Automated Identification of Amino Acid Sequence Variations in Proteins by HPLC/Microspray Tandem Mass Spectrometry

Christine L. Gatlin,[†,§] Jimmy K. Eng,[†] Stacy T. Cross,[‡] James C. Detter,[‡] and John R. Yates III[†,*]

Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, Washington 98195-7730, Red Cell Laboratory, Hematology Division, Department of Laboratory Medicine, University of Washington School of Medicine, Box 357110, Seattle, Washington 98195-7110

Amino acid sequence variations resulting from single-nucleotide polymorphisms (SNPs) were identified using a novel mass spectrometric method. This method obtains 99+% protein sequence coverage for human hemoglobin in a single LC-microspray tandem mass spectrometry (μLC-MS/MS) experiment. Tandem mass spectrometry data was analyzed using a modified version of the computer program SEQUEST to identify the sequence variations. Conditions of sample preparation, chromatographic separation, and data collection were optimized to correctly identify amino acid changes in six variants of human hemoglobin (Hb C, Hb E, Hb D-Los Angeles, Hb G-Philadelphia, Hb Hope, and Hb S). Hemoglobin proteins were isolated and purified, dehemed, (S)-carboxyamidomethylated, and then subjected to a combination proteolytic digestion to obtain a complex peptide mixture with multiple overlaps in sequence. Reversed-phase chromatographic separation of peptides was achieved on-line with MS utilizing a robust fritless microelectrospray interface. Tandem mass spectrometry was performed on an ion trap mass spectrometer using automated data-dependent MS/MS procedures. Tandem mass spectra were collected from the five most abundant ions in each scan using dynamic and isotopic exclusion to minimize redundancy. The spectra were analyzed by a version of the SEQUEST algorithm modified to identify amino acid substations resulting from SNPs.

As the sequencing of the human genome progresses, interest is shifting to the role of genetic variation in phenotype. To achieve this goal, studies have begun to target the identification of single-nucleotide polymorphisms (SNPs) at the genomic level as well as to identify specific polymorphisms in genes suspected of a role in disease.[1,2] Methods to identify SNPs include hybridization to sequence-specific oligonucleotides following enzymatic or chemical cleavage, denaturing gradient gel electrophoresis, single strand conformation polymorphism (SSCP), DNA sequencing, and hybridization to oligonucleotide arrays (DNA chips).[3−12] Mass spectrometry can also be used to identify SNPs in DNA. Mass-spectrometric-based methods have been demonstrated using solid-phase Sanger DNA sequencing and primer oligo base extension reactions (PROBE) followed by high resolution/ high mass accuracy detection by MALDI-TOF-MS.[13−17] These MS-based techniques can identify single-nucleotide polymorphisms (SNPs) in short DNA molecules (<100 bases). A principal advantage to the use of mass spectrometry is detection of sequence differences on the basis of m/z value rather than different migration times on a gel or hybridization and duplex melting temperatures.

Genetic variation influences disease susceptibility and prognosis as well as the response to drug treatments (pharmacoge-

[*] Corresponding author: (e-mail) jyates@u.washington.edu.

[†] University of Washington.

[‡] University of Washington School of Medicine.

[§] Present address: Large-Scale Biology Corporation, 9620 Medical Center Drive, Rockville, Maryland 20850.

(1) Wang, D. G.; Fan, J. B.; Siao, C. J.; Berno, A.; Young, P.; Sapolsky, R.; Ghandour, G.; Perkins, N.; Winchester, E.; Spencer, J.; Kruglyak, L.; Stein, L.; Hsie, L.; Topaloglou, T.; Hubbell, E.; Robinson, E.; Mittmann, M.; Morris, M. S.; Shen, N.; Kilburn, D.; Rioux, J.; Nusbaum, C.; Rozen, S.; Hudson, T. J.; Lipshutz, R.; Chee, M.; Lander, E. S. *Science* (*Washington, D.C.*) **1998**, *280*, 1077−82.

(2) Rieder, M. J.; Taylor, S. L.; Clark, A. G.; Nickerson, D. A. *Nat. Genet.* **1999**, *22*, 59−62.

(3) Saiki, R. K.; Bugawan, T. L.; Horn, G. T.; Mullis, K. B.; Erlich, H. A. *Nature* (*London*) **1986**, *324*, 163−6.

(4) Myers, R. M.; Larin, Z.; Maniatis, T. *Science* (*Washington, D.C.*) **1985**, *230*, 1242−6.

(5) Youil, R.; Kemper, B. W.; Cotton, R. G. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 87−91.

(6) Sheffield, V. C.; Cox, D. R.; Lerman, L. S.; Myers, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 232−6.

(7) Orita, M.; Suzuki, Y.; Sekiya, T.; Hayashi, K. *Genomics* **1989**, *5*, 874−9.

(8) Gibbs, R. A.; Nguyen, P. N.; McBride, L. J.; Koepf, S. M.; Caskey, C. T. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 1919−23.

(9) Leren, T. P.; Solberg, K.; Rodningen, O. K.; Rosby, O.; Tonstad, S.; Ose, L.; Berg, K. *Hum. Genet.* **1993**, *92*, 6−10.

