Identification of Microbial Pathogens Using Nucleic Acid Sequencing

By Peter C. Iwen, PhD, Associate Director, NPHL

For more than 100 years, Robert Koch's postulate that required in part the cultivation of a pathogen to show a disease/pathogen relationship, was seldom questioned and was considered the basic standard used in clinical diagnostics. Organism identification to taxon (species, genus) was subsequently accomplished by studying phenotypic characteristics such as Gram stain, morphology, culture requirements, and biochemical reactions along with a combination of intuition and stepwise analysis of the results. In today's laboratory, the ability to detect and identify pathogens has undergone major changes. The development of molecular methods that rely on the detection of genomic elements (DNA or RNA) with or without culture has led the way in this charge. Some of the main reasons for this change from phenotypic to molecular testing include such issues as the slow growth of pathogens, the detection of organisms that exhibit biochemical characteristics that do not fit patterns of known species, and the inability to detect non-cultivatable organisms. Although culture-based methods are still considered the gold standard for identification diagnosis, molecular methods have emerged as the confirmatory method for identification in many diagnostic applications.

The basic principle of any molecular test is the detection of a specific nucleotide sequence (signature sequence) within the organisms' genome which is then hybridized to a labeled complementary sequence followed by a detection mechanism. The first application of these methods in the clinical laboratory was in the development of labeled probes for culture confirmation testing. The original probes were designed to detect "problem" pathogens such as those that were historically difficult to identify using phenotypic methods. These original probes included tests for the culture confirmation of dimorphic fungal pathogens (*Blastomyces dermatitidis, Coccidioides immitis,* and *Histoplasma capsulatum*) and to identify the more common *Mycobacterium* species (*M. tuberculosis* complex and *M. avium* complex). Subsequently, direct detection probes were designed for high volume testing of STD pathogens e.g., *Chlamydia trachomatis* and *Neisseria gonorrhoeae* and for the testing of pathogens that were difficult to grow and identify in the laboratory e.g., *Legionella pneumophilia* and Human papillomavirus.

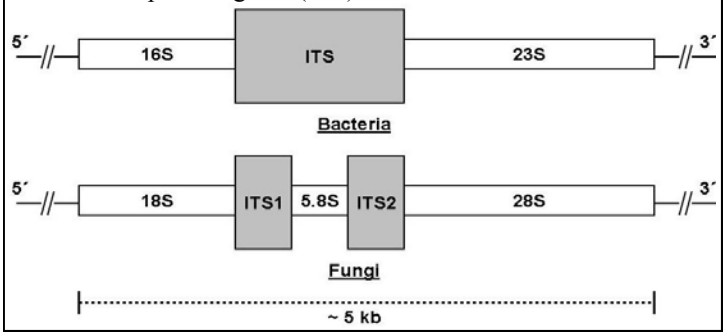
Although extensively used today, nucleic acid probing unfortunately has been shown to have limited selectivity and to lack sensitivity when testing from direct specimens. To overcome these problems, a process whereby the genomic target could be amplified using non-selective means was developed. The most widely used method for nucleic acid amplification is the polymerase chain reaction assay i.e., PCR. This assay includes a specific primer pair to amplify a unique genomic target nucleotide sequence for analysis. Following PCR, a variety of post-amplification methods are used to evaluate the product such as direct sequence analysis, use of genus or species specific probes, and utilization of restriction enzymatic analysis of the product, e.g., restriction fragment length polymorphism analysis (RFLP).

Even though all these post-amplification methods have been shown to be useful for the evaluation of microbes, sequence analysis is considered a particularly useful method for the identification of microbial species due to its wide range application to a variety of species. The basic steps involved in this technique are shown in **Table 1**. One drawback to this methodology is that access to sequencing facilities is not readily available for many laboratories, limiting the ability of most laboratories to conduct routine sequence analysis testing. To overcome this issue, commercial kits and some reference laboratories now offer low-cost sequencing for those instances where identification is required for diagnostic purposes.

Table 1. Sequential steps for the molecular identification of microorganisms using nucleic acid sequencing.
DNA or RNA extraction *In vitro* amplification e.g., PCR-based assays to detect specific DNA target Sequence determination i.e., analyze the PCR product Computer-aides sequence analysis e.g., BLAST search using the NCBI GenBank database[a]
[a] Basic Local Alignment Search Tool (BLAST) is a computational method for sequence comparison alignment which is available for public use

Sequence-based identification requires the recognition of a molecular target that is large enough to allow discrimination of a wide variety of microbes. One such target area that has been recognized is the rDNA gene complex which is present in all microbial pathogens. In bacteria, this complex is composed of a 16S rRNA gene and a 23S rRNA gene separated by a genomic segment called the internal transcribed spacer (ITS). Within fungi there are three genes (18S, 5.8S, and 28S) with spacers located between the genes (ITS1 and ITS2). **Figure 1** shows a representation of the rDNA gene complex in bacteria and fungi denoting gene order and position of the spacers. Located in the rDNA gene complex are highly variable sequences that provide unique signatures for the identification of species and also conserved regions that contain genomic codes for the structural restrains that are present within organism groups. It has been shown that the ITS regions contain the most variability and that these regions are useful under most circumstances for species recognition. The availability of these variable sequence regions (ITS) surrounded by conserved sequences (16S/23S and 18S/5.8S/28S) allows for the utilization of an amplification system using universal (or consensus) bacterial or fungal primers. Once amplification has occurred using the consensus primers, the sequence is determined and comparison analysis of the unknown sequence to known sequences contained within a large database (such as the National Center for Biological Information (NCBI), GenBank databases) can be done to determine similarity and subsequently may lead to species identification.

