Published in final edited form as:

J Mass Spectrom. 2009 April ; 44(4): 549–560. doi:10.1002/jms.1532.

GenoMass - a computer software for automated identification of oligonucleotide DNA adducts from LC-MS analysis of DNA digests

Qing Liao^a, Chang Shen^b, and Paul Vouros^{c,*}

^a Mass Spectrometry Center, FAS Center for System Biology, Harvard University, Cambridge, MA 02138, USA

^b Shenitech LLC, Woburn, MA 01801, USA

^c Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115, USA

Abstract

In the investigation of oligonucleotides, DNA and their adducts by LC-MS, a myriad of data are generated that make manual data processing quite difficult. This paper describes a ‘reversed pseudo-combinatorial’ approach for fragment identification and the software implementation of this approach. Combinatorial isomer libraries are generated *in silico* to represent the digestion products of oligonucleotides, DNA or DNA adducts of various sizes. The software automatically calculates ion masses of each isomeric segment of the library, searches for them in complicated LC-MS data, lists their intensities and plots extracted ion chromatograms (EIC). This customized new data analysis tool has enabled a study of the enzymatic behavior of a nuclease system in the digestion of normal and adducted DNA, and in the recognition of oligomers containing a carcinogen bound to a nucleobase. The software program potentially can be further expanded to postulate unknown DNA sequences and recognize the adduction sites.

Keywords

DNA enzymatic digestion; oligonucleotide DNA adducts; reversed pseudo-combinatorial data analysis; GenoMass software; ion-pairing LC-MS

Introduction

Advances in the accuracy, sensitivity and efficiency in mass spectrometric technologies and ancillary methods over the past several years have spurred an increased use of the techniques for the characterization of DNA and DNA adducts.[1,2] The general approach toward this goal starts with enzymatic digestion of the DNA followed by LC-MS analysis of the cleaved oligomeric fragments. From the MS data, one may be able to understand the specific enzyme digestion behavior, the oligomeric fragment structure, the binding preference of certain adducts to specific sequences, and quantify targeted species at the trace level.

Our current research is focused on the investigation of the selectivity of adduct formation on DNA using ion-pairing reversed-phase liquid chromatography mass spectrometry (IP-RP-LC-MS).[3–7] As a prerequisite, we first studied the enzymatic behavior of a nonspecific, random endonuclease system, benzonase/alkaline phosphatase.[5,6] It is envisioned that this enzyme

pair may prove to be of value, especially when working with an unknown DNA sequence. The information generated by MS usually needs preprocessing in order to improve the data quality as well as to reduce the data content in order to find the isomeric fragment(s) of interest. A 'reversed pseudo-combinatorial' approach is proposed in this paper for this purpose. The method initially considers all possible isomeric combinations of the DNA bases that might yield oligonucleotides of different lengths and searches the LC-MS data for appropriate matches. Thus, by examining each isomer of the combinatorial combinations against the MS data, one can find the isomers and adducts present in the original undigested oligonucleotide or DNA sequence.

The data analysis in studying model oligonucleotides, DNA and their adducts is quite complicated.[8–13] For a known model oligonucleotide sequence, it is necessary to find the masses of interest and quantify the peak areas. On the other hand, for an unknown DNA sequence, it is necessary to find and quantify all detectable masses of digested fragments. In MassLynx (Waters/Micromass), LCQ Navigator or XCaliber (ThermoFinnigan) and other data analysis software, this is usually done by searching for the mass and generating an extracted ion chromatogram (EIC). However, with the vast number of fragments produced from a typical DNA digest a program is needed to automate the LC-MS data analysis as efficiently as possible. For an unknown DNA sequence that can be digested into fragments of different sizes, the number of fragments will be $N = 4^n$, where n stands for the fragment size, or n -mer. Among those fragments there are $M = (4 + n - 1)!/n!/3!$ unique masses, and for each unique mass there may be numerous positional isomers. Adding another parameter, such as a DNA adduct which contains a carcinogen covalently bound to a site or different sites on the DNA base, introduces a further level of complexity to the data. Without automation, the data analysis and interpretation can be a daunting task. For example, to find a pentamer fragment in the MS data, one needs to compute the pentamer mass, enter the m/z value in a MS tool (such as MassLynx) and look at the output (usually a plot) to inspect whether the pentamer exists. In order to find all the pentamer fragments, this process needs to be repeated $4^5 = 1024$ times, because there are 1024 5-mer combinations of the four bases A, T, G and C. When multiply charged ions as well as adducts are taken into account, the number of combinations will be several times larger. For larger isomers, the possibilities are enormous, and it is impractical if not impossible, to process the data manually.

