Thulin CD, Taylor JA, Walsh KAMicroheterogeneity of human filaggrin: analysis of a complex peptide mixture using mass spectrometry. Protein Sci 5:1157-1164

ARTICLE *in* PROTEIN SCIENCE · JUNE 2008

Impact Factor: 2.85 · DOI: 10.1002/pro.5560050618 · Source: PubMed

3 AUTHORS, INCLUDING:

Craig D Thulin
Utah Valley University
34 PUBLICATIONS **1,124** CITATIONS
SEE PROFILE

Alex Taylor
Just Biothreapeutics
8 PUBLICATIONS **800** CITATIONS
SEE PROFILE

Protein Science (1996), *5*:1157–1164. Cambridge University Press. Printed in the USA.

Microheterogeneity of human filaggrin: Analysis of a complex peptide mixture using mass spectrometry

CRAIG D. THULIN, J. ALEX TAYLOR, AND KENNETH A. WALSH
Department of Biochemistry, University of Washington, Seattle, Washington 98195

(RECEIVED December 14, 1995; ACCEPTED March 28, 1996)

Abstract

Filaggrin is the product of posttranslational processing of the large, epidermal protein profilaggrin, which consists of 10 or more tandem filaggrin domains plus an amino and a carboxyl domain. According to fragmentary cDNA sequences, the filaggrin domains in the human protein vary at 40% of the amino acid positions; hence, mature filaggrin is a population of homologous but heterogeneous proteins, even within one individual. Available gene sequences give only a limited picture of the heterogeneity of human filaggrin protein because no complete human profilaggrin gene has been sequenced. Questions about the extent of heterogeneity of filaggrin within and between individuals have not been answered, nor have questions concerning the limited proteolytic cleavage of human profilaggrin that generates filaggrin in vivo.

In order to address these questions and to provide an analysis of the primary structure of human filaggrins, we employed various methods of mass spectrometry. The intact protein and a tryptic digest of the mixture of human filaggrins were examined by matrix-assisted laser desorption time-of-flight mass spectrometry. Tryptic digests of human filaggrin from single individuals were also separated and analyzed by liquid chromatography/mass spectrometry (LC/MS) (using electrospray mass spectrometry), and specific peptides were identified by tandem mass spectrometry (MS/MS). A robust data analysis program, Sherpa, was developed to facilitate the interpretation of both LC/MS and MS/MS. These experiments show that human filaggrin includes heterogeneity not yet seen in cDNA sequences, but that much structure is highly conserved. Interestingly, we found that the heterogeneity is conserved among individuals. An approximation of the regions linking filaggrins in human profilaggrin is developed. These investigations provide a unique test of the limits of tryptic mapping of complex mixtures using mass spectrometry.

Keywords: filaggrin; mass spectrometry; microheterogeneity; peptide mapping; profilaggrin

In the early days of protein sequence analysis, one of the great concerns among protein chemists was that sequence microheterogeneity would complicate the analysis of the covalent structures of proteins. These fears were laid to rest as it became known that polymorphisms in proteins are relatively rare. Moreover, contemporary methods of structure analysis require sufficiently small quantities of protein that one can often obtain all the material needed from a single individual, or even from a single organ or tissue. Because of its extensive heterogeneity, human filaggrin provides an exception to these generalities, and it has proven to be a challenge in terms of structural studies.

Filaggrin, the intermediate filament-aggregating protein of the epidermis, is the mature product of the precursor profilaggrin. In the various species studied, each profilaggrin molecule consists of 10–20 tandem filaggrin domains separated by short linker regions, all bordered by an amino-terminal and a carboxy-terminal domain that bear no homology to filaggrin (Presland et al., 1992). The protein is synthesized as profilaggrin, which is phosphorylated extensively, then posttranslationally processed, primarily by dephosphorylation and limited proteolysis, to produce the mature filaggrin protein (Dale et al., 1993). Although the tandem filaggrin domains are virtual repeats in rodents (97% identity in mouse [Rothnagel & Steinert, 1990] and 99.5% in rat [Resing et al., 1993]), human cDNA sequences coding for filaggrin repeats (more than a dozen from six different segments of

Reprint requests to: Craig D. Thulin, Vollum Institute for Advanced Biomedical Research, Oregon Health Sciences University, 3181 SW Sam Jackson Park Rd., Portland, Oregon 97201; e-mail: thulinc@ohsu.edu.

Abbreviations: LC/MS, liquid chromatography/mass spectrometry; MS/MS, tandem mass spectrometry; CID, collision-induced dissociation; PMSF, phenylmethylsulfonylfluoride; *m/z*, mass to charge ratio; MALDI-TOF, matrix-assisted laser desorption time-of-flight; MIM, multiple ion monitoring; BSA, bovine serum albumin; amu, atomic mass units; PSD, post-source decay.

profilaggrins) display variability at 40% of the amino acid residues (McKinley-Grant et al., 1989; Gan et al., 1990; Presland et al., 1992). Hence, filaggrin in a single human is really a population of homologous proteins. An additional source of heterogeneity is the presence of "ragged" (i.e., alternative) carboxy termini, observed in both rat and mouse filaggrin (Resing et al., 1985, 1993), and ragged amino termini, observed in rat and human filaggrin (Resing et al., 1993; Thulin & Walsh, 1995).

Resing et al. (1985) analyzed the covalent structure of mouse filaggrin using then standard techniques of proteolysis, HPLC separation, and Edman degradation. However, human filaggrin, even if isolated from a single individual, was expected to be a complex mixture of homologous, but heterogeneous proteins, making these methods inadequate (a theoretical digest based on 13 reported human filaggrin DNA sequences predicts 179 unique tryptic peptides, many of which have very similar sequences). In 1993, Resing et al. included the use of electrospray mass spectrometry to study the primary structure of rat filaggrin. We now report a similar strategy of liquid chromatography and on-line electrospray mass spectrometry to separate the tryptic peptides of human filaggrin and characterize them by mass. MS/MS was subsequently used to obtain sequence information from ions in the collected fractions. The two-dimensional nature of an LC/MS separation (in the chromatographic and m/z dimensions) allows analysis of much more complex digests than would be practical by conventional HPLC detection methods. The MS/MS technique offered the advantage that the sequence analysis was not limited by peptide purity or by the presence of blocking groups, in contrast to Edman sequencing techniques. Further information was gained using MALDI-TOF to analyze both the intact protein and the digest mixture.

