2554–2559 *Nucleic Acids Research, 1998, Vol. 26, No. 11*

DNA sequence analysis by MALDI mass spectrometry

Finn Kirpekar*, Eckhard Nordhoff, Leif K. Larsen, Karsten Kristiansen, Peter Roepstorff and Franz Hillenkamp[1]

Department of Molecular Biology, Odense University, Campusvej 55, DK-5230 Odense M, Denmark and [1]Institute for Medical Physics and Biophysics, University of Münster, Robert-Koch-Strasse 31, D-48149 Münster, Germany

Received March 10, 1998; Revised and Accepted April 16, 1998

ABSTRACT

Conventional DNA sequencing is based on gel electrophoretic separation of the sequencing products. Gel casting and electrophoresis are the time limiting steps, and the gel separation is occasionally imperfect due to aberrant mobility of certain fragments, leading to erroneous sequence determination. Furthermore, illegitimately terminated products frequently cannot be distinguished from correctly terminated ones, a phenomenon that also obscures data interpretation. In the present work the use of MALDI mass spectrometry for sequencing of DNA amplified from clinical samples is implemented. The unambiguous and fast identification of deletions and substitutions in DNA amplified from heterozygous carriers realistically suggest MALDI mass spectrometry as a future alternative to conventional sequencing procedures for high throughput screening for mutations. Unique features of the method are demonstrated by sequencing a DNA fragment that could not be sequenced conventionally because of gel electrophoretic band compression and the presence of multiple non-specific termination products. Taking advantage of the accurate mass information provided by MALDI mass spectrometry, the sequence was deduced, and the nature of the non-specific termination could be determined. The method described here increases the fidelity in DNA sequencing, is fast, compatible with standard DNA sequencing procedures, and amenable to automation.

INTRODUCTION

Milestones in genomic sequencing, e.g., the determination of the complete human genome, will be reached in a foreseeable future. In addition to a wealth of new information, many new questions with regard to the genetic variations amongst individuals will arise. The function of non-coding DNA may only be elucidated by comparing data from a number of individuals. Similarly, clarification of the causes underlying genetic disorders in clinical research or phenotypic differences in general is dependent on sequence information obtained from a large number of individuals. Therefore, an increasing need for sequencing technologies offering rapid analysis of many similar samples can be expected. Furthermore, high fidelity sequencing will be imperative because attempts to characterise diversities are only rational with essentially error-free data. The demands on sequence reading length and sensitivity will, in contrast, be moderate when using standard PCR products as template. Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-MS) (1) is a method that meets these requirements. MALDI-MS relies on a sample preparation in which the analyte is co-crystallised with an excess of a suitable matrix, most frequently an organic acid. The analyte/matrix mixture is desorbed/ionised by a short laser pulse; the generated ions are all accelerated to the same potential in an electrical field, and are subsequently separated in a field-free drift region according to their mass-over-charge ratio. Because singly charged nucleic acid ions are predominantly formed, spectra interpretation of multi-component samples is straightforward. The high vacuum of the instrument is ideal for size separation of sequencing products because the separation is independent of the structure of the analytes. Therefore, sequencing data relying on absolute molecular mass information rather than relative gel electrophoretic mobility will significantly increase the sequencing fidelity. Comparing signal intensities amongst four mass spectra performs the initial sequence determination. A dimension of sequence verification is introduced by checking the mass difference between adjacent signals. In the present work, we demonstrate that this additional verification is crucial for the fidelity of the sequence data whenever the sequencing enzyme performs less than ideal: the mass data can reveal the nature of dNTP chain termination and disclose the sequence at positions where essentially no termination occurs.

