Published in final edited form as:

J Mass Spectrom. 2012 April ; 47(4): 490–501. doi:10.1002/jms.2054.

GenoMass software: a tool based on electrospray ionization tandem mass spectrometry for characterization and sequencing of oligonucleotide adducts

Vaneet K Sharma^a, James Glick^a, Qing Liao^b, Chang Shen^b, and Paul Vouros^{a,*}

^aDepartment of Chemistry and Chemical Biology, Barnett Institute, Northeastern University, Boston, MA, 02115, USA

^bShenitech LLC, Acton, MA, 01720, USA

Abstract

The analysis of DNA adducts is of importance in understanding DNA damage, and in the last few years mass spectrometry (MS) has emerged as the most comprehensive and versatile tool for routine characterization of modified oligonucleotides. The structural analysis of modified oligonucleotides, although routinely analyzed using mass spectrometry, is followed by a large amount of data, and a significant challenge is to locate the exact position of the adduct by computational spectral interpretation, which still is a bottleneck. In this report, we present an additional feature of the in-house developed GenoMass software, which determines the exact location of an adduct in modified oligonucleotides by connecting tandem mass spectrometry (MS/MS) to a combinatorial isomer library generated *in silico* for nucleic acids. The performance of this MS/MS approach using GenoMass software was evaluated by MS/MS data interpretation for an unadducted and its corresponding *N*-acetylaminofluorene (AAF) adducted 17-mer (5'OH-CCT ACC CCT TCC TTG TA-3'OH) oligonucleotide. Further computational screening of this AAF adducted 17-mer oligonucleotide (5'OH-CCT ACC CCT TCC TTG TA-3'OH) from a complex oligonucleotide mixture was performed using GenoMass. Finally, GenoMass was also used to identify the positional isomers of the AAF adducted 15-mer oligonucleotide (5'OH-ATGAACCGGAGGCC-3'OH). GenoMass is a simple, fast, data interpretation software that uses an *in silico* constructed library to relate the MS/MS sequencing approach to identify the exact location of adduct on oligonucleotides.

Keywords

oligonucleotide adducts; DNA adducts; tandem mass spectrometry; GenoMass; software

INTRODUCTION

The establishment of liquid chromatography coupled with mass spectrometry (LC-MS/MS) as the method of choice for structural analysis in the fields of proteomics and genomics has been followed by a significant increase in the activity for the development of methodology for the interpretation of the vast amount of MS/MS data acquired. These efforts have

Copyright © 2012 John Wiley & Sons, Ltd.

*Correspondence to: Paul Vouros, Department of Chemistry and Chemical Biology, Barnett Institute, Northeastern University, Boston, MA 02115, USA., p.vouros@neu.edu.

Supporting Information

Supporting information may be found in the online version of this article.

attained a high degree of success in the field of proteomics with the development of several data mining software on the basis of database-driven approaches or the *de novo* sequencing approach. However, in contrast, progress in the field of genomics has been generally slow, and limited success has been achieved in the development of similar software packages.

One of the principal reasons for the relatively slower development in building software in genomics is due to the complexity of MS/MS fragmentation spectra along with inconclusive information about certain fragmentation pathways and the occurrence of various simultaneous and internal fragmentation processes. Despite the limitations in the development of software, there has been considerable progress in manual tandem mass spectrometry (MS/MS) analysis of oligonucleotides since the first report by McLuckey in 1992.^[1,2] In a typical electrospray ionization–liquid chromatography–tandem mass spectrometry (ESI-LC-MS/MS) measurement, an oligonucleotide is isolated according to the *m/z* value of its parent ion(s), and then the selected ions are analyzed using an MS^[2] scan by collision-induced dissociation. Collision-induced dissociation is the most commonly used MS/MS technique to produce sequence determining fragment ions, and typically for DNA, a_n – B_n and w_n -type fragment ions are the dominant ones whereas c_n - and y_n -type ions dominate the RNA spectra^[3,4] (Figure 1). These fragments are also routinely used to locate the exact position of modification in oligonucleotides.

In the last few years, a small number of algorithms for the computational interpretation of MS/MS data of nucleic acids have been reported. The first one was reported in 2002 by Rozenski and McCloskey when they presented an automated sequencing algorithm, simple oligonucleotide sequencer (SOS), a user-interactive local search algorithm for *ab initio* determination of unknown oligonucleotides for 20 bases or less.^[5] This SOS tool is based on building expected results by extending 5' (a–B-ions)- and 3' (w-ions)-end ion series using all the four possible nucleotide masses and searching for a best match in the experimental mass spectra. Modifications in sugar and backbone of oligonucleotides were identified using SOS algorithm, but the results were favorable for 10-mer long oligonucleotides and of increasingly lower effectiveness for oligomers of up to 20 nucleotides. In 2002, Oberacher *et al.*^[6] introduced a global strategy algorithm, comparative sequencing algorithm (COMPAS), for the sequencing verification as well as for the detection and localization of point mutations in 5- to 51-mer oligonucleotides. This proposed COMPAS was not applicable for *de novo* sequencing, and thus later in 2004, the same group presented a global *de novo* sequencing algorithm for 5- to 12-mer long oligonucleotides.^[7] A major limitation in the global *de novo* sequencing algorithm referred earlier is the longer calculation time even for small 9- to 12-mer oligonucleotides, and it was this limitation that was addressed by Oberacher and Pitterl^[8] more recently by incorporating simulated annealing as a stochastic optimization technique. Briefly, the strategy was evaluated for 5- to 24-mer long oligonucleotides and is based on finding among all possibilities the sequence, whose simulated tandem mass spectrum shows the highest degree of similarity to the measured spectrum. However, as none of the aforementioned sequencing algorithms have shown to be directly applicable to the characterization of base-modified oligonucleotides found in the LC-MS analysis of complex mixture of DNA digest, in 2007 our research group developed a “reversed pseudo-combinatorial” approach software package, GenoMass, for fragment identification from LC-MS analysis for oligonucleotide adducts in complex mixture produced from the enzymatic digestion of DNA.^[9] This GenoMass software simplifies the data analysis process and works on a local strategy of building a library of expected results on the basis of search query and looks for a match in raw mass spectral data for small oligonucleotide sequences.^[10]

