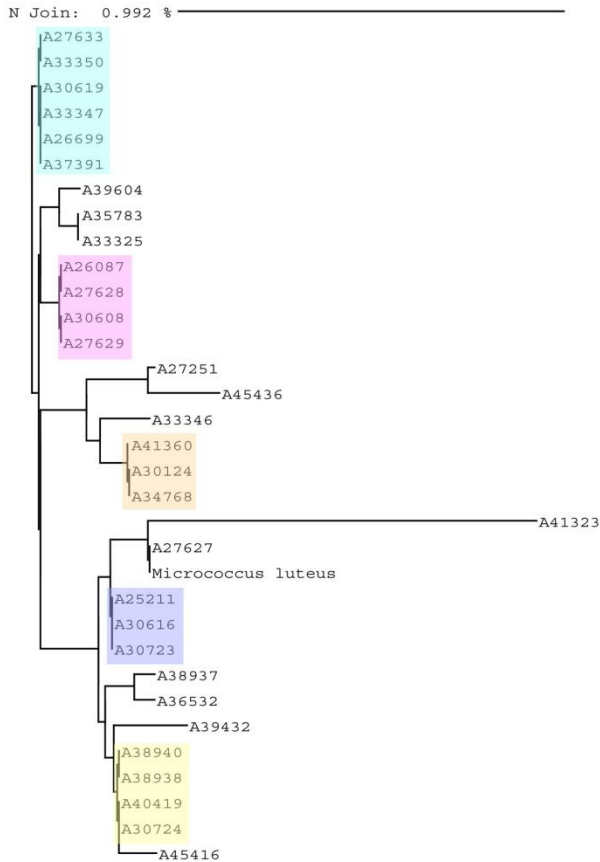


Figure 3. Samples identified by MicroSEQ® as *Micrococcus luteus*. Raw data files were processed through each method. The consensus sequences were then exported and shown using the identical Neighbor Joining Tree tool. Clearly tracking and trending is not possible with variable length sequences.

In Figure 4, the blue ellipse contains data points with scatter, due to the method of distance measurement utilized by MicroSEQ®. One example (circled data point from Figure 4) demonstrates how the distance measurement can affect result interpretation. Accugenix identified the organism “A42468” to the genus level (*Bacillus* sp.), with a similarity of 98.22% to its closest match, *Bacillus simplex*. MicroSEQ® identified the same raw sequence as *Bacillus simplex*, with a similarity of 99.67%. The end user of MicroSEQ® may mistakenly identify this organism to the wrong species level. Based on the resulting MicroSEQ® data, 29 mismatches out of 534 bases in the library entry, the % match should be approximately 95%. Figure 5 shows the neighbor joining tree and a concise alignment of both sequences. Because polymorphic positions are given a lower quality value, the % match calculated by MicroSEQ® is erroneous and misleading, resulting in incorrect distance calculations.

It would be expected that utilizing the same library for both methods of data analysis would result in similar identifications. However, 43 (10%) of the identification results did not agree. The MicroSEQ® system incorrectly identified 6% (28) sequences to the species level. The remaining 15 (4%) results provided inconclusive data for interpretation. The Applied Biosystems MicroSEQ® ID Analysis Software Version 2.0 Guide does not provide interpretation guidelines for results analysis.

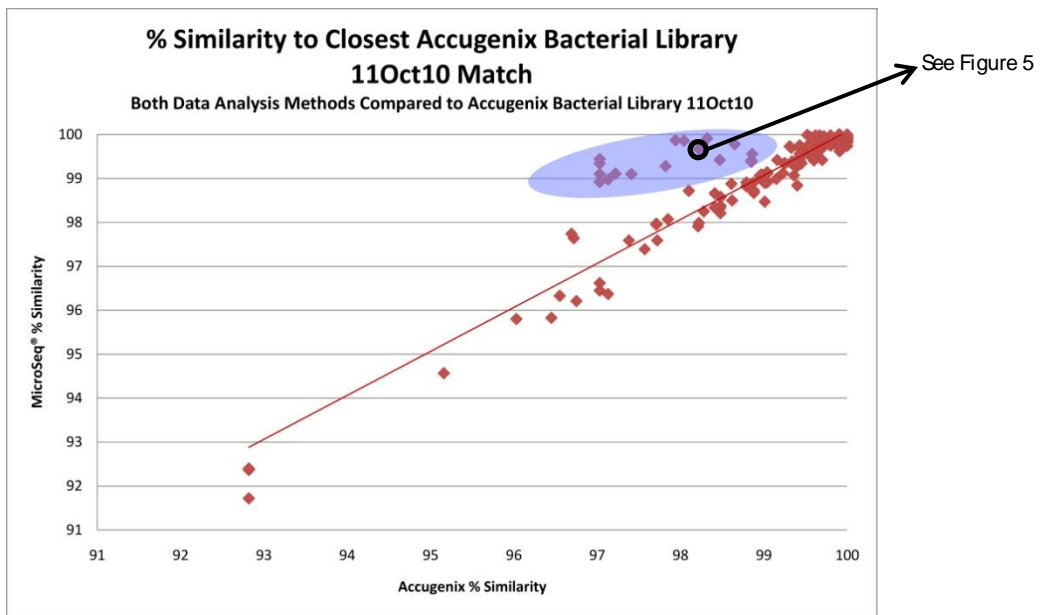


Figure 4. A comparison of % similarities to the closest match after 442 sequences were analyzed by both the Accugenix and MicroSEQ® methods, utilizing the Accugenix Bacterial Library 11Oct10. The blue ellipse contains data points indicating distance measurement discrepancies, which may lead to errors in data interpretation and tracking and trending.