In MALDI-MS, the high resolution needed to obtain the required mass accuracy is mainly compromised by ion fragmentation (2,3) and salt formation, primarily with alkali ions (4,5). We have implemented two means, an instrumental and a chemical one, in order to reduce the fragmentation of our analytes. One is the recently introduced delayed ion extraction (DE) in which a delay time between ion generation and ion acceleration is introduced (6). The use of DE strongly improves signal resolution and reduces ion fragmentation for nucleic acid ions (7). The other means was to perform the sequencing with the purine nucleotide derivatives 7-deaza-dATP and 7-deaza-dGTP that exhibit significantly higher ion stability than their natural analogues (8,9). Additionally, we have developed a simple and fast sample purification technique using miniaturised reverse phase columns with elution directly onto the

*To whom correspondence should be addressed. Tel: +45 6557 2414; Fax: +45 6593 2781; Email: f.kir@PR-group.ou.dk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

MALDI target. Salts, enzymes, and non-incorporated nucleotide triphosphates are efficiently removed from the sequencing reactions, significantly reducing salt formation and increasing the detection sensitivity in the subsequent MALDI-MS analysis.

The feasibility of using MALDI-MS in the separation of short sequence ladders has been demonstrated several times with model systems using single-stranded templates (10–13). Initial attempts to use the mass information for sequence determination has been described with single stranded synthetic template using MALDI-MS equipment without DE (13). Here we show the bona fide use of the mass spectrometric ladder-separation approach on PCR templates taking full advantage of the generated mass data. Three applications of this strategy are illustrated below.

MATERIALS AND METHODS

PCR

The murine PPARδ exon 1 fragment was generated by PCR in a 100 µl volume containing 50 pmol of the primers 5′-TTC CCA TTA ATC CCC AGG TC-3′ and 5′-GCG GCG CGG TCA CAT CAC-3′, 50 ng of a plasmid harbouring the promoter/exon 1 region of murine PPARδ gene, 200 µM dNTP, 2.5 U of *Pfu* DNA polymerase (Stratagene, USA) in the buffer supplied with the enzyme. The reaction was incubated at 95°C for 3 min and subsequently 30 cycles of 95°C/10 s; 50°C/60 s; 72°C/10 s were performed, ending with a 7 min incubation at 72°C. The PCR of the human LDL receptor gene, exon 9 and the human α_1 antitrypsin gene, exon 5 is described by Nissen *et al.* and Andresen *et al.* respectively (14,15).

Sequencing

Cycle sequencing of agarose gel purified PCR fragments from the human LDL receptor gene, exon 9 and the murine PPARδ, exon 1 was performed in 20 µl volumes containing 30 pmol sequencing primer, ~1.5 pmol PCR template, 30 mM Tris–HCl (pH 9.3), 3 mM $MgCl_2$, 0.15 mM DTE, 0.05 g/l n-octylglycosid, 100 µM 7-deaza-dATP, dCTP, 7-deaza-dGTP and dTTP, 3.3 µM of one of the four ddNTPs, and 8 U Thermosequenase (Amersham Life Science, UK). The reactions were overlaid with 25 µl paraffin oil, heated to 95°C for 2 min, subjected to 30 cycles of 55°C/30 s, 95°C/60 s, and finally incubated at 55°C for 2 min. Subsequently, the paraffin oil was removed. The human α_1 antitrypsin gene, exon 5 PCR fragment was sequenced as above except that only 0.5 pmol PCR fragment and 6 pmol of primer were employed per reaction.

Sample preparation and mass spectrometry

EDTA and triethyl ammonium acetate, pH 7.0 (TEAA) were added to the sequencing reactions to final concentrations of 5 and 300 mM, respectively. Micro-purification columns were prepared as follows: a GelLoader tip (Eppendorf) was flattened at the tip of the extended outlet with forceps, reducing the inner diameter to <50 µm. An aliquot of 50 µl of methanol was filled into the GelLoader tip and 3 µl of a suspension of Poros 50R2 (PerSeptive Biosystems, USA) material in methanol were added to the methanol fill. After the Poros material had settled, the GelLoader tip was mounted onto a 1.25 ml Combitip (Eppendorf), and the methanol was pressed through. Hereby, the Poros material forms a small column with a volume of ~1 µl in the GelLoader tip. The column was washed once with 20 µl methanol, and equilibrated by adding 20 µl of 10 mM TEAA/3% acetonitrile. Hereafter, the sample was loaded onto the micro column, washed with 200 µl of 10 mM TEAA/3% acetonitrile, and eluted directly onto the MALDI target with 2 µl of 10 mM TEAA/15% acetonitrile. An aliquot of 0.7 µl of matrix solution (22.5 mg/ml 3-hydroxypicolinic acid and 2.5 mg/ml picolinic acid in 20% acetonitrile) was added and the solvent was allowed to evaporate. All measurements were performed in the linear TOF mode of a Voyager Elite MALDI mass spectrometer (PerSeptive Biosystems, USA) with the detection of positively charged ions. To improve the signal-to-noise ratio, typically 100 single-shot spectra were averaged before data transfer to the computer. Data processing was performed using the software package GRAMS/386, version 3.04 (Galactic Industries Corporation).

