Application of Bioinformatic Tools for the Identification and Characterization of Microbes in the Medical Microbiology Laboratory

DANIEL GOLEMBOSKI

ABSTRACT

Genome sequencing technologies have provided information that enables a faster, more precise characterization of bacteria, according to DNA and RNA sequences. Additionally, comparison of genomes from closely related bacteria makes it possible to determine why some strains are pathogenic, to predict clinical outcomes of infections, and to develop therapeutic strategies. Application of this technology to the clinical laboratory provides reliable and accurate means of reducing turnaround time and identification of pathogenic bacteria that might be incorrectly or not readily identified by traditional methods. In this class unit, undergraduate Medical Laboratory Science students used genome databases and online computer algorithms with simulated sequence data to identify a bacterium, as an alternative to traditional biochemical analysis. In addition, students used the genome sequence to search for virulence genes and antibiotic resistance genes. These analytical processes are applicable to molecular protocols currently in use. The students gained familiarity with bioinformatics analysis and deepened their understanding of genome structure and function.

ABBREVIATIONS: DNA – deoxyribonucleic acid, RNA – ribonucleic acid, MLS – medical laboratory science, rRNA – ribosomal ribonucleic acid, NCBI – National Center for Biotechnology Information, BLAST – Basic Local Alignment Search Tool, RAST – Rapid Annotation using Subsystem Technology, IDNS – Integrated Database Network Service, FASTA – a text based nucleotide and/or protein sequence format, RefSeq – reference sequence, NIH – National Institutes of Health

INDEX TERMS: Computational biology/Education, Genomics/Education, Microbiology, Medical Laboratory Science/ Education, Instructional strategies

Clin Lab Sci 2015;28(1):19

Daniel Golemboski, PhD, *Department of Medical Laboratory Science, Bellarmine University, 2001 Newburg Rd., Louisville, KY*

Address for Correspondence: Daniel Golemboski, PhD, *Department of Medical Laboratory Science, Bellarmine University, 2001 Newburg Rd., Louisville, KY 40205, dgolemboski@bellarmine.edu*

INTRODUCTION

Traditionally, Medical Laboratory Science (MLS) students have learned to identify bacteria using criteria such as morphology, immunological markers, and biochemical characteristics. Some bacteria provide obstacles to identification such as long generation times, strict growth requirements, uncharacteristic metabolic biochemistry, and unusual morphology. Genome sequencing technologies provide information that enables faster, more precise characterization of bacteria according to DNA, ribosomal RNA (rRNA), and other gene sequences. Additionally, comparison of genomes from closely related bacteria makes it possible to determine why some strains are pathogenic, predict clinical outcomes of infections, and develop therapeutic strategies. The development of high-throughput DNA sequencing technology and subsequent completion of the human genome project has led to the availability of a vast amount of eukaryotic and prokaryotic genomic data ready for analysis. An understanding of bioinformatics is quickly becoming an important skill in the clinical setting. The Institute of Medicine has identified use of information technology as an essential competency for healthcare professionals; in the laboratory, information technology in the genomic era will include application of bioinformatics.[1]

New teaching methods incorporating bioinformatics are

actively being developed for many of the life sciences. These approaches are based on the recognition that involvement and ownership enhance student interest and performance.[2] Therefore, it is becoming increasingly important for Medical Laboratory Science educators to integrate an introduction of genomic analysis methods into curricula to prepare students to interpret this data as it becomes available and applicable to the clinical laboratory.

In this exercise, students used computer databases and bioinformatics to analyze bacterial proteins and genes as an alternative to the traditional methods of identification and characterization. The assignment focused on the ability to rapidly identify an organism and to determine the presence of antibiotic resistance genes in that organism. This procedure provides a no-cost simulation of an expensive process; by using nucleotide sequence data which is publicly available, students still experience a real-life scenario. To fully understand and perform these functions, students should possess basic computer skills and be familiar with the microbial genome structure. Classroom experiences such as this one, using online bioinformatics software and databases, will develop familiarity with the use of genomics in organism identification and characterization. While the intention of this paper is to provide a bioinformatics exercise for educators to utilize in their classroom, it could also have educational value for practicing laboratory professionals.