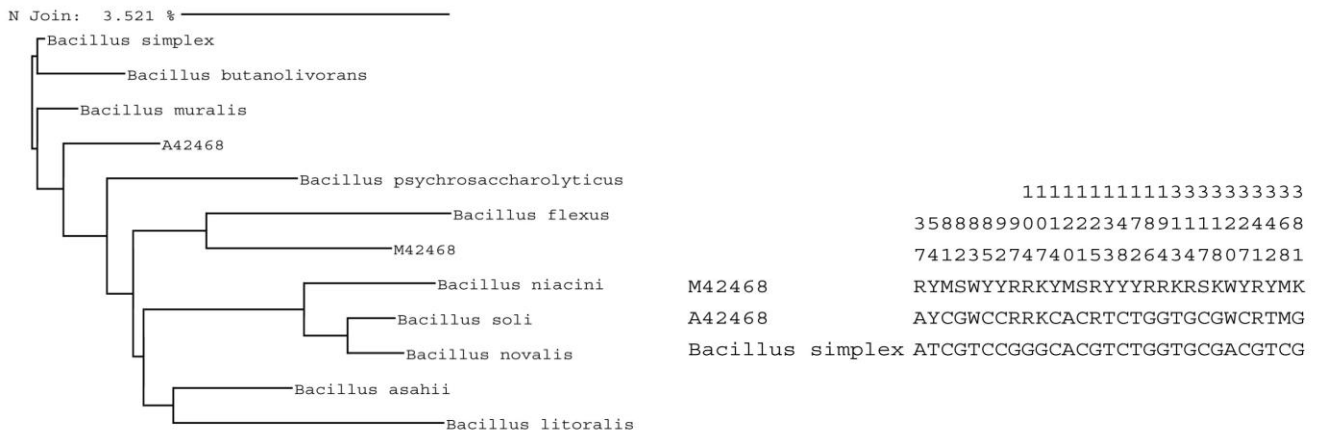


Figure 5. The Neighbor Joining Tree and concise alignment of consensus sequences from identical raw data files after data analysis with both methods. "A42468" and "M42468" are Accugenix analyzed and MicroSEQ® analyzed sequences, respectively. Polymorphic positions (bases other than A, C, T or G) were incorrectly assigned and calculated for M42468, resulting in the wrong identification.

Conclusion

Accugenix provides a superior method of data analysis, incorporating relevant, up-to-date libraries over the MicroSEQ® ID Analysis Software Version 2.0 (with Bacterial Library v 2.2). The inherent problems with the MicroSEQ® library and the method used to determine distance calculations are seen in its performance. At best MicroSEQ® can identify 74% of organisms to the species level, due to invalid and missing library entries. MicroSEQ® also lacks the ability to provide sequences of a quality high enough to accurately trend and track relevant organisms, due to truncated consensus sequences and distance measurement calculations.

Utilizing a method that includes manually assisted automated assembly and continuously updated, proprietary libraries will allow accurate trending and tracking of strains, as well as a high confidence in result interpretation.

About Us

Accugenix, Inc. provides leading-edge technology in microbial identification and characterization services. Our FDA-registered lab is cGMP compliant and maintains rigorous standards competitive at the global level. We specialize in testing, analyzing and interpreting data from environmental isolates commonly found in pharmaceutical, biotechnology, medical device, nutraceutical, personal care and other manufacturing industries.

For more than 20 years, Accugenix has provided the fastest, most accurate and reliable microbial identification services to over 400 facilities around the world. Accugenix updates its validated, proprietary DNA sequence libraries annually to reflect current taxonomy and newly described relevant species. We have the industry's first Fungal Library based on the ITS region. Since inception, we have tested more than 400,000 microorganisms – more than any other service laboratory in the industry, while maintaining an on-time delivery of over 99%.

History of Accugenix

1990.

Accugenix, Inc. began as Acculab, Inc., a reference laboratory specializing in microbial identification for industry and research clients. At the time we were one of only a few service laboratories in the world offering cellular fatty acid analysis, beginning a tradition of bringing cutting-edge microbiology methods to full commercial potential and utilization.

1999.

To reflect the addition of comparative DNA sequencing to our menu of validated methods, we created Accugenix, A Division of Acculab, Inc. Since then we have sequenced hundreds of thousands of environmental isolates from over 1000 pharmaceutical and biotechnology production facilities around the world, allowing us to build the largest and most unique industry database for bacteria and fungi that often occur in cleanroom manufacturing environments.

2005.

Our official name changed to Accugenix, Inc. on February 25, 2005.

2008.

Accugenix GmbH, our European subsidiary, was launched in Spring 2008.

Today.

Dedicated to being the industry leader for providing the most progressive microbiology methods available, Accugenix has invested in the technology, instrumentation and expertise to conquer genetic-based testing methods, their process validation, cGMP compliance, and other rigorous regulatory standards at the global level. Accugenix continues to staff its ranks with scientists and experts to guide and/or fast-forward your transition to genotypic microbial identification.



Visit www.accugenix.com or call +1 302.292.8888/ +49 (0)621 3709 556.