Results

Human filaggrin is heterogeneous even within an individual (due to the variation among the tandem filaggrin domains of profilaggrin and variations at its termini). We attempted to define this microheterogeneity in filaggrins purified from foreskins of single individuals.

Molecular weight of human filaggrin

Electrospray MS of intact human filaggrin from a single individual fails to yield an interpretable mass measurement. LC/MS experiments with our preparations of human filaggrin show that the various isoforms of this protein are sufficiently numerous and variable to thwart attempts at mass measurement by a quadrupole but sufficiently similar to be unseparable by C_{18} reversed-phase HPLC (data not shown).

In contrast, proteins subjected to MALDI-TOF MS are only mono- or diprotonated and singly charged molecules of well over 100,000 amu can be analyzed (Hillenkamp & Karas, 1990). MALDI-TOF analysis of purified human filaggrin from a single individual displays a broad peak centered at m/z 34,144 ± 1,560 at half-height (Fig. 1). BSA, which was used to calibrate the MALDI-TOF instrument immediately prior to acquiring the spectrum of human filaggrin, had a width of approximately 800 m/z units at half maximal height (data not shown). The greater breadth of the filaggrin signal is indicative of the heterogeneity of the protein.

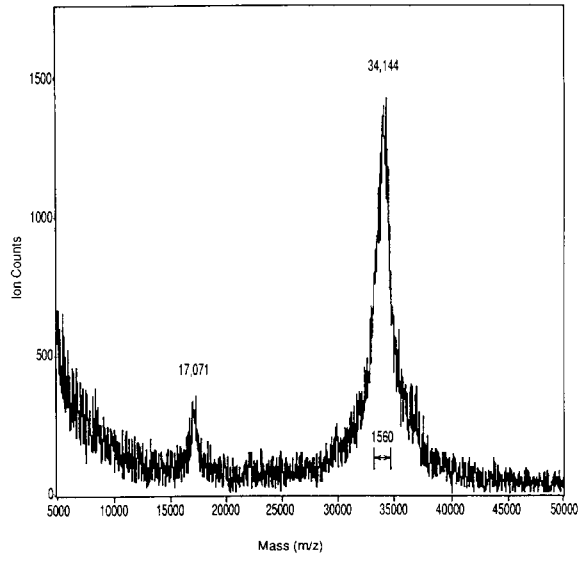

Fig. 1. MALDI-TOF mass spectrum of the mixture of human filaggrins. The protein is seen to give an unusually broad peak, due to heterogeneity.

Amino acid sequence of human filaggrin

In order to analyze the primary structures of the human filaggrin as an unseparated mixture, the protein was digested by trypsin. The resulting peptide mixture was then separated by reversed-phase and the effluent split so that 10% was analyzed in the mass spectrometer and 90% was diverted to fractions for later analysis. As seen in the two-dimensional display of the LC/MS data (Fig. 2), the peptides overlap greatly in the chromatographic dimension, and peptide purity in the fractions is therefore inadequate for interpretation of Edman degradation results. MS/MS, on the other hand, does yield sequence data on peptides in mixtures; however, these spectra are often difficult to interpret de novo, due to the incompleteness and/or complexity of fragmentation. If, however, one compares CID data with candidate sequences, it is often possible to correlate the spectrum with a given sequence with a reasonably high level of confidence. In this study, tentative identifications of the ion signals from an LC/MS separation of trypsinized human filaggrin were first made by using the Sherpa data analysis program (Taylor et al., 1996) to correlate them to masses of candidate peptides predicted from the available cDNA sequences. Collected fractions containing specific parent ions to be analyzed by MS/MS were then chosen based on these identifications. The fractions were infused into the mass spectrometer and CID spectra were collected (mass resolution of parent ion is 2 m/z). The MS/MS analysis feature of Sherpa was then employed to aid in the correlation of CID spectra with potential sequences. Figure 3 shows an example of a CID spectrum and its assignment to a filaggrin sequence.

Figure 4 illustrates a consensus human filaggrin sequence and the nomenclature for the tryptic peptides of human filaggrin. The 42 human filaggrin peptides tentatively identified by LC/MS and confirmed by MS/MS are listed in Figure 5. Of the 32 ion groups (ions in two or more consecutive charge states that coelute chromatographically) that Sherpa tentatively correlated to predicted filaggrin peptide sequences, more than two thirds