Mass difference calculations

Masses of incorporated nucleotidylate units relevant for this work are: 7-deaza-dAMP: 312.2 Da, ddAMP: 297.2 Da, dCMP: 289.2 Da, ddCMP: 273.2 Da, 7-deaza-dGMP: 328.2 Da, ddGMP: 313.2 Da, dTMP: 304.2 Da, ddTMP: 288.2 Da. The mass difference between a dideoxy-pyrimidine nucleotide terminated product and the directly succeeding termination product equals the mass of the latter terminating nucleotidylate unit plus 16.0 Da. The mass difference between a dideoxy-purine nucleotide terminated product and the directly succeeding termination product equals the mass of the latter terminating nucleotidylate unit plus 15.0 Da. The mass difference between any deoxy-NTP-terminated product and a subsequent neighbouring termination product matches the mass of the latter terminating nucleotidylate unit. If the peaks are not neighbouring, the mass of the intervening (7-deaza-)deoxy-nucleotidylate units should be added.

RESULTS AND DISCUSSION

The initial experiment was performed on a 97/98 bp PCR fragment from exon 9 of the human LDL receptor gene from an individual harbouring a single deletion in one allele (14). Employing the 5′-primer also used for the PCR, sequence ladders were generated and subsequently analysed by MALDI-MS (Fig. 1a). The primer-proximal mass region of the four spectra is expanded with the peak intensities brought into proper scale (Fig. 1b). The first two signals after the primer signal are present in all four mass spectra, demonstrating ddNTP-independent termination events in the DNA elongation reaction close to the primer. From the mass differences around 304 Da between the primer signal and the subsequent termination signal, it is clear that a deoxythymidylate is incorporated. This is substantiated by the presence of a double-peak in the ddTTP termination spectrum, corresponding to a concomitant incorporation of ddTMP (mass difference relative to primer signal: 288.4 Da). Similarly, the second peak after the primer corresponds to an incorporation of 7-deaza guanylate (plus 328.2 Da), with a minor signal showing the addition of ddGMP (plus 313.4 Da) in the ddGTP termination spectrum. At position 30 after the primer [marked (C) in Fig. 1a], peaks appear in both the ddCTP and in the ddTTP termination spectra, reflecting the deletion of a C in one strand. In addition to the concurrent presence of termination signals in two mass spectra, the deletion is confirmed by the emergence of doublets of signals (neighbouring peaks marked by N′ and N in Fig. 1a),

