Reference Assisted Nucleic Acid Sequence Reconstruction from Mass Spectrometry Data

Gabriel Ilie[1], Alex Zelikovsky[2], and Ion Măndoiu[1]

[1] Computer Science & Engineering Department, University of Connecticut
371 Fairfield Way, Storrs, CT 06269
{gsi12001,ion}@engr.uconn.edu
[2] Computer Science Department, Georgia State University
University Plaza, Atlanta, Georgia 30303
alexz@cs.gsu.edu

Abstract. Mass spectrometry (MS) of target nucleic acid sequences fully digested using base-specific cleavage reactions has emerged in the past decade as a cost-effective method for performing single nucleotide polymorphism discovery and DNA methylation analysis of targeted genomic regions in large numbers of samples. MS based assays are particularly attractive as an alternative to more expensive sequencing approaches for studying heterogeneity of viral populations, typically performed on a short hypervariable region of the viral genome under study. In this abstract we describe a novel algorithm for reference assisted reconstruction of nucleic acid sequences from MS data, and present preliminary experiments assessing reconstruction accuracy as a function of target sequence length and its distance from the reference.

1 Introduction

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 (Fig. 1) 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 [3, 5, 6, 13] and microbial identification [7, 11, 15] to DNA methylation analysis [4, 12, 14] and non-invasive prenatal genetic testing [8, 9]. 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 [10] and differentiation between acute and chronic forms of the disease [2].

In this abstract we describe 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. Preliminary experimental results on simulated data show that the true target sequence is almost always ranking the highest among the generated candidate sequences. In a significant percentage of testcases (that decreases with target length and distance from the reference) the true target is the unique candidate with highest rank, resulting in unambiguous reconstructions.

2 Problem formulation

Let $\Sigma = \{A, C, G, T\}$ denote the DNA alphabet. We define a *compomer* to be the base composition (the multiset of bases) of a fragment obtained from performing any of the four cleavage reactions, and refer to a compomer obtained by cleaving the sequence at base $\alpha \in \Sigma$ as an *α-compomer*. The *compomer spectrum* of a given sequence refers to the multiset of compomers obtained by performing the four base-specific cleavage reactions in silico. We denote by $\mathcal{CS}_\alpha(s)$ the compomer spectrum of a sequence s digested at cut base $\alpha \in \Sigma$, and by $\mathcal{CS}(s) = (\mathcal{CS}_\alpha(s))_{\alpha\in\Sigma}$ the compomer spectra obtained by performing all four cleavage reactions.

We assume that the biological sample consists of copies of a single DNA sequence, referred to as the *target*. The target sequence is typically obtained from a more complex biological sample via PCR. Thus, we will assume that short prefix and suffix sequences of the target (corresponding to PCR primers) are known. The remaining target sequence is unknown but assumed to be within a small edit distance of a known *reference* sequence. We denote by $\mathcal{MS}_\alpha$ the experimental mass spectrum obtained by base-specific cleavage of the target at cut base $\alpha \in \Sigma$, and by $\mathcal{MS} = (\mathcal{MS}_\alpha)_{\alpha\in\Sigma}$ the MS spectra obtained by perform-

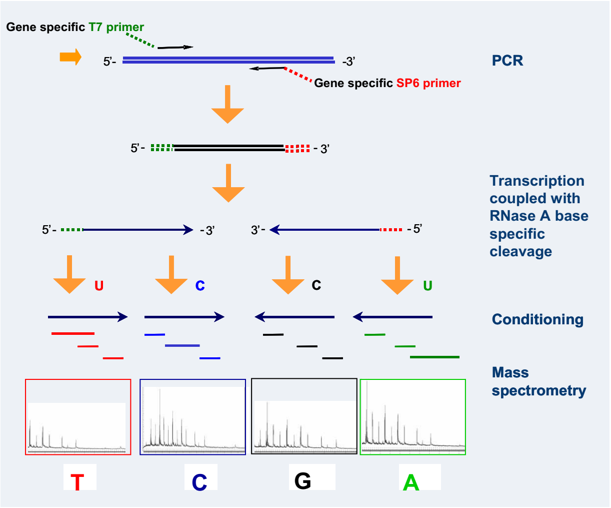

Fig. 1. MassCLEAVE assay for MS-based nucleic acid sequence analysis

ing all four cleavage reactions. Due to limitations of current mass spectrometers some fragments, e.g., cleavage products with mass smaller than some minimum detection threshold m_0, may not be detected by the instrument. Furthermore, detected masses are noisy. We assume that the signed relative errors follow a normal distribution with mean 0 and known standard deviation, e.g., $\sigma = 0.0001$.