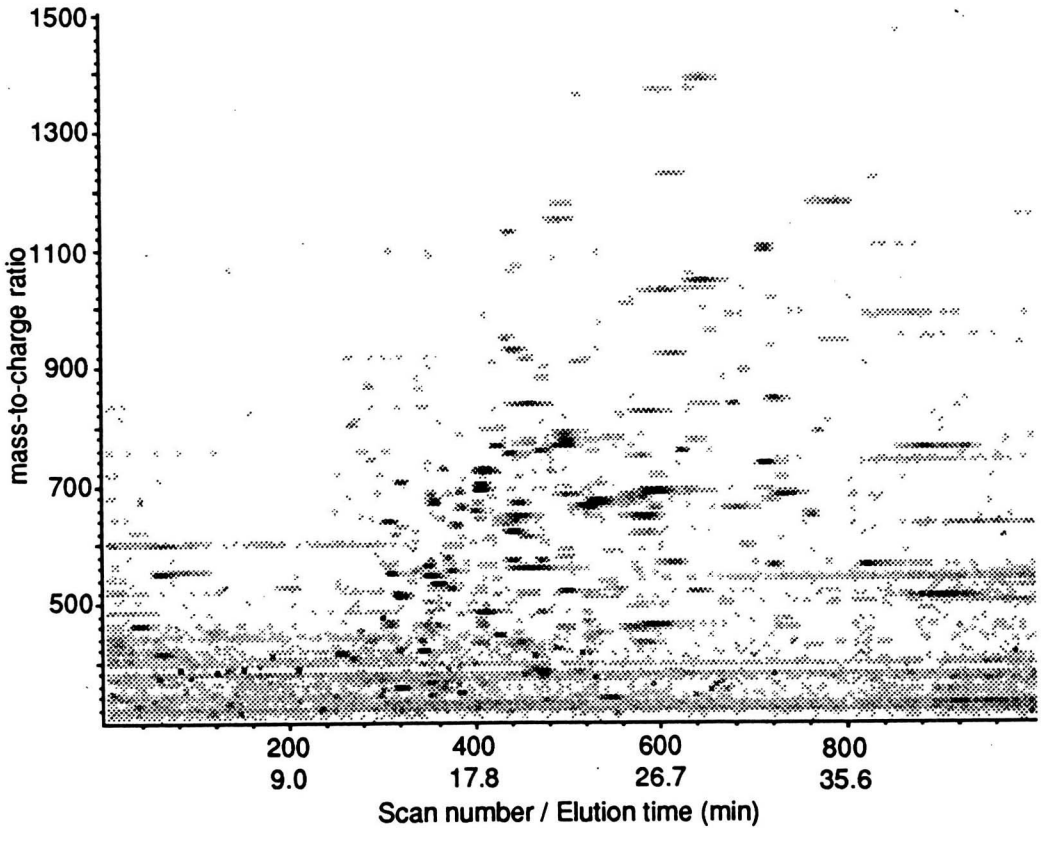

Fig. 2. Two-dimensional contour plot of an LC/MS data set from a tryptic digest of human filaggrin. The *y* axis displays the mass/charge ratio of ions, the *x* axis is time of elution (lower number) or scan number (upper number). Note that, at any given time, more than one ion is eluting.

were confirmed by MS/MS. However, of the peptides matched by Sherpa to single ions in the data set, 73% were false-positive identifications. In addition to identifying peptides predicted from human filaggrin cDNAs, six novel sequences, deduced by manual CID interpretation, are displayed in Figure 5, none of which correspond exactly to any peptide predicted from the cDNA sequences. Homologous relationships between these novel peptides and human filaggrin tryptic peptides predicted from the cDNA's are illustrated in Figure 6. The discovery of these unique sequences in more than 14% of the filaggrin peptides indicates that the diversity of expressed human filaggrin is even richer than had been suggested by the existing cloned cDNA sequences.

Four of the tryptic peptides were each recognized as a single sequence variant (F2, F5, F11, and F18; see Figs. 4, 5). Although three of these are moderate in length (less than 10 residues), one (F18) covers a region of 40 amino acids. Three ions coeluted at 28.5 min, corresponding to the $(M+H_3)^{3+}$, $(M+H_4)^{4+}$, and $(M+H_5)^{5+}$ forms of a peptide with a mass of 4,142.2 amu, which corresponds well to the calculated mass of 4,142.0 for the peptide GYSGSQASDNEGHSEDSDTQSVSAHGQAGSHQQ SHQESAR (F18, see Fig. 4). None of the signals for these ions were very intense, although at least two of them were found in all human filaggrin digests examined. As shown in Figure 7, this peptide gave poor CID spectra. No fragments were detected from the largest of these ions, *m/z* 1,382.3. Three b series ions seen in the CID spectrum of the 829.5 parent ion (see Fig. 7) were in accord with an N-terminus of GYSGS, and some other fragments are consistent with the identification of this peptide as F18. Most of the fragments generated from the other ion at *m/z* 1,037.5 are not identified easily, although a fragment may represent the doubly charged y_{10} ion.

Of the 100 most intense ions in the LC/MS digests of human filaggrin, 71 of them (comprising nearly 77% of the total ion signal) correspond to the filaggrin peptides shown in Figure 5. There are other ions that we were unable to identify with filaggrin or as autolysis products of trypsin, even by analysis of CID spectra. Some may represent novel filaggrin peptide sequences, although they may be minor contaminants of the filaggrin digest.

MALDI-TOF analysis of the digest

The MALDI-TOF mass spectrum of the digest mixture is shown in Figure 8. Many of the peptides identified in the LC/MS experiments were found in the MALDI-TOF spectrum. Both the resolution and the mass accuracy of MALDI-TOF in the linear mode (without reflectron) are less than those of electrospray/quadrupole analysis. Nonetheless, some species that are apparently refractory to electrospray MS may be observable by MALDI-TOF. For example, an ion signal was seen at *m/z* 2,934.7, which may correspond to the peptide QGYHHEHSVD SSGHSGSHHSHTTSQGR (calculated MW = 2,936.9), a variant of tryptic fragment F4 predicted by a human filaggrin cDNA

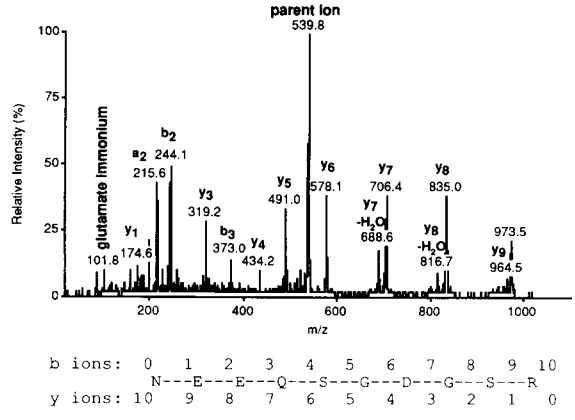

Fig. 3. A CID spectrum for the filaggrin peptide NEEQSGDGSR, shown as an example of MS/MS data and its interpretation. The doubly charged parent ion, as well as the b, y, and immonium ions, are all indicated (for ion nomenclature see Biemann, 1990). The Sherpa program assigned a score of 0.719 to this spectrum. This score is the fraction of the total intensity (disregarding the parent ion and sequence-independent parent derivatives) that can be accounted for by CID fragments predicted for the peptide. Sherpa scores for other peptides identified in this study range from 0.470 to 0.916.