Student Learning Outcomes

At the conclusion of this activity, students should be able to:

1. Access and interpret DNA sequence data to identify bacterial species.
   a. Students are provided with bacterial genome sequence data for a clinically-relevant organism (from publicly available genetic databases such as the National Center for Biotechnology Information, NCBI).
   b. Students are instructed in analysis of sequence data using an online Basic Local Alignment Search Tool (BLAST) to identify bacteria using the provided genomic sequence data.

2. Access and interpret genome annotation data to correlate bacterial genotype and phenotype.

Students are instructed in the use of online Rapid Annotation using Subsystem Technology (RAST) to predict the phenotype of the organism (e.g., antibiotic resistance or virulence factors) based on genomic information.

METHODS

1. Access and interpret DNA sequence data to identify bacterial species
   a. The instructor selects a group of organisms for investigation (i.e., respiratory pathogens) and obtains bacterial 16S rRNA sequence data for the selected organism(s) from a publicly-available database (Table 1).

Table 1. Retrieval of 16s rRNA sequence.

- Access the National Center for Biotechnology Information website (http://www.ncbi.nlm.nih.gov/gene/)
- Enter or select the desired organism(s)
- Select "Limits" to filter results
- In the search box enter "16S rRNA"
- A single organism name may be entered
- Or scroll down further to the section titled "Limit by Taxonomy", check the appropriate box and click on "Search"
- Locate the name of the organism on the right side of the screen and click to access the 16S rRNA gene sequence

The 16S rRNA gene sequences consist of highly conserved sequences that flank regions of hypervariability; the hypervariable regions provide the basis for speciation.[3,4] Other genes of interest can also be searched and identified in the same way. The software used for this exercise is free and web-based; only internet access is required. Ribosomal RNA sequence files for microbial genomes were obtained from either the National Center for Biotechnology Information website (http://www.ncbi.nlm.nih.gov/) or the Ribosomal Database Project (http://rdp.cme.msu.edu/html/). Other available databases with the same or similar information include the Ribosomal Differentiation of Medical Microorganisms (http://www.ridom.com/) Ribosomal Database Project European Molecular Biology Laboratory (http://www.ebi.ac.uk/embl/) and Smart Gene IDNS (http://www.smartgene.ch).

The NCBI GenBank database includes publicly

available DNA sequences submitted from individual laboratories and large-scale sequencing projects which may include multiple copies of a gene sequence from a particular organism. The first step in locating these sequences from the NCBI site is to access http://www.ncbi.nlm.nih.gov/gene/ and then enter into the search box "16S ribosomal RNA" and the name (genus and species) of the organism to be used in the exercise. This search performed without designating a particular organism will yield over 16,000 16S rRNA gene sequences; including the name of an organism in the search will reduce this significantly. Alternatively, a taxonomic tree can be used if the instructor wishes to select from a list of a variety of bacteria instead of designating a single organism. On the right-hand side of the screen there is a column headed "Top Organisms [Tree]" and any one of the organisms displayed can be selected. By clicking on "[Tree]", organisms can be selected by virtue of their location in the taxonomic tree. If the phylum Proteobacteria is being examined, then double clicking on that link will display the classes within that phylum from which the genus of interest may be selected. From the resulting list of rRNA gene sequences, as many organisms as desired may be selected to create a list for future use which may be saved as a file by returning to the top of the page and clicking the arrow next to "Send to", revealing a drop-down box of options.

After the organisms to be identified are selected, the 16S rRNA gene sequences can be retrieved as follows, and are then given to the students as a simulation of data that is typically obtained from a DNA sequencing process. Clicking on the name of the organism will bring up a full report on the 16S rRNA gene for that organism, and the Summary data confirms the gene being displayed. Select the link FASTA, located under the heading "Genomic regions, transcripts, and products" to display the actual nucleotide sequence; this is then copied and saved as a ".txt" file which is provided to the student. The first line of the FASTA sequence contains the name of the organism; it is useful to copy this information along with the sequence for reference, but it should be deleted from the sequences given to the students, since the organism is an "unknown". It is best to email the FASTA sequence to the student, as a .txt file attachment.

b. Students use an online BLAST tool to identify their organism using the rRNA sequence data (Table 2).

