Similarity Index Calculation

for the Sherlock MIS

Understanding Similarity Index

Many traditional microbiology systems, such as biochemical tests, present results as a "probability". For example, the system may report a 98% probability for the identification of an isolate. The basic assumption behind these "probabilities" is that species are well-defined groups of organisms with little variation in how they react to certain biochemical tests. Comparisons have traditionally been made between two or more biochemical test systems and as a result these comparisons were nothing more than how well the systems perform similar enzyme assays. Even when the identification is incorrect, the "probability" of the identification may be quite high and may be "confirmed" using a similar enzyme assay system.

More recently developed techniques such as DNA homology and fatty acid analysis (Sherlock Microbial Identification System) use other means to express the identification of microorganisms. The current standard for DNA homology employs a cutoff of 70% DNA homology to indicate that two strains are of the same species.

The technique used by the Sherlock Microbial Identification System (MIS) is based on Similarity Index. The Similarity Index of the MIS is a numerical value, which expresses how closely the fatty acid composition of an unknown compares with the mean fatty acid composition of the strains used to create the library entry or entries listed as its match. The database search presents the best matches and the associated similarity indices. This index value is a computer-generated calculation of the distance, in multi-dimensional space, between the profile of the unknown and the mean profile of the most similar library entry. Thus, it is not a "probability" or percentage, but an expression of relative distance from the population mean. An exact match of the fatty acid makeup of the unknown and the mean of a library entry would result in a Similarity Index of 1.000. As each fatty acid varies from the mean percentage, the Similarity Index will decrease in proportion to the cumulative variance between the composition of the unknown and the library entry.

The Similarity Index of the MIS assumes that species of microorganisms have normal Gaussian Distributions (classical 'bell shaped curve'). The mean of the population in any series of traits (eg. fatty acid percentages) characterizes the group. Most of the population falls somewhere near the mean, but individuals will differ in composition and thus may show considerable variance from the mean.

Looking at a chart of the fatty acid composition of an unknown and observing the distance from the mean of each fatty acid can help visualize the cumulative variance of the fatty acid composition of an unknown strain. For example, Figures 1 and 2 show the library comparisons for two strains of *Staphylococcus epidermidis* that have quite different Similarity Index numbers. Strain A (Figure 1) has a fatty acid comparison much more similar to the mean of the population than does Strain B (Figure 2).

In Figure 1 and 2, all of the fatty acids found in a sample and in the library entry are listed in elution order on the left side of the chart. A scale of percentages is printed across the bottom of the chart. For each fatty acid, a symbol is placed on the line opposite the name indicating the amount of that acid. The library entry mean value for an acid is indicated with a vertical line. The horizontal bar gives a ±2 standard deviation window around the mean value for the library entry. The unknown's value is indicated by a dot on the chart.

Figure 1: MIS Library Comparison Chart For A Good Match

Strain A:

CLIN40 Staphylococcus 0.762

S. epidermidis 0.762

For this high-match sample, the unknown matches the mean for 15:0 ISO and is within two standard deviations for 15:0 ANTEISO