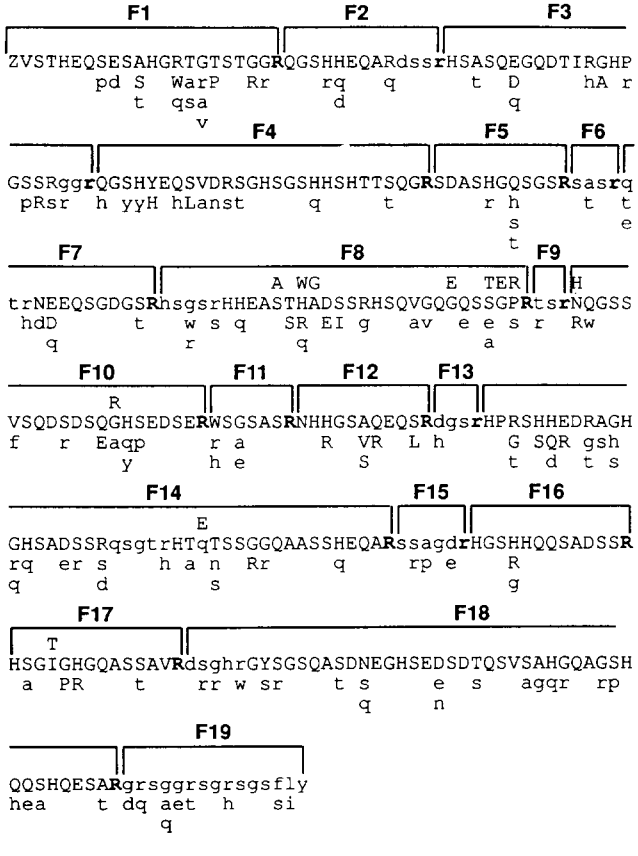

Fig. 4. A consensus sequence of human filaggrin (see text for references for cDNA sequences), with all alternative amino acids listed below. The peptide nomenclature scheme (F1–F19) used in these studies is based on the conserved (boldface) arginine residues (there is no lysine in filaggrin). Lower case residues are reported in cDNA sequences but not observed in these studies. Alternate residues predicted by cDNAs are displayed below the consensus sequence. Novel residues not reported in cDNA data are displayed above the consensus. Observed peptides account for 82.7% of the sequence.

peptide	version	sequence	MW calc.	MW obs.	# of clones
F1	F1.1	ZVSTHEQSESAHGR	1535.6	1535.3	9
	F1.1-2a	ZVSTHEQSESSHGWTGPSTR	2181.2	2181.0	2
	F1.1-2b	ZSESSHGWTGPSTR	1499.5	1500.0	2
	F1.2	rTGTSTGGR	735.8	735.6	1
F2	F2	QGSHHEQAR	1049.1	1048.4	5
F3	F3.1a	HSASQDGQDTIR	1314.3	1313.8	2
	F3.1b	HSASQEGQDTIR	1328.4	1327.8	8
	F3.2a	rGHPGSSR	696.7	696.4	7
	F3.2b	rAHPGSR	623.3	623.3	2
F4	F4.1a	QGSHHEQSVDR	1279.3	1278.6	1
	F4.1b	QGSHYEQSVDR	1305.3	1305.3	3
	F4.1c	QGSHYEQLVDR	1331.4	1331.2	2
	F4.2	rSGHSGSHHSHTTSQGR	1659.7	1659.5	10
F5	F5	SDASHGQSGSR	1088.1	1087.8	2
F6	***				
F7	F7.2a	NDEQSGDGSR	1063.4	1063.4	3
	F7.2b	NEEQSGDGSR	1077.4	1077.4	3
F8	F8.1a	HHEAASWGESSR	1353.4	1352.0	Novel
	F8.1b	HHEASTHADISR	1360.4	1359.8	1
	F8.1c	HHEASSR	822.8	822.8	2
	F8.2a	rHSQVGQGQSSGPR	1324.4	1324.0	Novel
	F8.2b	rHSQVGQEQSTER	1385.0	1386.7	Novel
F9	***				
F10	F10.1a	rQGSSVSQDSDSEGHSEDSER	2123.0	2122.8	2
	F10.1b	HQGSSVSQDSDSER	1518.5	1518.0	Novel
	F10.1c	NQGSSVSQDSDSQGHSEDSER	2236.1	2235.4	1
F11	F11	WSGSASR	749.8	750.0	5
F12	F12.1a	NHHGSAQEQLR	1276.3	1275.8	3
	F12.1b	rGSVQEQSR	889.9	889.8	1
	F12.1c	NHHGSSR	793.4	793.4	1
	F12.1d	rGSAQEQSR	861.4	861.6	5
F13	***				
F14	F14.1a	HPR	408.5	408.3	4
	F14.1b	HPGSSHR	776.4	777.0	1
	F14.2a	rSHHEDR	779.8	778.0	1
	F14.2b	rSHQEDR	770.8	770.0	3
	F14.3	rAGHGHSADSSR	1081.1	1080.6	4
	F14.4a	...rGQAASSHEQAR	1141.2	1141.2	1
	F14.4b	...rHTETSSGGQAASSHEQAR	1840.8	1840.2	Novel
F15	***				
F16	F16.1a	HGSHHQQSADSSR	1433.4	1432.7	5
	F16.1b	rHQQSADSSR	1015.0	1014.4	1
F17	F17.1a	HSGIPR	665.8	664.2	1
	F17.1b	HSGIGHGQASSAVR	1363.5	1363.0	7
	F171.c	HSGTGHGQASSAVR	1352.0	1351.1	Novel
F18	F18	GYSGSQASDNEGHSEDSDTQSVSAHGQAGSHQQSHQESAR	4142.0	4142.3	2
F19	***				

Fig. 5. The 42 human filaggrin peptides identified in these studies. Peptide nomenclature is shown in Figure 4. The peptide sequence is shown, as well as the number of times this peptide is predicted among the 13 reported human filaggrin domain cDNAs. Note the six novel peptides. Lower case "r" indicates an implied arginine residue immediately preceding the tryptic peptide found (there is no lysine in filaggrin). Triple asterisks indicate peptides not recovered. N-terminal Z residues in F1 denote pyrrolidone carboxyl groups identified by Thulin and Walsh (1995).

(P. Fleckman, P.V. Haydock, W. Nirunsuksiri, B.A. Dale, F. Grant, W. Kindsvogel, R.B. Presland, C. Blomquist, & S.G. Brumbaugh, in prep.). PSD cannot be done on an ion of this size in our instrument, hence it was not possible to confirm the identity of the peptide.

