Reference Assisted Nucleic Acid Sequence Reconstruction from Mass Spectrometry Data

Gabriel Ilie*, Alex Zelikovsky†, Ion Măndoiu*
*CSE Department, University of Connecticut, 371 Fairfield Way, Storrs, CT 06269
†CS Department, Georgia State University, University Plaza, Atlanta, Georgia 30303
E-mail: {gsi12001,ion}@engr.uconn.edu, alexz@cs.gsu.edu

While tandem MS has long been the main technique used for protein and small molecule identification in proteomics and metabolomics, MS-based protocols for nucleic acid analysis have only gained acceptance in the past decade. Commercially available from Sequenom, assays such as MassCLEAVE start by PCR amplification of one or more regions of interest using primers tagged with two different promoters (T7 and SP6). PCR amplification is followed by four in vitro transcription and RNA cleavage reactions that generate molecules corresponding to fragments ending at each occurrence of specific nucleotides in the original DNA template. Subjecting these fragments to matrix assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS results in four base-specific mass spectra. Depending on instrument precision, peak masses can be matched with one or more fragment base compositions, or compomers. This information can be used for performing a number of nucleic acids analyses ranging from polymorphism discovery and genotyping and microbial identification to DNA methylation analysis and non-invasive prenatal genetic testing. MS-based assays are also becoming increasingly popular in molecular epidemiology due to the very low cost and relatively high throughput (384 reactions in less than one hour for a single MassARRAY system) compared to next-generation sequencing. They are a particularly good fit for studying heterogeneity of viral populations, since virus genomes are small and have even smaller hypervariable regions of common interest. For example, most studies of the Hepatitis C Virus (HCV) focus on a single viral amplicon containing the ≈290bp long Hypervariable region 1 (HVR1). This region is sufficient for estimating several population genetics parameters of interest. Analysis of genetic heterogeneity of hepatitis viruses is useful for tracing the route of transmission and the geographical migration of hepatitis carriers [1] and is essential for outbreak analysis [2] and differentiation between acute and chronic forms of the disease [3].

In this work, we propose a novel algorithm for reference assisted reconstruction of nucleic acid sequences from MS data. Our algorithm has three main stages. In the first stage we identify fragments of the reference sequence that are unambiguously supported by MS data and thus are very likely to be present in the unknown target sequence. In the second stage we use a branch-and-bound approach to fill in remaining gaps and generate a set of candidate sequences consistent with the MS data. Finally, in the third stage we rank candidate sequences based on the total relative error of matches between masses in the experimental MS data and compomers in theoretical spectra, efficiently computed via linear programming. The linear program can additionally take into account the fit between experimental peak intensities and compomer multiplicity in theoretical spectra.

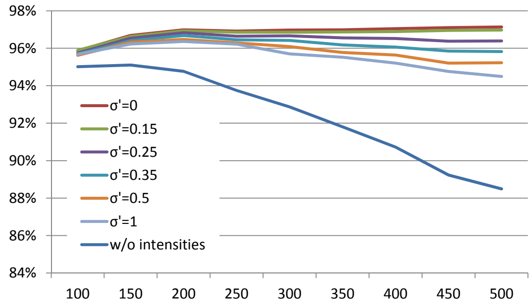

Fig. 1. F-measure (harmonic mean of precision and recall) with and without consideration of peak intensities in experiments with target sequences that differ from the given reference sequence by a single substitution. MS data was simulated with 0-mean normally distributed relative mass (intensity) errors with standard deviation of $\sigma = 0.0001$ (respectively σ' between 0 and 1).

Preliminary experimental results on simulated data show that the true target sequence is almost always ranked highest among generated candidate sequences. In a majority of testcases the true target is the unique candidate with highest rank, resulting in unambiguous reconstructions. Reconstruction accuracy is significantly improved by incorporating peak intensity data in candidate scoring (Fig. 1) even when intensities have high noise levels. In ongoing work we are validating the algorithm on real MS datasets and developing methods for accurate estimation of compomer multiplicity from peak intensity data.

Acknowledgments

This work has been partially supported by Collaborative Research Grant AG110891 from Life Technologies, awards IIS-0916401 and IIS-0916948 from NSF, and Agriculture and Food Research Initiative Competitive Grant no. 201167016-30331 from the USDA National Institute of Food and Agriculture.

References

[1] A. Alexopoulou and S. Dourakis, "Genetic heterogeneity of hepatitis viruses and its clinical significance," *Current drug targets-inflammation & allergy*, vol. 4, no. 1, pp. 47–55, 2005.

[2] P. Patel, A. Larson, A. Castel, and *et al.*, "Hepatitis C virus infections from a contaminated radiopharmaceutical used in myocardial perfusion studies." *JAMA*, vol. 296, no. 16, pp. 2005–2011, 2006.

[3] I. Astrakhantseva, D. Campo, A. Araujo, C.-G. Teo, Y. Khudyakov, and S. Kamili, "Differences in variability of hypervariable region 1 of hepatitis C virus (HCV) between acute and chronic stages of HCV infection." *In Silico Biol.*, vol. 11, no. 5, pp. 163–73, 2012.