

## Evaluating Library Databases for Microbial Identification: Critical Aspects and Recommendations

### *Introduction*

Accuracy, reproducibility and cost are all directly impacted by the method of data analysis and the library database used for determining a reliable identity. The attributes and limitations of different methodologies in analyzing results to determine an identification are important considerations. Equally important in a highly regulated industry, the cGMP compliance and validation strategies for library database entries and maintenance should be explored to ensure appropriate requirements are met. The focus of many system providers is *in vitro* diagnostics, therefore, recommendations for “Fitness For Use” in aseptic and sterile environments is a very critical component in any technology evaluation.

### *The Evaluation Process*

The evaluation process for methods and services to identify microbes typically includes a technical and financial review of available commercial systems for internal testing and outsourcing solutions. Before choosing a system, one starts by creating a detailed list of specifications required to meet the Microbiology Laboratory’s objective: Providing accurate, reliable, reproducible, timely and cost effective identification results. A thorough historical review to determine the gaps in your current process, the technology advances available and understanding what can be practically implemented for routine use that demonstrates a positive return on investment are all components of this process.

### *Tangible Specifications*

A significant amount of time in this evaluation process will include applicable regulatory guidance requirements that must be fulfilled, matching your specification requirements against the features and benefits of the systems and services you are investigating. Most often this means comparing tangible features such as: sample throughput, time to first result, the cost of reagents and consumables, the labor requirements, the skill set required for reliable operation and results interpretation, the extent of automated processes, compatibility with the existing workflow, interfacing with laboratory information systems, capital expenses and facility requirements.

### *Importance of Libraries and Analytical Methods*

Tracking microorganisms found in the manufacturing areas of pharmaceutical and biotech facilities is an important element of any company’s environmental monitoring program. This information is used to create trending reports on a regular basis, which provide information regarding the state of environmental control within the manufacturing area and product safety. An increase in the number of microorganisms recovered in certain areas of the facility may indicate potential breaches in the HVAC system, water, cleaning procedures or other sources of microbial contamination. Historically, most manufacturing facilities have used information regarding the identity of the microorganisms to help aid in the tracking process. Ideally, a genus and species name would be used, but in some cases where a species name is not possible, either a genus name is used, or in others, a Gram stain result. The identity of organisms recovered in an investigation can assist in finding the source of the contamination. One of the primary areas of an evaluation that often is not given to a thorough

### *Keywords*

Microbial Identification, Evaluation of Microbial Libraries, Microbial Databases, MALDI-TOF, Genotypic Identification, Phenotypic Identification

investigation is the microorganism identification database; the analytical methods employed and database content to determine an identity. The critical feature of the system you will routinely utilize is that it performs identifications consistently and accurately for tracking purposes. According to the FDA's Guidance for Industry<sup>1</sup>, "The goal of microbiological monitoring is to reproducibly detect microorganisms for purposes of monitoring the state of environmental control. Consistent methods will yield a database that allows for sound data comparisons and interpretations". Methods that provide inconsistent identifications or no identification to the same organisms are not useful for tracking isolates to their source or for generating trending reports. For a database to be reliable, inquiries into the process used to build and test new entries and the frequency of validated library updates should be a considered as part of your evaluation. As an example, in May 2010 approximately 40 papers were published in the International Journal of Systematic and Evolutionary Microbiology describing new bacterial species and taxonomic name changes. Therefore, the system that can reliably report microbes down to a species level over time will be influenced by the frequency of validated updates to the database. In addition, although bacterial names may change, the true identity of any organism is its genetic sequence.

### ***Inherent Weakness of Phenotypic Identification Systems***

What are the important factors that you need to consider, aside from the number of entries in the database itself in your evaluation? Commercially available phenotypic methods rely on judicious attention to detail; media selection, culture conditions, age of the culture, cellular morphology, Gram staining and selective enzymatic and biochemical reactions for identification against the system's reference library. Errors in identifying microorganisms isolated from the pharmaceutical manufacturing environment are not uncommon using phenotypic systems given that many of these organisms are aberrant strains and physiologically stressed to the point of not fully expressing their phenotypic characteristics. This inherent metabolic variability will lead to inaccurate species identification in a biochemical based system because the microorganisms do not consistently express their expected phenotypic profiles.

