Accurate detection of SNPs using base-specific cleavage and mass spectrometry

Ruimin Sun* ‡, Xiang Gao† ‡, Nanyu Han†, Qiong Wu*, Yuguang Mu†, Kai Tang† § and Xin Chen* §

*School of Physical and Mathematical Sciences
†School of Biological Sciences
Nanyang Technological University
{rsun1, gaox0005, ha0001yu, qiongwu, ygmu, tkai, chenxin}@ntu.edu.sg
‡Co-first authors, §Co-corresponding authors

***Abstract*—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. In this paper, we present a new SNP detection algorithm, which in particular permits an efficient and effective integration of the information in four complementary base-specific mass spectra. The new algorithm was implemented in a program called SNPMS. Comparative evaluation has been carried out on both simulated and real biological datasets, where experimental results clearly demonstrated the high ability of SNPMS as a tool to accurately detect SNPs.**

***Keywords*-algorithm, single-nucleotide polymorphism, mass spectrometry, base-specific cleavage, and genotyping**

I. INTRODUCTION

The MALDI-TOF mass spectrometry-based approach for SNP detection proceeds with the following typical data-acquisition procedure [1], [2], [3]. Polymerase chain reaction (PCR) is first employed to amplify the target sample DNA sequence, with a T7 promoter tag incorporated to the $5'$ end of either a forward or a reverse primer. The PCR product is then subjected to the shrimp alkaline phosphatase (SAP) treatment, which should degrade the unused dNTP. After the SAP treatment, the PCR product is *in vitro* transcribed with mutant T7 transcriptase to generate single-strand RNA transcripts. Two independent transcription experiments may be performed, in which one uses dCTP instead of rCTP and the other uses dTTP or dUTP instead of rUTP. The resulting transcripts are subsequently digested with the endonuclease RNase A, which cleaves the transcripts after every rC or rU. The substitution of rNTPs by non-cleavable dCTP or dTTP/dUTP nucleotides during the transcription of either forward or reverse strand of the sample DNA sequence enables cleavage reactions specific to each of four bases. The cleavage fragments are finally assayed by MALDI-TOF mass spectrometry to generate four base-specific mass spectra, in which each signal peak corresponds to a single mass value of cleavage fragments.

For a reference DNA sequence, we may generate its theoretical mass spectra by performing *in-silico* base-specific cleavage reactions and mass spectrometry analysis. SNPs are the base substitutions that necessarily account for the discrepancies between the experimentally measured mass spectra of the sample sequence and the *in-silico* predicted mass spectra of the reference sequence. To detect SNPs, visual interpretation of mass spectra is often employed [1], [2], [3], which is very labor-intensive and time-consuming. To facilitate the automatic detection of SNPs from the mass spectrometry data, two software packages have been previously developed. The first one is called RNaseCut [2], which is available at http://www.vetmed.uni-muenchen.de/gen/forschung.html. RNaseCut aims only to compute all the possible mutation candidates that can interpret a differing mass peak, for which manual validation is still needed to confirm true mutations. The second one is the proprietary MassARRAY^{TM} SNP Discovery software package from Sequenom, Inc. It goes a step further, in which a scoring and thresholding procedure is applied to all the mutation candidates. Thus, it provides a fully automatic process for SNP detection. Basically, this software implemented Böcker's algorithm in [4].

In this paper, we present a new algorithm for accurate detection of SNPs from mass spectrometry data. Compared to Böcker's algorithm, it allows for a more effective way to integrate the information in four complementary base-specific mass spectra. As mentioned above, Böcker's algorithm employs a two-step procedure which first generates all the mutation candidates and then scores them. In the first step, the *additional* peaks in the measured mass spectra are examined independently rather than collectively, which consequently produces a large number of spurious mutations as candidates. These spurious mutations will inevitably confound the scoring analysis in the second step, making the true mutations less likely to be detected. In contrast, our algorithm adopts an iterative and progressive procedure. It repeatedly identifies SNPs that have most likely occurred in the sample sequence, while at the same time it progressively updates the reference sequence by correcting these mutations. As a result, the earlier a mutation is detected,

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

the more likely it is true. Moreover, the mutations detected earlier may largely determine the mutations that would be detected later, thereby avoiding many spurious mutations to be evaluated.