In this abstract we further assume that (a) each target compomer with mass above the minimum detection threshold m_0 is detected (no missing peaks), and thus must be explained by a mass in $\mathcal{MS}$, and (b) all masses in $\mathcal{MS}$ represent compomers of the target (no extraneous peaks). If a compomer c with mass $m(c) \geq m_0$ is matched to a mass m in $\mathcal{MS}$, the *relative error* is defined as

$$\eta(c, m) = \left| \frac{m}{m(c)} - 1 \right| \qquad (1)$$

Using minimization of the total relative error as optimization objective, we formulate the reconstruction problem as follows:

Reference assisted sequence reconstruction from MS data
Given: reference sequence r including position of PCR primers, mass spectra $\mathcal{MS}$, instrument parameters m_0 and σ, and maximum edit distance D
Find: Target sequence t flanked by the known PCR primers that is within edit distance D of r and yields a matching of compomers of $\mathcal{CS}(t)$ to masses of $\mathcal{MS}$ with minimum total relative error.

3 The algorithm

A naïve algorithm for solving the reference assisted sequence reconstruction problem would be to generate all sequences within an edit distance of D of the reference and compute the minimum total relative error for matching the compomers of each of these sequences to the masses in $\mathcal{MS}$. As shown in Section 3.3, computing the minimum total relative error over all possible matchings of compomers to experimentally determined masses can be done efficiently by using linear programming. However, the number of sequences within edit distance D of the target grows exponentially with D, so the naïve algorithm becomes impractical for all but very small values of D. Below we present a more scalable algorithm that 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 their total relative error computed via linear programming. The following subsections detail each stage of the algorithm.

3.1 Finding strongly supported regions of the reference

Under the assumption that the signed relative errors are normally distributed with mean 0 and standard deviation σ, by Chebyshev's inequality we get that

$$\mathcal{P}\left(\eta(c,m) \geq \varepsilon\right) = \mathcal{P}\left(\left|\frac{m}{m(c)} - 1\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2} \tag{2}$$

A detectable compomer $c \in \mathcal{CS}_\alpha(r)$ is *strongly matched* to mass $m \in \mathcal{MS}_\alpha$ if

$$\eta(c,m) = \left|\frac{m}{m(c)} - 1\right| < \varepsilon \tag{3}$$

where $\varepsilon = \frac{\sigma}{\sqrt{\tau}}$ is set based on a user specified parameter τ, called *tolerance*, representing the probability upperbound in Chebyshev's inequality (2). A strong match between compomer $c \in \mathcal{CS}_\alpha(r)$ and mass $m \in \mathcal{MS}_\alpha$ is called *unambiguous* if (i) c has multiplicity of 1 in $\mathcal{CS}_\alpha(r)$, (ii) c can be strongly matched in $\mathcal{MS}_\alpha$ only to m, and (iii) m can be strongly matched in $\mathcal{CS}_\alpha(r)$ only to c. The set M_α of unambiguous matches of α-compomers of r can be found efficiently by binary search. Let these matches, indexed in non-decreasing order of their relative errors, be $(c_1, m_1), \ldots, (c_n, m_n)$. We iteratively apply Chebyshev's inequality with tolerance τ to the running means of signed relative errors,

$$X_i = \left(\left(\frac{m_1}{m(c_1)} - 1\right) + \cdots + \left(\frac{m_i}{m(c_i)} - 1\right)\right)/i$$

which are normally distributed with mean 0 and standard deviation $\sigma/\sqrt{i}$. If the Chebyshev's inequality fails for index i, i.e., if

$$|X_i| \geq \frac{\sigma}{\sqrt{i\tau}} \tag{4}$$

then the match (c_i, m_i) is removed from M_α. Finally, a position in the reference sequence is marked as having *strong support* if all detectable compomers overlapping it can be strongly matched and at least one of these matches is in $\cup_{\alpha \in \Sigma} M_\alpha$. Positions within PCR primers are automatically marked as having strong support.