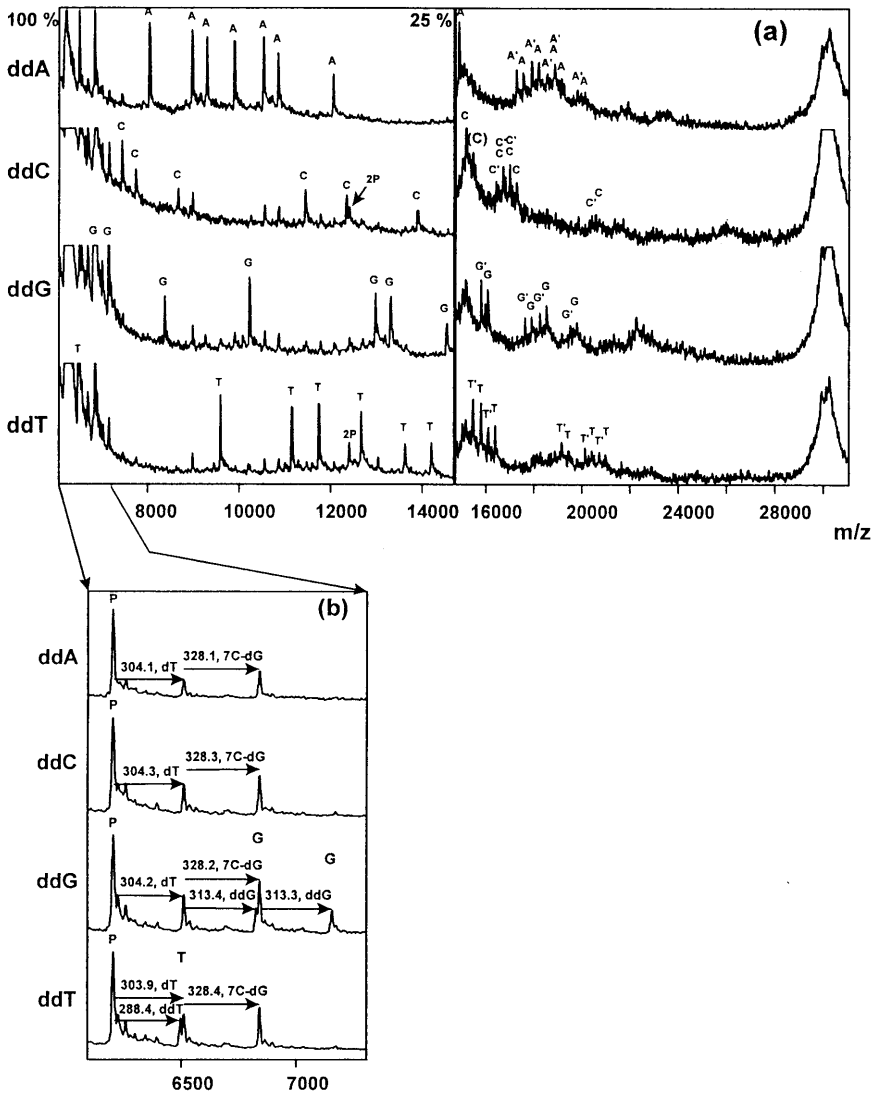

Figure 1. MALDI-MS analysis of a cycle sequencing reaction on a PCR template. The PCR product was obtained from an individual harbouring a single bp deletion in one exon 9 allele of the human LDL receptor gene (14). **(a)** Spectra of sequencing reactions with the four different ddNTPs. Above 15 000 m/z, the full-scale signal intensity is 25% of the intensity below 15 000 m/z. (C) corresponds to the deoxycytidylate at position 30 after the primer, present in the wild type allele only. Peaks marked with a prime (′) originate from the allele with the deletion. 2P indicates the signal of the dimer of the sequencing primer. **(b)** Magnification of the primer-proximal mass range demonstrating terminations with the incorporation of deoxynucleotidylate units.

always separated by ~289 Da, equal to the mass of the missing deoxycytidylate in the mutant copy strand.