(10) Rieder, M. J.; Taylor, S. L.; Tobe, V. O.; Nickerson, D. A. *Nucleic Acids Res.* **1998**, *26*, 967−73.

(11) Chee, M.; Yang, R.; Hubbell, E.; Berno, A.; Huang, X. C.; Stern, D.; Winkler, J.; Lockhart, D. J.; Morris, M. S.; Fodor, S. P. *Science* (Washington, D.C.) **1996**, *274*, 610−14.

(12) Hacia, J. G.; Brody, L. C.; Chee, M. S.; Fodor, S. P.; Collins, F. S. *Nat. Genet.* **1996**, *14*, 441−7.

(13) Murray, K. K. *J. Mass Spectrom.* **1996**, *31*, 1203−15.

(14) Koster, H.; Tang, K.; Fu, D. J.; Braun, A.; van den Boom, D.; Smith, C. L.; Cotter, R. J.; Cantor, C. R. *Nature Biotechnol.* **1998**, *1*, 1.

(15) Fu, D. J.; Tang, K.; Braun, A.; Reuter, D.; Darnhofer-Demar, B.; Little, D. P.; O'Donnell, M. J.; Cantor, C. R.; Koster, H. *Nature Biotechnol.* **1998**, *16*, 381−4.

(16) Little, D. P.; Braun, A.; O'Donnell, M. J.; Koster, H. *Nat. Med.* (*N.Y.*) **1997**, *3*, 1413−6.

(17) Haff, L. A.; Smirnov, I. P. *Genome Res.* **1997**, *7*, 378−88.

10.1021/ac991025n CCC: $19.00 © 2000 American Chemical Society
Published on Web 01/14/2000

nomics).[18] The most direct mechanism by which sequence variation can affect phenotype is through variations in promoter and coding regions of a gene. Variations in promoter regions may lead to over- or underexpression of a particular gene and potentially abnormal levels of translated protein. Recent studies suggest protein and gene expression may not correlate well, and thus, measurement of mRNA abundance may not be indicative of protein levels.[19,20] Sequence variations in exons (coding regions) can affect the activity of proteins more directly. Variations in coding regions have been associated with susceptibility to diseases such as atherosclerosis, and links between anemias and hemoglobin variations are well-established.[21,22] Sequence variations in enzymes involved in drug metabolism may be responsible for slow clearance and toxicity—a relationship of interest to pharmacogenomic studies.[23]

Several problems exist for the analysis or identification of sequence variations at the protein level. First, proteins can be difficult to obtain or be relatively rare, requiring highly sensitive methods of detection. Antibody-based ELISAs can be used to target specific sequence variations, but different antibodies are required for each sequence variation. The sensitivity of mass spectrometry analysis for peptides and proteins is continually improving and is currently in the attomole range.[24] Mass spectrometry techniques such as tandem mass spectrometry can be used for de novo identification of sequences or for the identification of proteins of known sequences. Software to automatically match tandem mass spectra to sequences has been developed, but is not effective for unanticipated sequence variations or does not specifically indicate the type and site of the amino acid variation.[25−27] An additional requirement to determine amino acid sequence variations is the ability to generate complete sequence coverage of proteins. Complete sequence coverage ensures that fragmentation data, and thus amino acid sequence information, is obtained for every position in the protein. In this paper, we report a method to generate complete sequence coverage for hemoglobin using automated tandem mass spectrometry data acquisition and software to identify the site and type of variations resulting from SNPs. No efforts were made to determine the limits of detection in this study.

EXPERIMENTAL PROTOCOL

Materials. Deionized water from a MilliQ RG ultrapure water system (Millipore, Bedford, MA) was used at all times. HPLC grade methanol, acetonitrile, and glacial acetic acid were furnished from Fisher Scientific (Fair Lawn, NJ). Endoproteinase Glu-C and trypsin were purchased from Boehringer-Mannheim (Mannheim, Germany). All other chemicals, human hemoglobin (Hb A), sickle-cell Hb (Hb S), and subtilisin, were obtained from Sigma (St. Louis, MO) and used without further purification. All other hemoglobin variants were obtained from patient samples remaining after routine laboratory analysis at the University of Washington Medical Center.

Hb Isolation and Purification. Blood samples were drawn by venipuncture into EDTA-containing tubes. Whole blood was washed three times with 0.85% NaCl to remove serum proteins. The remaining red cells were lysed in four volumes of H_2O and vortexed. Cell membranes were removed by spinning for 4 min at 13 400*g*. Isoelectric focusing (IEF) was performed using the methods of Basset et al. on agarose plates (ampholines pH 6−8, Isolab, Akron, OH) and of Schneider and Barwick on citrate agar (pH 6, Helena Laboratories, Beaumont, TX).[28,29] Variant hemoglobins were cut out of the gel and eluted into H_2O. The isolated variant was filtered through a 0.45-μm Millipore Ultrafree-MC filter by spinning for 4 min at 13 400*g* to remove particulates. Humectant (triethylene glycol) was removed with a 10 000 NMWL Millipore Ultrafree-MC filter (8 min, 13 400g). Hemoglobin was resuspended in 80 μL of H_2O.