In this article, we report on a further extension of the GenoMass software, which also incorporates the principal dissociations pathways of polyanionic oligonucleotides to

interpret the MS/MS data for unadducted and base-modified oligonucleotides. Oligonucleotides of varying length (12- to 17-mer) and base composition modified with a carcinogen were initially used, and results for a 17-mer long unadducted and *N*-acetylaminofluorene (AAF)-modified oligonucleotide are reported here.

Subsequently, a further aspect of the software demonstrated here is its use as high-throughput screening algorithm capable of efficient data mining of MS/MS information from a complex mixture to locate in this example, a particular AAF adducted 17-mer oligonucleotide. GenoMass software applicability to determine and characterize the positional isomers of a 15-mer long AAF adducted oligonucleotides is also presented in this article.

The LC-MS/MS method developed to investigate the utility of the GenoMass software is based on an ion pair reversed phase high-performance liquid chromatography (IP-RP-HPLC) via an in-house produced monolithic poly(styrene-divinylbenzene) (PS-DVB) capillary column coupled to electrospray ionization ion trap mass spectrometry (ESI-IT-MS).

EXPERIMENTAL

Chemicals and reagents

Acetonitrile, methanol, and water (all HPLC grade) were obtained from Fisher Scientific (Pittsburgh, PA). Triethylammonium bicarbonate (TEAB), DVB (synthesis grade), styrene (synthesis grade), decanol (synthesis grade), tetrahydrofuran (analytical reagent grade), and azobisisobutyronitrile (synthesis grade) were purchased from Sigma-Aldrich. ($\pm$)-Anti-7r,8t-dihydroxy-9t,10-epoxy-7,8,9,10-tetrahydrobenzo[*a*]pyrene [($\pm$) BPDE] and *N*-acetoxy-*N*-acetylaminofluorene [AAAF] were obtained from the National Cancer Institute Chemical Carcinogen Reference Standard Repository (Midwest Research Institute, Kansas City, MO).

The 17-mer synthetic oligodeoxynucleotides were obtained from the midland certified reagent company, and the 12- and 15-mers were obtained from Sigma life sciences. (–) (1*R*, 2*S*, 3*S*, 4*R*)-1,2-epoxy-3,4-dihydroxy-1,2,3,4-tetrahydrobenzo[*c*] phenanthrene [(–)-(*R,S,S,R*)-BcPh DE-2] adducted 12-mer (TAG TCA AGG GCA) was a generous gift from Dr Donald M. Jerina, National Institutes of Health.^[11] The oligonucleotides were dissolved in water without further purification and were used as stock solutions. The concentration of the stock solution was 1 $\mu\text{g}/\mu\text{l}$. The oligonucleotide adducts were synthesized according to procedure described previously.^[6,8,9]

ESI-LC-MS/MS conditions

An HP1100 series liquid chromatography was used to generate solvent flow at a rate of 0.3 ml/min (Agilent Technologies, Palo Alto, CA). A microTee (Upchurch Scientific, Oak Harbor, WA) and a polyimide-coated fused-silica capillary tubing (360 μm o.d., 75 μm i.d.; Polymicro Technologies, Phoenix, AZ) were used to split the flow rate to 5 $\mu\text{l}/\text{min}$ (360 μm o.d., 25 μm i.d.). A six-port micro-bore valve (VICI, Valco, Houston, TX) equipped with an external sample loop was used for manual sample introduction. The PS-DVB monolithic capillary column, 0.25 $\times$ 100 mm, was used, which was prepared by an *in situ* polymerization procedure.^[12,13] Our group have previously reported that use of 25 mM TEAB as the ion-pairing reagent with PS-DVB monolithic capillary column in the negative ionization mode was found to shift the charge state of oligonucleotide to (–3) with low abundance of the lower or higher charge states.^[8] Modified oligonucleotides were analyzed using a linear gradient consisting of 25 mM TEAB, pH 8.3, as mobile phase A, and methanol as mobile phase B.

ESI-MS was performed on an LCQ Deca quadrupole ion trap mass spectrometer (Finnigan MAT, San Jose, CA) equipped with an electrospray ion source operated in the negative ionization mode. The instrument was tuned in the negative ion mode by infusion at 5 $\mu\text{l}/\text{min}$ of 20 pmol/ μl solution of single-strand 15-mer oligonucleotide 5'-TGTTTTGCCAACTGG-3' in 50:50 (v/v) 25 mM TEAB/methanol using a syringe pump (Harvard Apparatus, Holliston, MA) equipped with a 250- μl glass syringe (Hamilton, Reno, NV). The PS-DVB monolithic capillary column was directly connected to the spray capillary (fused silica, 105 μm o.d., 40 μm i.d., Polymicro Technologies) using a microtight ZDV union (Upchurch Scientific, Oak Harbor, WA). For oligonucleotide adduct analysis, the ESI-MS was operated in negative ion detection mode, an electrospray voltage of 5.5 kV, and a nitrogen sheath gas flow of 15 to 20 arbitrary units (LCQ). The temperature of the heated capillary was set to 210 $^{\circ}\text{C}$. Total ion chromatograms and mass spectra were recorded on a personal computer using Xcalibur software version 1.4 (Thermo Finnigan, San Jose, CA). The LC-MS experiments were conducted for the scan range m/z 700 to 2000, and data-dependent MS/MS spectra were generated at a collision width of 2 Da and relative collision energy of 30%.