II. Preliminaries

Consider a cleavage reaction with respect to the cut base x. If a cleavage fragment f has the base composition of $\mathrm{A}_i\mathrm{C}_j\mathrm{G}_k\mathrm{T}_l$, then we can compute its *in-silico* predicted mass value $m_x(f)$ as the following

$$m_x(f) = i \cdot m(\mathrm{A}) + j \cdot m(\mathrm{C}) + k \cdot m(\mathrm{G}) + l \cdot m(\mathrm{T}) + m_0$$

where $m(\cdot)$ is the mass value of the respective base and m_0 is *an experiment-specific mass intermediate*. For instance, if the endonuclease RNase A is used in the cleavage reaction, then we have $m_0 = 18$ which accounts for an H at the $5'$ terminus and an OH at the $3'$ phosphate.

To detect SNPs, MALDI-TOF mass spectrometry is applied to the products of a cleavage reaction, resulting in a sample spectrum that correlates mass and signal intensity of the cleavage fragments [5]. The sample spectrum is then analyzed to extract a list of signal peaks whose attributes include mass, relative intensity, and signal-to-noise ratio. The above mass spectrometry assay is applied to the cleavage reactions specific to all four bases, resulting in four complementary base-specific mass spectra. Below, we use $\mathcal{M}_\Sigma$ to denote the set of signal peaks from the four complementary mass spectra (after peak calling and mass calibration). The mass value and signal-to-noise ratio of a peak p can be retrieved by using the functions $m(p)$ and $r(p)$, respectively.

We say a cleavage fragment f can *explain* (*interpret* or *yield*) a measured mass peak p (under the same cut base x) if the *in-silico* predicted mass value of f is equal to the measured mass value of p up to a small precision (e.g., $\pm 0.01\%$ for a reflection TOF instrument). Furthermore, we say a reference sequence s can *explain* (or *interpret*) a measured mass peak p if there exists a cleavage fragment in s that can explain p.

Given a reference sequence s and four complementary measured mass spectra $\mathcal{M}_\Sigma$ (experimented for an unknown sample sequence), let $\mathcal{M}_\Sigma(s)$ be the maximum-cardinality subset of $\mathcal{M}_\Sigma$ in which every mass peak can be yielded only by a *unique* cleavage fragment of s. With this subset $\mathcal{M}_\Sigma(s)$, we next define a score that can be used to reflect how well the reference sequence s can *explain* the measured mass spectra $\mathcal{M}_\Sigma$. That is,

$$r(s, \mathcal{M}_\Sigma) = \sum_{p \in \mathcal{M}_\Sigma(s)} r(p),$$

where $r(p)$ is the signal-to-noise ratio value of a measured mass peak p. Note that the higher the score $r(s, \mathcal{M}_\Sigma)$ is, the better the reference sequence s would explain the measured mass spectra $\mathcal{M}_\Sigma$. This score plays an important role in the algorithm presented in the next section.

III. Algorithm

Our iterative greedy algorithm is summarized below in Algorithm 1. It begins with an initialization procedure, in which we first find all the bases in the reference sequence s that are necessary for s to explain some peaks in the mass spectra $\mathcal{M}_\Sigma$. Precisely, they are the bases of those cleavage fragments that yields peaks in the mass spectra subset $\mathcal{M}_\Sigma(s)$, plus the cut bases located at both ends of the cleavage fragments. These bases are then labeled as being the status of *fixed*, simply indicating that they will not be subject to any further mutation. Meanwhile, we update the mass spectra $\mathcal{M}_\Sigma$ by deleting those mass peaks of $\mathcal{M}_\Sigma(s)$ from $\mathcal{M}_\Sigma$, that is, $\mathcal{M}_\Sigma := \mathcal{M}_\Sigma \setminus \mathcal{M}_\Sigma(s)$. This update can be performed because no SNPs are needed in order for the reference sequence s to explain any mass peak of $\mathcal{M}_\Sigma(s)$.