Preparation of Isolated Variants for Proteolytic Digestion. *Apohemoglobin (apoHb)*. Hb A and Hb S (obtained from Sigma) and hemolysates containing known human hemoglobin variants were dehemed by adding, dropwise, 50 μL of protein solution into 20 vol of ice-cold acetone/0.6% HCl.[30] Precipitated globin was washed three times with cold acetone and dissolved in 100 μL of H_2O (10 μM). *Reduction/Alkylation*. Solutions of 10 μM apohemoglobin, 10 mM DTT (dithiothreitol), and 50 mM iodoacetamide were sonicated for 15 min and flushed with N_2 to remove dissolved oxygen. A sample of 0.4 μL of DTT (4 mg/mL) was added to 40 μL of apohemoglobin and incubated at 37 °C in the dark for 30 min. A sample of 0.4 μL of iodoacetamide (9 mg/mL) was added (incubated at 37 °C in the dark for 30 min) to the reduced sample. The reaction was quenched with 0.4 μL of DTT with another 30-min incubation at 37 °C in the dark.

Combination Proteolytic Digestion. Three separate proteolytic digests were performed on each protein. They were then combined into one sample tube for subsequent LC/MS/MS analysis. *Trypsin*. Trypsin was added to CAM-apoHb (4 μM) at an enzyme-to-substrate ratio of 1:50 (w/w) in 1 mM $CaCl_2$/100 mM NH_4HCO_3 buffer, pH 8.5. The protein was digested for 5 h at 37 °C in the dark. The reaction was stopped by lowering solution pH with 1% HOAc to pH 3. *Endoproteinase Glu-C*. Endoproteinase Glu-C (Glu-C) was added (1:40) to CAM-apoHb (5 μM) in 50 mM Na_2HPO_4. After a 12-h incubation at 37 °C in the dark, the reaction was quenched with 1% HOAc (to pH 3). *Subtilisin*. Subtilisin was added (1:60) to CAM-apoHb (2μM) in 100 mM Tris-HCl pH 8.8/5.6 M urea. The reaction proceeded for 1 h (37 °C in the dark) and was immediately followed by quenching with acetic acid to pH 3. Twenty picomoles of each hemoglobin digest was added to a single microfuge tube. Twenty picomoles of CAM-apoHb was

[18] Kleyn, P. W.; Vesell, E. S. *Science* (*Washington, D.C.*) **1998**, *281*, 1820−1.
[19] Anderson, L.; Seilhamer, J. *Electrophoresis* **1997**, *18*, 533−7.
[20] Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R. *Mol. Cell Biol.* **1999**, *19*, 1720−30.
[21] Emmert-Buck, M. R.; Bonner, R. F.; Smith, P. D.; Chuaqui, R. F.; Zhuang, Z.; Goldstein, S. R.; Weiss, R. A.; Liotta, L. A. *Science* (*Washington, D.C.*) **1996**, *274*, 998−1001.
[22] Van Gelder, R. N.; von Zastrow, M. E.; Yool, A.; Dement, W. C.; Barchas, J. D.; Eberwine, J. H. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 1663−7.
[23] Vatsis, K. P.; Martell, K. J.; Weber, W. W. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 6333−7.
[24] Gatlin, C. L.; Kleemann, G. R.; Hays, L. G.; Link, A. J.; Yates, J. R., III *Anal. Biochem.* **1998**, *263*, 93−101.
[25] Eng, J. K.; McCormack, A. L.; Yates, J. R., III *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.
[26] Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−9.
[27] Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426−1436.
[28] Basset, P.; Beuzard, Y.; Garel, M. C.; Rosa, J. *Blood* **1978**, *51*, 971−82.
[29] Schneider, R. G.; Barwick, R. C. *Hemoglobin* **1982**, *6*, 199−208.
[30] Witkowska, H. E.; Bitsch, F.; Shackleton, C. H. *Hemoglobin* **1993**, *17*, 227−42.

added to the same tube for a final concentration of 1 μM of digested and 1 μM of intact Hb in 20 μL.

Fritless Electrospray Interface. The microelectrospray interface has been described previously by Gatlin et al.[24] Briefly, the interface utilizes an Upchurch PEEK Microtee (Upchurch Scientific, Oak Harbor, WA) with microfingertight fittings. A 0.025" gold wire was inserted into one stem of the tee to supply the electrical connection. A fused silica capillary (FSC) transfer line from an HP1100 HPLC pump was inserted into the second stem, and into the last stem was inserted the fritless microcapillary column. The column was prepared by taking a 100-μm × 365-μm FSC needle (pulled with a Sutter Instruments P-2000 laser puller) and packing 10-μm reversed-phase POROS particles to a depth of 10 cm at 300−400 psi using a stainless steel high-pressure bomb. A 150-μL/min flow from the HPLC solvent delivery system (HP 1100, Hewlett-Packard, Palo Alto, CA) was reduced using a splitting tee to achieve a flow rate of 400 nL/min. Spray voltage was 1.8 kV. Sample was loaded onto the column through pneumatic infusion of the sample from an eppendorf tube. The volume loaded was monitored by measuring the volume of solution displaced from the column.

