Supplemental Material

Organism identification using the VITEK MS v2.0 system

Description of the Knowledge Base

The reference database for the VITEK MS system (Knowledge Base) includes data representing 755 taxa, including 645 bacteria and 110 fungi. Each species or species group is represented by an average of 10 isolates (range 2 - 475). In order to capture the degree of acceptable variation within spectra from the same species, each reference isolate was grown on multiple media types under several growth conditions. The raw spectra were then acquired by more than one technician using multiple instruments. This process resulted in an average of 40 reference spectra per species (>30,000 in total).

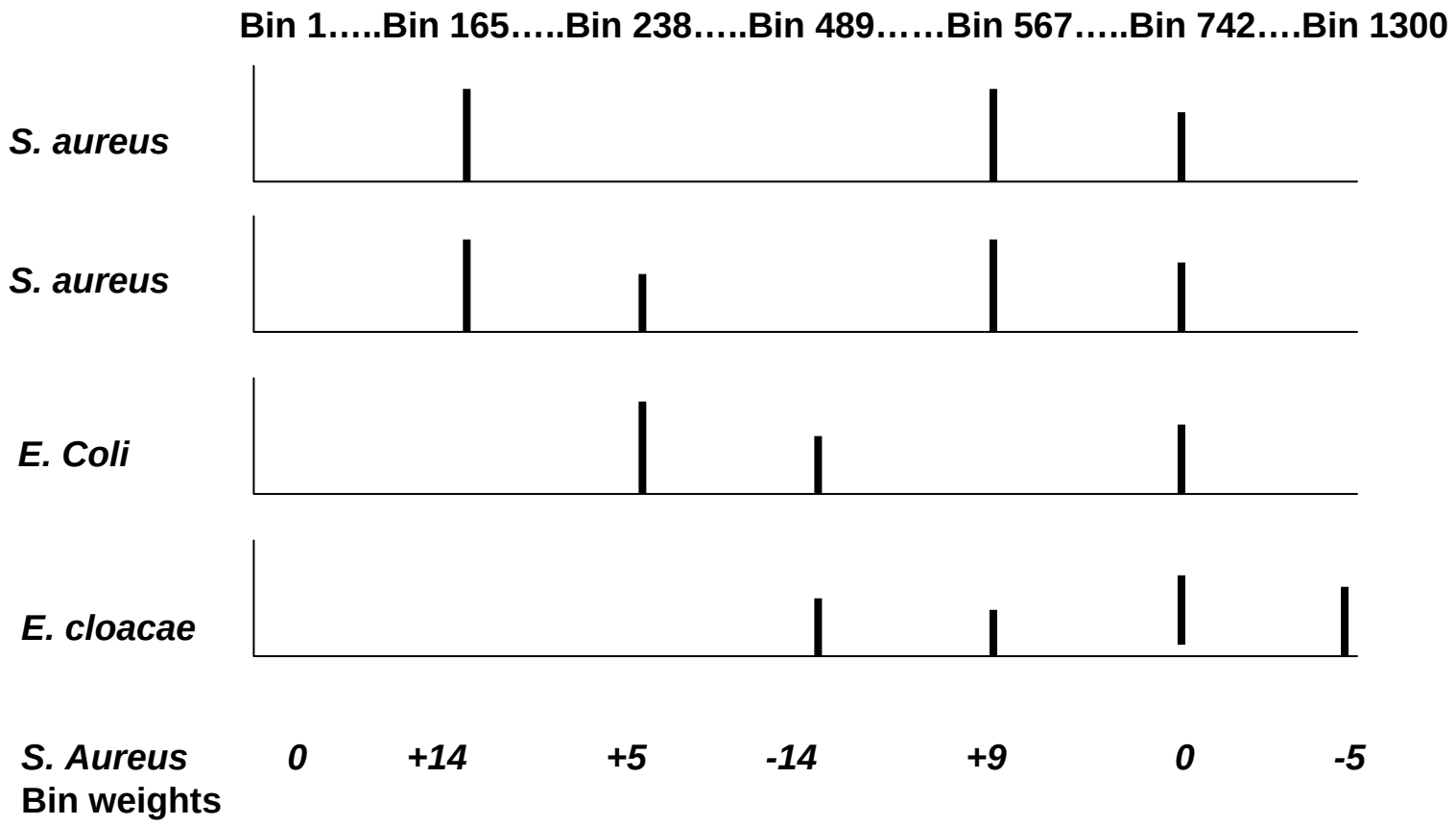
For each reference species in the database, baseline correction and de-noising was performed on the raw spectrum and peak detection was performed to identify well-defined (significant) peaks. The list of significant peaks was then subjected to a proprietary process called "mass binning." In this process, the spectrum between 3,000 and 17,000 Daltons was divided into 1300 pre-defined intervals called “bins”. An algorithm based on supervised machine learning, known as the “Advanced Spectrum Classifier”, was then used to determine how informative each bin was in differentiating that species from all other species in the database. For example (Supplemental Figure 1, using hypothetical numbers), if there were multiple reference spectra for *S. aureus* and nearly all of them contained a peak in bin 165, but few strains of other species contained a peak in