In agreement with the LC/MS data, an ion signal was observed at *m/z* 4,142 in the MALDI-TOF spectrum (although this peak was broad) and attributed to the F18 peptide. The breadth of the peak may indicate some heterogeneity in the sequence of this peptide. In addition to the signal found for the F18 peptide at *m/z* 4,142, there are two other components between *m/z* 4,000 and 5,000 that were not observed in the LC/MS data. The

```
F8.1b   HHEAASWGESSR
        ....STQADI..
        ....STHADI..
        ....S.RAD...
        ....STRA....

F8.2a   rHSQVGQGQSSGPR
F8.2b   rHSQVGQEQSTER
        ....AV.G..EGSR
        G......G..EGPR
        .......GE.SGSR
        .......G..AGSR
        G...A..GE.SGSR

F10.1b  HQGSSVSQDSDSER
        R............G
        NW...F......QG
        N....F...R..QA
        N........R..EG
        N....F......QG
        N...........QG
        R...........EA

F14.4b  rHTETSSGGQAASSHEQAR
        ..A.................
        ...Q.................
        ...Q...RR............
        H.A.N................
        ....S..R.............

F17.1c  HSGTGHGQASSAVR
        ...IPR........
        ...I..........
        ...I.R........
        ...I.R....T...
        .A.I..........
```

Fig. 6. Six novel human filaggrin peptides identified in these experiments (shown in boldface) are compared with homologous sequences reported in various cDNAs (see text for cDNA sequence references). Residues identical to novel sequences are shown as a period.

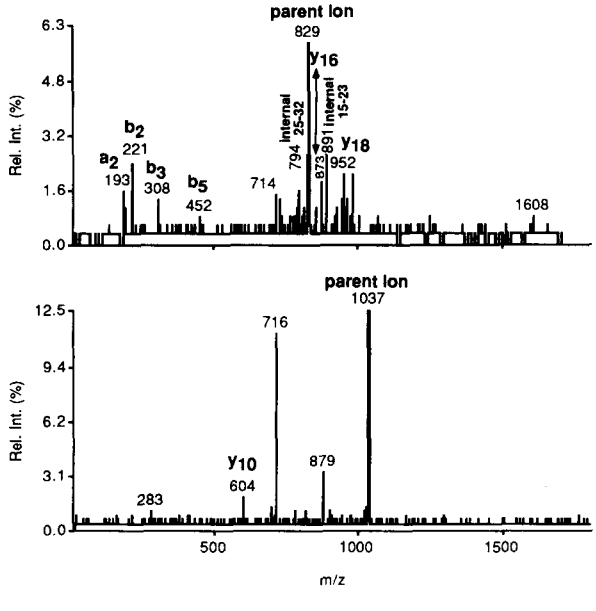

Fig. 7. CID spectrum of two ions (quintuply charged of m/z 829, and quadruply charged of m/z 1,037) believed to represent peptide F18 (of mass 4,142 amu). Indicated are the three b ions that identify the N-terminus of this peptide as GYSGS, as well as other ions believed to represent a or y series ions or internal fragments (y10, y16, and y18 fragments are doubly charged).

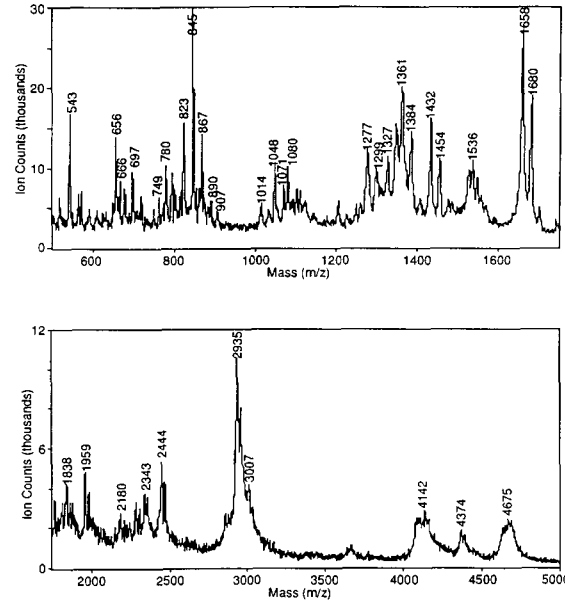

Fig. 8. MALDI-TOF spectrum of the unfractionated peptide mixture in a tryptic digest of human filaggrin. Indicated are the peptides that are also seen in the LC/MS data, listed in Figure 5 (except the peak at m/z 2,935, which is discussed in the text). The three ions of $m/z > 4{,}000$ are not seen in the corresponding LC/MS (Fig. 2) and their significance is discussed in the text.

small, broad peak at m/z 4,374 may represent the predicted peptide HWGSSGSQASDSEGHSEESDTQSVSGHGQAGPHQQ SHQESAR, a variant of tryptic peptide F18 (human cDNA; P. Fleckman, P.V. Haydock, W. Nirunsuksiri, B.A. Dale, F. Grant, W. Kindsvogel, R.B. Presland, C. Blomquist, & S.G. Brumbaugh, in prep.) that has a predicted mass of 4,372.3 amu. The third peak in this region of the spectrum, centering at m/z 4,675, does not correlate with any predicted tryptic peptide of any known human filaggrin variant. It may represent a variant of the F18 peptide or a product of incomplete digestion of an unreported variant of human filaggrin.

Comparison of filaggrin from single individuals

Unexpectedly, in view of the microheterogeneity of the filaggrins, the LC/MS patterns of tryptic digests of filaggrin purified from single individuals were virtually superimposable. All of the peptides identified were found in filaggrins from eight human foreskins analyzed. Quantitative comparison of specific ion signals were compared among several digests. Although the signal strengths varied, much of the variability was believed to be due to imprecision of detection during the 0.5-ms dwell time (time of detector residence at a particular m/z) necessary to scan the mass range frequently enough to enable on-line analysis of the effluent.