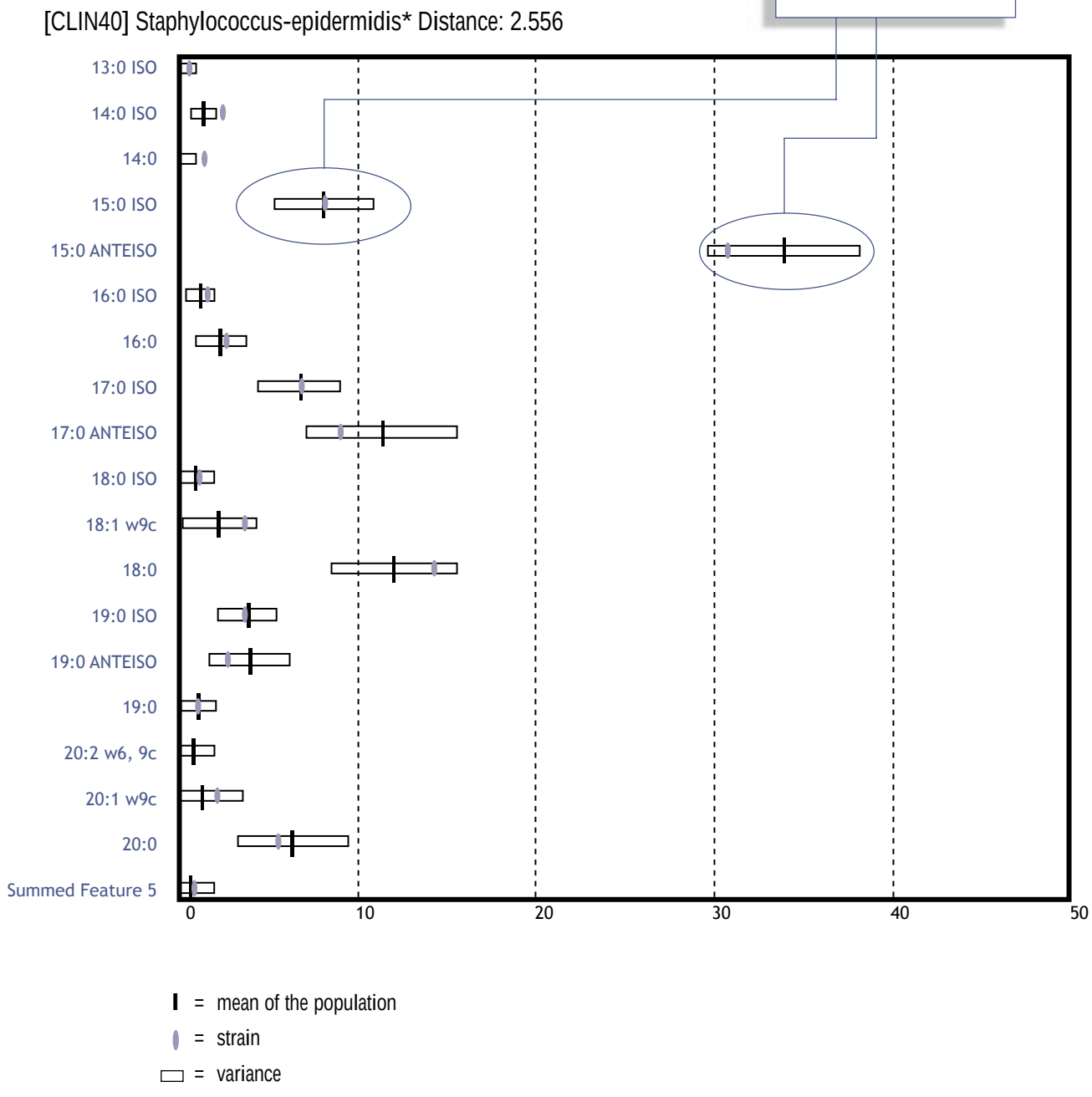


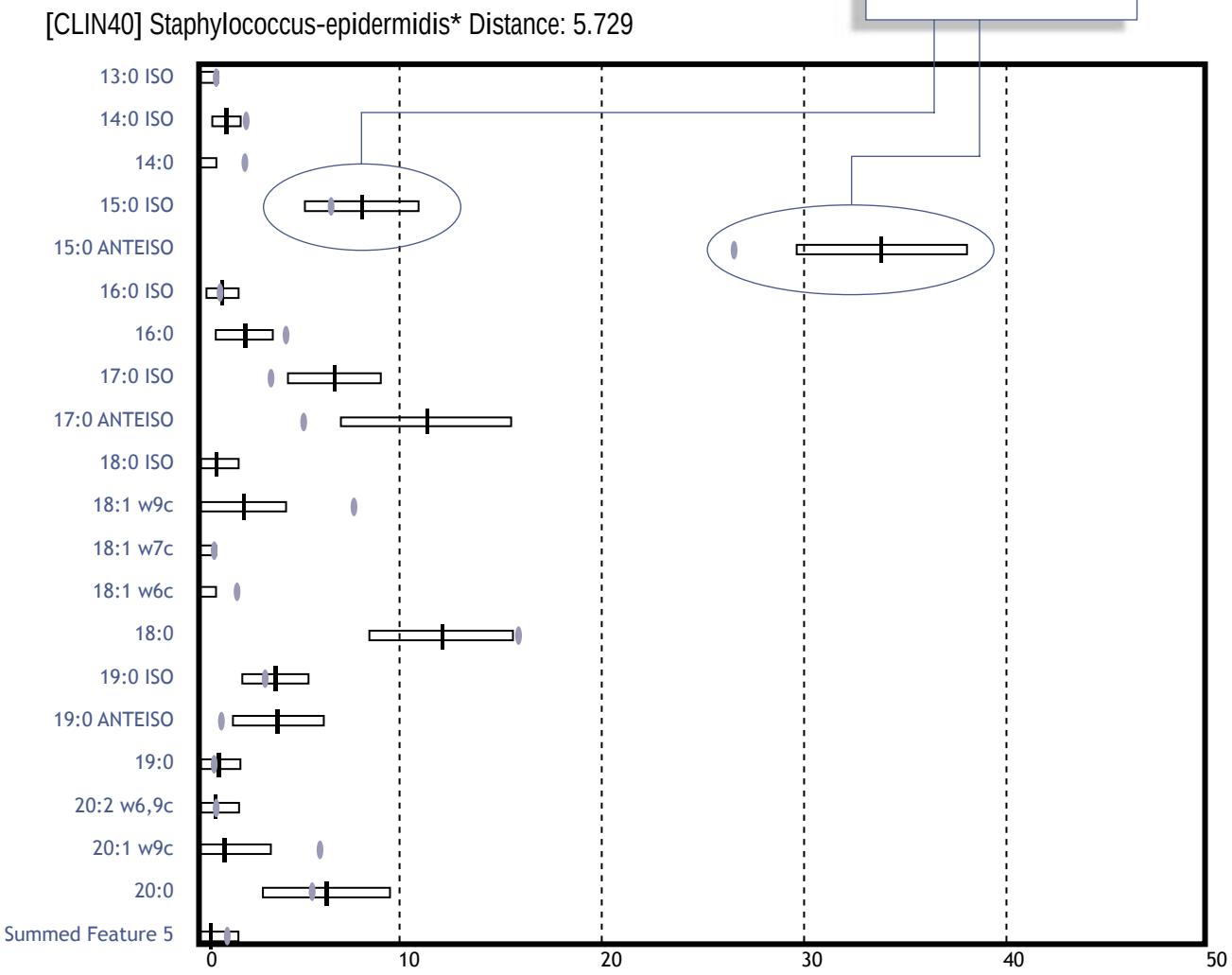
Figure 2: MIS Library Comparison Chart For An Atypical or Low Match