3.2 Generating candidate targets by branch-and-bound

The target sequence is assumed to match the reference at all positions identified in first stage to have strong support from the MS data. In the second stage we use a branch-and-bound approach to fill in remaining gaps one base at a time, in from left to right order. Since most of the time we expect the target sequence to match the reference, we first try using the reference base to fill in the current position. When backtracking from the first choice we try all possible mutations, up to a total of D, first substitutions, then deletion and insertions. For each choice we check for support of newly created detectable compomers, using a Chebyshev test with tolerance τ on the running means of signed relative errors of closest matches, similar to Section 3.1. If the test fails the search is pruned, resulting in a significant speed-up and a reduced number of candidate sequences compared to exhaustively generating all sequences within edit distance D of the reference.

3.3 Scoring candidates by linear programming

For each candidate target sequence t we compute the matching with minimum total relative error via linear programming. Under the assumption that there are no missing or extraneous peaks, in a feasible matching each detectable compomer $c \in \mathcal{CS}_\alpha(t)$ must be matched to exactly one of the masses in $\mathcal{MS}_\alpha$, and each mass $m \in \mathcal{MS}_\alpha$ must be matched to at least one detectable compomer $c \in \mathcal{CS}_\alpha(t)$. For each $c \in \mathcal{CS}_\alpha$ and $m \in \mathcal{MS}_\alpha$, let $x_{c,m}$ be a variable that is set to 1 if compomer c is matched to mass c and to 0 otherwise. Then, the minimum total relative error of a feasible matching is given by $\sum_{\alpha \in \Sigma} z_\alpha$, where z_α is the optimum objective value of the following linear program (integrality of the solution follows from total unimodularity):

Minimize:

$$z_\alpha = \sum_{c \in \mathcal{CS}_\alpha} \sum_{m \in \mathcal{MS}_\alpha} \eta(c,m) x_{c,m}$$

Subject to:

$$\sum_{m \in \mathcal{MS}_\alpha} x_{c,m} = 1, \quad \forall c \in \mathcal{CS}_\alpha$$

$$\sum_{c \in \mathcal{CS}_\alpha} x_{c,m} \geq 1, \quad \forall m \in \mathcal{MS}_\alpha$$

$$0 \leq x_{c,m} \leq 1, \quad \forall c \in \mathcal{CS}_\alpha, m \in \mathcal{MS}_\alpha$$

4 Experimental Results

4.1 Simulation setup

We generated reference sequences uniformly at random, varying the length between 100bp and 500bp with an increment of 50bp. Target sequences were generated by inserting at random one or two mutations. For single mutation experiments we generated ten references for each length, and for each of these references we exhaustively generated all target sequences which are different by one deletion, one substitution, or one insertion. For two mutation experiments we generated 100 references for each length, and for each reference we generated one target sequence by inserting two random mutations. Consistent with current parameters of Sequenom technology, the simulated MS data was generated using a minimum detection threshold m_0 of 1400Da and a standard deviation on relative errors σ of 0.0001. For comparison, we also ran single mutation experiments with error free MS data ($\sigma = 0$). The tolerance parameter τ used by our reconstruction algorithm was set to 0.01 for noisy data and to 0 for error free data.

4.2 Results

Table 1 gives the percentage of testcases for which the target is included among the list of candidate sequences with minimum total relative error, which we refer to as *sensitivity*. For single substitutions and deletions sensitivity is 100% for reconstruction from both error free and noisy MS data generated with $\sigma = 0.0001$. For single insertions, sensitivity is at least 99.97% for any fixed reference sequence length, with an average of 99.99% over all experiments. In experiments with $D = 2$, sensitivity is 98% or higher over testcases generated with any fixed reference sequence length and mutation type, with an overall average of 99.78%.