Figure 1. Representation of the rDNA gene complex in bacteria and fungi denoting gene order and position of the internal transcribed spacer regions (ITS).



Though public databases such as GenBank are useful, the lack of quality sequences and the absence of sequence information on a large number of species as well as the availability of computational tools to reliably analyze the results are drawbacks to this technology. Additionally, strain variability within species also has not been fully evaluated and has proven to be problematic when evaluating species.

Even with these challenges however, nucleic acid sequence analysis has proven to be a valuable asset for organism identification in a number of applications. Some of the most interesting applications of this technology are for the identifications of variant strains of know species, the identification of un-cultivatable organisms in clinical samples and the recognition of new species.

Identification of variant strains of known species. The utilization of phenotypic identification methods classically requires a probability-based analysis to determine identity. In cases where identification probabilities are ≥98% with known species, the identification is generally considered acceptable. The lower the probability percentage however, the less accurate the identification becomes, frequently resulting in supplemental testing to resolve discrepancies among test results. It is not unusual for the laboratory to be unable to identify variant strains of known species using phenotypic methods. DNA sequencing now allows the laboratory a means to resolve those instances where phenotypic testing cannot differentiate among closely related organisms.

Identification of non-cultivatable pathogens. The etiological agents for a variety of diseases continue to elude current diagnostic testing. The inability to perform *in vitro* culture of microbial pathogens is not a new concept. For example, *Treponema pallidum* even though recognized as the cause of syphilis, continues to be non-cultivatable in the laboratory. Other organisms such as *Bartonella* species, *Legionella* species, *Ehrlichia* species, and *Helicobacter pylori* were only recently cultivatable once the nutritional requirements were recognized. Fortunately, DNA sequencing now allows for the direct detection of microbial genomic material in tissues suspected to contain a microbial pathogen.

Table 2 gives examples of human pathogens that were first examined in clinical material using a molecular approach. The bacterial pathogens of this group were all detected in tissue using universal bacterial primers followed by sequence analysis of the 16S rDNA gene complex. Expectations are that other microbial pathogens will be recognized in the future using this technology.

Table 2. Human pathogens first identified in clinical specimens using molecular approaches.

Disease	Causative agent
Non-A, non-B hepatitis	Hepatitis C virus
Bacillary angiomatosis	*Bartonella henselae*
Whipple's disease	*Tropheryma whipplei*
Hantavirus pulmonary syndrome	Sin nombre virus
Kaposi's Sarcoma	Human herpesvirus 8
Disseminated infection in AIDS	*Mycobacterium genavense*

Identification of new species. The recognition of a species that does not match known schemes for phenotypic identification may represent a previously unrecognized species. Sequencing of areas within the rDNA complex may be useful to suggest a new species when there is a < 98% of the sequence similarity with known species. The ability to separate a new species from an atypical strain of a known species is however, difficult. The first approach to recognition of a new species is to determine the phylogenetic position of the suspect new species compared to closely related known species. Phylogenetic trees using the 16S gene for bacteria and the 18S gene for fungi are commonly used for this type of analysis. A degree of high degree of phenotypic consistency and rDNA sequence similarity as well as, a significant degree of DNA-DNA hybridization, is suggestive of a new species. Similar approaches were recently used in a research laboratory at UNMC to describe a previously unrecognized species subsequently named *Mycobacterium nebraskense* (see separate article on page 5)

In closing, researchers at the UNMC in collaboration with the clinical lab scientists at The Nebraska Medical Center continue to study ways to apply molecular detection techniques to enhance disease diagnosis. One recent improvement was the development of a molecular assay in combination with a computational algorithm (called MycoAlign) for the identification of *Mycobacterium* species. This prototype system is currently undergoing evaluation at multiple off-site locations throughout the United States and at this time being considered for international distribution. Additionally, an algorithm and database for the identification of fungal pathogens is also being developed. Personnel at the Nebraska Public Health Laboratory will continue to conduct research in this area as the technology evolves.

Any questions concerning sequence analysis testing can be directed to Dr. Peter Iwen at 402-559-7774.

References

Iwen, P. C., S. H. Hinrichs, and M. E. Rupp. 2002. Utilization of the internal transcribed spacer regions as molecular targets to detect and identify human fungal pathogens. Med. Mycol. **40:87-109.**

Mohamed, A. M., D. J. Kuper, H. H. Ali, P. C. Iwen, D. R. Bastola, and S. H. Hinrichs. 2004. Computational approach for the identification of *Mycobacterium* species using the internal transcribed spacer-1 region. 104th General Meeting of the American Society for Microbiology, New Orleans, LA, Abstract #U-074

Relman, D.A. 1999. The search for unrecognized pathogens. Science **284: 129-131.**