Table 2. Organism identification using BLAST

- Access BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi)
- Under the Basic BLAST section, select "nucleotide blast" to open a search page
- Enter the FASTA sequence by pasting under "Enter Query Sequence"
- In the "Choose Search Set" section, select "Others" and then "16S ribosomal RNA sequences (Bacterial and Archaea)" from the drop-down box
- Choose "Optimize" for "Highly similar sequences (megablast)"
- Hover over the first line with the computer mouse to display the name of the organism, or click on the line to show the name of the organism

Students identify the organism(s) associated with their 16S rRNA sequence by submitting the assigned sequence to the Basic Local Alignment Search Tool (BLAST; http://blast.ncbi.nlm.nih.gov/Blast.cgi). Under the Basic BLAST section, select "nucleotide blast". This will open a new page where the students enter the FASTA sequence, under "Enter Query Sequence", by copying and pasting the complete sequence they were given. In the "Choose Search Set" section, the database to search is "Others" and from the drop-down box "16S ribosomal RNA sequences (Bacterial and Archaea)" should be selected; then, choose "Optimize" for "Highly similar sequences (megablast)". It may take a few minutes for the results to appear as the software searches for identical sequences. The results are presented as a graphical illustration (see Figure 1), the colored lines representing different organisms that have sequences similar to the sequence the student entered. Each of