### ***The Potential of Ribosomal Protein Databases***

Matrix-assisted laser desorption ionization based on time of flight (MALDI-TOF) is the newest technology available for bacterial ID utilizing ribosomal proteins. Ribosomal proteins, which are constitutively expressed and present in very high numbers, play a critical role in all cells. A unique protein fingerprint or spectra based on these

ribosomal proteins is compared against the database for identification. MALDI-TOF has been shown to have increased accuracy and reproducibility over phenotypic expressions systems and results are less complicated to analyze than with genotyping systems. Since the commercial availability of the technology is new and first introduced for clinical use, the database coverage for environmental species is in the building process. Therefore, it is imperative in your decision making process to include another system with a complementary database or a service provider that offers supplemental full species coverage testing, such as 16S sequencing.

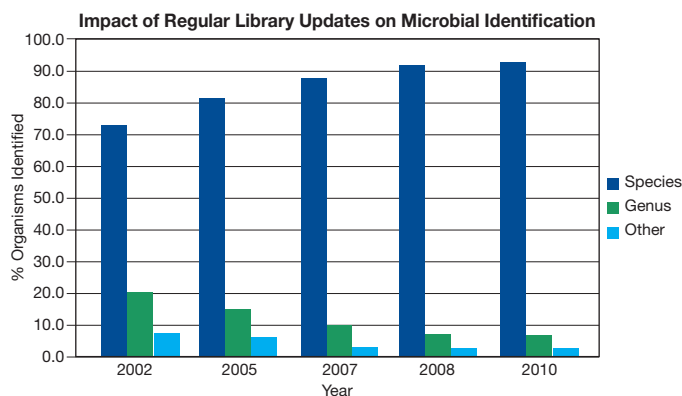
### ***Genotypic Data Analysis***

One approach to genotypic systems is to amplify and sequence the 16S gene to reveal the identity of the organism. The questions to ask in evaluating these systems and services are: Does the data analysis of the unknown microbe include a comparison to the full 500 base pairs of the 16S genome region within the library, or less, which may lead to an inaccurate identity? How are the raw data files and consensus sequences analyzed to determine the true identity of an unknown? Taxonomists routinely classify a bacterium by sequencing the entire length of the 16S rRNA PCR product, approximately 1500 base pairs. For routine identifications, typically the first 500 base pair region is generated from a PCR product and should be subsequently analyzed without omitting any nucleotides. In the case of automated analysis, some programs will truncate nucleotides from the ends of the sequence until a certain quality limit is reached. The most accurate systems generate data employing the entire 500 base pair region for comparison against the library reference entries. Equally important is the alignment of the sequence against these references, checking and correcting for base caller errors and obtaining accurate distance measurements to determine an identity. This approach to data analysis is the most accurate and reliable method to account for deletions, insertions and polymorphic changes in the region. Methods that utilize this systematic approach provide the most accurate measurements for determining the genetic distance between the unknown sample and the reference entries. How much distance between the unknown and reference sequences determines the match confidence level. Even when a species name is not achieved, using the sequence data provides specific information for tracking and trending purposes. However, the level of complexity required for data review steps and interpretation reflect the need for a resource with a higher skill set.



## Library Coverage and Maintenance Program

Other measures to consider are the frequency of the library database updates and if the coverage and capacity of the database reflects your work environment (i.e. manufacturing, clinical). Asking your vendor to provide trending data after the release of new libraries should provide very clear evidence that timely updates to libraries greatly increase the probability of a Species level of identification. With each release, the number of organisms that are not identified to the Species level should decrease significantly. The absence of frequently occurring organisms that reflect your working environment will reduce the chances of Species level identifications and cost you more time and money in repetitive testing. Regular library updates encompassing taxonomic changes and novel organisms are imperative for superior performance, continued reliability and relevance. The chart below illustrates the improvement of species level identifications following regular library updates and concurrent decrease in genus level ID.



## Validation Process

Working in a highly regulated environment requires an initial audit of the library validation procedures to insure that the manufacturer or outsourcing provider follows a rigorous cGMP compliance program. When auditing a library database, find out if old and new library entries undergo identical development, documentation and testing processes. What practices are utilized to document and validate library entries with each new library release? Does the manufacturer or service provider go back to older library entries, reacquiring strains and testing sequences, protein spectra arrays and/or biochemical enzymatic reactions?