LC-MS/MS data processing

The maximum number of possible sequences for n -mer-long oligonucleotides can be calculated according to $N = 4^n$, where n stands for the number of bases. Among these possible permutations, there are $M = (4 + n - 1)!/n!/3!$ unique masses, and each unique mass may have numerous isomers. Moreover, the presence of an adduct bound to the oligonucleotide introduces a further level of complexity to the possible permutations. This vast LC-MS/MS data pool contains useful information from all the constituents of the sample along with irrelevant information. However, as the complexity of data processing increases progressively with the length of oligonucleotide, without automation, the data analysis and interpretation becomes a daunting task.

GenoMass software is based on a “reversed pseudo-combinatorial” approach to identify MS/MS fragments and works backward from the expected results to the MS/MS data. The expected results are in the form of an isomer library generated *in silico* for all the possible combinations of the modifications composed of the four bases (A, T, G, and C), and once generated, GenoMass searches through the experimental MS/MS spectra to identify a match. The search output lists the isomer information, including its sequences, molar mass, charge state, and peak intensity. However, along with useful and irrelevant information, one has to also contend with background noise generated in the LC-MS/MS data. To make the corresponding data more reliable, proper filtering is needed. Thus, GenoMass is programmed to perform the following filtering: background subtraction, noise threshold processing, integration over scan, calibration correction, time interval, integration over mass interval, total ion current threshold processing, and retention windowing (Figure 2).

GenoMass is written in visual basic 6.0 under Windows XP. The installation packaging was performed with Inno Setup 5.2.3 software. The recommended minimum requirements to run this software are 1.0-GHz CPU, 512-KB memory, and 1-GB hard drive, running Windows XP or Windows 2000 and MassLynx 3.5 software. GenoMass software can accept input data files from many MS systems, such as those manufactured by Micromass, Thermo, Applied Biosystem, and Agilent. The data file formats such as Xcaliber (*.raw), ICIS (*.dat), GCQ (*.ms), Magnum (*.ms), ANDI (*.cdf), Automas (*.spa), Masslab2 (*.raw), and Lasermat (*.*) can be converted into the standard MassLynx raw data file format used by GenoMass software by using data converter in XCaliber or other data conversion tools such as DataBridge in MassLynx.

GenoMass software implementation

It has been reported previously that the MS/MS spectra for oligonucleotides are dominated mostly by (a-B)-type and w-type fragment ion series along with low-intensity peaks from other fragment ions or peaks from internal fragmentation. A computer-aided MS/MS data interpretation software should be able to identify most of these peaks and hence determine the type of fragment ions present and, by doing this, ascertain the exact location of adduction on the modified oligonucleotide. The GenoMass software is able to determine not only the characteristic, (a-B)-type, and w-type fragment ion but also the low-intensity peaks from other fragment series, which would otherwise take a lot of time if performed manually. The code for GenoMass software is such that it considers all the oligonucleotides as 5'OH- and 3'OH-end sequence; thus, the *in silico* library, which is generated for expected fragment type, results in y-type or b-type ion series. To generate an *in silico* library for other fragments, such as w-type or (a-B)-type fragments, an adjustment in terms of molar mass has to be made in the adduction box of the software.

For example, a search is performed using GenoMass for the w_n -type ion as follows:

- i. In the gene sequence box on the GenoMass graphical user interface (GUI), the known oligonucleotide sequence is written, and GenoMass software considers all the oligonucleotides as 5'OH- and 3'OH-end sequence.
- ii. To perform search for w_n -type ion, "n" is written in isomer type box on GUI, that is, for w_3 -type fragment, 3 is written in the isomer type box.
- iii. For unadducted w_n -type ion, in the adduction box 79.9 is written corresponding to molecular mass of HPO_3^- (5' HPO_3^- -end, molecular mass 79.9) (Figure 3a). In this case, an *in silico* library will be generated for all possible combinations of unadducted w_n -type ion.

If the search is to be performed for "AAF" adducted w_n -oligonucleotides, in the adduction box on GUI 300.99 is written corresponding to molecular mass of HPO_3^- (5' HPO_3^- -end, molecular mass 79.9) and AAF adduct (molecular mass 221.09) (Figure 3b). The generated *in silico* library incorporate all possible w_n -results AAF adduct.

Similarly, for (a_n-B_n) -type ion, a search is performed as follows:

- i. In the gene sequence box on the GenoMass GUI, the known oligonucleotide sequence is written, GenoMass software considers all the oligonucleotides as 5'OH- and 3'OH-end sequence.
- ii. To perform search for (a_n-B_n) -type ion, $(n - 1)$ is entered in the isomer type(mer) box, that is, for (a_3-B_3) -type fragment, 2 is written in the isomer type box.
- iii. If the search is to be performed for AAF adducted (a_n-B_n) -oligonucleotides, on GUI in the adduction box 382.172 is written corresponding to molecular mass of $(a-b_n)^{\text{th}}$ abasic site (molecular mass 161.082) and molecular mass of AAF adduct (molecular mass 221.09) (Figure 3c). The generated *in silico* library is for all possible (a_n-B_n) -type fragments having AAF adduct.
- iv. If a search is to be conducted for an unadducted (a_n-B_n) -type ion, in the adduction box 161.082 is written corresponding to the $(a-b_n)^{\text{th}}$ abasic site; hence, the resulting library is for all possible (a_n-B_n) -type ion fragments (Figure 3d).