Algorithm 1 SnpMs$(s, \mathcal{M}_\Sigma)$

Input: A reference sequence s and the four complementary mass spectra $\mathcal{M}_\Sigma$ of an unknown sample sequence
Output: A list Δ of potential SNPs that might have taken place in the sample sequence

```
1: Δ ← null
2: Calculate M_Σ(s).
3: Fix bases in s needed to explain peaks of M_Σ(s).
4: M_Σ ← M_Σ \ M_Σ(s).
5: repeat
6:    δ ← null
7:    r(δ) ← 0
8:    for each permissible base substitution δ' do
9:       Calculate r(δ').
10:      if r(δ') > r(δ) then
11:         δ ← δ'
12:         r(δ) ← r(δ')
13:      end if
14:   end for
15:   if δ ≠ null then
16:      Add δ to the set Δ.
17:      Update s by applying δ to it.
18:      Calculate M_Σ(s).
19:      Fix bases in s needed to explain peaks of M_Σ(s).
20:      M_Σ ← M_Σ \ M_Σ(s).
21:   end if
22: until δ == null
23: return Δ
```

An iterative greedy procedure is then invoked. At each iteration, we first identify a *potential* SNP from all the *permissible* base substitutions that could be made to the reference sequence s. Here, a base substitution is permissible if it is applied to a base of s that is not yet labeled as being in the fixed status. For each permissible base substitution δ,

we calculate a score $r(\delta)$ as

$$r(\delta) = r(s', \mathcal{M}_\Sigma) = \sum_{p \in \mathcal{M}_\Sigma(s')} r(p),$$

where s' is the reference sequence s after the base substitution δ is applied to it. As we can see, this score can offer a rough estimate on how much a base substitution could aid in the explanation of the mass peaks in $\mathcal{M}_\Sigma$. Therefore, a reasonable choice of the potential SNP is the base substitution with the highest score $r(\delta)$. Once the potential SNP is chosen, we apply it to s to obtain a new reference sequence (still denoted as s). Then, like we already did in the initialization step, find all the bases in the new reference sequence s that are necessary for s to explain some peaks in $\mathcal{M}_\Sigma(s)$ and label them as being in the status of fixed. Meanwhile, we update the mass spectra $\mathcal{M}_\Sigma$ by deleting those mass peaks of $\mathcal{M}_\Sigma(s)$ from $\mathcal{M}_\Sigma$, that is, $\mathcal{M}_\Sigma := \mathcal{M}_\Sigma \setminus \mathcal{M}_\Sigma(s)$. The above procedure is iterated until no more potential SNP can be found. At that time, the reference sequence s can no longer explain any mass peaks in $\mathcal{M}_\Sigma$ (if it is still not empty), even after a single base substitution is applied to s.

We implemented the above algorithm in a program called SNPMS using the C++ programming language. It is freely available at http://www1.spms.ntu.edu.sg/~chenxin/SnpMs.

IV. RESULTS

As mentioned in the introduction, there are two software for SNP discovery using basic-specific cleavage and mass spectrometry in the literature. The first one is called RNaseCut [2], which can be found at http://www.vetmed.uni-muenchen.de/gen/forschung.html. The second one is the proprietary MassARRAY$^{\text{TM}}$ SNP Discovery software package from Sequenom, Inc. Its algorithmic details were presented in the reference [4]. Unfortunately, we were not able to obtain a copy for our experiments in this study.

A. Simulated data

We carried out several tests on simulated data to assess the effectiveness of our iterative algorithm for SNP detection. In two of these tests, a randomly generated DNA sequence of length 653 bp is used as the reference sequence, followed by either five or ten random SNPs added into the reference sequence to obtain a sample sequence. We further simulate the four base-specific cleavage reactions to generate the *in-silico* predicted mass spectra for the sample sequence. Only the mass peaks that corresponds to cleavage fragments of at least 3 bp are included in the mass spectra because smaller cleavage products are not mass-specific. Both SNPMS and RNaseCut will take the reference sequence and the *in-silico* predicted mass spectra as input for SNP detection. Their detection results are then validated with the true SNPs using the following three performance measures—recall, precision and F-measure. Finally, we generate 100 random data instances as above, and compute the means and variances of the respective performance measures.