Present software for mass spectrometric characterization of nucleic acids can fulfill a variety of functions but none that addresses the specific tasks described above. For example, the Mongo Oligo Mass Calculator developed by McCloskey and coworkers is a handy tool for calculating masses of oligonucleotides, electrospray series, fragments obtained by collision-induced dissociation (CID), and fragments of enzymatic digests by endonuclease and exonuclease. [14] The program can also calculate the masses of modified residues. An accompanying tool is Oligo Composition Calculator which finds oligonucleotide composition when the user inputs a mass. Mongo Oligo Mass Calculator is by far the most comprehensive program for the analysis of genetic mass spectrometric data with multiple functionalities. Simple Oligonucleotide Sequencer (SOS) developed by the same group, is an interactive tool for the *de novo* determination of short nucleic acid sequences.[15] The sequence algorithm works by extending from both the 5' (a-B-ions) and 3' (w-ions) ends ion series encoding the complete DNA sequence. Mass ladders are identified by sequentially adding each of the four possible nucleotide masses and searching for the best match of expected ions. The method compares the fragment ion spectra generated by CID of multiply charged oligonucleotide ions to the predicted m/z values employing established fragmentation pathways from a known reference sequence. Sequences of 5- to 12-mer oligonucleotides were successfully determined. GeMS is a user-friendly, advanced commercial software package for gene design developed by Jayaraj and others at KOSAN Biosciences.[16] The software comprises an integrated suite of programs that automatically perform the following tasks in gene design: restriction site prediction, codon

optimization for any expression host, restriction site inclusion and exclusion, separation of long sequences into synthesizable fragments, T_m and stem-loop determinations, optimal oligonucleotide component design and design verification/error checking. The software has been extensively tested in the design and successful synthesis of over 400 kb of genes, many of which exceeded 5 kb in length. Links and MS2Links were developed by Young and coworkers at Sandia National Laboratories as a program to analyze mass spectrometric data generated from native, modified and crosslinked protein and nucleic acid digests.[17] Links calculates the theoretical crosslinking and modification possibilities for single or multiple protein, RNA and/or DNA sequence(s) given information about the crosslinkers and proteases/nucleases used and the expected amino acid or base modifications. Links then returns putative assignments within a defined mass error threshold for a list of input mass (MH^+) values. MS2Links is an analogous program of Links that is capable of assigning MS/MS peak lists generated from the fragmentation of crosslinked, modified or unmodified peptides, proteins and/or nucleic acids. MS2Links calculates the theoretical MS/MS fragment library given information about the identity of the base ion, crosslinkers (if applicable), desired ion types, and amino acid modifications. It then returns assignments within a defined mass error threshold for the list of input mass (MH^+) values. GNU polyxmass is a software framework developed by Filippo Rusconi for mass spectrometric simulations of linear (bio-)polymeric analytes that performs common (bio-)chemical simulations along with simultaneous mass spectrometric calculations for DNA, RNA, saccharides or proteins.[18] Using the program, the user can (1) define brand new polymer chemistries, (2) perform quick mass calculations using a desktop calculator paradigm, (3) graphically edit polymer sequences and perform (bio-) chemical/mass spectrometric simulations.

While of considerable value in several respects, none of these programs can search through a mass chromatogram, deconvolute the spectra, analyze the peaks, and make the intelligent call on the identity and the sequence of the ion peak based on the parent ion and the fragment ions. This type of data interpretation has to be done manually. The GenoMass software tool presented here is designed to solve these problems. It can automatically calculate ion masses of digestion products of DNA, DNA adducts of various sizes, search for them in complex LC-MS data, list their intensity and plot their EICs. It can also generate prospective n-mer fragments in defined enzymatic cutting patterns. In addition, GenoMass provides specialized data processing and reduction algorithms to increase the confidence in the results as well as to increase the computation speed. The utility of the software is illustrated here in the investigation of the enzymatic digestion of oligonucleotides and DNA with benzonase/alkaline phosphatase and in the recognition of oligomeric N-acetylaminofluorene (AAF) adducts formed in the reaction of the carcinogen N-acetoxy-N-acetylaminofluorene (AAAF) with DNA.[6]

GenoMass's main functionalities at the current stage are automation of the data analysis tasks such as ion extraction/plotting and streamlining the complex data. It is particularly handy in finding the adducted oligonucleotide fragments after enzymatic digestion and predicting the mass ions of digestion products of unmodified and modified oligo/DNA fragments. All the functions are performed on an intuitive, user-friendly program. More functions can be added to the ones already in place, according to the needs of the researchers. This 'reversed pseudo-combinatorial' approach has not been used in any of the aforementioned software.

Methods

Experimental conditions

Sample preparation

(1) Stock Solution Preparation: The tubes containing x nmol (x is the actual yield of each standard which varies), ss and ds oligonucleotide standards were centrifuged prior to opening.