Similar adjustments in terms of molar mass also need to be made in the adduction box of GenoMass software for x-type fragment ion and c-type fragment ion type as depicted in Figures 3e and 3f. As indicated, no adjustment is needed in the software to search for b-type

and y-type fragments ions because the code for GenoMass software considers all the oligonucleotides as 5'OH- and 3'OH-end sequence (Figures 3g and 3h).

An important feature of the GenoMass software is the capability to incorporate base modifications and fragment series ion type while building an *in silico* isomer library as compared with a pregenerated library. The advantage of incorporating these modifications into GenoMass helps in the rapid identification of all the fragments ion series for LC-MS/MS analysis.

The challenge to identifying false-positives

One of the major challenges for any MS/MS-based software is to identify false-positives and false-negatives in high-throughput MS/MS experiments. Their identification becomes even more difficult in a typical LC-MS/MS analysis where LC-MS/MS data are embedded with those acquired in LC-MS. For a complex mixture of isomeric oligonucleotides, which have different adducts and might end up having similar MS/MS peaks at various retention times, this is further compounded by the problem of distinguishing b-series peaks from y-series peaks. To identify these false-positives and hence improve the performance of GenoMass, various strategies were incorporated to identify fragments series:

- i. Peak selection: Retention times are defined, such that a cutoff on the GenoMass is applied to look for results within the defined retention time only.
- ii. Signal threshold: For the selected peak, limits are defined for noise intensity level and signal to baseline ratio to make data more trustworthy.
- iii. Comparison of fragment and parent retention time: GenoMass extracts and plots the resulting fragment ion sequence in Masslynx 3.5, and thus a manual comparison can be made between Masslynx 3.5 result from GenoMass and extracted ion chromatogram from XCalibur software. A common retention time in both removes the ambiguity about the origin of fragment ion in GenoMass software.

By performing these tasks, a reliable computationally derived data set using GenoMass software is envisioned.

RESULTS AND DISCUSSION

To evaluate the effectiveness and efficiency of the GenoMass software for the characterization of oligonucleotide adducts, a series of oligonucleotides base modified with *N*-acetylaminofluorene (AAF) were used. AAF is a potent bladder carcinogen, which does not need metabolic activation and *in vitro* binds primarily to G-base in the oligonucleotide to form primarily *N*-acetyl-*N*-(guan-8-yl)-2-aminofluorene (C8-AAF-dGuo) adduct and thus is a good model to study positional isomers as well an excellent model for other carcinogens that exercise their action mostly through forming a covalent bond to the C-8 position of guanine-like nitroaromatic compounds.^[13,14]

To demonstrate the capabilities of GenoMass software, all the types of fragment ions identified using GenoMass from the LC-MS/MS data for a mixture of unadducted and AAF adducted 17-mer (5'OH-CCT ACC CCT TCC TTG TA-3'OH) are listed (Figure 4). In the process of identifying the fragment ions, the exact adduct location was also determined for the modified 17-mer oligonucleotides using GenoMass software. The process of identifying the MS/MS peaks using GenoMass takes just few minutes as compared with time consuming manual interpretation. Further, GenoMass was applied for the rapid identification of individual modified oligonucleotides in complex oligonucleotide mixtures using ESI-MS/MS data. As a proof of concept, the rapid screening for a 5'OH-CCT ACC

CCT TCC TTGTA-3'OH-modified oligonucleotide in complex oligonucleotide mixture (Table 1) is presented using GenoMass. The program consistently identifies the correct modified oligonucleotide from a mixture by mapping the LC-MS/MS data (Figure 4d). In the second case, the performance of GenoMass software was confirmed by correctly determining the various positional isomers for a 5'OH-ATGAACCGGAGGCC-3'OH modified with AAF (Figure 7). The presence of base modifications in oligonucleotides especially by carcinogens is an unpredictable event such that it does not follow a routine and definite pattern; thus, there is a chance of its presence on every available nucleobase although it has been reported that "G" has a high affinity for AAF. Thus, AAF may bind to all the available "G" nucleobases and hence various singly adducted positional isomers as well as positional isomers, containing two and or more adducts. All of these were correctly identified by GenoMass.

Data interpretation of an unadducted and AAF adducted 17-mer oligonucleotide (5'OH-CCT ACC CCT TCC TTG TA-3'OH)

In accordance with the objectives of the design of the data processing, the performance of the GenoMass software was first evaluated for the computational data interpretation of an unmodified and its corresponding AAF adducted 17-mer oligonucleotide (5'OH-CCT ACC CCT TCC TTGTA-3'OH) using MS/MS spectra.

Using GenoMass, the entire process of data interpretation was performed within a few minutes. This mixture of an unadducted 17-mer and an AAF adducted 17-mer was separated on PS-DVB monolithic column using TEAB as ion pairing agent in mobile phase A, methanol as mobile phase B, and a linear gradient of 1% B per minute. The chromatogram shows two peaks, the first one corresponds to unadducted 17-mer (1677.23, $[M-3H]^{3-}$, $t_r = 17.53$) and the second one corresponds to the AAF adducted 17-mer (1751.13, $[M-3H]^{3-}$, $t_r = 24.57$). The general scheme for finding the various fragments remains the same as already defined earlier in the section on GenoMass software implementation. This is based on the principles of manual spectral identification, such as the search for each and every possible fragment type series but more importantly dominant with w-type or (a-b)-type fragment series. The results of the data interpretation of the unmodified and AAF-modified 17-mer obtained through GenoMass software are described in Figure 4. An example showing how the search was performed for w_5 -type and $(a_{12}-b_{12})$ -type ions is depicted in Figure 5.