## Fitness for Use

A good quality system must include studies that demonstrate the effectiveness of the design, method and processes employed in delivering a library database and identification methodology that fits the operational conditions of the pharmaceutical environment or “Fitness for Use”. Evaluate the Fitness for Use studies for accuracy, robustness and reproducibility when applied to an extensive panel of known culture collection strains and frequently observed organisms found in the sterile and aseptic manufacturing environments. These results will assist in determining if the technology and database will meet your specified performance expectations. In addition to panels containing the typical Gram negative, Gram positive, aerobic and anaerobic bacteria, consider challenging the system with the organisms that are frequently encountered in your manufacturing environment. The short list below represents additions to a panel that would demonstrate fitness for use for your facility:

- 1) Isolates detected in your starting materials and in-process testing
- 2) Isolates detected through your environmental monitoring of your facility
- 3) Isolates from your production areas that represent low nutrient and high stress growing conditions
- 4) Microorganisms reported in the literature to be common isolates from a particular product type that you manufacture

## Conclusion

When evaluating a system or service, assessment of the library database to be used for your microbial identifications has the greatest impact on the accuracy, reproducibility and cost. Database libraries that are deficient will lead to errors in species level identification or inconclusive results that will add directly to costs attributed to a higher than expected repeat testing rate. Additionally, intangible costs may rise due to time delays in contamination investigations and mitigation of risk, production interruptions and product safety.

### Resources and References:

1. FDA Guidance for Industry – 09/2004 - Sterile Drug Products Produced by Aseptic Processing - Current Good Manufacturing Practice



The content presented in this paper originally appeared in the July 14, 2010 issue of BioSciences Quality Testing Forum, an on-line platform of interactive communication, exchange of information, resources for R&D, Quality Control/Assurance, Manufacturing professionals in the Bio-Pharma and Cosmetics industries, focusing on analytical tools & methods. Access the latest Industry News, Product Information, Events, Opinion Polls, etc. at [www.BioQTFforum.com](http://www.BioQTFforum.com)

## About Accugenix

As the world's leading provider for microbial identifications, Accugenix is committed to providing the most accurate identifications to the industries we serve. This commitment requires that we continuously update our validated libraries to stay current with an evolving microbial world, where taxonomic changes and new species are described daily. The direct benefit is that Accugenix is able to identify more samples to the species level for our customers. Outsourcing your samples to Accugenix allows you to gain access to the most relevant, up-to-date and validated Bacterial and Fungal Libraries in the industry.

We routinely evaluate our sample identification rate, for bacterial and fungal customer samples. Since our objective is to achieve the highest percentage of Species level identifications, we compare sample sequence data that do not result in a Species level match against a variety of publicly available sequence databases. If a newly described organism is a good match for those sample sequences, and is validly named and published, the sequence proceeds to our validated library entry process. Building up-to-date libraries is not a trivial task; each entry goes through an extensive testing, review and approval process that is governed by our Standard Operating Procedures.

As an independent contract laboratory service provider offering multiple technology solutions, we routinely compare competitor databases against our gold standard reference libraries. The Library Comparison document discusses the results of our comparison against six commercially available identification systems and in this study you'll see confirmation that the current Accugenix bacterial database surpasses all other systems.



### Corporate Office (US):

Accugenix, Inc.  
223 Lake Drive  
Newark, DE 19702 USA  
Phone: +1.302.292.8888  
Toll Free: 800.886.9654  
Fax: +1.302.292.8468

### Email:

Literature Requests: [marketing@accugenix.com](mailto:marketing@accugenix.com)  
Technical Support: [technicalsupport@accugenix.com](mailto:technicalsupport@accugenix.com)  
Sales Department: [sales@accugenix.com](mailto:sales@accugenix.com)  
Sales Europe (Accugenix GmbH): [europe@accugenix.com](mailto:europe@accugenix.com)

### Europe (sales inquiries only):

Accugenix, GmbH  
Schimperstraße 1  
68167 Mannheim Germany  
Phone: +49 (0)621 3709 556  
Fax: +49 (0)621 3709 023

**Identification Request Forms** and additional information about our services and capabilities can be found on-line at [www.accugenix.com](http://www.accugenix.com)