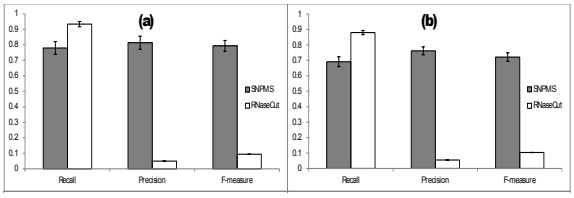

Figure 1. Performance evaluation on simulated datasets. (a) Five random SNPs in a randomly generated sample sequence, and (b) ten random SNPs in a randomly generated sample sequence.

As we can easily see from Figure 1, SNPMS achieved a significantly higher precision rate as well as the F-measure score than RNaseCut at a small cost of the recall rate. The extremely low precision score of RNaseCut should be attributed to the fact that RNaseCut's aim is to "*find possible locations for mutations that are detected by a differing mass peak*" [2] without any attempt to tell which mutations are really true mutations.

B. Biological data

Influenza A H1N1 virus was the most common cause of human influenza during recent years, especially responsible for the flu pandemic in 2009. In our experiments, the influenza A H1N1 viral strain WSN/33 was used and the comparative analysis was mainly focused on the hemagglutinin (HA) gene. The reference sequence that we used was CY009604, taken from NCBI database (http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/multiple.cgi). Due to natural accumulated mutations, it is commonly expected that the WSN HA gene samples kept in the lab would have base differences from the reference sequence in the database.

Hemagglutinin (HA) is an elongated trimeric transmembrane glycoprotein. It plays a central role in both the viral infection process and in the production of the antibodies that neutralize the virus. The HA gene is about 1750 bp in length. In the experiments, we designed four pairs of PCR primers to amplify four (overlapping) fragments from the HA gene sequence and then performed a separate comparative analysis for each fragment. Below we report the experimental results for the fragment which has incurred the largest number of base mutations (among the four amplified fragments).

We performed the base-specific cleavage and MALDI-TOF assay to the sample fragment under examination. The resulting four complimentary base-specific mass spectra were then input to our algorithm SNPMS for automatic mutation detection. The reference sequence is the corresponding DNA sequence segment in gene CY009604 from position 410 to position 920, plus a 26-bp PCR primer added at the 5' end. Finally, SNPMS predicted a total of 18 mutations, and they are summarized in Table I.

Table I
MUTATION DETECTION RESULTS OF SNPMS ON A SAMPLE SEQUENCE FROM THE INFLUENZA A H1N1 VIRAL STRAIN WSN/33

#	mutation	position	peaks (mass / SNR)	remarks
1	C / T	88	1907.30 / 302.51	true positive
2	A / G	462	2363.16 / 151.14	partially true positive
3	G / A	102	2942.15 / 147.00	true positive
4	T / C	292	1633.06 / 79.90 2925.21 / 45.85	true positive
5	C / T	12	1578.22 / 107.78	in T7 promoter region
6	C / A	235	2252.35 / 80.62 1978.06 / 24.11	false positive
7	A / C	147	3287.60 / 97.97	different base change
8	A / G	247	3271.63 / 81.71	partially true positive
9	G / T	197	2002.10 / 76.36	false positive
10	C / A	354	2832.13 / 76.77	partially true positive
11	C / T	142	2910.49 / 34.18	false positive
12	G / T	201	1328.03 / 34.26	false positive
13	T / G	269	1673.05 / 32.72	partially true positive
14	A / G	265	4219.86 / 28.23 2965.14 / 20.36	true positive
15	T / C	4	1601.21 / 30.21	in T7 promoter region
16	A / C	107	1985.22 / 30.25	false positive
17	A / T	77	1689.04 / 26.71	different base change
18	G / A	55	1649.01 / 21.09	false positive

To validate the above prediction, we sent the influenza A H1N1 viral strain WSN/33 sample for direct Sanger sequencing. The direct sequencing revealed ten mutations that have occurred in the sample sequence. In the following, we consider these ten mutations as 'ground truth' to evaluate the predictive performance of SNPMS.

