Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography B, 787 (2003) 63–76
JOURNAL OF CHROMATOGRAPHY B
www.elsevier.com/locate/chromb

Review

Proteomic database of wool components

Jeffrey E. Plowman*

Wool Research Organisation of New Zealand, Private Bag 4749, Christchurch, New Zealand

Abstract

The separation, classification and identification of wool fibre proteins has been of interest for many years. The purposes of this review are to summarise past work in this area and to evaluate the application of modern proteomic techniques to the identification and characterisation of wool proteins. The current state of knowledge of the wool proteome will also be presented.

Keywords: Reviews; Proteomics; Wool fibre proteins; Keratins

Contents

1. Introduction

Wool is composed largely of proteins, and these proteins are responsible for the major structural and mechanical properties of wool fibres. Since the first attempt to fractionate them in 1935 [1], the identification and characterisation wool keratin proteins have presented protein chemists with an enormous challenge. This is because they are the products of several gene families, each with a number of closely related members, and there is considerable compositional similarity between the different families. Nevertheless a great deal of progress has been made in the past 20 years to improve our knowledge and understanding of their structure and role in the fibre.

This review will discuss earlier research efforts to identify and characterise wool keratin proteins, and how the knowledge generated from this research relates to current efforts to identify and characterise

*Tel.: +64-3-325-2421; fax: +64-3-325-2717.
E-mail address: plowman@wronz.org.nz (J.E. Plowman).

1570-0232/02/$ – see front matter
doi:10.1016/S1570-0232(02)00211-8

these proteins using modern proteomic techniques. It will then examine the currently available approaches of matrix-assisted laser desorption ionisation time-of-flight (MALDI-TOF) mass spectral peptide fingerprinting and electrospray ionisation-tandem mass spectrometry (ESI-MS–MS) and the difficulties encountered in the identification of proteins from some of the strongly homologous wool keratin families.

2. Wool fibre structure

Wool fibres are generally composed of three different types of spindle-shaped cortical cells surrounded by a sheath of overlapping, roughly rectangular cells known as the cuticle, which forms the external layer of the fibre (Fig. 1) [2]. Approximately 90% of the cortical cell type is made up of longitudinally arrayed intermediate filaments (IFs) with accompanying matrix, the remainder being membranes and remnants from the nucleus and cytoplasm [2].

The basic building blocks of wool fibres are the fibrous, low-sulfur α-keratins that are part of the IF superfamily of proteins [3]. IF proteins (IFPs) are almost completely helical, consisting of four α-helical segments, linked by short stretches of sequences predicted to be non-helical [4]. The helical regions display a heptad (seven amino acid residue) substructure in which the first and fourth sites are generally occupied by apolar residues, resulting in a stripe of apolar residues winding its way around the right-handed α-helix in a left-handed manner. It is this apolar stripe that allows two different IFPs to line up with their helical segments in register to form dimers. From electron microscope studies it is considered that 16 of these dimers assemble to form an IF in which the dimers are arranged in a ring around a hollow core [5].

In contrast to the rod-like structure of the central region of the IFPs, the end domains are considered to have little or no tertiary structure [4]. These domains, which are rich in cysteine residues, are thought to project out of the IFs [6] where they are free to interact with the proteins found in the matrix of the fibre.

Matrix proteins are noted for their high content of either cysteine residues or glycine and tyrosine residues. The ones high in sulfur are referred to as either high sulfur proteins (HSPs) or ultra-high sulfur proteins (UHSPs), depending on their cysteine content, while those high in glycine and tyrosine are referred to as high glycine–tyrosine proteins (HGTPs). The matrix proteins are thought to surround the IFs at a later stage in the development of the follicle and to interact with them through inter-

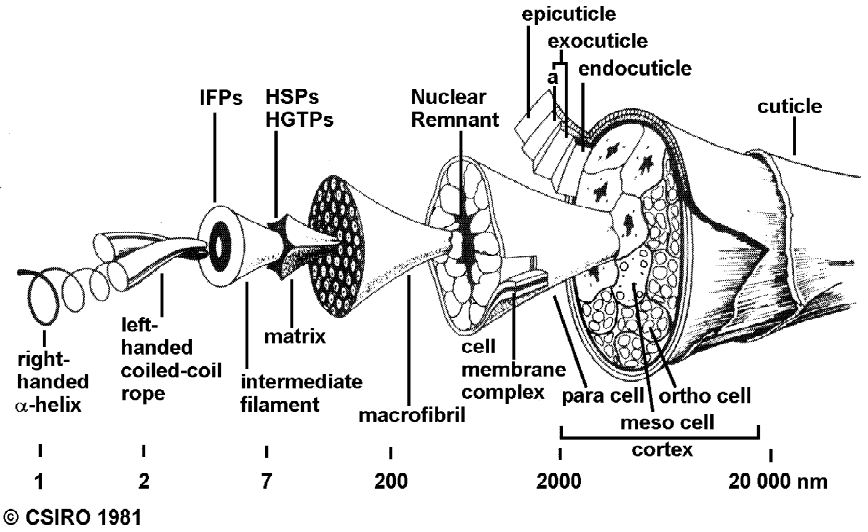

Fig. 1. Schematic diagram of a wool fibre showing the major structural features found in the cortical cells. (Reproduced with the permission of CSIRO Textile and Fibre Technology).

molecular disulfide bonds. Bundles of these IFs combined with matrix proteins form macrofibrils within the cortical cells [4]. While matrix proteins have little or no discernable effect on IF structure, their effect on IF assembly into larger arrays is considered to be crucial. HSPs, such as the B2 family, could be considered as large bifunctional cross-linking agents. It is, in fact, the formation of the cross-linked IF–matrix protein composite that gives the α-keratins their special mechanical attributes of strength, inertness and rigidity [4].