Then x μ l DI water was added to the NarI 21-mer, and 80-mer oligonucleotides respectively to create stock solutions of concentrated 1 nmol/ μ l for each sample. The NarI 21-mer was resuspended with x μ l RNase-Free Duplex Buffer to make a 100 μ M solution. These mixtures were all heated to 94 °C for 2 min, removed from heat and allowed to cool slowly to room temperature. This resulted in a stable, double-stranded form of oligomers, and they were stored at –20 °C. Next, 1 mg calf thymus DNA (ctDNA) was weighed out, and 1 ml 10 mM citrate buffer (pH = 6) was added to a concentration of 0.1 nmol/ μ l. All samples were then vortexed down.

(2) AAAF Modification: For this, 35.5 μ l of 1 nmol/ μ l 4-mer oligonucleotide standards, NarI 21-mer, and ss 80-mer oligonucleotides were centrifuged down. Then, 355 μ g ctDNA was added to 355 μ l of 10 mM citrate buffer (pH 6) to a concentration of 0.1 nmol/ μ l DNA. The above oligonucleotide solutions and DNA were mixed with the AAAF stock solution of 3.55 nmol/ μ l in acetonitrile in 1 : 1 molar equivalence. The reaction was allowed to progress at room temperature. The mixtures were then incubated on a heat block at 37 °C for 18 h.

(3) Enzymatic Digestion: The modified DNA was extracted 3 times with 1 ml ethyl acetate to remove the unreacted AAAF. The aqueous fractions were then dried down using a Speedvac from Savant (Pleasanton, CA) for 2 h at medium heat. The modified oligomers and DNA were then reconstituted in 200 μ l of 5 mM Tris/10 mM MgCl₂ buffer (pH 8). After being vortexed down, 5 μ l benzonase and 5 μ l alkaline phosphatase were added to the oligomers and DNA in buffer solution. The mixtures were vortexed well and incubated for 24 h at 37 °C.

(4) Solid Phase Extraction: The digest mixture then went through solid phase extraction (SPE) to concentrate the DNA adducts and remove the unadducted DNA fragments. SPE was performed by using a vacuum manifold fitted with disposable Teflon needles to prevent sample cross-contamination. Unmodified oligomers were washed out while modified oligomers were eluted. The optimized SPE conditions were as follows. (1) Conditioning with 1 ml of methanol. (2) Conditioning with 1 ml water. (3) Load the samples. (4) Wash with 1 ml 5% (v/v) methanol/water. (5) Elute with 75% (v/v) methanol/water. The adducts-containing fraction was dried down and reconstituted in 100 μ l water for analysis.

Chromatographic conditions—The LC system employed was an Agilent Technologies 1090 HPLC with an autosampler (Wilmington, DE). The HPLC system was controlled by HP ChemStation software running on an IBM-compatible PC. The reverse phase RP-1 PSDVB (polystyrene divinylbenzene) column (2.1 mm $\times$ 100 cm $\times$ 3.5 μ m) was purchased from Hamilton Chromatography (Reno, NV). The column was kept at ambient temperature (25 °C) during analysis. The mobile phase consisted of 25 mM tetraethylammonium bicarbonate (TEAB) as A, and 25 mM TEAB with 20% acetonitrile as B. The gradient went from 0%B to 100%B in 20 min and was held at 100%B for 5 min. The flowrate was 0.2 ml/min and the injection volume was 20 μ l.

Mass spectrometric conditions—MS data were acquired on a ThermoFinnigan LCQ Classic mass spectrometer (San Jose, CA). The mass spectrometer was tuned on a series of standard oligonucleotides. The samples were analyzed by negative ion electrospray mass spectrometry, first in full scan mode, and subsequently by MS/MS. The MS conditions were as follows. Spray voltage: 3.8 kilovolts; capillary voltage: –4 volts; capillary temperature: 200 °C; sheath gas: 25 arbitrary units; tube lens voltage: –25 volts. Data were acquired and processed with Navigator 1.2 software. Mass spectra were recorded in continuum mode.

Oligonucleotides, oligonucleotide-adducts and DNA adducts were first digested to suitable lengths by benzonase/alkaline phosphatase. The digestion products were then subjected to ion-pairing LC-MS and LC-MS/MS analysis. The raw data files were converted to a standard

MassLynx raw data file format upon which GenoMass will operate. The data were then processed with several filtering algorithms to reduce the interference and noise as well as to reduce the data. Subsequently, a 'reversed pseudo-combinatorial' approach was applied to analyze the data, identify the isomers, and extract the information of interest.

In the study of ctDNA, the data obtained with the LCQ MS did not show adequate detection on the modification products, which indicated the LC-MS condition was not sensitive enough to be applied to an *in vivo* system. Later, we improved the ion-pairing reversed-phase LC-MS method by switching to nanoLC- μ ESI with a monolithic column, albeit the solvent composition and ion-pairing reagent were kept the same. The nanoLC system is an Eksigent ExpressLC-100. The LC system was controlled by the MassLynx 4.1 software. The monolithic column (75 μ m i.d., 360 μ m o.d., 15 cm) was made inhouse with a PSDVB stationary phase. The flow rate was 200 nl/min. The gradient initially stayed at 98%A, 25 mM TEAB for 3 min, then went from 98%A to 98%B, 25 mM TEAB with 20% acetonitrile in 40 min. Then the gradient was held at 98%B for 20 min. The MS was conducted on a Waters Q-ToF micro system at the following conditions: sample cone voltage at -55 volts, capillary voltage at 1500 volts, extraction cone voltage at 4.5 volts, cone gas at 11 l/hr, desolvation gas at 15 l/hr, and source temperature at 100 °C.