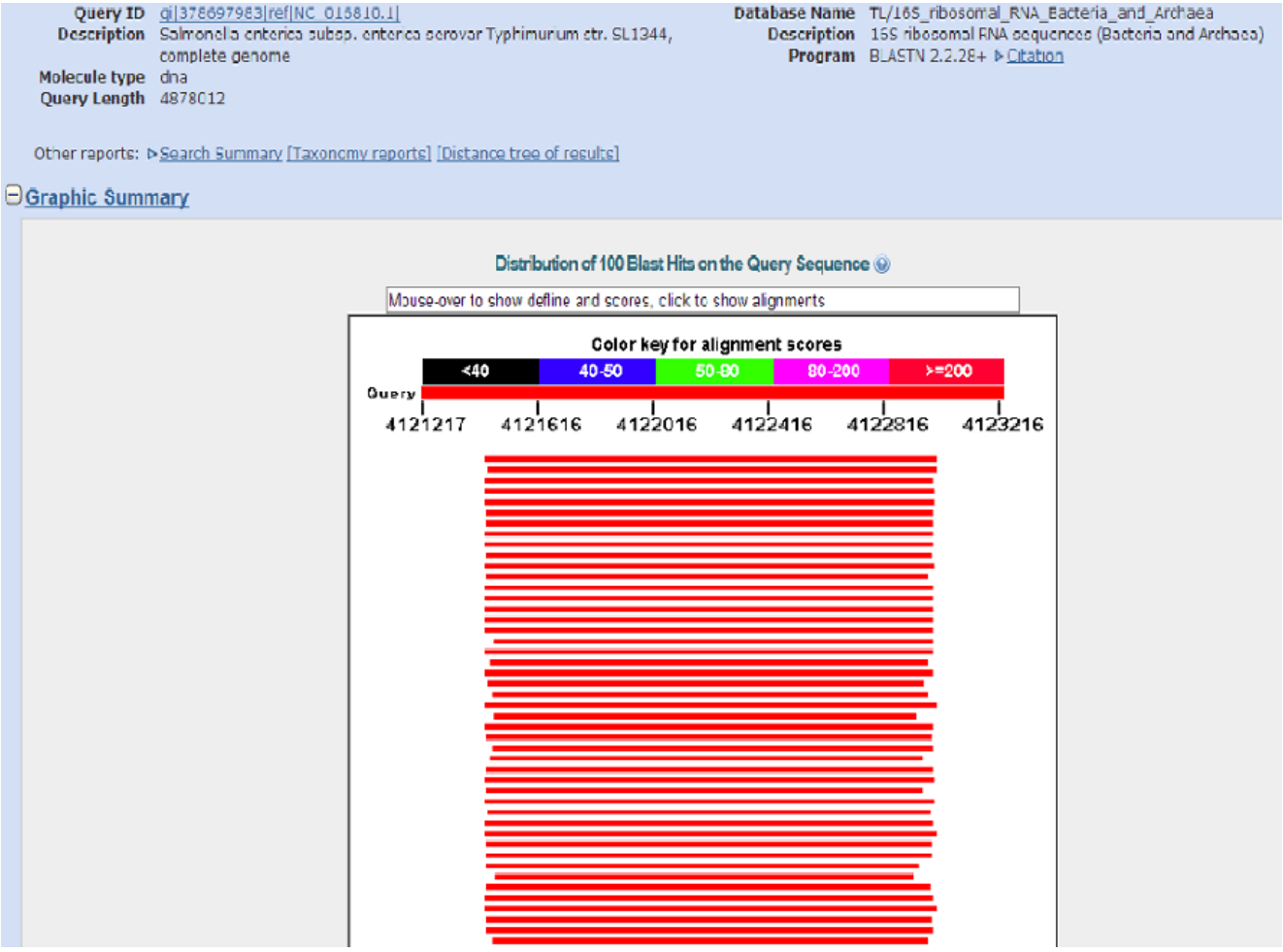

Figure 1. The results of the BLAST search are displayed graphically where a color coded line is used to represent a nucleic acid sequence that matches the original rRNA gene sequence being identified by the student. The length and location of the line indicates the region that is similar to the query sequence and the color is indicative of degree of relatedness.

the organisms shown here have varying degrees of homology with the original rRNA sequence. For the purposes of this exercise, it was determined to be sufficient for the student to choose the first organism (i.e., the first line) as their identified unknown. Hovering over this line with the computer mouse will display the name of the organism, or clicking on the line will show the name of the organism and the actual alignment of the two sequences.

It is important for the student to understand that, due to the highly conserved nature of rRNA genes, a database search does not always result in an absolute unambiguous identification of an unknown organism. It has been shown that low sequence variation can limit discrimination between bacterial variants (or genera in some cases).[5] When 16S rRNA sequence analysis alone does not enable identification, other genes can also be used to discern closely related species. For example, the *rpoB* gene, which codes for the β-subunit of RNA polymerase, has been used to differentiate *Mycobacterium chelonae* and *M. abscessus*, which have identical 16S rRNA gene sequences but <97% similarity in the *rpoB* sequence.[6] Other genes that have been used successfully in cases where 16S rRNA genes are identical are the *tuf* gene (elongation factor Tu), *sodA* (superoxide dismutase) and the gyrase genes *gyrA* and *gyrB*.[4,6] The *tuf* gene has been shown to be particularly effective in the identification of clinical isolates of coagulase-negative *Staphylococci*.[8]

Students then need to obtain the complete genome for the organism they identified, in order to identify other genes involved in the pathogenicity of the organism

(Table 3). To download the complete nucleotide sequence of the bacteria's chromosome, go to http://www.ncbi.nlm.nih.gov/genome and search by entering the name of the identified organism into the search box at the top of the page. These search results may not provide the data for the exact strain or variant of the species identified by the 16S rRNA BLAST analysis. Many more ribosomal RNA genes have been sequenced than complete genomes, so this search provides a reference sequence for that particular species, which is not necessarily the exact strain that was identified; however, for the purpose of this exercise, it provides the necessary sequence data. Clicking on the "+" next to any of the Reference genomes will open a comparative genome analysis data table which shows the accession numbers and size(s) of the organism's chromosome and plasmid(s), if present, %GC, and other relevant information. The number in the "RefSeq" column, in the "Chr" (chromosome) row, is a link that opens a detailed description of the genome and provides the sequence in FASTA format for the entire chromosome (the "FASTA" link is found directly below the name of the organism). Students can either copy and save the genomic sequence as a ".txt" file or downloaded it directly as a "FASTA" file by clicking on the arrow next to "Send", selecting "Complete Record" and setting the "Destination" as "File".

Table 3. Retrieval of the complete organism genome

- Go to http://www.ncbi.nlm.nih.gov/genome
- Enter the organism name the search box at the top of the page
- Click on the number in the "RefSeq" column, in the "Chr" (chromosome) row to open a detailed description of the genome and the entire chromosomal sequence in FASTA format

2. Students access and interpret genome annotation data to correlate bacterial genotype and phenotype (Table 4).

Students use the automated annotation service of the Rapid Annotation and Subsystem Technology (RAST)[9] to predict the phenotype of the organism they have identified (e.g., antibiotic resistance or virulence factors) based on the complete genome information.