This limitation can be overcome during LC/MS by limiting the observations to preselected m/z windows and MIM. In this mode, m/z values (each within a narrow window) were preselected and the dwell time for data collection was increased to 10 ms. Figure 9 shows the results of such analyses in which human filaggrin digests from three individuals were each analyzed by LC/MS and MIM for seven ions corresponding to previously

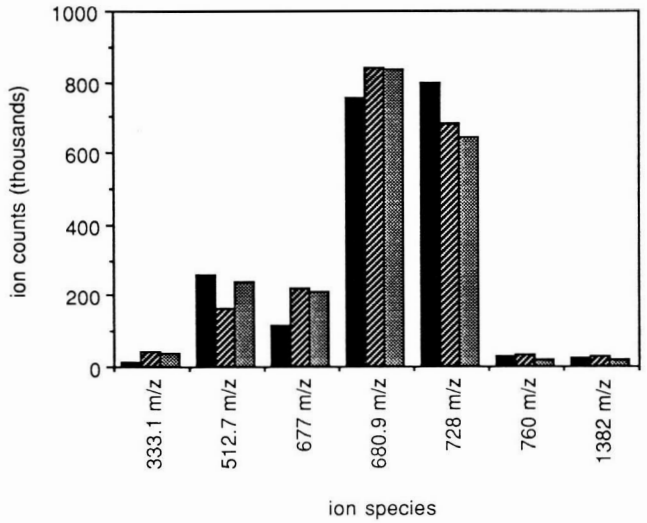

Fig. 9. Comparison of the normalized yields of selected molecular ions of specific human filaggrin peptides from single individuals. This comparison was made from LC/MS experiments using MIM with a 10-ms dwell time.

identified tryptic fragments (MIM limits the number of *m/z* values that can be selected). There is very little variation among the three runs (after normalization standard deviation ranges between 3.8 and 80.5 [units = thousands of ion counts] with an average of 33.5, see Fig. 9), but there is slightly more variability than seen with a similar comparison of a control tryptic digest of myoglobin (range of standard deviation following normalization is 1.5–63.9 with an average of 26.0; data not shown).

Analysis of human profilaggrin

An LC/MS analysis of a tryptic digest of human profilaggrin was also attempted. Because it is not possible to purify sufficient profilaggrin from a single foreskin sample to analyze by these methods, epidermal extracts from the tissues of several individuals were pooled. The similarity of filaggrin as isolated from different individuals (above) gave us confidence that profilaggrin purified from a population would not be significantly more heterogeneous than that of any given individual. All of the peptides found in filaggrin digests were also found in the digest of profilaggrin, except for the amino-terminal peptides of filaggrin. In addition to the filaggrin domains, profilaggrin has been reported to contain an amino-terminal domain that is homologous to the S100-like family of calcium-binding proteins (Presland et al., 1992). This region of the molecule is present at only one 10th to one 12th the amount of the filaggrin repeats, and the yield of its tryptic peptides thus provides a test of the sensitivity of the LC/MS experiments. Sherpa correlated eight potential ion groups to masses of tryptic fragments predicted from this region. One of these was confirmed by MS/MS, whereas CID spectra showed conclusively that five of the remaining seven were false-positives (data not shown). The one peptide that was identified positively has the sequence HENTSQVPLQESR, and it was found at 5–15% of the signal intensity of the most prominent filaggrin peptides. A curious and unexpected finding, however, was the identification and confirmation by MS/MS of the incompletely digested peptide SRHENTSQVPLQESR, which is found to be relatively abundant (on the order of other filaggrin ions) in the profilaggrin LC/MS. Several other ions that were observed in the LC/MS of the profilaggrin digest do not correspond to known filaggrin peptides or to peptides predicted from the profilaggrin amino-terminal calcium binding domain. These may be trace contaminants or filaggrin peptides that are phosphorylated or modified in other ways.

Discussion

Analyses of the primary structure of human filaggrin purified from single individuals confirm that filaggrin is a heterogeneous collection of highly similar proteins that are generated by the limited proteolytic excision of tandem domains from the profilaggrin precursor. Novel sequences, homologous to human filaggrin cDNAs, indicate that the diversity of this molecule has not been exhaustively defined in the published partial cDNA sequences. Despite this diversity, the heterogeneity is not sufficient to allow the reversed-phase separation of differing filaggrin molecules, or to clearly resolve them by MALDI-TOF MS. A single, broad peak at *m/z* 34,144 is seen in the MALDI-TOF spectrum of intact human filaggrin. Mass analysis of intact human filaggrin by electrospray methods is uninterpretable. Electrospray LC/MS of a tryptic digest of human filaggrin, however, yielded the masses of 42 filaggrin peptides that were identified and confirmed by MS/MS (see Fig. 5).

Electrospray mass spectrometry is not a quantitative method, because peptides and proteins vary in efficiency of ionization and detection in the instrument (Downard & Biemann, 1995). It can be asked whether the peptides identified represent all of the filaggrin peptides in the mixture. If a given peptide sequence is found in only one of the 10–12 tandem filaggrin repeats on one of the profilaggrin alleles in an individual, it would be at least 20-fold less abundant than a peptide that occurred in all filaggrins. Unfortunately, ion intensities do not correspond well to the abundance of a molecular species. For instance, there is more than a 20-fold difference in the intensity of the least and most abundant ions representing tryptic fragments in a digest of horse skeletal muscle myoglobin (data not shown), where the various molecular species should be present in the spray droplet in roughly equal amounts. Analysis of the LC/MS of a tryptic digest of human profilaggrin indicates that the majority of tryptic peptides from the amino-terminal calcium-binding domain (which is represented only once per allele) are not detected. We infer that a filaggrin peptide variant that is represented only once among the filaggrin loci within an individual might not be detectable; hence, most of the sequence variants that are reported in Figure 5 probably occur more than once in the tandem domains of each profilaggrin molecule.

Notably, no single reported cDNA sequence for the human filaggrin domain is represented by a full complement of peptides in these studies. Instead, the predominant peptides may reflect those most conserved despite the heterogeneity of the protein. Several of the peptides found correspond to gene sequences that are conserved in more than one of the cloned cDNAs (see Fig. 5); one, F4.2, is found in 10 of the 13 sequences available and is the only sequence variant of the region identified in these analyses. Perhaps the conserved features elucidated by these data will assist with the identification of structure/function relationships

in the filaggrin domain. Two other regions of the filaggrin domain (peptides F8.2 and F14.4) are represented only by peptides that are novel in sequence when compared with the cDNAs, indicating that these may be the least-conserved segments of the filaggrin molecule.