To search for w_5 -type fragment ions in the first chromatographic peak (Figure 5a),

1. On the GenoMass GUI "basic settings," the corresponding retention times ($t_r = 17.00$ – 21.00 min) and S/N settings were entered for the chromatographic peak being searched.
2. In the gene sequence box, the oligonucleotide sequence CCT ACC CCT TCC TTG TA was entered.
3. As the search was being conducted for w_5 -fragment ion type in the isomer(mer), Box 5 was written.
4. To search for unadducted w_5 -type fragment, in "adduction box" 79.9 is written corresponding to molar mass of $5' \text{HPO}_4^-$.
5. GenoMass analysis is started by hitting the start button, and all the possible results for w_5 -fragment ions are generated within 2 to 3 s.

The first result is of highly intense $5' \text{P-TTGTA}$ peak ($1572.15 \text{ } m/z$) and accepted correct as compared with the other three results listed by GenoMass because of low ionization at $3'$ -end for thymine nucleobase.

A similar search was also performed by looking for w_5 -type fragment ion in the second chromatographic peak (Figure 5b), which corresponds to the AAF adducted 17-mer ($t_r = 24.00$ – 27.00 min). Accordingly, on GenoMass GUI 300.99 is written in the adduction box corresponding to combined molecular mass of 5' phosphate end plus AAF adduct. The analysis results list high-intensity TTGTA ($1793.24\ m/z$) as its first result. GenoMass also predicts presence of unadducted 5' P-TA ($634.39\ m/z$); thus, it was determined that the w_5 -type fragment has an AAF adduct at G “nucleobase”.

Similarly, a search was carried out for the other half of this w_5 -fragment, that is, the unadducted (a_{12} – b_{12})-type ion fragment,

1. On GenoMass GUI “basic settings” the corresponding retention times ($t_r = 20.00$ – 26.00 min) and S/N settings were entered for the chromatographic peak being searched.
2. In the gene sequence box, the oligonucleotide sequence CCTACC CCT TCC TTG TA was entered.
3. As the search was being conducted for (a_{12} – b_{12})-fragment ion type in the isomer(mer), Box 11 was written.
4. To search for unadducted (a_{12} – b_{12})-fragment ion type, in “adduction box” 161.082 was added corresponding to (a_{12} – b_{12})th abasic site.
5. GenoMass analysis is started by hitting the start button, and all the possible results for (a_{12} – b_{12})-fragment ion type are generated within 2 to 3 s.

The presence of peak (1673.56) confirms the unadducted (a_{12} – b_{12})-type fragment, which was further verified by plotting the results in MassLynx 3.5 (Figure 5c). Further, GenoMass listed the peak at 1292.99 as unadducted (a_5 – b_5)-type ion fragment, thus ruling out the possibility of AAF adduction onto “A” nucleobase.

There are a large number of peaks that are unassigned in an MS/MS spectrum if performed manually, and it is generally believed that a large fraction of these peaks are from other types of fragment ions or more often from internal fragmentation, thus very time consuming to identify. In addition to the characteristic peaks described earlier, which help in correctly identifying the exact location of AAF adduct onto this 17-mer oligonucleotide, an entire spectrum of peaks were identified originating from y -, c -, x -, a -, and b -fragmentation series as well as internal fragmentation series using GenoMass within minutes.

Computational screening of AAF adducted 17-mer oligonucleotide adduct (5'OH-CCT ACC CCT TCC TTG TA-3'OH) from a complex oligonucleotide mixture using GenoMass

It has been proposed that mass spectrometry could potentially play an important role in the future as a rapid screening platform for oligonucleotides markers obtained from circulating cell free DNA or digested cellular DNA.^[15–17] During the past couple of years, there have been significant advance in the comprehensive analyses of genomes^[18] for cancers such as small-cell lung^[19] and melanoma.^[20] The proposed GenoMass software could act as a rapid screening tool for the regular monitoring of the adduction activity in “hot spots” of genes, thus operating as an early warning system. The need is to have a software that can answer a specific question, *where is this oligonucleotide adduct present on the DNA?*

The mapping LC-MS/MS chromatogram of a complex oligonucleotide mixture is shown here as a demonstration of the utility of GenoMass to screen for a particular oligonucleotide adduct. As a first step toward developing such a platform, the GenoMass software was used for the LC-MS/MS data obtained for a synthetic oligonucleotides complex mixture (Table 1). In total, the mixture consisted of more than 17 synthetic oligonucleotides of various

concentrations (20–40 pmol), different adducts (AAAF, BPDE, and BnzPDE), and their corresponding positional isomers. This complex mixture was efficiently separated over a 0.25×100 -mm PS-DVB monolithic column using 25 mM TEAB as mobile phase A, 100% methanol as mobile phase B for a linear gradient 0.6% B per minute (Table 1, *Figure 6*). The resulting LC-MS/MS data collected were processed by GenoMass to screen AAAF adducted 17-mer oligonucleotide (5'-OH-CCT ACC CCT TCC TT TA-3'-OH) using the rules outlined in previous section, and in the process, the location of AAAF adduct was correctly determined.