As we can see in Table I, SNPMS was able to correctly detect four of the ten true mutations. They are mutations 1, 3, 4 and 14 (i.e., ranked the 1st, 3rd, 4th and 14th) in the output of SNPMS. All these mutations are supported by strong signal peaks in the measured mass spectra. For example, detection of mutation 1 is due to the mass peak with relative intensity of 76.07% and signal-to-noise ratio of 302.51.

For another four true mutations (i.e., mutations 2, 8, 10 and 13), SNPMS can actually determine their correct base substitutions but only fail to unambiguously localize them. This happened because there exist multiple occurrences of a mutated base in the respective cleavage fragment but no signal peaks exist in the measured mass spectra that allow us to pinpoint which occurrence has actually mutated. Therefore, we indicate such putative mutations as "partially true positives" in Table I.

For the remaining two true mutations, SNPMS can still detect mutations at their (exact or nearby) positions but with different base changes. For example, there is a true mutation A/C at position 77, but SNPMS detected a mutation A/T at the same position. In Table I, we indicate such putative mutations as "different base changes".

Among the eight putative mutations that are considered as false positives, two are located inside the T7 promoter regions. Indeed, most of the signal peaks that were used to support these false positive mutations are quite weak. For example, the detection of mutation 11 is due to the mass peak that has low relative intensity of 13.25% and signal-to-noise ratio of 34.18. Mutation 6 is a noticeable exception.

For comparison, we also ran the program RNaseCut on the same biological dataset above. It reported 1377 potential mutations. Apparently, there are too many false positives to be useful for any downstream analysis.

V. CONCLUSION

In this paper we presented an iterative and progressive algorithm for accurate detection of SNPs using base-specific cleavage and mass spectrometry. It works mainly by repeatedly identifying the SNPs that have potentially occurred in the sample sequence while progressively updating the reference sequence by correcting these mutations. Unlike Böcker's algorithm [4], it allows detection of SNPs in close vicinity without increasing the sequence variation cost. We implemented the proposed algorithm in a program called SNPMS. Comparative evaluation has been carried out on both simulated and real biological datasets, where experimental results demonstrated the high ability of SNPMS to accurately detect SNPs. In particular, it achieved significantly higher precision scores than RNaseCut, the only other publicly available program to date. In future work, we shall further improve the SNP detection performance of SNPMS by optimizing the algorithm and incorporating more effective peak calling and mass calibration.

ACKNOWLEDGMENT

This work was supported in part by a collaborative research award from NTU CoS, the MOE AcRF Tier 1 grant RG78/08, and an NMRC grant (NMRC 1244/2010).

REFERENCES

[1] R. Hartmer, N. Storm, S. Boecker, C. P. Rodi, F. Hillenkamp, C. Jurinke, and D. van den Boom, "RNase T1 mediated base-specific cleavage and MALDI-TOF MS for high-throughput comparative sequence analysis," *Nucleic Acids Research*, vol. 31, no. 9, 2003.

[2] S. Krebs, I. Medugorac, D. Seichter, and M. Förster, "RNaseCut: a MALDI mass spectrometry-based method for SNP discovery," *Nucleic Acids Research*, vol. 31, no. 7, 2003.

[3] P. Stanssens, M. Zabeau, G. Meersseman, G. Remes, Y. Gansemans, N. Storm, R. Hartmer, C. Honisch, C. P. Rodi, S. Böcker, and D. van den Boom, "High-throughput MALDI-TOF discovery of genomic sequence polymorphisms," *Genome Research*, vol. 14, no. 1, pp. 126–133, 2004.

[4] S. Böcker, "SNP and mutation discovery using base-specific cleavage and MALDI-TOF mass spectrometry," *Bioinformatics*, vol. 19 Suppl 1, pp. i44–53, 2003.

[5] K. Tang, P. Oeth, S. Kammerer, M. Denissenko, J. Ekblom, C. Jurinke, D. van den Boom, A. Braun, and C. Cantor, "Mining disease susceptibility genes through SNP analyses and expression profiling using MALDI-TOF mass spectrometry," *Journal of Proteome Research*, vol. 3, no. 2, pp. 218–227, 2004.