Data preprocessing

The raw data from mass spectrometry may contain much irrelevant information as well as background noise. Proper filtering was employed to make the data more reliable and compact in size. The data preprocessing implemented in the GenoMass software includes background subtraction, calibration correction, noise threshold processing, integration over scan time interval, integration over mass interval, total ion current (TIC) threshold processing, and retention windowing. The following section highlights some of these techniques.

Initially, data from different applications software are converted into standard MassLynx raw data file format by using Data Converter in XCaliber or other data conversion tools such as DataBridge in MassLynx. The original data could come from any common MS systems, such as those manufactured by Micromass, Thermo, Applied Biosystem, Agilent, etc. Specifically, the following data file formats can be accommodated by the GenoMass program: Xcaliber (*.raw), ICIS (*.dat), GCQ (*.ms), Magnum (*.ms), ANDI (*.cdf), Automas (*.spa), Masslab2 (*.raw), Lasermat (*.*)).

Subsequently, a retention window is applied to the raw data so as to reduce the processing time and remove irrelevant information. The start time and stop time of the retention window are normally known and can be configured through the user interface.

The next step is to apply a noise filter on the data which are within the retention window. The filter first computes the baseline of the intensity data by a moving-average operator with a 2D triangle weighting-window. It then calculates the signal-to-baseline ratio of each data point and compares them with a preset threshold.

Two criteria must be satisfied in order to declare a raw data point as a valid mass data. One is that the intensity of the raw data must be over a preset S/N level. Another is that the signal-to-baseline ratio of the raw data must be above a preset ratio threshold. Both noise level and ratio thresholds can be set by the user from the graphic user interface. Raw data not satisfying the above criteria are considered to be noise and are dropped immediately.

The qualified data are used to build a mass list on which each mass (m/z) is unique in value and has a peak intensity associated with it. Since the raw MS acquired data are in terms of m/z values, the data on the mass list are also displayed accordingly. The mass list contains all the

expected isomer combinations. GenoMass uses COM objects (.dll) to connect to Micromass MassLynx software so that all the MassLynx graphic functionalities are available for the user. This means that a user who is familiar with MassLynx tool does not need to spend any time to learn a new graphic tool. Raw data generated by different machines can be converted to the format that can be processed in GenoMass and MassLynx.

The 'reversed pseudo-combinatorial' approach

In oligonucleotide and DNA analysis by LC-ESI-MS, negative ion detection is normally used due to the phosphate presence in the compounds. For LC-MS analyses employing the most common MS systems of relatively low mass range and/or moderate resolution (e.g. quadrupoles, ion traps), the size of the fragments which can be comfortably characterized is limited to somewhere around 10-mers–15-mers.[18,19] In the digestion of oligonucleotides or DNA by benzonase/alkaline phosphatase, the fragments observed are typically of short length occurring in the range of 2-mers–12-mers. According to our LC-MS data, at least for this enzyme system, the occurrence of fragments larger than 10-mers–12-mers is rare and their abundances are near or below noise level.

On the basis of the above criteria, this investigation focused on the consideration of MS data representative of 2-mers–12-mers fragments for the purpose of their detection, identification and quantification. In order to achieve those objectives, a 'reversed pseudo-combinatorial' approach is proposed here. This approach works from the expected results back to the MS data (thus, the term 'reversed'). The expected results are the isomers belonging to an isomer library which is comprised of all the 2-mer–12-mer isomers, i.e. all the combinations of the four bases A, T, G and C with segment lengths from 2 to 12 bases (hence our use of the term 'pseudo-combinatorial'). For each isomer in the library, the program searches through the mass list obtained in the previous step to see if there is a mass value identical to the computed isomer mass. If a match is found, the presence of the isomer in the original oligonucleotide or DNA sequence is claimed. The isomer information, including its name, structure, mass and peak intensity (ion abundance), are registered on the result list. This procedure is repeated until all the isomers in the isomer library have been processed.

For small oligomers, the isomer library is generated in the software on the fly, while for relatively larger isomers (>8-mer), the library is pregenerated and saved in a database in order to speed up the computation. Figure 1 demonstrates an example of the generation and display of a tetramer library.

The above considerations are applicable for an unknown oligonucleotide or DNA sequence. For a known oligonucleotide sequence, however, the library is generated differently to improve the computing efficiency. The sequence is first cut into n -mers orderly or randomly from 5'-end to 3'-end, and then from 3'-end to 5'-end. The cut segments are sorted according to their masses.

