Automated Confidence Scores for DNA Base Calls and their Application to Forensic Sequencing

Max A. Karlovitz, Ph.D.
Daniel H. Wagner, Associates, 40 Lloyd Avenue, Suite 200, Malvern, Pennsylvania 19355

Wagner Associates has developed algorithms (and software) that quantify the accuracy of DNA sequence base calls produced by the current standard gel electrophoretic technology. The algorithms assign a confidence score to each DNA base call by evaluating the quality of the fluorescent trace evidence that supports the call. A confidence score corresponds to the probability that the base call is correct.

Base call confidence scores have important applications to forensic DNA sequencing. Confidence scores may be used to automate trace editing operations. In many forensic laboratories, human editors review trace data in order to identify incorrect base calls and to identify 5’ and 3’ trims that exclude noisy data. We find that 5’ and 3’ trims can be entirely automated based on confidence scores and by encoding knowledge of sequencing and amplification primers. Additionally, confidence scores can be used to flag calls of low quality. Flagging poor quality calls can simultaneously speed up the process of trace editing and ensure that incorrect calls are not missed by a human editor. Confidence scores can also be used to automatically identify poor sequences that are unlikely to contain useful information.

We distinguish primary DNA sequences (as derived from a single gel lane) and consensus DNA sequences. A consensus sequence is derived from multiple primary sequences in which both strands of the original sample are typically represented. Consequently, a consensus sequence will be much more accurate than any of the original primary sequences. Clearly the accuracy of the consensus sequence is highly relevant in forensic sequencing. The primary sequence confidence scores that we have developed play an important role in understanding consensus confidence levels - a consensus call that is supported by three highly confident primary calls is more confident than one supported by five primary calls, each of which has a low confidence. Wagner Associates currently has a NIH grant to research consensus sequence confidence (NIH Phase I SBIR grant 1 R43 HG01848-01). Our work on primary sequence confidence level was supported by two previous NIH grants (DOC grants 50-DKNB-5-00118 and 50-DKNB-6-90158) and is currently supported by a contract with the Armed Forces DNA Identification Laboratory under PO 027321.