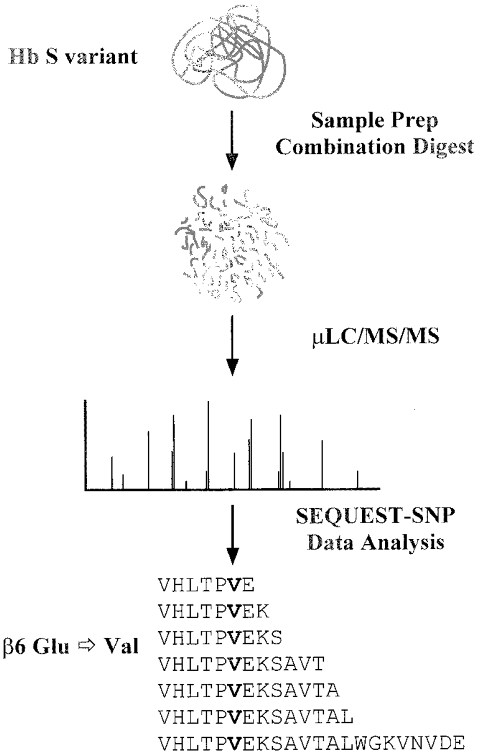
Microcapillary HPLC-Mass Spectrometry. A thirty-minute gradient of 0−60% solvent B (A: 0.5% HOAc, B: 80% MeCN/0.5% HOAc) was selected for separation of Hb peptides from the combination digest. Peptide analyses were performed on a Finnigan LCQ ion trap mass spectrometer (Finnigan MAT, San Jose, CA). The heated desolvation capillary was held at 200 °C, and the electron multiplier was set to 1.0 kV. Spectra were acquired in automated MS/MS mode with a relative collision energy (RCE) for CID preset to 35%. During an automated run, if an ion was present in a scan above a specified threshold, a product-ion spectrum was acquired. The mass spectrometer continued to alternate between MS (3 μscans per scan) and MS/MS mode (5 μscans per scan) until the ion intensity dropped below the threshold. The scan range for MS mode was set at m/z 375−1200. A parent ion default charge state of +2 is used to calculate the scan range for acquiring tandem mass spectra. Further details are provided in the Results.

SEQUEST-SNP Data Analysis. Automated analysis of peptide tandem mass spectra was performed with a modified version of the SEQUEST computer algorithm, which allows for identification of amino acid substitutions.[25] A database, comprising one or many genes, is selected to be searched. The SEQUEST-SNP program reads this database, dynamically generates all possible single-nucleotide polymorphisms (SNPs), translates these sequences to peptides, and analyzes these peptides using the process of Eng et al.[25] The program is run on a DEC Alpha 5000 computer and requires 1−2 s to search all SNPs for the hemoglobin α and β genes.

RESULTS

Amino acid sequence variations were identified by using μLC-MS/MS to generate tandem mass spectra to cover 99+% of the protein's sequence. Tandem mass spectra were analyzed using a modified version of the SEQUEST computer algorithm to determine the site and type of sequence variation (Scheme 1). To obtain consistent 100% sequence coverage for hemoglobin, three processes were optimized: sample preparation, chromatographic separation of the digested protein mixture, and data collection on an LCQ ion-trap mass spectrometer.

Scheme 1. Approach Used to Identify Amino Acid Sequence Variations in Proteins[a]



[a] The protein is first digested in three separate aliquots with two proteases that cut at specific amino acid residues and then a third protease with minimal specificity. The collections of peptides are combined into a single fraction and then analyzed using microLC and tandem mass spectrometry. The resulting collection of tandem mass spectra is analyzed using a modified version of the SEQUEST computer algorithm to identify the sites of amino acid sequence variation.

	1 10 20 30 40 50
α-globin	VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKL LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
β-globin	VHLTPEEKSAVTALWGNVNVDEVGGEALGRLLVVYPWTQRFFESFGDLST PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDP ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH

Figure 1. Primary amino acid sequences of the α and β chains of hemoglobin.

Sample Preparation. Hemoglobin proteins from human blood samples were isolated and purified as described in the Experimental Section. The amino acid sequences of the α and β chains of Hb are shown in Figure 1. Simple tryptic cleavage followed by μLC-MS/MS rarely or never covered certain areas of the sequence (highlighted in the text). For example, two tryptic fragments never observed were AHGK (at position β62−65) and GHGK (α57−60). These short His-containing peptides act as coordinating ligands to the Fe−heme complex and were most likely washed off the LC column during loading of the sample and not detected by MS. Removal of the porphoryrin group by precipitating the proteins using cold acidic acetone followed by digestion and μLC-MS/MS analysis led to detection of these tetramers.

Hemoglobin contains three cysteines (two on the α chain, one on the β chain). Sequence coverage in these regions was difficult to obtain since disulfide bonds can prevent protein denaturation, leading to incomplete proteolytic digestion. As expected, reduction and alkylation of the apohemoglobin significantly improved recovery of Cys-containing peptides. The reduced and (S)-carboxyamidated form of apohemoglobin was used for all further experiments.