Genome annotation is the process of interpreting sequence data using biological information. Detecting characteristic sequences within the genome makes it possible to identify the regions that may yield functional gene products. Analysis of these putative genes by comparison to other known gene sequences can provide additional evidence of their function.

Table 4. Correlate genotype and phenotype using RAST

- RAST is accessed at http://rast.nmpdr.org/
- Register for a RAST user account before submitting a genome for analysis
- Login and select "Upload New Job" from the "Your Jobs" menu, and browse for the FASTA genomic sequence saved after BLAST identification
- Select number 11 in step 2
- Proceed to step 3, leave the optional information blank, and click "Finish" (results typically take several hours)
- Log in to view the results in your user account
- Select "view details" to open the "Job Details" page
- Choose "Browse annotated genome in SEED Viewer" to view the data
- Click on the gene link to open a Subsytem overview; select the spreadsheet tab to display a list of all organisms with identical genes

RAST is accessed at http://rast.nmpdr.org/. Each student must register for a RAST user account before submitting a genome for analysis. After email account confirmation, students can login and begin by selecting "Upload New Job" from the "Your Jobs" menu. The sequence (in FASTA format, as described above) is uploaded by browsing for the sequence file saved earlier. It is important that the sequence contains all of the identifying information, including that which is included in the first line of the FASTA sequence. Step 2 asks for information about the sequence origin, such as whether the organism is bacteria or archaea, genus and specie, and genetic code (select number 11). Proceed to step 3, leave the optional information blank, and click "Finish". Detailed instructions for using RAST can be found at http://www.nmpdr.org/FIG/wiki/pub/Main/RAST/RASTtutorial.pdf.

Analysis of the complete genome sequence with the RAST server is typically an overnight process. An email is sent to each user when the annotation is complete; the results can be viewed after logging-in to the RAST site. Completed jobs are listed in each user's account. By selecting the "view details" under the Annotation Progress heading, the "Job Details" page will open.

Choosing "Browse annotated genome in SEED Viewer" will open a new tab, which displays the curated data by gene subsystems. Here the student can observe the genome organized according to biological function, allowing for easy selection of genes that further characterize the organism (i.e., antibiotic resistance) without additional laboratory testing. A pie chart of all the subsystems identified in the genome can be seen on the genome overview page, including antibiotic resistance, metabolic characteristics, and virulence mechanisms. The subsystem categories which are listed to the right of the chart can be expanded to reveal the subcategories and subsystem names, along with the number of proteins assigned to each (see Figure 2).

Additional information can be obtained here about the gene and other organisms which have the same gene. When the searched organism possesses a particular gene, a link is provided that will open a Subsystem overview page; selecting the Subsystem Spreadsheet tab then displays the gene(s) and a list of all organisms with identical genes. The original organism will be listed first. (Clicking on the name of the gene will open the Annotation Overview window where the nucleotide sequence in FASTA format can either be copied or downloaded directly for further analysis if desired).

Although not necessarily directly applicable to the clinical laboratory, another interesting continuation of this analysis is the comparison of the students' identified microbial genome with other genomes in the database, particularly if the 16S rRNA gene sequence analysis was not definitive. At the top of the Organism Overview page, under the Comparative Tools tab, selecting "Sequence Based Comparison" will open a new window where the student's bacterial genome can be selected and compared to any other selected genome. This section could be used to simulate comparison of genomes as might be done when looking at isolate-to-isolate variability among clinical specimens for epidemiological studies during an outbreak of a bacterial illness.[11]