As shown above, a characteristic of the MALDI-MS analysis is that sequence determination close to the primer is readily possible, the use of which is demonstrated in Figure 2. A 97 bp PCR-fragment derived from exon 5 of the human α_1-antitrypsin gene was partially sequenced. The PCR-fragment was obtained from an individual harbouring one wild type allele and one allele with a G^{9989} to A transition (15); the point mutation is located in the fourth position after the sequencing primer. Signals at position four in both the ddATP and the ddGTP spectra clearly arise. The measured mass difference of 15.7 Da between the two signals is in agreement with the expected difference (16.0 Da) between guanine and adenine. At the same time, this mass difference excludes that the position-four signal in the ddGTP spectrum originates from a termination event with incorporation of 7-deaza-dAMP: in that case, the expected mass difference would be 15.0 Da. Making mass distinctions like the above requires a mass accuracy better than 0.5 Da. Such a mass accuracy is currently obtainable in our laboratory up to approximately the 30mer level (9000 Da) when using the primer signal for internal calibration. The point mutation at position four is confirmed by the doublet nature of the succeeding peaks originating from the extensions of two ‘substrates’ with a mass difference of 16 Da. Thus, as in the example with the deletion, the nature of the mutation is confirmed ‘horizontally’, in addition to the ‘vertical’ primary identification by termination signals in two spectra. The dominating peak at position 3 is in the ddCTP spectrum as expected from the published sequence (15). The additional signal in the ddTTP spectrum is the result of a termination occurring with the incorporation of dCMP as confirmed by the mass difference 16.2 Da (calculated value: 16.0 Da). The presence of

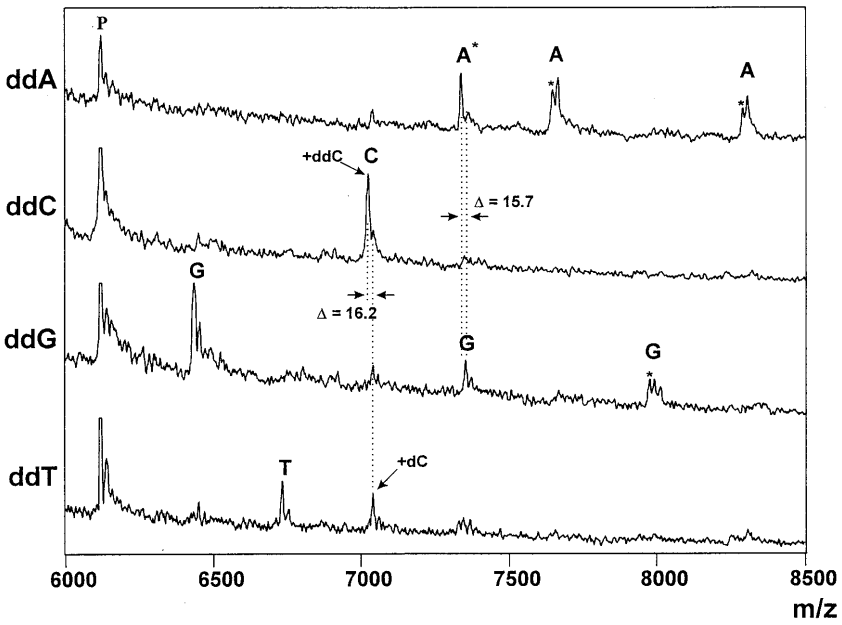

Figure 2. Cycle sequencing on a PCR product obtained from an individual harbouring a point mutation in one exon 5 allele of the human α_1 antitrypsin gene (15). Analysis by MALDI-MS. Only the mass window around the point mutation is shown. The asterisks indicate the 16 Da peak heterogeneity resulting from the point mutation. See text for further details.

a second point mutation (C to T) at position three is simultaneously excluded because a mass difference of nominally 15.0 should have been observed.