Identification of positional isomers of AAAF adducted 5'-OH-ATGAACCGGAGGCC-3'-OH

In oligonucleotides containing more than one G, the high reactivity of AAAF may lead to the formation of a multitude of adducts comprised both of positional isomers and a mixture of diverse products containing a combination of more than one adducted and/or nonadducted sites. The 15-mer long 5'-OH-ATGAACCGGAGGCC-3'-OH oligonucleotide, which contains 5 G nucleobases and presents a more challenging system, was selected for the evaluation of the performance of the GenoMass software. These positional isomers gave rise to multiple chromatographic peaks with identical masses and were separated by a linear gradient of 5% to 50% methanol in 30 min over a monolithic column using IP-RPLC as described earlier. In addition to the positional isomers of singly modified oligonucleotides, di- or tri-modified oligonucleotides isomers were formed and separated in the same chromatographic run (*Figure 7*). All the positional isomers formed due to adduct modification have the same backbone sequence, so the goal for the GenoMass software was to design an effective computational approach for efficient mining of MS/MS data so as to identify these different positional isomers and to determine the exact location of an AAAF adduct (*Figure 8*). The GenoMass software accurately identified the three positional isomers for singly modified oligonucleotide, the three positional isomers for di-modified oligonucleotide, and the three positional isomers for tri-modified oligonucleotide (*Figure 7*).

For a singly adducted AAAF 15-mer (1602, $[M-3H]^{3-}$), there were three distinct peaks, with a shoulder peak in the second peak. To search for various fragments and hence locate the exact position of adduct on this 15-mer, strategies similar to those discussed in detail in previous sections were used.

To begin with, a search was conducted for chromatographic peaks of single adducted 15-mer (1602, $[M-3H]^{3-}$) in GenoMass ($R_t = 13.00$ – 16.00 min), and various combinations of w-type fragment ions and (a–b)-type fragments ions were probed as one might do manually, but the entire process was completed rapidly using GenoMass. Finally, it was ascertained that the presence of G*CCC (1434.796, AAAF adducted w_4 -type fragment) and unadducted ATGAACCGGA (1453.426, $(a_{10}-b_{10})$ -type fragment) as well as ATGAACCGGAGG* (1774.634, $(a_{12}-b_{12})$ -type fragment) peaks confirmed that the chromatographic peak at retention time 13.77 corresponded to ATGAACCG-GAGG*CCC positional isomer. Similarly, the presence of ATG*A (1265.649, AAAF adducted (a_4-b_4) -type fragment), ATG*AA (1578.858, AAAF adducted (a_5-b_5) -type fragment), ATG*AAC (1892.067, AAAF adducted (a_6-b_6) -type fragment), CCGGAGGCC (1558.011, unadducted w_{10} -type fragment), and GAGGCC (1092.21, unadducted w_7 -type fragment) confirmed that that the chromatographic peak at retention time 15.26 corresponds to the ATG*AACCGGAGGCC positional isomer. The center peak at retention time 14.22 was found to be ATGAACCGG*AGGCC positional isomer, and it was based on the presence of G*AGGCC peak (1202.71, adducted w_7 -type fragment), GGCC (1543.004, unadducted w_5 -type fragment), and ATGAACCGG*A (1563.926, AAAF adducted $(a_{10}-b_{10})$ -type fragment) as well as absence of ATGAACCGG* (AAAF adducted (a_9-b_9) -type fragment) peaks (*Figure 7*). All of these peaks were verified by plotting them into MassLynx 3.5 and by comparing the retention time with the one obtained from total ion current. The

search using GenoMass was performed to identify more singly adducted positional isomers, but it was found to be challenging as G nucleobases are situated adjacently, although GenoMass did provide individual results pointing to the existence of AAF adduct onto second G and fourth G from 5'-end. However, the software was unable to locate the corresponding unadducted peaks, or on plotting them on MassLynx 3.5, the origin of these peaks were not traced to the parent ion retention time of singly adducted 15-mer; thus, they are not being labeled in Figure 8. Similarly, GenoMass software was used to identify the positional isomers for this 15-mer oligonucleotide, which was di-AAF modified and tri-AAF modified and for various 12-mer, 15-mer, and 17-mer adducted oligonucleotide using similar set of rules as described in this report (data not shown).

CONCLUSIONS

To ascertain the relationship between the “hot mutational” spots and their predilection for carcinogen adducts, an efficient LC-MS/MS-based analytical platform along with fast, reliable software is required to correctly identify the location of adduct onto the oligonucleotides. To realize this goal, progress has been made as reported in this article by developing a combination of GenoMass software tool and LC-MS/MS. By using GenoMass software, the exact location of carcinogenic adduct has been correctly identified for up to 17-mer long oligonucleotides. This approach could be further extended for oligonucleotides >20-mer long by using a high-resolution mass spectrometer having higher dynamic mass range and higher mass accuracy. Another necessary feature required to extend it for oligonucleotides >20-mer is the expansion of computing power needed by GenoMass software. Future improvements of GenoMass software will also incorporate the *de novo* sequencing and full automation. One of the most potential applications of GenoMass could be as a tool to monitor the adduction activity in “hot mutational” spots in the genes, by rapidly screening for presence of the relevant oligonucleotide adducts.

It is absolutely necessary to not only constantly improve the existing software but also develop new software in the field of genomics, which would be connected to the human genome database similar to the field of proteomics. The approach described in this article represents a significant step toward accomplishing such goals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health (grant nos. 1RO1 CA69390 and 1RO1 CA112231). This is Contribution No. 996 from the Barnett Institute.