In certain cases, the user may be interested in a specific group of isomers. For example, AAAF adduction occurs primarily on the G-base. In this case, the search may be limited to identifying fragments which have at least one G-base. To serve this purpose, a sublibrary is generated on the fly by simply rescaling the complete library according to the user's specification. To find fragments which contain a covalently bound adduct, the isomer library is modified by adding the mass of the adduct to each isomer mass of the library. Similarly, to find doubly charged fragments or even adducted doubly charged fragments, the isomer library is modified accordingly. The modified library is then used to search for the isomer presence in the MS data. Due to the convenient structure of the 'reversed pseudo-combinatorial' approach, it is relatively easy to adapt the software to those application variations.

The input and basic interface settings are self-explanatory. As shown in Fig. 2, first, the user loads the raw data file from the main window by clicking the button right next to the Raw Data File box. If the background needs to be subtracted, load the background data file as well. Click the Subtract button to conduct the background subtraction process. Then, if it is known, enter the gene sequence from which the raw data was generated. For an unknown gene sequence, just leave the Gene Sequence box blank. Then, enter the criteria to describe what fragments should be searched in the data. For example, one may want to see if there are any 4-mer fragments in the data which contains G- with AAF-adduct. One just enters '4' in the Isomer-Type box, G, in Isomer-Containing box, and 221.09 in the Adduct box.

Some parameters related to data processing should be set in the Basic Settings window as well. When those are done, a click of the Start button initiates the data analysis process. The output box lists all the found fragments and the corresponding peak intensity values. The results may be sorted according to mass, intensity or category by checking the options below the output box. They can be exported to Excel program for further analysis by using the Save utility under File menu. Since the GenoMass software is connected to MassLynx by COM objects, one can easily view the graphic presentations of the EIC and spectra of those found isomers by simply clicking the Plot button of the interface.

Software implementation

The software utilizes client-server architecture for implementing the 'reversed pseudo-combinatorial' approach. As shown in Fig. 3, the system consists of a client, a search engine server, and an isomer library server. The client first conducts the data preprocessing and data reduction on the raw MS data. This generates a mass list consisting of a collection of unique mass-intensity pairs. The search engine uses this list as its database. Upon receiving a query from the client, it searches the database to see if there is a match within a certain mass range and returns the found mass-intensity pair accordingly. The default range is ± 0.1 (m/z) and can be changed from the Basic Settings window under item 'Integration Mass Interval' (Fig. 3).

The isomer library server generates an isomer mass table as described previously and returns it to the client whenever requested. The client examines each isomer in the mass table and queries the search engine to look for a match in the MS data. When a match is found, the corresponding isomer and its related information are registered on the output list.

GenoMass software was written in Visual Basic 6.0 under Windows XP. The installation packaging was done with Inno Setup 5.2.3 software. The recommended minimum requirements to run this software are 1.0GHz CPU, 512KB memory and 1GB hard drive, running Windows XP or Windows 2000.

Results and Discussion

As an example of the utility of the GenoMass software, we present, in this section, results from the processing of LC-MS data of enzymatically digested model DNA systems. An enzyme cocktail consisting of benzonase/alkaline phosphatase which is known to cleave DNA in a nonspecific manner was used in all these studies. Since the ultimate objective of our work is the application of the GenoMass software to the study of carcinogen-modified oligonucleotides, the 12-mer NarI gene sequence which is a known mutational target for the carcinogen AAF was incorporated in the model systems. Initially, the software was tested with the enzymatic digests of two unmodified oligonucleotides, a double-stranded 21-mer and a single-stranded 80-mer. Following confirmation of the feasibility of the general approach, both the 80-mer and ctDNA, whose structure is unknown, were reacted with AAF and the software was used to search the LC-MS profiles of the respective digests for oligonucleotide

adducts generated by enzymatic cleavage. The model sequences tested are shown in Table 1 with the NarI gene sequence segment highlighted in blue and the hotspot in red.

Benzonase/alkaline phosphatase digests of unmodified oligonucleotides

In earlier studies of oligonucleotide digests by benzonase/alkaline phosphatase it was reported that, in the generation of 4-mer fragments from short DNA segments, the enzyme pair proceeded by cleaving four bases at a time from either the 3'-end or the 5'-end.[6] For the purposes of this study, we define this pattern as 'orderly cut'. On the other hand, 'random cut' is defined as the process whereby the enzyme pair cuts anywhere in the sequence and generates *n*-mer fragments, some of which may be isomeric to those produced from an orderly cut, or consist of entirely different base composition (referred to here as 'random cut only' fragments). In the case of the benzonase/alkaline phosphatase digestion of the NarI 21-mer Duplex, which is an asymmetric sequence

5'-ATCCCAACACCGGCGCCACAT-3'

3'-TAGGGTTGTGGCCGCGGTGTA-5,

orderly cut and random cut fragments were found in the experimental data, from both the upper strand and the bottom strand, and their abundances were compared. GenoMass computed the library of fragments from all possible cuts, as shown in Fig. 4. This information was then used to search the LC-MS data for the presence of such fragments in the digested sample.