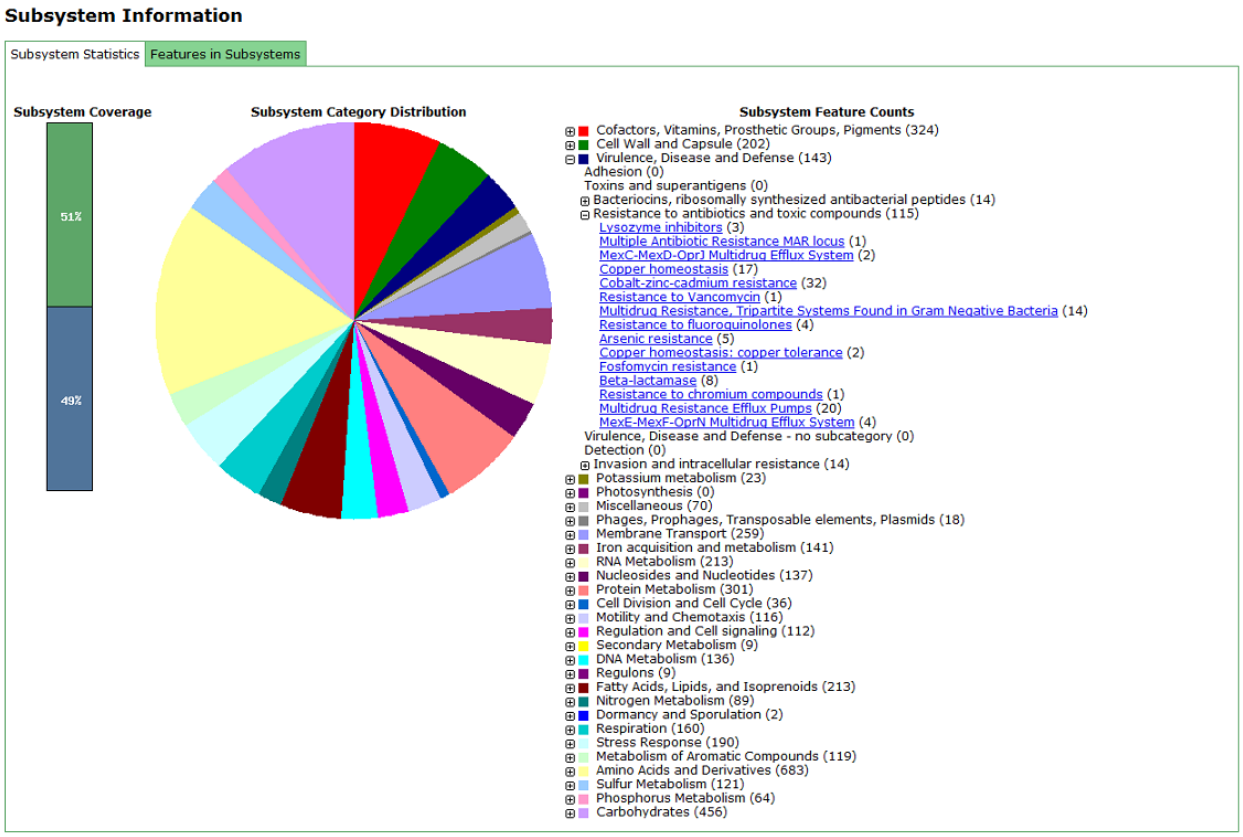

Figure 2. Selecting the subsystem "Virulence, Disease and Defense" for *Pseudomonas aeruginosa* PAO1, will expand it to show "Adhesion (0); Toxins and superantigens (0); Bacteriocins, ribosomally synthesized antibacterial peptides (14); Resistance to antibiotics and toxic compounds (115); Virulence, Disease and Defense - no subcategory (0); Detection (0); Invasion and intracellular resistance (14); the number in parentheses corresponds to the number of relevant genes that exist in that particular organism.

	Excellent (22 – 25 points)	Good (18 – 21 points)	Needs improvement (14 – 17 points)	Score
Students use provided sequence data in an online Basic Local Alignment Search Tool (BLAST) to accurately identify bacteria.	☐ Followed directions to use database correctly ☐ Correctly identified organism	☐ Was unable to follow directions OR Did not correctly identify organism	☐ Was unable to follow directions AND Did not correctly identify organism	
Students access and interpret genome annotation data to accurately detect presence of antibiotic resistance genes.	☐ Followed directions to use database correctly ☐ Correctly identified gene	☐ Was unable to follow directions OR Did not correctly identify gene	☐ Was unable to follow directions AND Did not correctly identify gene	
Students accurately correlate bacterial genotype and phenotype.	☐ Correctly identified organism's antibiotic resistance ☐ Correctly predicted resistance to additional antibiotics which might be targeted by the same resistance mechanism	☐ Was unable to identify organism's antibiotic resistance OR Was unable to predict resistance to additional antibiotics which might be targeted by the same resistance mechanism	☐ Was unable to identify organism's antibiotic resistance AND Was unable to predict resistance to additional antibiotics which might be targeted by the same resistance mechanism	
			Total Score	