Strain B:

CLIN40 Staphylococcus 0.454

S. epidermidis 0.454

For this low-match sample, the unknown is off one standard deviation for 15:0 ISO and more than two standard deviations for 15:0 ANTEISO



| = mean of the population

⬮ = strain

▭ = variance

Gaussian Distribution

Another way of visualizing the Similarity Index is by looking at the Gaussian Distribution of a population trait (in this case, fatty acid composition). In Figure 3, the perfect mean percentage for all fatty acids in a single species entry (no variance on any fatty acid) is the line at the center. The Similarity Index for a strain that falls on this line is 1.000. As the variance increases, the strain falls farther and farther from the line and the Similarity Index drops. As you can see in Figure 3, a strain with a Similarity Index of 0.600 fall three standard deviations from the mean.

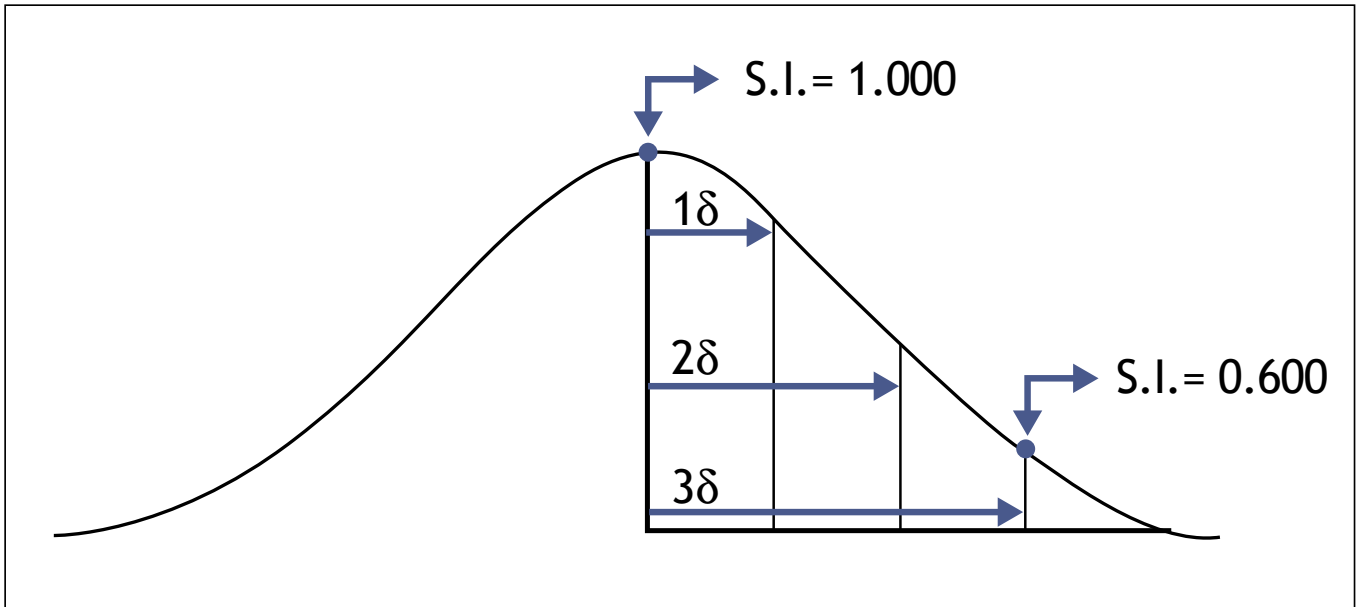


Figure 3: Gaussian Distribution

Interpreting Similarity Index

You should use the following guidelines when interpreting the MIS Similarity Index. Strains with a similarity of 0.500 or higher and with a separation of 0.100 between the first and second choice are considered good library comparisons. If the Similarity Index is between 0.300 and 0.500 and well separated from second choice (> 0.100 separation), it may be acceptable but an atypical strain (falling very low on the normal distribution curve in Figure 3). Values lower than 0.300 suggest that we do not have the species in the database, but indicate most closely related species.