In this case, if cut orderly from either the 3'-end or 5'-end, ten 4-mer fragments with different base sequences were generated. However, two sets of fragments were isomers: CAAC and CACA are both of *m/z* 1141.8; CCGG and GCGC are both of *m/z* 1173.8. If cut randomly only, meaning starting anywhere except from the beginnings of the sequences, then eight 4-mer fragments were generated. There were also two sets of isomers: CCAC, CACC, CCCA are all of *m/z* 1117.8; ACAC and CCAA are both of *m/z* 1141.8. Note the display in Fig. 4 of the *m/z* or the number of occurrences for the fragment(s) of a given *m/z* value (isomer counts) derived from either orderly cut or random cut. For example, a frequency number, '2', is given for the isomer counts of *m/z* 1141.8, when considered in terms of an orderly cut. This means that there were two fragments of such *m/z* value among all the orderly cut fragments. Similarly, we found seven fragments of this *m/z* among all the randomly cut results. The exclamation mark '!' indicates that this *m/z* appeared in both orderly and randomly cut results.

The oligo 80-mer presented an opportunity to test the effectiveness of the software and the action of the enzyme in a more realistic and challenging setting than that of the shorter 21-mer. The GenoMass searched the LC-MS profile for all possible 2-mers–12-mers that could be produced from this oligonucleotide of known sequence and listed their intensities (data not shown). The program then found and compared all the possible ion masses of significant abundance up to 10-mers and determined that the enzyme system cut the 80-mer predominantly into 3-mers and 4-mers, and progressively to a lesser extent to the larger 5-mers, 7-mers, etc. In addition, the GenoMass search revealed that there was a higher abundance of orderly cut than random cut 4-mer fragments.[6] As an example, Fig. 5 summarizes the output generated by the GenoMass software in the search for 5-mer fragments, and the reconstructed ion chromatogram plots of the six most abundant 5-mers in the digested sample.

Recognition of AAF-oligonucleotides in modified DNA

Of paramount importance in our work is the development of the capacity to recognize oligonucleotide as opposed to monomeric DNA adducts in order to develop an understanding of the sequence dependence in the reaction of carcinogens with DNA bases. Accordingly, the GenoMass software was used to analyze the LC-MS data of the benzonase/alkaline phosphatase digest of both the AAF-modified 80-mer and ctDNA. The enzymatic digestion generated

complex LC-MS data marked by the presence of a broad range of concentrations of AAF-modified fragments produced from all possible combinations of the four nucleobases A, T, G, C. In this specific case, since the favored target of AAAF is the guanine base, the software was directed toward the search of oligonucleotide adduct masses containing at least one G-unit. For example, a search for G-containing AAF 5-mer fragments from the 80-mer and the ctDNA was conducted and their EICs were selected and plotted with a click of the mouse as shown in Figs 6 and 7, respectively.

In comparing the results from the digestion of the adducted 80-mer and ctDNA, it was interesting to find that five out of the six most abundant AAF-5-mer masses (m/z) were common to both systems. The data are summarized in Table 2 and partially denoted in the chromatograms of Figs 6 and 7. While presently the search generated information strictly on the base composition of the adducts, it should be noted that full sequence can be obtained by MS/MS in a subsequent LC-MS analysis.

Ultimately, the ability to relate sequence selectivity of adduct formation in specific domains of DNA or of short oligomers produced in the partial digestion of the DNA, would depend on the ability of the software to reconstruct chromatograms of longer fragments. However, when the length of the fragments increases, so does the complexity of the isomer library. Moreover, due to the nature of ESI-MS, larger fragments tend to produce a significantly higher abundance of doubly charged, as opposed to singly charged ions. As a consequence, the database is much larger and requires far more computing power. The utility of the GenoMass software to screen the data for longer sequences is demonstrated in the search for 9-mers as doubly charged ions and is illustrated in the chromatogram of Fig. 8 which shows the abundant presence of the 9-mers in the digest of ctDNA reacted with AAAF.

Conclusions

The ‘reversed pseudo-combinatorial’ approach-based GenoMass software is a powerful computer program that can simplify and automate the data analysis process in genomic research. It can be utilized towards the investigation of enzyme behavior and is also directly applicable towards the streamlining of the complex LC-MS data generated in studies associated with the characterization of DNA adducts.