Tryptic digestion of Hb, chromatographic separation, and data collection (see below) would typically yield 30−35 peptides resulting in 92−96% sequence coverage. This would be more than adequate if the goal were simple protein identification, for which one to five peptides are usually needed. However, complete sequence coverage is required to identify all possible amino acid changes resulting from SNPs. A tryptic cleavage pattern resulting in 95% sequence coverage would not guarantee unambiguous determination of amino acid substitutions. In addition, there are usually areas of individual peptide CID spectra which do not have the necessary fragment ions to unambiguously assign an amino acid to a specific site. By using molecular weight alone to identify a sequence variation, the specific site will not be identified if there are two or more amino acid changes that result in the same mass deviation.

To ensure complete and overlapping sequence coverage, a combination digest was devised. Three proteolytic digestions (trypsin, endoproteinase Glu-C, and subtilisin) were performed separately, and the resulting peptides combined into one tube. Trypsin cleavage was performed for 5 h. Less sequence coverage was obtained using a Glu-C digest in ammonium bicarbonate then was observed with trypsin cleavage. Glu-C digestion produced longer peptides, which often contained internal basic amino acid sites resulting in the formation of ions of charge state greater than 2. Fragmentation of highly charged peptides is often more difficult, resulting in poorer correlation of tandem mass spectra with sequences. However, Glu-C digestion carried out in phosphate buffer will cleave at both Glu and Asp residues. By performing the Glu-C digest overnight in sodium phosphate buffer (pH 7.4), sequence coverage was greatly improved (from approximately 38/33% in nonphosphate buffer, to 51/74% in the α and β chains, respectively). This 51/74% (α/β) coverage was from 22 tandem mass spectra matching to sequences with good correlation scores. Subtilisin is a nonspecific highly active protease, which produces very small fragments (1−5 amino acids in length) that are not as informative for on-line μLC-MS/MS sequencing. To produce longer peptides, a time-limited subtilisin digest (1 h) was performed in a Tris/2M urea buffer. Because of the nonspecific nature of the protease, many overlapping peptides were obtained (91/88% coverage from 59 peptides) with an average length of 11 amino acids. Interestingly, the number of peptides obtained on average from a subtilisin digest was greater than peptides obtained from trypsin and Glu-C digests combined. When these digests were combined into one tube, followed by a single μLC-MS/MS analysis, 100% sequence coverage was achieved about half of the time, with 97% or greater coverage obtained in all attempted runs. Approximately 75−80 peptides with overlapping sequences were generated per run using automated SEQUEST database searching. The ability to obtain complete or nearly complete sequence coverage for hemoglobin is greatly improved with this procedure.

Chromatographic Separation. A microelectrospray interface was utilized for on-line microcapillary-HPLC-MS/MS analysis of Hb peptides. Figure 2 shows the chromatographic trace of a combination digest of the hemoglobin variant Hb-Hope. In addition to the 4 pmol of digested Hb, 10 pmol of intact hemoglobin protein (dehemed and alkylated) was loaded on-column to simultaneously measure the molecular weight of the intact subunit. In a 30-min gradient of 0−60% buffer B with a microspray flow rate of 400 nL/min, peptides eluted over a 14-min window. There was a lag of about 2 min before the intact α and β chains eluted. These chains were conveniently separated from one another to obtain molecular weight (MW) information. This information added additional data to support amino acid substitution patterns (inset in Figure 2). The molecular weight data indicate that the substitution is localized to the β chain since the charge state envelope results in a MW of 16 041 u (±2 u). This molecular weight is 58 Da (±2 u) higher than that of the normal hemoglobin β chain (MW 15 983 u, alkylated), and the MS/MS data identified a β136Gly→Asp transition. Toggling between MS and MS/MS data results in a spiky appearance in the chromatographic trace because the total ion current decreases in MS/MS mode.

Data Collection. Normal human hemoglobin (Hb A) was utilized for the optimization of data-collection parameters. Automated tandem mass spectrometry was carried out on a Finnigan LCQ ion-trap mass spectrometer. A computer-controlled MS/MS experiment allows for automated toggling between MS and MS/MS modes of operation on a time scale conducive to acquiring fragmentation data as ions elute from an LC column. To increase data-acquisition efficiency, the additional parameters of "*n*th most intense ions," dynamic exclusion, and isotopic exclusion were incorporated into the autoMS/MS procedure.