The power of obtaining data consisting of absolute molecular masses rather than relative gel mobility was ultimately demonstrated by sequencing a murine PPARδ exon 1 PCR fragment. A stretch of ~10 bp had previously proven impossible to determine by gel-based sequence ladder separation because severe band compression phenomena and non-specific termination arose, despite sequencing attempts on both template strands with a series of different primers and different DNA polymerases. A 91 bp PCR fragment harbouring the stretch in question was sequenced from both ends and MALDI-MS analysed, generating a 25 nt sequence overlap. The MALDI-MS data were in agreement with the gel data for the regions that could be unambiguously determined by the latter method (data not shown). MALDI mass spectra of the problematic stretch, located 8 nt from the primer, are shown in Figure 3. The first 2 nt are clearly identified by strong signals in the ddCTP termination spectrum, whereas the peaks at the third position reveal an unexpected nucleotide polymerisation event: a clear signal in the ddTTP spectrum at a mass corresponding to the incorporation of a dideoxythymidylate unveils the identity at that position, but distinct signals originating from the termination with the incorporation of dGMP and to a lesser extent also dCMP also appear. The phenomenon is particularly prominent in the ddCTP termination reaction, and clearly shows a DNA elongation reaction that does not rely on Watson–Crick base pairing with the template. The mass difference between the signals at the second and the fifth position in the ddCTP termination spectrum is 921.2 Da, proving that the extension products harbour a deoxythymidylate at position three (calculated value: 921.6 Da). Thus, a mis-incorporated nucleotide leads to termination with no detectable elongation products. It is plausible that particular structures, in this case an extreme G/C-richness, of the template strand promote mis-incorporations. A similar observation has been reported for the *Taq* DNA polymerase (16), of which 'Thermosequenase' is a genetically engineered derivative (17). However, it is unclear why the mis-incorporation is especially pronounced in the reaction containing ddCTP. It is noteworthy that the 'C' at position five is easily determined by MS, whereas it was highly ambiguous in the gel electrophoretic analysis due to odd sequencing product mobility. At position six in Figure 3, a strong stop signal is observed in all four spectra, but mass differences, e.g. between the peaks at position five and six in the ddCTP termination spectrum (344.2 Da calculated, 344.6 Da found), reveal the incorporation of a 7-deaza-deoxyguanylate. The appearance of a double peak in the ddGTP termination reaction displays the additional incorporation of the dideoxyguanylate, ascertaining a 'G' at this position.

The intensity of the termination signals around m/z 11 500 and 11 800 was very low, a problem also observed in the gel-electrophoretic separation of the sequencing ladders. By MALDI-MS, sequence information could, nevertheless, be derived, again by making use of the accurate mass information. The mass difference between the two unambiguous termination signals, flanking the positions in question, in the ddGTP and ddATP spectra was determined to 906.3 Da. It can therefore be concluded that one 'C' and one 'T' must be present (905.6 Da calculated). The 'C' can conceivably be assigned to the m/z 11 800 peak: the mass difference from the 5′-side (904.1 Da observed, 905.6 Da calculated) in the ddCTP termination spectrum between the position eight deoxycytidylate terminated product and the product at m/z 11 800, together with the 3′-side mass difference of 311.8 Da between this putative 'C' peak and the following peak in the ddATP spectrum suggest that a dideoxycytidylate terminated product is present at m/z 11.800. The 'T' is consequently bound to be at the m/z 11 500 position, which is also inferred by the mass difference of 305.7 Da between the 5′-flanking ddGTP termination signal and the subsequent small signal in the ddTTP spectrum.