Upon consideration of the length of fragment(s) cut by a given enzyme, one enters the relevant isomer and mass composition into a dialog box, defines the threshold intensity for the selected ion masses, and sorts the LC-MS data in a variety of ways. For a known oligomer or DNA sequence, with defined enzymatic cutting pattern, GenoMass can automatically generate corresponding masses of all digestion products. It then searches through the LC-MS raw data file for the masses, outputs their m/z values and their intensities and plots EICs of these found masses. For an unknown oligomer or DNA sequence, GenoMass can automatically search for all masses of possible enzymatic digestion products, output their m/z and intensity, and can plot their EICs. When searching for oligonucleotide adducts, the software incorporates the carcinogen mass into the search and is then able to map the base sequences and compositions of the modified oligos. In studies of genomic DNA, the use of this ‘reversed pseudo-combinatorial’ approach can possibly postulate sequence-relevant information and indicate sequential preference of adduction. The computer program automates the processing of data that are too complex to be accomplished by manual analysis currently performed with most mass spectrometry application software such as MassLynx and LCQ Navigator. The program is especially useful when handling complex DNA data whose sequence is unknown.

In a broad sense, GenoMass is potentially a high-throughput computer software capable of fast and efficient data mining of complex mass spectrometric data for nucleic acid analysis. It draws

some analogy from software used in proteomic research with the 'bottom up' approach, yet is quite different due to the difference in the makeup of protein and nucleic acid. While there are 20 amino acids in peptide and protein samples, there are only four (five) nucleotides in DNA (RNA) samples. While *a priori* it would appear that the development of software for the analysis of spectral data from genetic samples would be relatively simpler than for proteins, there are actually fewer than a dozen popularly used software in this arena. This is in contrast to the proteomics area where there are nearly 600 different data analysis software available. The reason behind it is that mass spectrometry itself has been used far less in nucleic acid analysis than in proteomics, as array and chip-based technologies are more widely employed. However, mass spectrometry has been gaining a stronger footing in genomic research in recent years with the advancement in instrumentation. Particularly in the area of nucleic acid modification, mass spectrometry now can make a much bigger contribution with improved sensitivity, mass accuracy and dynamic range.

Although GenoMass is at its very initial stage, the functionalities are very different from the few existing software for mass spectrometric characterization of nucleic acids. In combining some of the features in programs such as Mongo and SOS, i.e. the MS/MS sequencing and sequence assignment based on both molecular ion and fragment ions, GenoMass can, over time, evolve into a full functional data analysis tool for mass spectrometric characterization of genomic samples including oligonucleotides, DNA, RNA and their modification products, and be made available to the public potentially as an open source software.

In the near future, acquisition of data with a higher mass resolution will aid in narrowing the search window and introduce additional selectivity to the GenoMass reconstruction process by improving the quality and accuracy of the chromatographic data. Moreover, GenoMass program could be developed to have real-time instrument control to automatically start MS/MS with found ions of interest.

Acknowledgements

This work was supported by grants from the National Institutes of Health (1ROCA169390 and 1RO1CA112231). This is publication No. 932 from the Barnett Institute.

References

1. Vrkic AK, O'Hair RAJ, Foote S, Reid GE. Fragmentation reactions of all 64 protonated trimer oligonucleotides and 16 mixed base tetramer oligonucleotides via tandem mass spectrometry in an ion trap. *International Journal of Mass Spectrometry* 2000;194:145.
2. Singh R, Gaskell M, Le Pla RC, Kaur B, Azim-Araghi A, Roach J, Koukouves G, Souliotis VL, Kyrtopoulos SA, Farmer PB. Detection and quantitation of benzo[a]pyrene-derived DNA adducts in mouse liver by liquid chromatography-tandem mass spectrometry: comparison with 32P-postlabeling. *Chemical Research in Toxicology* 2006;19:868. [PubMed: 16780367]
3. Oberacher H, Huber CG. Capillary monoliths for the analysis of nucleic acids by ion-pair reverse-phase chromatography-electrospray ionization mass spectrometry. *Trends in Analytical Chemistry* 2002;21:166.
4. Dickman MJ, Matin MM, Hornby DP. High-throughput analysis of nucleic acid modification reactions using ion-pair reverse-phase high-performance liquid chromatography. *Analytical Biochemistry* 2002;301(2):290. [PubMed: 11814299]
5. Andrews CL, Harsch A, Vouros P. Analysis of the in vitro digestion of modified DNA to oligonucleotides by LC-MS and LC-MS/MS. *International Journal of Mass Spectrometry* 2004;231(2-3):169.
6. 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. *Analytical Chemistry* 2007;79(5):1907. [PubMed: 17261027]