For a complex mixture of peptides, coelution is often a problem. Less abundant peptides can be missed for selection to obtain CAD (collisionally activated dissociation) spectra when more abundant peptides are eluting. The "*n*th most intense ion" settings in the LCQ software help to circumvent this problem by collecting CAD spectra from a preselected number of the most abundant ions in an MS scan. For collection of Hb data, 6 scan events were chosen. The first scan event acquires an MS scan following which the five most abundant ions in the MS scan are acquired. This cycle was repeated throughout the acquisition. Dynamic exclusion was also used to prevent reacquisition of tandem mass spectra of ions once a spectrum had been acquired for a particular m/z value. Four dynamic exclusion parameters were optimized for μLC separation of peptides: repeat count, repeat duration, exclusion duration, and exclusion mass width. Values chosen were 2, 0.35 min, 1 min, and 3 u, respectively. Very often, the second most abundant peak in an MS scan is the ^{13}C isotope of a predominant peptide. The isotopic exclusion function allows for dismissal of the ion associated with the ^{13}C isotope of peptides from the list of ions slated for MS/MS. A 3-u mass width window was selected for this purpose. Altogether, acquiring auto-MS/MS data with the additional three parameters as described above ("top 5 ions", dynamic and isotopic exclusion) leads to a dramatic increase in the number of new peptide ions selected for CAD. These changes translated into a protein sequence coverage

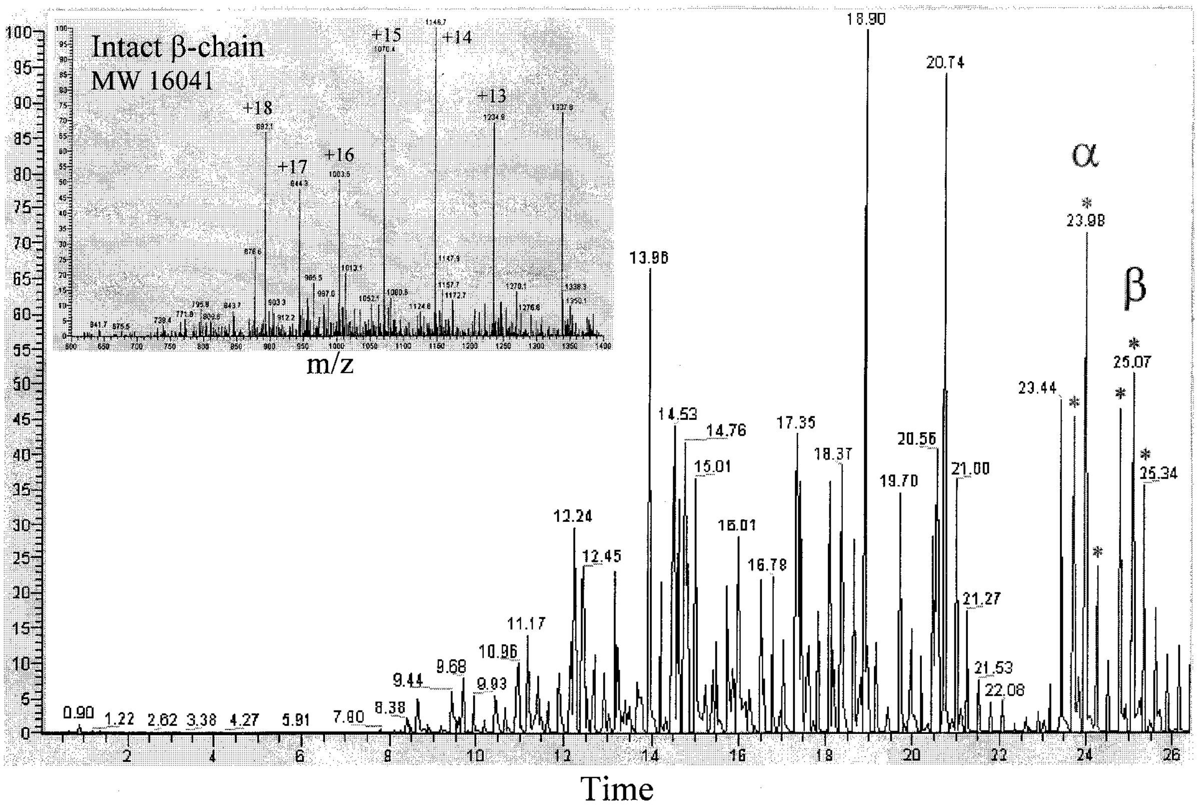


Figure 2. Reconstructed ion chromatogram (RIC) of Hb Hope combination digest and intact Hb with μLC-MS/MS analysis. Inset shows charge state profile of intact β-chain at RT 25.07 min.

Table 1. Percent Sequence Coverage of Hb Variants

variant	α chain (percent)	β chain (percent)
Hb C	100	100
Hb E	100	100
Hb D-Los Angeles	100	100
Hb G-Philadelphia	100	100
Hb Hope	100	100
Hb S	97	100
Hb S[a]	100	100

[a] From Sigma Chemical Co., Inc.

Table 2. Hemoglobin Variants from Human Patients

variant	substitution	MW α chain	MW β chain
Hb A[a]	normal	15183	15983
Hb C	β6(A3)Glu→Lys (−1 u)	Same[b]	Same[b]
Hb E	β26(B8)Glu→Lys (−1 u)	Same[b]	Same[b]
Hb D−Los Angeles	β121(GH4)Glu→Gln(−1 u)	Same[b]	Same[b]
Hb G−Philadelphia	α68(E17)Asn→Lys (+14 u)	15199 ± 2	Same[b]
Hb Hope	β136(H14)Gly→Asp (+58 u)	Same[b]	16041 ± 2
Hb S	β6(A3)Glu→Val (−30 u)	Same[b]	15952 ± 2
Hb S[a]	β6(A3)Glu→Val (−30 u)	Same[b]	15952 ± 2