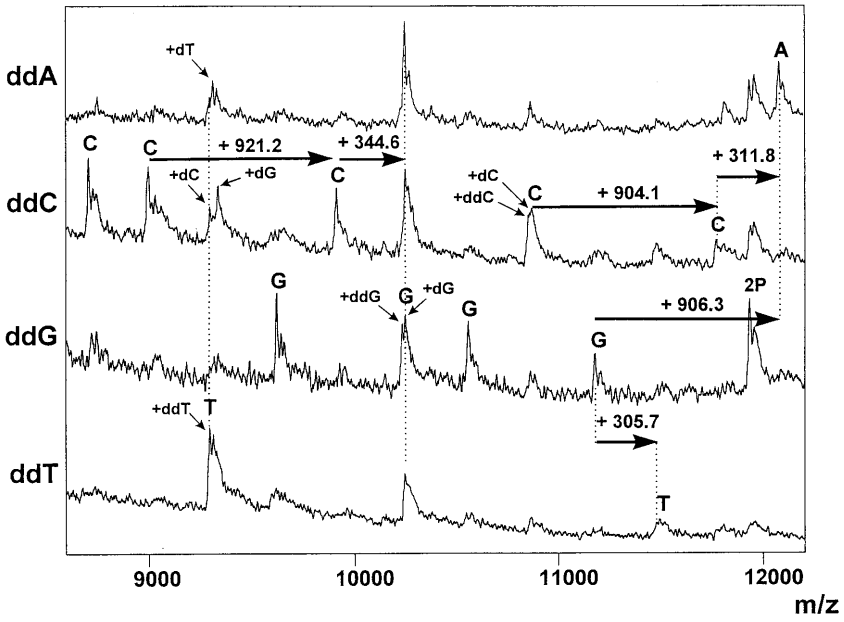

Figure 3. Generation of *de novo* sequence information by MALDI-MS product separation. Cycle sequencing was performed on a murine PPARδ, exon 1 PCR fragment. Vertical arrows show the mass differences used for sequence determination discussed in the text. '2P' marks the signal of the dimer of the sequencing primer.

In summary, the sequence (including the PCR primers written in bold) of the murine PPARδ exon 1 PCR fragment was: 5′-**TTC CCA TTA ATC CCC AGG TC**C GCC GCG GCC TGC GGC GTC ACA CGC TCC CAG CCC CGG CGC CGC CCC CAG CCC C**GT GAT GTG ACC GCG CCG C**-3′. The underline portion is deduced from the data presented in Figure 3.

The work presented here suggests that sequence ladder analysis by MALDI mass spectrometry may become a realistic alternative to electrophoresis in, e.g., clinical research, where large-scale analysis of short sequence motifs is frequently desirable. The time from completion of the sequencing reactions to the generation of the mass spectra was <1 h with the described procedures. The MALDI-MS approach may readily be speeded up by automation, including sample clean up, sample preparation and the subsequent MALDI-MS analysis (18). The spectra interpretations exemplified here can be implemented in computer software, thereby automating the sequence deduction. Such a software should, in addition to signal intensities, compare the measured mass differences between a given peak and its two neighbours with the possibly encountered values in order to determine/verify the nucleotide identity.

The use of the method for the characterisation of different types of point mutations, as often encountered in clinical research, is also documented in this work. Furthermore, the accuracy of the obtained mass data allowed acquisition of *de novo* sequence information that could not be acquired by gel based sequence ladder separation. One obvious reason is that problems associated with aberrant migration of certain fragments do not exist in high vacuum. Another reason is that dideoxynucleotide terminated products can be distinguished from those carrying a deoxynucleotide at their 3′-end. Mass spectrometric analysis of sequencing reactions may prove to be particularly useful in cases where G/C-rich templates are encountered, e.g. in thermophilic organisms. All templates used were obtained by standard PCR, demonstrating that the sensitivity of the method is compatible with standard laboratory routines. Sensitivity can additionally be increased, since recent research has shown that MALDI-MS sample preparations of sequencing reactions can be scaled down to low nanoliter volumes by the use of a piezoelectric pipetting device; this also results in an increased reproducibility of the mass spectrum acquisition (19).

DNA ion fragmentation is a major cause for the present limitations in reading length, substantiated by the observation that RNA, being mass spectrometrically much more stable, is easily detected above the 100mer level with a sensitivity that is significantly higher than for DNA (20,21). Though the use of 7-deaza purine nucleotides in the sequencing reaction results in a dramatic improvement in stability at the corresponding positions in DNA, the cytosine base is still prone to fragmentation. In a recent report, it was demonstrated that the presence of 2′-F-cytidine in oligodeoxynucleotides significantly reduces ion fragmentation (22). The 2′-F-cytidine triphosphate is readily accepted jointly with the 7-deaza purine nucleotides by the T7 DNA polymerase and derivatives thereof (F.K. and E.N., unpublished). This may lead to very stable sequencing products and should, in turn, further improve the applicability of MALDI mass spectrometry in DNA sequencing.

ACKNOWLEDGEMENTS

This work was supported by the Commission of the European Communities [grants No. GENE-CT93-0082 (SSMA) and TMR ERBFMBI-CT950446] and by the Danish Biotechnology Program. PerSeptive Biosystems, USA, generously provided the Voyager Elite mass spectrometer. We would like to thank B.S.Andresen, Skejby Sygehus, Denmark and H.Nissen, Odense Universitets-hospital, Denmark for kindly providing us with PCR fragments from the α_1-antitrypsin gene and the LDL receptor gene, respectively.