References

1. McLuckey SA, Van Berker GJ, Glish GL. Tandem mass spectrometry of small, multiply charged oligonucleotides. *J Am Soc Mass Spectrom.* 1992; 3:60–70.
2. Wu J, McLuckey SA. Gas-phase fragmentation of oligonucleotide ions. *Int J Mass Spectrom.* 2004; 237:197–241.
3. McLuckey SA, Habibi-Goudarzi S. Decompositions of multiply charged oligonucleotide anions. *J Am Chem Soc.* 1993; 115:12085–12095.
4. Holz G, Oberacher H, Pitsch S, Stutz A, Huber CG. Analysis of biological and synthetic ribonucleic acids by liquid chromatography–mass spectrometry using monolithic capillary columns. *Anal Chem.* 2005; 77:673–680. [PubMed: 15649070]

5. Rozenski J, McCloskey JA. SOS: a simple interactive program for ab initio oligonucleotide sequencing by mass spectrometry. *J Am Soc Mass Spectrom.* 2002; 13(3):200–3. [PubMed: 11908799]
6. Oberacher H, Wellenzohn B, Huber CG. Comparative sequencing of nucleic acids by liquid chromatography-tandem mass spectrometry. *Anal Chem.* 2002; 74(1):211–8. [PubMed: 11795796]
7. Oberacher H, Mayr BM, Huber CG. Automated de novo sequencing of nucleic acids by liquid chromatography-tandem mass spectrometry. *J Am Soc Mass Spectrom.* 2004; 15(1):32–42. [PubMed: 14698553]
8. Oberacher H, Pitterl F. Tandem mass spectrometric de novo sequencing of oligonucleotides using simulated annealing for stochastic optimization. *Int J Mass Spectrom.* 2004; 304:124–129.
9. Liao Q, Chiu NHL, Shen C, Chen Y, Vouros P. Investigation of Enzymatic Behavior of Benzonase/Alkaline Phosphatase in the Digestion of Oligonucleotides and DNA by ESI-LC/MS. *Anal Chem.* 2007; 79(5):1907–1917. [PubMed: 17261027]
10. Liao Q, Shen C, Vouros P. GenoMass—a computer software for automated identification of oligonucleotide DNA adducts from LC-MS analysis of DNA digests. *J Mass Spectrom.* 2009; 44(4):549–560. [PubMed: 19053159]
11. Harsch A, Sayer JM, Jerina DM, Vouros P. HPLC–MS/MS Identification of Positionally Isomeric Benzo[. c]phenanthrene Diol Epoxide Adducts in Duplex DNA. *Chem Res Toxicol.* 2000; 13(12):1342–1348. [PubMed: 11123977]
12. Xiong W, Glick J, Lin Y, Vouros P. Separation and Sequencing of Isomeric Oligonucleotide Adducts Using Monolithic Columns by Ion-Pair Reversed-Phase Nano-HPLC Coupled to Ion Trap Mass Spectrometry. *Anal Chem.* 2007; 79(14):5312–5321. [PubMed: 17579370]
13. Premstaller A, Oberacher H, Huber CG. High-Performance Liquid Chromatography–Electrospray Ionization Mass Spectrometry of Single- and Double-Stranded Nucleic Acids Using Monolithic Capillary Columns. *Anal Chem.* 2000; 72(18):4386–4393. [PubMed: 11008774]
14. Gao L, Zhang L, Cho BP, Chiarelli MP. Sequence Verification of Oligonucleotides Containing Multiple Arylamine Modifications by Enzymatic Digestion and Liquid Chromatography Mass Spectrometry (LC/MS). *J Am Soc Mass Spectrom.* 2008; 19(8):1147–1155. [PubMed: 18524623]
15. Laken SJ, Jackson PE, Kinzler KW, Vogelstein B, Strickland PT, Groopman JD, Friesen MD. Genotyping by mass spectrometric analysis of short DNA fragments. *Nature Biotechnol.* 1998; 16:1352–1356. [PubMed: 9853618]
16. Qian GS, Kuang SY, He X, Groopman JD, Jackson PE. Sensitivity of Electrospray Ionization Mass Spectrometry Detection of Codon 249 Mutations in the p53 Gene Compared with RFLP. *Cancer Epidemiol Biomarkers Prev.* 2002; 11:1126–1129. [PubMed: 12376521]
17. Sharma VK, Vouros P, Glick J. Mass spectrometric based analysis, characterization and applications of circulating cell free DNA isolated from human body fluids. *Int J Mass Spectrom.* 2011; 304(2–3):172–183. [PubMed: 21765648]
18. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordóñez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, Gormley NA, Grocock RJ, Haudenschild CD, Hims MM, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, Mudie LJ, Ning Z, Royce T, Schulz-Trieglaff OB, Spiridou A, Stebbings LA, Szajkowski L, Teague J, Williamson D, Chin L, Ross MT, Campbell PJ, Bentley DR, Futreal PA, Stratton MR. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature.* 2009; 463:191–196. [PubMed: 20016485]
19. Pleasance ED, Stephens PJ, O’Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordóñez GR, Mudie LJ, Latimer C, Edkins S, Stebbings L, Chen L, Jia M, Leroy C, Marshall J, Menzies A, Butler A, Teague JW, Mangion J, Sun YA, McLaughlin SF, Peckham HE, Tsung EF, Costa GL, Lee CC, Minna JD, Gazdar A, Birney E, Rhodes MD, McKernan KJ, Stratton MR, Futreal PA, Campbell PJ. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature.* 2009; 463:184–190. [PubMed: 20016488]
20. Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, Davis S, Stemke-Hale K, Davies MA, Gershenwald JE, Robinson W, Robinson S, Rosenberg A, Samuels Y. Exome sequencing

identifies GRIN2A as frequently mutated in melanoma. Nat Genet. 2011; 43:442–446. [PubMed: 21499247]

Figure 1.
The principal dissociation pathways of polyanionic oligonucleotide MS/MS fragmentation scheme.