this bin, then a high positive weight (e.g., +14) would be assigned to bin 165 in the *S. aureus* row of the bin matrix, indicating a peak here is highly specific for *S. aureus*. Similarly, if few of the *S. aureus* strains contained a peak in bin 489, but many strains of other species contained a peak in this bin, then a strongly negative weight (e.g., -14) would be assigned to bin 489 of the *S. aureus* row in the bin matrix, indicating the absence of a peak here is highly specific for *S. aureus*. A neutral (near zero) weight was assigned when the presence or absence of a peak was not specific to any one species (i.e. bin 1 and bin 742). This process was replicated for each of the reference species thus creating a matrix with a species-specific weight for each of the 1300 bins (Supplemental Figure 2). Finally, as shown by example in Supplemental Figure 3, for each reference species, a mathematical function (blue curve) was derived from the distribution of weighted bin scores from all spectra for the given species (red curve) compared to the distribution of weighted bin scores from all other spectra in the database (green curve). This function can then take the summed bin score for an unknown (for example, +1.11) and provide the confidence value for how similar this score is to the given reference species in the Knowledge Base (in this case, 99.9%). By examining this function for all claimed species, it was determined that a threshold of 60% indicates that an unknown isolate's overall score is within the range of scores generated by known examples of that species, but is outside the range of scores generated by every other species in the database.

Spectral analysis for unknown isolates

Once an unknown's raw spectrum is acquired by the mass spectrometer it goes through pre-processing and mass binning as described above. The bin scores then go through an iterative process whereby the score within each bin is multiplied by the weighted bin value for each reference species in the Knowledge Base. The sum of the weighted bin scores is then calculated and used to determine the confidence value of the unknown relative to each reference species. After confidence values are obtained, a decision analysis is performed to retain only the significant organisms and create the final report to be returned to the user. This begins by eliminating all species for which the unknown isolate produced a confidence value less than the 60% threshold. A confidence level tolerance is then applied to eliminate species with a confidence value "too far" from species identifications with the highest confidence value. Finally, the resulting organism(s) is reported. In the event that there are more than four species on this list or if no species are on the list, a result of "no identification" is reported. As a hypothetical example (Supplemental Figure 4), given an unknown with a score of 0.56 in bin 165, 0.3 in bin 238, 0.8 in bin 489, 0.23 in bin 567, and 0 in bin 742, the overall weighted bin score relative to *S. aureus* would be -0.59 giving a confidence value of 36%. Similarly, the overall weighted bin score relative to *E. coli* would be +1.11 giving a confidence value of 99.9%. Given that *E. coli* is the only reference species that exceeds the 60% cutoff, the resulting identification for this unknown would be *E. coli*.

Supplemental Figure 1. Hypothetical Example of creating species specific bin weights using the Advanced Spectral Classifier