These data are not complete in terms of the identification of all of the peptide sequences in human filaggrin in spite of the application of the analytical power of LC/MS, the ability of MS/MS to examine peptide sequences in mixtures, and the data analysis capability of the Sherpa program. The experiments do demonstrate the power and the limits of LC/MS and MS/MS for the analysis of exceedingly complex mixtures in protein chemistry. As indicated in Figure 4, 83% of the consensus human filaggrin sequence was recognized among ions in the LC/MS tryptic map. Most of the predicted sequences that were not recovered as peptides are less than 700 amu, and their hydrophilicity may have prevented their retention on the reversed-phase matrix. The only missing predicted peptide with some hydrophobic character (SGSFLY) corresponds to the six residues in the linker segment SGSFLYQVSTHEQSESAHGR, which is recovered in abundance during LC/MS analysis of human profilaggrin. The glutamine residue in the QVS sequence is found (cyclized as PCA) at the amino terminus of filaggrin (Thulin & Walsh, 1995). Thus, the SGSFLY peptide may have been excised or shortened during proteolytic processing in vivo (see below).

An important structural determination made in these experiments is the delimitation of the linker region between filaggrin domains of human profilaggrin. We recently reported variations in the amino termini of human filaggrin (Thulin & Walsh, 1995), but the carboxy terminus of this protein has resisted identification. We applied four independent methods in an attempt to identify the carboxy terminus, namely: (1) treatment with carboxypeptidases (carboxypeptidase A, B, and Y were tried) to remove the carboxy-terminal amino acids, thus modifying the carboxy-terminal peptide; (2) chemical modification of the carboxy terminus (both amidation and methylation); (3) digestion in 60% ^{18}O-water, to label all but the carboxy-terminal peptide with this heavier isotope prior to mass spectrometry; and (4) chromatography of the tryptic digest on anhydrotrypsin immobilized on Sepharose, to scavenge all peptides except the C-terminus; yet none of these methods elucidated the carboxy terminus of the protein (Thulin, 1995). Three results from the present studies provide a limit to the size of the linker region and an approximate location of the carboxy terminus of human filaggrins within the profilaggrin sequence. First, by identifying the peptide F18 in filaggrin ending with Arg 309, the human profilaggrin linker region is limited to 15 amino acids. Second, the absence of peptides from any variant of the sequence GRSGGR SGRSSGSFLY suggests that a segment of this sequence ending in -FLY may be excised in the conversion of profilaggrin to filaggrin. Third, MALDI-TOF analysis shows the protein to have an average mass of 34,144 amu, which is just 221 amu less than the average mass predicted from the 13 cloned filaggrin domains (through -FLY), suggesting that the excised linker segment may be as short as a dipeptide. However, pulse-chase experiments by Gan et al. (1990) show that this Phe is absent from mature filaggrin, indicating that the missing linker must be at least the tripeptide FLY. The cDNA sequences each encode proteins of 324 amino acids, and calculated masses of these proteins range from 34,023 to 34,841 amu (it should be noted that it is unknown how representative the cDNA sequences are of expressed filaggrin proteins). An upper limit on the size of the excised region is 15 residues, because no part of the sequence DRSGGRSGRSGS FLY (or any F19 homologue) was identified among the tryptic peptides.

In view of the marked heterogeneity among human filaggrin domains, an unexpected finding of this study was that tryptic maps determined by LC/MS were invariant among individuals. This may be simply because the peptides observed are those that vary least from one individual to another. Additionally, there is no a priori reason why tandem sequences that differ greatly within an individual must differ as much between individuals. For example, globin loci display a tandemly repeated structure that varies a great deal from one repeated unit (a globin gene, be it a pseudogene or an expressed gene) to another, but the entire set of genes is relatively well conserved from one individual to another. This relates to the evolutionary history of these sequences, and the issue of the order of appearance of the variability and the duplication. As an alternative interpretation, perhaps a great deal of variability among human filaggrin sequences is allowed because of their redundancy, and selection constrains these sequences enough to result in only a relatively small number of significant peptide sequences among the products of any profilaggrin gene. It must be realized that the extent of conservation of the tryptic peptides does not speak directly to the conservation of whole filaggrin domains. It is even conceivable that the most abundant tryptic peptide sequences are shuffled about by unequal crossing-over events among highly heterogeneous filaggrin molecules. Judged by the heterogeneity of the clone data, this may be the case.

Materials and methods

Human filaggrin was purified and trypsinized as has been described previously (Thulin & Walsh, 1995).

Human profilaggrin was purified as follows: the epidermis from 30 to 40 foreskins (a gift from Dr. Philip Fleckman, U.W. Department of Dermatology) were separated as described previously (Thulin & Walsh, 1995) and homogenized two at a time in approximately 0.5 mL 9 M urea 50 mM Tris, pH 8.0, with a crystal of PMSF. The homogenates were combined and spun at 12,000 × *g* for 30 min. The supernatant was loaded on a DE52 ion exchange column that had been equilibrated with 9 M urea 50 mM Tris, pH 8.0, and a gradient of 0–0.4 M NaCl in this urea buffer was applied, collecting 1-mL fractions. Fractions containing profilaggrin were pooled. A metal chelating column of fresh iminodiacetic acid-sepharose 6B Fast Flow resin (Sigma) was preloaded with zinc using 20 mL of 50 mM $ZnSO_4$, then equilibrated with the urea buffer. The DE52 profilaggrin pool was then loaded and the column washed with 10 mL of 9 M urea at pH 8, followed by 10 mL of 9 M urea at pH 5. The profilaggrin was then eluted with 9 M urea at pH 3, and 1-mL fractions were collected. The purification was monitored by SDS-PAGE (4–15% gradient gel), and Western blots with an anti-human filaggrin monoclonal antibody, AKH1 (a gift of Dr. Beverly Dale). After zinc affinity chromatography, SDS-PAGE showed only a single protein band (by Coomassie staining), which migrated at >200 kDa; it reacted with the AKH1 antibody.

To digest human profilaggrin, 50 μL of the protein (approximately 1 mg/mL) in 9 M urea was first precipitated by adding 250 μL of 10 mM $MgCl_2$ for 10 min at 0 °C. The supernatant was discarded and the procedure was repeated in the same tube

four times. An additional 35 μL of the protein solution was added, as was 125 μL of 100 mM Tris, 1 mM $CaCl_2$, pH 8.0, and 5 μL of bovine trypsin (0.1 mg/mL in 1 mM HCl). After incubation at 37 °C for 5 h, some insoluble material was removed by centrifugation and the supernatant examined by LC/MS.