Figure 2. GenoMass software: (a) Graphical User Interface detailed description, (b) basic settings description, and (c) schematic drawing describing the working principle. All values for panels a and b can be set by the user from graphical user interface.

Figure 3.

Adjustments in terms of molar mass to be made in the adduction box of GenoMass software: (a) for an unadducted oligonucleotide, w-type ion; (b) for an AAF base-modified oligonucleotide, w-type ion; (c) for an AAF base-modified oligonucleotide, (a-B)-type ion; (d) for an unadducted oligonucleotide, (a-b)-type ion; (e) for an unadducted oligonucleotide, x-type ion; (f) for an unadducted oligonucleotide, c-type ion; (g) for an unadducted oligonucleotide, b-type ion, no adjustment needed; (h) for an unadducted oligonucleotide, y-type ion, no adjustment needed.

Figure 4.

Data interpretation of MS/MS peaks using GenoMass software. (a) LC-MS separation of an unadducted and an AAF added 17-mer oligonucleotide (5' OH CCT ACC CCT TCC TTG TA 3' OH) using PS-DVB monolithic capillary column, linear gradient 5% to 50%, 50 min; mobile Phase A, 25 mM triethyl-ammonium bicarbonate; mobile phase B, 100% methanol. (b) MS/MS peaks interpretation of 17 mer long oligonucleotide, unadducted $[M-3H]^{3-}$, 1677.33. (c) MS/MS peaks interpretation of 17 mer long oligonucleotide, AAF added $[M-3H]^{3-}$, 1751.07.

Figure 5.

Examples of how the search was performed for w_5 -type and $(a_{12}-b_{12})$ -type ions using GenoMass. The arrows point to the molar mass adjustment done as discussed in text, and the fragment mass is further plotted in MassLynx 3.5 using GenoMass. (a) GenoMass search performed for unadducted w_5 -ion type. (b) GenoMass search performed for AAF adducted w_5 -type ion. (c) GenoMass search performed for $(a_{12}-b_{12})$ -type fragment from AAF adducted oligonucleotide. The origin of false-positive at 23.11 min in panel c is discussed in supplementary information.

Figure 6.

(a) LC-MS separation of complex synthetic oligonucleotide mixture (Table1) using PS-DVB monolithic capillary column, linear gradient 5% to 50%, 80 min; mobile Phase A, 25 mM triethylammonium bicarbonate; mobile phase B, 100% methanol. (b) High throughput screening of 17-mer AAF adducted oligonucleotide (*CCT ACC CCT TCC TTG TA*) using GenoMass, using the basic parameters discussed in the text.

Figure 7.

(a) LC-MS/MS separation of AAF adducted 15-mer oligonucleotide (ATGAACCGGAGGCC) over PS-DVB monolithic column, linear gradient 5% to 50%, 35 min. (b) Positional isomers for singly adducted 15-mer oligonucleotide. (c) Positional isomers for bi-adducted 15-mer oligonucleotide. (d) Positional isomers for tri-adducted 15-mer oligonucleotide. (e) Positional isomers for tetra-adducted 15-mer oligonucleotide. The nucleotide "G" in red implies an adducted position.

Figure 8.

MS/MS spectra of singly AAF adducted 15-mer oligonucleotide (*ATGAACCGGAGGCC*) positional isomers. The nucleotide “G” in red implies an adducted position. The w-type and (a-b)-type fragments identified by the GenoMass software are marked in the corresponding plots.

Table 1

Complex mixture of synthetic oligonucleotides

Retention time (min)	Sequence	Adduct	Measured molecular mass
10.58	ATG ACC GGA GGC CC	–	4587
11.70	CCG CGT CCG CC	<i>N</i> -2-acetylaminofluorene (AAF)	3804.09
15.24	CCC CGA GCA ATC CA AT	–	5099.93
18.68	TG TTT TGC CAA CTG G	–	4574
23.69	TAG TCA AGG GCA ^a	Benzo[c]phenanthrene diol epoxide (BzPhDE)	3971.15
24.12	ATG AAC CGG AGG CCC ^a	Benzo[a]pyrene diol epoxide((±)- <i>anti</i> -BPDE)	4888
32.28	TG TTT TGC CAA CTG G ^a	Benzo[a]pyrene diol epoxide((±)- <i>anti</i> -BPDE)	4795
32.35	CCT ACC CCT TCC TTG TA	–	5031.39
30.31	CCC CGA GCA ATC TCA AT ^a	<i>N</i> -2-acetylaminofluorene (AAF)	5321.02
33.30	CCT ACC CCT TCC TTG TA	<i>N</i> -2-acetylaminofluorene (AAF)	5253
40.65	CCC CGA GCA ATC TCA AT	<i>N</i> -2-acetylaminofluorene (AAF)	5542.11
42.29	CCG CGT CCG CGC ^a	<i>N</i> -2-acetylaminofluorene (AAF)	4025.18
46.80	CCG CGT CCG CGC ^a	<i>N</i> -2-acetylaminofluorene (AAF)	4264.27
60.05	ATG AAC CGG AGG CCC ^a	<i>N</i> -2-acetylaminofluorene (AAF)	5471.36
63.70	TG TTT TGC CAA CTG G ^a	<i>N</i> -2-acetylaminofluorene (AAF)	5458.36
67.42	CCG CGT CCG CGC	<i>N</i> -2-acetylaminofluorene (AAF)	4467.36
68.32	ATG AAC CGG AGG CCC ^a	<i>N</i> -2-acetylaminofluorene (AAF)	5692.45

^aPositional isomers present.