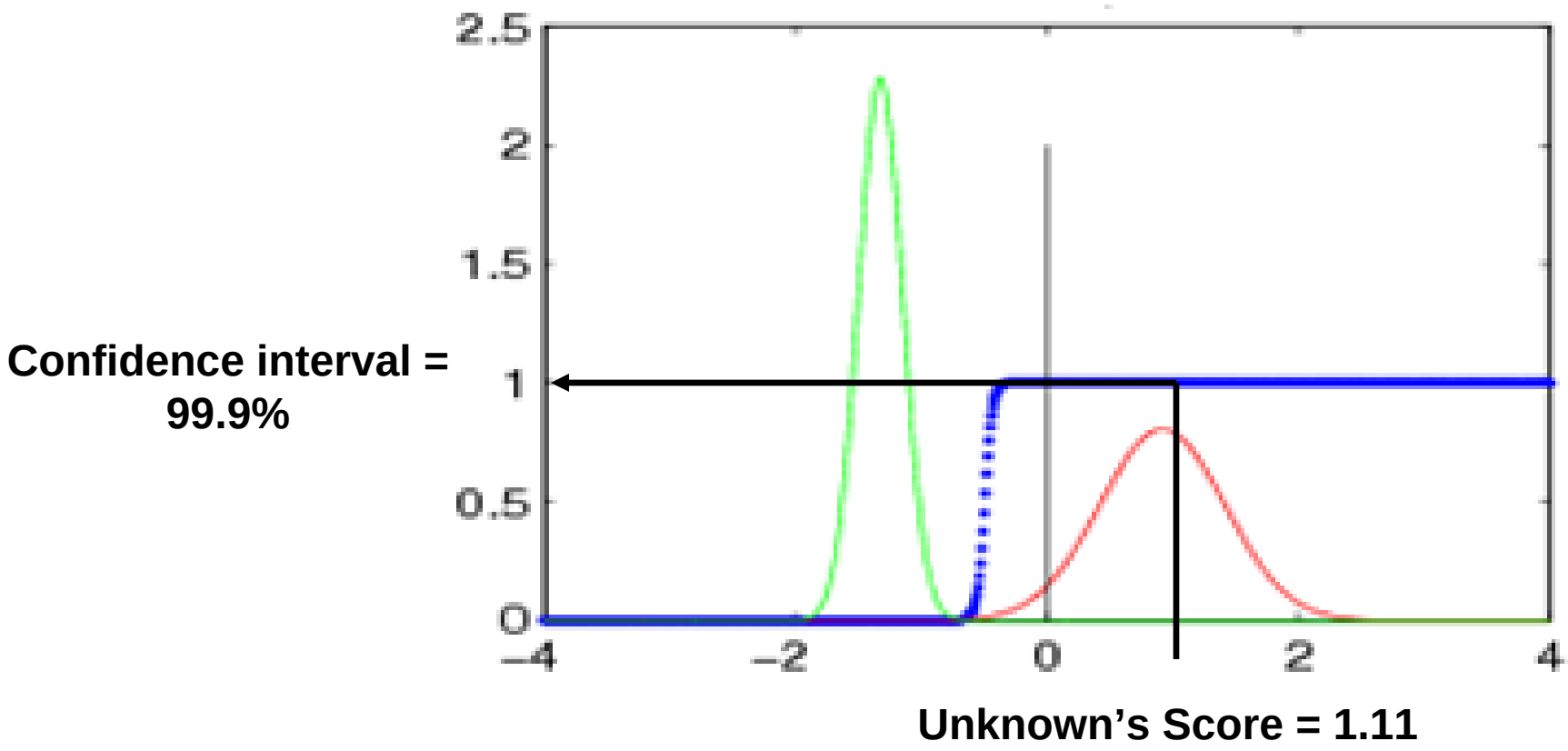
A hypothetical example of how weighted bin scores are calculated for each reference species in the Knowledge Base. Bin weights are calculated based on whether a bin contains (positive weight) or fails to contain (negative weight) peaks for the given reference species. A neutral weight is assigned when the presence or absence of a peak is not specific for any one species Adapted from: van Nuenen, Marc. VITEK MS: Taking Microbial MALDI-TOF ID from Research into the Routine. Presentation to CDC on August 13, 2012.

Supplemental Figure 2. Hypothetical Weighted Bin Matrix

	Bin 1	…	**Bin 165**	…	**Bin 238**	…	**Bin 489**	…	**Bin 567**	…	**Bin 742**	…	**Bin 1300**
A. defectiva	0		-12		6		-1		8		0		-3
.													
.													
.													
E. cloacae	0		-20		-3		7		1		0		20
E. coli	0		-13		9		12		-17		0		-2
S. aureus	0		14		5		-15		9		0		-5
.													
.													
.													
Y. pseudotuberculosis	0		17		7		-12		-17		0		6

A hypothetical example of the bin matrix, which is made up of rows containing each reference species in the database and columns containing the calculated weight of the bin for that species. Adapted from: van Nuenen, Marc. VITEK MS: Taking Microbial MALDI-TOF ID from Research into the Routine. Presentation to CDC on August 13, 2012.

Supplemental Figure 3. Hypothetical Confidence Interval Function



A hypothetical example of the confidence interval function, which is associated with each reference species in the Knowledge Base. The red curve is the distribution of all weighted bin scores from spectra within the given species. The green curve is the distribution of all weighted bin scores from all other spectra in the database. The blue curve is the mathematical function derived from these two curves that allows a confidence value to be determined relative to the summed bin score of an unknown. Adapted from: van Nuenen, Marc. VITEK MS: Taking Microbial MALDI-TOF ID from Research into the Routine. Presentation to CDC on August 13, 2012.

Supplemental Figure 4. Hypothetical Spectral Analysis Results

Unknown	**0.56**		**0.3**		**0.8**		**0.23**		**0**		**Sum**	**Confidence**
	…	**Bin 165**	**…**	**Bin 238**	**…**	**Bin 489**	**…**	**Bin 567**	**…**	**Bin 742**	**…**	
A. defectiva		-12(0.56)		6(0.3)		-1(0.8)		8(0.23)		0(0)		**-3.88**	**<1%**
.													
.													
.													
E. cloacae		-12(0.56)		-3(0.3)		7(0.8)		1(0.23)		0(0)		**-1.79**	**<1%**
E. coli		-13(0.56)		9(0.3)		12(0.8)		-17(0.23)		0(0)		**1.11**	**99.9%**
S. aureus		14(0.56)		5(0.3)		-15(0.8)		9(0.23)		0(0)		**-0.59**	**36%**
.													
.													
.													
Y. pseudotuberculosis		17(0.56)		7(0.3)		-12(0.8)		-17(0.23)		0(0)		**-1.89**	**<1%**

A hypothetical example of spectral analysis results. In this case, the bin scores of the unknown go through an iterative process to determine a score and confidence value relative to each reference species in the database. Adapted from: van Nuenen, Marc. VITEK MS: Taking Microbial MALDI-TOF ID from Research into the Routine. Presentation to CDC on August 13, 2012.