However, the algorithm's reconstruction is not always unique since there can be multiple candidate sequences with minimum LP score $\sum_{\alpha \in \Sigma} z_\alpha$. Figure 2 shows the percentage of testcases for which the algorithm's reconstruction is

Table 1. Percentage of testcases for which the true target has minimum total relative error among all candidates.

Target length	100	150	200	250	300	350	400	450	500
1 substitution, $\sigma = 0$	100	100	100	100	100	100	100	100	100
1 deletion, $\sigma = 0$	100	100	100	100	100	100	100	100	100
1 insertion, $\sigma = 0$	100	100	99.98	99.98	99.97	99.98	99.99	99.99	99.99
1 substitution, $\sigma = 0.0001$	100	100	100	100	100	100	100	100	100
1 deletion, $\sigma = 0.0001$	100	100	100	100	100	100	100	100	100
1 insertion, $\sigma = 0.0001$	100	100	99.98	99.98	99.97	99.98	99.99	99.99	99.99
2 deletions, $\sigma = 0.0001$	99	100	100	99	99	100	100	100	100
2 insertions, $\sigma = 0.0001$	100	100	99	100	100	100	100	100	100
2 substitutions, $\sigma = 0.0001$	98	100	100	100	100	100	100	100	100

unique. As expected, this percentage is higher for shorter sequences, and gradually decreases with target sequence length for any fixed mutation type and distance threshold D. Targets that are the result of single insertions in the reference sequence have a percentage of unique reconstructions from noisy MS data varying from 75% for sequence lengths of 100bp to 56% for sequence lengths of 500bp. This is significantly lower than the percentage of unique reconstructions for targets obtained by single substitution (90% to 79%), respectively deletions (90% to 75% for the same sequence length range). For single mutation experiments, the percentage of unique reconstructions from noisy data generated with $\sigma = 0.0001$ is only slightly lower than that of reconstructions from error free data, suggesting that the algorithm is robust to levels of noise typical of current technology. On the other hand, reconstruction uniqueness drops significantly for experiments with two mutations.

5 Conclusions

In this abstract we presented a novel algorithm for reference assisted reconstruction of nucleic acid sequences from MS data. The algorithm combines a heuristic preprocessing step that identifies regions of the reference sequence unambiguously supported by MS data with a branch-and-bound strategy to fill in remaining gaps and an LP-based algorithm for selecting candidate target sequences with minimum total relative error. Preliminary experiments on simulated MS data show that our algorithm has very high sensitivity (98% or higher) as measured by the percentage of testcases for which the target is included among the list of candidate sequences with minimum total relative error. Although the percentage of unique reconstructions is high for targets that differ from the reference by single substitutions or deletions (90% to 75%, depending on sequence length), it drops significantly for single insertions and multiple mutation events. In ongoing work we are extending the algorithm to relax the assumptions of no missing/extraneous peaks and to use peak intensities for further reducing reconstruction ambiguity.

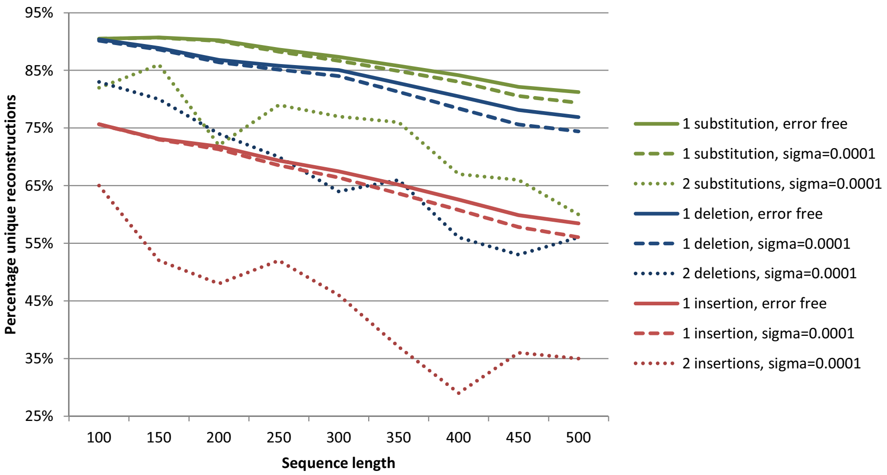

Fig. 2. Percentage of unique reconstructions as a function of target sequence length.