7. Xiong W, Glick J, Lin Y, Vouros P. Separation and characterization of diastereoisomeric oligonucleotide adducts using monolithic columns by ion-pair reversed-phase nano-HPLC coupled to ion trap mass spectrometry. *Analytical Chemistry* 2007;79:5312. [PubMed: 17579370]
8. Lexa M, Valle G. PRIMEX: rapid identification of oligonucleotide matches in whole genomes. *Bioinformatics* 2003;19(18):2486. [PubMed: 14668240]
9. Hail ME, Elliott B, Anderson J. High-throughput analysis of oligonucleotides using automated electrospray ionization mass spectrometry. *American Biotechnology Laboratory* 2004;22(1):12.
10. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)* 2005;21(18):3674.
11. Fan, JB.; Bibikova, M. Specific DNA methylation patterns for diagnosing lung cancer. 2005. Patent 2004-973783, 2005164246, 20041025
12. Gaddis Sara S, Wu Q, Thames Howard D, DiGiovanni J, Walborg EF, MacLeod MC, Vasquez KM. A web-based search engine for triplex-forming oligonucleotide target sequences. *Oligonucleotides* 2006;16(2):196. [PubMed: 16764543]
13. Barry JP, Vouros P, Van Schepdael A, Lay SJ. Mass and sequence verification of modified oligonucleotides using electrospray mass spectrometry. *Journal of Mass Spectrometry* 1995;30:993.
14. Rozenski J, Crain P, McCloskey JA. The RNA Modification Database. *Nucleic Acids Research* 1999;27:196. [PubMed: 9847178]
15. Rozenski J, McCloskey JA. SOS: a simple interactive program for ab initio oligonucleotide sequencing by mass spectrometry. *Journal of the American Society for Mass Spectrometry* 2002;13:200. [PubMed: 11908799]
16. Jayaraj S, Reid R, Santi DV. GeMS: an advanced software package for designing synthetic genes. *NucleicAcids Research* 2005;33:3011.
17. Kellersberger KA, Yu E, Kruppa GH, Young MM, Fabris D. Top-down characterization of nucleic acids modified by structural probes using high-resolution tandem mass spectrometry and automated data interpretation. *Analytical Chemistry* 2004;76(9):2438. [PubMed: 15117181]
18. Rusconi F. GNU polyxmass: a software framework for mass spectrometric simulations of linear (bio) polymeric analytes. *BMC Bioinformatics* 2006;7:226. [PubMed: 16643644]
19. McLuckey S, Berkel G, Glish G. Tandem mass spectrometry of small, multiply charged oligonucleotides. *Journal of the American Society for Mass Spectrometry* 1992;3:60.

Figure 1. The tetramer library of an unknown oligonucleotide sequence with AAF-adducts. The formula for calculating MW of anhydrous oligonucleotide sequence is: Mass = S (individual Base Mass) – 63.980 + 2.016. The Mass of AAF = 221.09.

Figure 1. The tetramer library of an unknown oligonucleotide sequence with AAF-adducts. The formula for calculating MW of anhydrous oligonucleotide sequence is: Mass = S (individual Base Mass) – 63.980 + 2.016. The Mass of AAF = 221.09.

Figure 2.
The interfaces used by GenoMass to search for the 6 most abundant 5-mers in the digest of oligo 80-mer.

Figure 3. Client-server architecture for implementing the 'reversed pseudo-combinatorial' approach.

Figure 4.
The 4-mer enzymatic digestion products of the NarI 21-mer duplex, categorized as orderly and random cut fragments.

Figure 5. EICs of identified 6 most abundant 5-mers in the digest of oligo 80-mer. Automatically generated and plotted by GenoMass. Brackets indicate base composition rather than sequence.

Figure 6. EICs of the 6 most abundant AAF-modified 5-mers identified in the digest of oligo 80-mer following exposure to AAF. Automatically generated and plotted by GenoMass.

Figure 7.

EICs of identified 6 most abundant AAF-modified 5-mers in the digest of ctDNA following exposure to AAF. Automatically generated and plotted by GenoMass. A 2-minute offset in retention times of adducts common to Figs 6 and 7 is due to solvent delay.

Figure 8.

EICs of identified 6 most abundant AAF-modified 9-mers in the digest of ctDNA following exposure to AAAF. Chromatograms were automatically generated and plotted by GenoMass.

Table 1

Sequences studied in the investigation of enzyme behavior and the selectivity of DNA-adduct formation

Oligonucleotide ID	Sequence
NarI 12-mer-1	5'-ACCGGCGCCACA-3'
NarI 21-mer Duplex	5'-ATCCCAACACCGGCGCCACAT-3' 3'-TAGGGTTGTGGCCGCGGTGTA-5
Oligo 80-mer	5'-GGGGAGGGTGGGAAGGGGGCATGAAAGTTGGAGGCAATG TGGCATTGAAGCTTTGGGCCATGCTTGCAGCCTGCCGCCC-3'
Calf thymus DNA	Unknown

Table 2

Comparison of the 6 most abundant modification products (doubly charged ions) of the 5-mer fragments [x] found on the ss oligonucleotide 80-mer and the ds ctDNA. The blue colors represent the same composition found in both samples. [x] denotes the compositional isomers

[M - 2H] ²⁺ (da)	828.9	860.9	872.9	848.9	868.9	888.9
80-mer	TGCCC	ATGGC	AATGG	TGGCC	TGGGC	TGGGG
DNA	TGCCC	ATGGC	AATGG	TGGCC	TGGGC	AATGC