LC/MS and MS/MS followed procedures reported previously (Thulin & Walsh, 1995) using a PE Sciex API-III Plus triple quadrupole instrument with an Ionspray source. The MacSpec software provided by Sciex was used to visualize the data in the form of contour plots (see Fig. 2) in which ion signals are displayed in both the chromatographic and mass-to-charge dimensions. Many of the CID spectra for filaggrin peptides were taken from digests of human filaggrin pooled from a number of individuals. LC/MS MIM followed the same method, except that the dwell time was increased to 10 ms (from 0.5 ms) and specific ions were scanned with a window of 3 *m/z* as follows: *m/z* 333.1, 512.7, 677.0, 680.9, 728.0, 760.0, and 1,382.0. A standard digest of horse skeletal muscle myoglobin was monitored for comparison, using a tryptic digest of 100 pmol, and *m/z* values of 316.2, 636.3, 395.9, 471.5, 606.3, 790.5, 992.0, and 1231.7. Ion signals among LC/MS MIM experiments were normalized to one another using the average of all ions monitored.

MALDI-TOF data were collected on a PerSeptive Biosystems Voyager Elite in linear mode with the sample embedded in a sinapinic (3,5-dimethoxy-4-hydroxy cinnamic) acid matrix for the intact protein, or α-cyano-4-hydroxy cinnamic acid matrix for the peptide mixture. Accelerating voltages were 25,000 volts for the intact protein and 30,000 volts for the peptide mixture. External calibration used horse skeletal myoglobin and bovine serum albumin for protein and bradykinin for peptides.

LC/MS and MS/MS data were analyzed using the MacSpec software from PE Sciex, as well as an in-house program, Sherpa, written by one of us (J.A. Taylor, see Taylor et al., 1996). Sherpa was designed to aid in the analysis of LC/MS and MS/MS data from peptides derived from a protein of known sequence. To facilitate LC/MS interpretation, Sherpa identifies and relates *m/z* values observed in a data set to masses predicted from a theoretical digest of a given protein sequence. These identifications can then be verified by subsequent MS/MS analysis of an ion in the corresponding chromatographic fraction. Sherpa was also used to identify the most likely sequence among known filaggrin sequences that would account for the ions of a given CID spectrum, and to generate lists of the theoretical ion fragments predicted from known filaggrin sequences. Final analysis of each CID spectrum was completed by visual inspection. All masses reported here, both calculated and observed, are average isotope distribution (rather than monoisotopic masses). Preliminary beta versions of Sherpa are available on the World Wide Web <http://128.95.12.16/Sherpa.html>.

Acknowledgments

C.D.T. and J.A.T. were supported by individual Public Health Service National Research Service Awards T32 GM07270 from the National Institute of General Medical Sciences. We acknowledge helpful discussions on electrospray mass spectrometry from Dr. R.S. Johnson and Lowell Ericsson; on MALDI-TOF mass spectrometry from Dr. J.A. Kowalak; on filaggrin biology from Drs. R.B. Presland, E. Kam, P. Fleckman, and B.A. Dale; on analysis of variability from Dr. D.C. Teller and the University of Washington Department of Biostatistics; on the genetics of repeated tandem sequences from Dr. J. Felsenstein; and on mass spectrometry, filaggrin chemistry, and biology from Dr. K.A. Resing.

References

Biemann K. 1990. Sequencing of peptides by tandem mass spectrometry and high-energy collision-induced dissociation. *Methods Enzymol 193*:455-480.

Dale BA, Presland RB, Fleckman P, Kam E, Resing KA. 1993. Phenotypic expression and processing of filaggrin in epidermal differentiation. In: Darmon M, Blumenberg M, eds. *Molecular biology of the skin; the keratinocyte*. San Diego: Academic Press. pp 79-106.

Downard KM, Biemann K. 1995. Charging behavior of highly basic peptides during electrospray ionization. A predilection for protons. *Intl J Mass Spectr and Ion Processes 148*:191-202.

Gan SQ, McBride OW, Idler WW, Markova N, Steinert PM. 1990. Organization, structure, and polymorphisms of the human profilaggrin gene [erratum published in 1991 *Biochemistry 30*:5814]. *Biochemistry 29*(40): 9432-9440.

Hillenkamp F, Karas M. 1990. Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization. *Meth Enzymol 193*:280-295.

McKinley-Grant LJ, Idler WW, Bernstein IA, Parry DA, Cannizzaro L, Croce CM, Huebner K, Lessin SR, Steinert PM. 1989. Characterization of a cDNA clone encoding human filaggrin and localization of the gene to chromosome region 1q21. *Proc Natl Acad Sci USA 86*(13):4848-4852.

Presland RB, Haydock PV, Fleckman P, Nirunsuksiri W, Dale BA. 1992. Characterization of the human epidermal profilaggrin gene. Genomic organization and identification of an S-100-like calcium binding domain at the amino terminus. *J Biol Chem 267*(33):23772-23781.

Resing KA, Dale BA, Walsh KA. 1985. Multiple copies of phosphorylated filaggrin in epidermal profilaggrin demonstrated by analysis of tryptic peptides. *Biochemistry 24*(15):4167-4175.

Resing KA, Johnson RS, Walsh KA. 1993. Characterization of protease processing sites during conversion of rat profilaggrin to filaggrin. *Biochemistry 32*(38):10036-10045.

Rothnagel JA, Steinert PM. 1990. The structure of the gene for mouse filaggrin and a comparison of the repeating units. *J Biol Chem 265*(4): 1862-1865.

Taylor JA, Walsh KA, Johnson RS. 1996. Sherpa: A Macintosh-based expert system for the interpretation of ESI LC/MS and MS/MS of protein digests. *Rapid Commun Mass Spectrom*. Forthcoming.

Thulin CD. 1995. Posttranslational processing and human profilaggrin [dissertation]. Seattle, Washington: University of Washington.

Thulin CD, Walsh KA. 1995. Identification of the amino terminus of human filaggrin using differential LC/MS techniques; implications for profilaggrin processing. *Biochemistry 34*(27):8687-8692.