Figure 3. Assessment Rubric

RESULTS

Student performance in this exercise was assessed using a rubric developed from the learning outcomes (Figure 3). All of the students successfully identified the organism on the first attempt. Some needed additional assistance to interpret annotation data. After completing the assignment, students demonstrated a more concrete grasp of the connection between genes and virulence.

DISCUSSION

The simulation exercise described here is not intended as an endorsement of the GenBank, the NIH genetic sequence database, as the preferred or sole criterion for bacterial identification. When used in conjunction with phenotypic characteristics, the likelihood of an accurate identification increases greatly.[10] However, even in the absence of cultivated organisms or a definitive biochemical characterization, rRNA sequence analysis can provide accurate and reproducible identification of bacteria.[11] A profile of the entire microbiome, including organisms that are non-cultivatable, can be constructed using this type of analysis.[12,13] Proprietary commercial databases may contain more accurate rRNA sequences, but unfortunately these are more limited collections of representative organisms and are not always freely available to the public.

As molecular technology is increasingly incorporated into clinical laboratory practice, future practitioners will be expected to access and interpret genomic data. This exercise provided students the opportunity to further understand the rationale and process of genotypic identification, as well as an introduction to the use of online bioinformatics software.

REFERENCES

1. Morris S, Otto C, Golemboski K. Improving patient safety and healthcare quality in the 21st century: competencies required of future Medical Laboratory Science practitioners. Clin Lab Sci. 2013;26(4):200-4.
2. Levis-Fitzgerald M, Denson N, and Kerfeld, CA. Undergraduate students conducting research in the life sciences: Opportunities for connected learning. 2005. Available from: http://eric.ed.gov/?id=ED491735 .
3. Clarridge JE. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev 2004;17:840-62.
4. Petti CA, Polage CR, Schreckenberger P. The role of 16S rRNA gene sequencing in identification of microorganisms misidentified by conventional methods. J Clin Microbiol 2005;43:6123-5.

5. La Scola B, Gundi VAKB., Khamis A, Raoult D. Sequencing of the *rpoB* gene and flanking spacers for molecular identification of *Acinetobacter* species. J Clin Microbiol. 2006;44:827-32.
6. Adekambi T, Drancourt M. Dissection of phylogenetic relationships among 19 rapidly growing *Mycobacterium* species by 16S rRNA, *hsp65*, *soda*, *recA*, and *rpoB* gene sequencing. Int J Syst Evol Microbiol. 2004;54:2095-105.
7. Poyart C, Quesne G, Boumaila C, Trieu-Cuot P. Rapid and accurate species-level identification of coagulase-negative staphylococci by using the *soda* gene as a target. J Clin Microbiol. 2001;39:4296-301.
8. Heikens E, Fleer A, Paauw A, Florijn A, Fluit AC. Comparison of genotypic and phenotypic methods for species-level identification of clinical isolates of coagulase-negative staphylococci. J Clin Microbiol. 2005;43:2286-90.
9. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75.
10. Fenolar F, Roux V, Stein A, Drancourt M, Raoult D. Analysis of 525 samples to determine the usefulness of PCR amplification and sequencing of the 16S rRNA gene for diagnosis of bone and joint infections. J Clin Microbiol. 2006;44:1018-28.
11. Justa SA, Knudsena JB, Uldumb SA, Holtc HM. Detection of *Legionella bozemanae*, a New Cause of Septic Arthritis, by PCR Followed by Specific Culture. J. Clin. Microbiol. 2012;50(12):4180-2.
12. Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, et al. Correction: Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing. PLoS Genet. 2008 http://dx.doi.org/10.1371/annotation/3d8a6578-ce56-45aa-bc71-05078355b851.
13. Huse SM, Ye Y, Zhou Y, Fodor AA. A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters. PLoS ONE 2012;7(6):e34242.