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. Alexopoulou, A., Dourakis, S.: Genetic heterogeneity of hepatitis viruses and its clinical significance. Current drug targets-inflammation & allergy 4(1), 47–55 (2005)
2. Astrakhantseva, I., Campo, D., Araujo, A., Teo, C.G., Khudyakov, Y., Kamili, S.:
3. Böcker, S.: SNP and mutation discovery using base-specific cleavage and MALDI-TOF mass spectrometry. In: Proc. ISMB. pp. 44–53 (2003)
4. van den Boom, D., Ehrich, M.: Mass spectrometric analysis of cytosine methylation by base-specific cleavage and primer extension methods 507 (2008)
5. Ehrich, M., Běcker, S., van den Boom, D.: Multiplexed discovery of sequence polymorphisms using base-specific cleavage and MALDI-TOF MS. Nucleic Acids Res 33(4), e38 (2005)
6. Krebs, S., Medugorac, I., Seichter, D., Förster, M.: RNaseCut: a MALDI mass spectrometry-based method for SNP discovery. Nucleic Acids Res 31(7), e37 (2003)
7. Lefmann, M., Honisch, C., Böcker, S., Storm, N., von Wintzingerode, F., Schlötelburg, C., Moter, A., van den Boom, D., Göbel, U.: Novel mass spectrometry-based tool for genotypic identification of mycobacteria. J Clin Microbiol 42(1), 339–46 (2004)

8. Palomaki, G.E., Deciu, C., Kloza, E.M., Lambert-Messerlian, G.M., Haddow, J.E., Neveux, L.M., Ehrich, M., van den Boom, D., Bombard, A.T., Grody, W.W., Nelson, S.F., Canick, J.A.: DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. Genetics in medicine 14, 296–305 (2012)
9. Palomaki, G.E., Kloza, E.M., Lambert-Messerlian, G.M., Haddow, J.E., Neveux, L.M., Ehrich, M., van den Boom, D., Bombard, A.T., Deciu, C., Grody, W.W., Nelson, S.F., Canick, J.A.: DNA sequencing of maternal plasma to detect Down syndrome: an international clinical validation study. Genetics in medicine 13, 913–20 (2011)
10. Patel, P., Larson, A., Castel, A., *et al.*: Hepatitis C virus infections from a contaminated radiopharmaceutical used in myocardial perfusion studies. JAMA 296(16), 2005–2011 (2006)
11. Sauer, S., Kliem, M.: Mass spectrometry tools for the classification and identification of bacteria. Nature Reviews Microbiology 8(1), 74–82
12. Schatz, P., Dietrich, D., Schuster, M.: Rapid analysis of CpG methylation patterns using RNase T1 cleavage and MALDI-TOF. Nucleic Acids Res. 32(21), e167 (2004)
13. Stanssens, P., Zabeau, M., Meersseman, G., Remes, G., Gansemans, Y., Storm, N., Hartmer, R., Honisch, C., Rodi, C., B Böcker, S., van den Boom, D.: High-throughput MALDI-TOF discovery of genomic sequence polymorphisms. Genome Res 14(1), 126–33 (2004)
14. Tost, J., Schatz, P., Schuster, M., Berlin, K., Gut, I.: Analysis and accurate quantification of CpG methylation by MALDI mass spectrometry. Nucleic Acids Res 31(9), e50 (2003)
15. von Wintzingerode, F., Böcker, S., Schlötelburg, C., Chiu, N.H.L., Storm, N., Jurinke, C., Cantor, C.R., Göbel, U.B., van den Boom, D.: Base-specific fragmentation of amplified 16s rRNA genes analyzed by mass spectrometry: a tool for rapid bacterial identification. Proc. Natl. Acad. Sci. 99(10), 7039–44 (2002)