REFERENCES

1 Karas,M. and Hillenkamp,F. (1988) *Anal. Chem.* **60**, 2299–3301.
2 Wu,K.J,. Shaler,T.A. and Becker,C.H. (1994) *Anal. Chem.* **66**, 1637–1645.
3 Nordhoff,E., Kirpekar,F., Karas,M., Cramer,R., Hahner,S., Hillenkamp,F., Kristiansen,K., Roepstorff,P. and Lezius,A. (1994) *Nucleic Acids Res.* **22**, 2460–2465.
4 Pieles,U., Zürcher,W., Schär,M. and Moser,H.E. (1993) *Nucleic Acids Res.* **14,** 3191–3196.
5 Nordhoff,E., Ingendoh,A., Cramer,R., Overberg,A., Stahl,B., Karas,M., Hillenkamp,F. and Crain,P.F. (1992) *Rapid Commun. Mass Spectrom.* **6**, 771–776.
6 Vestal,M.L., Juhasz,P. and Martin,S.A. (1995) *Rapid Commun. Mass Spectrom.* **9**, 1044–1050.
7 Juhasz,P., Roskey,M.T., Smirnov,I.P., Haff,L.A., Vestal,M.L. and Martin,S.A. (1996) *Anal. Chem.* **68**, 941–946.
8 Kirpekar,F., Nordhoff,E., Kristiansen,K., Roepstorff,P., Hahner,S. and Hillenkamp,F. (1995) *Rapid Commun. Mass Spectrom.* **9**, 525–531.
9 Schneider,K. and Chait,B.T. (1995) *Nucleic Acids Res.* **23**, 1570–1575.
10 Shaler,T.A., Tan,Y., Wickham,J.N., Wu,K.J. and Becker,C.H. (1995) *Rapid Commun. Mass Spectrom.* **9**, 942–947.
11 Roskey,M.T., Juhasz,P., Smirnov,I.P., Takach,E.J., Martin,S.A. and Haff,L.A. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 4724–4729.
12 Mouradian,S., Rank,D.R. and Smith,L.M. (1996) *Rapid Commun. Mass Spectrom.* **10**, 1475–1478.
13 Köster,H., Tang,K., Fu,D.-J., Braun,A., van den Boom,D., Smith,C.L., Cotter,R.J. and Cantor,C.R. (1996) *Nature Biotechnol.* **14**, 1123–1128.
14 Nissen,H., Hansen,A.B., Guldberg,P., Petersen,N.E., Larsen,M.L., Haghfelt,T., Kristiansen,K. and Hørder,M. (1994) *Aterosclerosis* **111**, 209–215.
15 Andresen,B.S., Knudsen,I., Jensen,P.K.A., Rasmussen,K. and Gregersen,N. (1992) *Clin. Chem.* **38**, 2100–2107.
16 Loewen,P.C. and Switala,J. (1995) *Gene* **164**,59–63.
17 Tabor,S. and Richardson,C.C. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 6339–6343.
18 Jensen,O.N., Mortensen,P., Vorm,O. and Mann,M. (1997) *Anal. Chem.* **69**, 1706–1714.
19 Little,D.P., Cornish,T.J., O'Donnell,M.J., Braun,A., Cotter,R.J. and Köster,H. (1997) *Anal. Chem.* **69**, 4540–4546.
20 Nordhoff,E., Cramer,R., Karas,M., Hillenkamp,F., Kirpekar,F., Kristiansen,K. and Roepstorff,P. (1993) *Nucleic Acids Res.* **21**, 3347–3357.
21 Kirpekar,F., Nordhoff,E., Kristiansen,K., Roepstorff,P., Lezius,A., Hahner,S., Karas,M. and Hillenkamp,F. (1994) *Nucleic Acids Res.* **22**, 3866–3870.
22 Tang,W., Zhu,L. and Smith,L.M. (1997) *Anal. Chem.* **69**, 302–312.