The accurate detection of single-nucleotide polymorphisms (SNPs) is crucial for the success of many downstream analyses such as clinical diagnosis, virus identification, genetic mapping, and association studies. Among many others, one valuable approach for SNP detection is based on the base-specific cleavage of single-stranded nucleic acids followed by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) analysis.

Figure 1. Schematic outline of SNP detection using base-specific cleavage and mass spectrometry.

To exploit the full potential of this biotechnology for accurate SNP detection, we formulated two new combinatorial optimization problems. While both problems are aimed at reconstructing the sample sequence that would attain the minimum number of SNPs, they search over different candidate sequence spaces. The first problem limits its search to sequences whose in silico predicted mass spectra have all their signals contained in the measured mass spectra, while the second problem limits its search to sequences whose in silico predicted mass spectra instead contain all the signals of the measured mass spectra. We believe that an efficient computational solution to either problem above could offer a more seamless integration of information in the four complementary base-specific mass spectra than previously done, thereby improving the capability of the underlying biotechnology for sensitive and accurate SNP detection.

For practical use, we developed an efficient and effective algorithm to detect SNPs from mass spectra data. The implemented software is called SNPMS. It works mainly by repeatedly identifying the SNP mutations that have potentially occurred in the sample sequence while progressively updating the reference sequence by correcting those mutations. Comparative evaluation has been carried out on both simulated and real biological datasets, where our experimental results clearly demonstrated the high ability of SNPMS as a tool to accurately detect SNP mutations.

Figure 2. Mass peaks that support two putative SNP mutations. (Left) The putative mutation is a true positive. Its supporting peak has very high relative intensity of 76.07% and signal-to-noise ratio of 302.51. (Right) The putative mutation is a false positive. Its supporting peak has low relative intensity of 13.25% and signal-to-noise ratio of 34.18.

Selected Publications


Dr. Xin Chen is an assistant professor at the Division of Mathematical Sciences. He received his PhD in 2001 from Peking University, China.