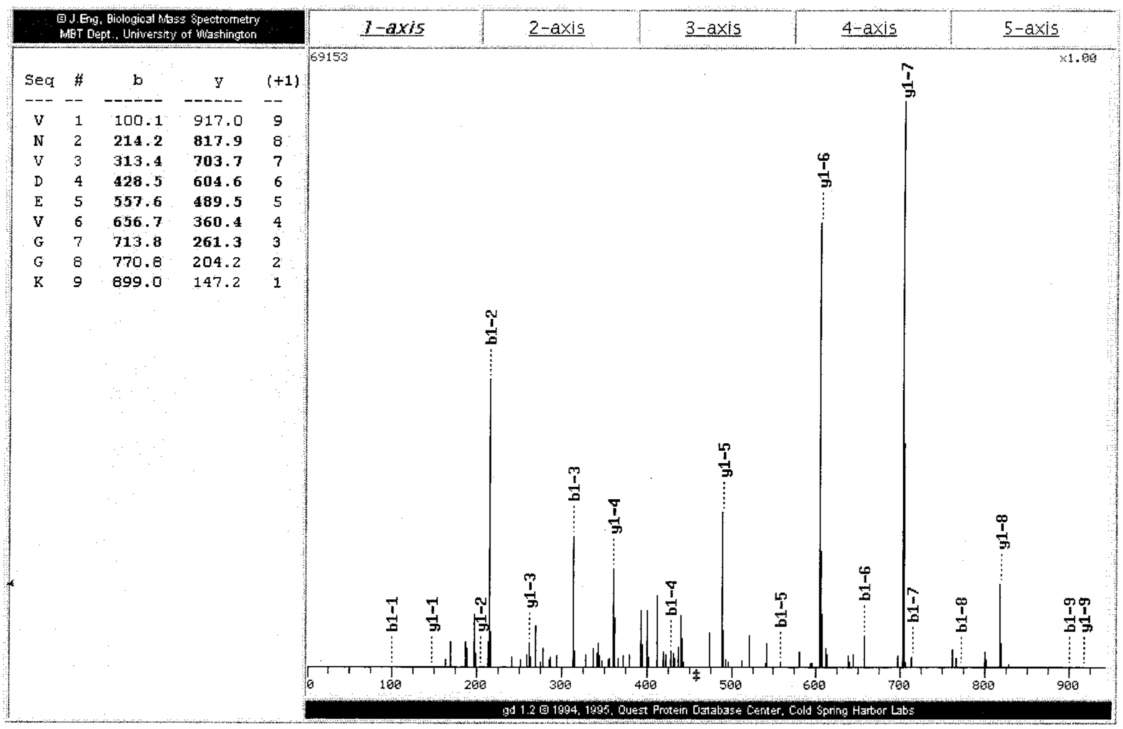
[a] Exceptions from Sigma. [b] Molecular weight of intact Hb chains, "same" as Hb A to within ± 2 u.

increase for trypsin-digested hemoglobin from 70−75% to 90−95%.

Data Analysis. The SEQUEST-SNP data analysis program identifies amino acid substitutions by searching against a gene or genes of interest. All possible SNPs are generated, translated to protein sequences, and searched using SEQUEST. Six human Hb variants were obtained as blind samples from the Red Cell Laboratory, University of Washington School of Medicine (Table 1). Following sample preparation and μLC-MS/MS analysis, the data was analyzed with SEQUEST-SNP. One hundred percent sequence coverage was obtained for all variants except the β chain of Hb S, although a Hb S (sickle cell) sample obtained from Sigma Chemical Co. did, however, give 100% coverage in both globin chains. Correct identification of the sequence variations was achieved for all variants (Table 2). The molecular weights of the α and β chains listed in Table 2 gave additional information on the location of amino acid variation except in the case of Hb C, Hb D, and Hb E since mass accuracy is only ±2 u for proteins >30 kD.

For Hb S, seven peptides (shown in Scheme 1) confirmed the identity of a β6Glu→Val substitution. The combination digest resulted in extensive overlap in this region of sequence. One peptide resulted from tryptic cleavage, two from Glu-C digest, and five from subtilisin. Three peptides were identified for Hb Hope (YQKVVA**D**VA, VVA**D**VANALAHK, and VVA**D**VANALAHKYH). These peptides, supported by the +58 u mass shift in the β chain, localized the transition to β136Gly→Asp. The Hb G-Philadelphia variant has a substitution of Asn for Lys at position 68 in the α chain. This created an extra tryptic site in the protein, resulting in the observed peptides VADALT**K** and KVADALT**K** (α62−68 and α61−68, respectively). The presence of (**K**)AVAHVDDMP-NALSALSDLHAHK (α69−90) provided indirect confirmation since this peptide is never detected from a combination digest of normal hemoglobin.

Hb E β26 Glu→Lys VNVDEVGGK



© J.Eng, Biological Mass Spectrometry
MBT Dept., University of Washington

```
Seq  #     b        y     (+1)
---  --  ------   ------   --
 V   1   100.1    917.0    9
 N   2   214.2    817.9    8
 V   3   313.4    703.7    7
 D   4   428.5    604.6    6
 E   5   557.6    489.5    5
 V   6   656.7    360.4    4
 G   7   713.8    261.3    3
 G   8   770.8    204.2    2
 K   9   899.0    147.2    1
```

Figure 3. Tandem mass spectrum of Hb E peptide indicating β26Glu→Lys transition.

The mass accuracy of the LCQ ion trap is sufficient to differentiate amino acid substitutions with a one unified atomic mass unit difference. Nevertheless, additional evidence is recommended before assigning a substitution site. For Hb E, substitution of Lys for β26Glu (loss of 1 u) in the tryptic peptide VNVDEVGG**K** (β18−26) was determined. The CAD spectrum is shown in Figure 3 with SEQUEST-SNP output. Indirect evidence is provided by the lack of detection of the normal VNVDEVGGEALGR (β18−30), which is routinely detected in the analysis of Hb A. The same logic was used to confirm the one unified atomic mass unit mass difference substitutions for Hb C and Hb D-Los Angeles, i.e., presence of certain peptides and lack of others. Further evidence can come from what is predicted from IEF gel migration or patient symptoms.

DISCUSSION

A rapid, automated method has been developed for obtaining 100% sequence coverage (in the majority of cases and over 97% coverage in all attempted analyses) in hemoglobin protein in a single LC-microspray tandem mass spectrometry experiment. Achieving nearly consistent 100% coverage ensures the identification of an amino acid substitution arising from a SNP. The amino acid substitution site was correctly identified for all of the six human Hb variants analyzed in this study. Hemoglobin was chosen as a model since ~90% of the over 700 Hb abnormalities known are due to single amino acid substitutions.[31] In addition, clinical screening of newborns for hemoglobin-based illnesses has been increasing. Worldwide, an estimated 150 million people carry Hb variants. Soon, molecular biology and mass spectrometry will play an expanding role for routine Hb diagnosis as recently reviewed by Shackleton and Witkowska.[32] This automated method for rapid determination of human hemoglobin variants represents an attractive approach for clinical screening.

Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF)[33] and electrospray methods of mass spectrometry have been used successfully for Hb variant characterization.[30,34,35] The high-mass accuracy of MALDI-TOF (0.01%) allows for accurate mass measurement of intact Hb chains. However, for a truly unknown sample lacking any patient information or corroborating evidence, the exact location of an amino acid substitution in a protein can only be estimated. Fragmentation information from ES-MS/MS analysis of Hb tryptic peptides can correctly pinpoint a substitution location as has been demonstrated.[30,34,35] These previous reports use more than one ES-MS analysis to pinpoint the substitution site: mass measurement (MS) of tryptic peptides to locate the one abnormal peptide, then acquiring fragmentation data (MS/MS) of the one peptide to locate the site of substitution in the sequence.

In this study, to decrease analysis time, amino acid substitutions were determined in a *single* μLC-MS/MS run with optimized conditions of sample preparation, chromatography, and automated data collection. Two major factors contributing to the attainment of nearly consistent 100% sequence coverage were the development of a combination digest allowing for multiple overlaps in

[31] Center, I. H. I. *Hemoglobin* **1997**, *21*, 507−602.

[32] Shackleton, C. H.; Witkowska, H. E. *Anal. Chem.* **1996**, *68*, 29A−33A.

[33] Houston, C. T.; Reilly, J. P. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1435−9.

[34] Deon, C.; Prome, J. C.; Prome, D.; Francina, A.; Groff, P.; Kalmes, G.; Galacteros, F.; Wajcman, H. *J. Mass Spectrom.* **1997**, *32*, 880−7.

[35] Lippincott, J.; Hess, E.; Apostol, I. *Anal. Biochem.* **1997**, *252*, 314−25.

sequence and the “top 5 ions” option in the automated data collection parameters of LCQ software which dramatically increased the number of peptides selected for CAD. Data analysis was simplified and enhanced by the development of a modified version of the SEQUEST algorithm, SEQUEST-SNP. This program rapidly identifies amino acid substitutions in a gene of interest by dynamically creating and translating a nucleotide database of SNPs. Future work will extend this method for the analysis of other proteins to acquire 100% sequence coverage from other proteins and to determine the limits of detection for the acquisition of complete sequence coverage of rare proteins. Full sequence coverage would guarantee not only the position of amino acid substitution but the site of any posttranslational modifications as well.

The strategy outlined in this paper should be general and applicable to the rapid identification of amino acid sequence variations in any protein. Along with the high level of interest in establishing the phenotypic impact of polymorphisms at the nucleotide level, the effect of nonsilent coding polymorphisms on function, expression level, and turnover rate of the corresponding protein would need to be considered. This method provides a strategy to identify amino acid sequence variations at the protein level, but additional efforts are needed to establish the limits of detection of the method and to develop strategies to quantitate protein expression levels. The functional and physiological impact of SNPs in hemoglobin is well-established, and it will be of interest to study the consequences of sequence polymorphisms in proteins associated with increased risk of disease.

Received for review September 3, 1999. Accepted November 16, 1999.

AC991025N