3. Protein composition of wool

The earliest attempt to identify and classify wool proteins involved the fractional salting out of s-carboxymethylated derivatives with ammonium sulfate into two classes of extractable proteins—class A and B—which were, respectively, lower and higher in sulfur than the original wool [7]. Application of urea polyacrylamide gel electrophoresis (PAGE) resulted in a further fractionation of the low sulfur class A into three components, 5, 7 and 8, of which component 7 was found to consist of three subcomponents (7a, 7b, 7c), while component 8 was originally thought to consist of five subcomponents (8a, 8b, 8c-1, 8c-2, 8c-3) [8]. However, a subsequent study showed that at least one component, 8c-3, was in fact an artifact of the separation system used [9]. Since then there have been a number of attempts to revise the classification system for keratin proteins. The proteins of component 8, which range in size from 392 to 416 amino acids, were later classified as the acidic Type I IFPs, while those of components 5 and 7, which range in size from 479 to 506 amino acids [9], were classified as members of the neutral basic Type II IFPs. More recently there has been an attempt to unify the nomenclature system, first for wool and hair proteins [10] (Table 1), and then to include keratin proteins from other animals such as mice and rabbits [11]. Thus the Type I IFP component 8c-1 is known as K1.1 in this new system. Vestiges of the previous naming systems still persist today, particularly in the web-based databases, where the IFPs are often referred to as microfibrillar proteins. Thus there are inherent dangers in trying to fit current knowledge of keratin proteins to these original studies, unless there is some appreciation of how keratin protein nomenclature developed and how the current system relates to the original.

While separation by PAGE resulted in the resolution of IFPs into discrete bands, the application of DEAE–cellulose chromatography was less successful, resulting in a crude separation of these proteins for which, in most cases, there was no baseline separation [12]. Nevertheless, sufficient quantities of pure protein material were isolated from some of these bands to enable the protein sequence determination of components 5 [13], 7c [14] and 8c-1 [15].

Table 1
Nomenclature of wool keratin families (Powell and Rogers [11])

New nomenclature	Abbreviation	Old nomenclature
Keratin IF protein		
KRT1.n	K1.n	IF Type I (LS 8a, 8b, 8c1, 8c2)
KRT2.n	K2.n	IF Type II (LS 5, 7a, 7b, 7c)
Keratin IF associated protein		
KRTAP1.n	KAP1.n	HS B2
KRTAP2.n	KAP2.n	HS BIIIA
KRTAP3.n	KAP3.n	HS BIIIB
KRTAP4.n	KAP4.n	UHS Cortex
KRTAP5.n	KAP5.n	UHS Cuticle
KRTAP6.n	KAP6.n	HGT Type II
KRTAP7	KAP7	HGT Type I C2
KRTAP8	KAP8	HGT Type I F
KRTAP10	KAP10	UHS Cuticle

Furthermore, component 8a was also partially sequenced, though no sequence information has been published [16]. More recently it has been established that the wool IFP families comprise four proteins in each of the Type I and II subfamilies and close linkage of genes has been observed within but not between the Type I and II gene families [11]. One gene sequence of a Type I IFP (K1.2) is also available and, while it was considered to be similar to component 8c-1, it has not been assigned to any of the known components from the Type I IFP subfamily [17]. In addition, a gene has been located for a Type II IFP family that is expressed in the cortical cells in the early stages of follicle development [18]. The full gene sequence has been determined for one protein (K2.9) [18] and partially determined for another (K2.11) [19]. These studies have also confirmed that K2.10 is component 5 and K2.12 is component 7c [18,19]. A number of partial sequences also exist for IFPs, including a partial amino acid sequence for a Type II IFP [20], a complete gene sequence for a Type II IFP not expressed in the follicle [21] and a partial gene sequence for another IFP [22].

As mentioned before, the matrix proteins have been further classified under the headings of high sulfur, ultra-high sulfur or high glycine/tyrosine proteins [10,11], though this is considered inadequate by some, as proteins rich in cysteine and glycine have also been found, and there are now known to be at least eight families of sulfur-rich proteins with cysteine contents ranging from 12 to 41 mol% [10,11]. Once again the classification system applied to the protein families of the HSPs has its origins in the method of separation employed. Fractional precipitation with ammonium sulfate at pH 6.2 and 4.0 of the alkylated wool protein component B resulted in the isolation of two components, B1 and B2, respectively [12]. Subsequent application of ion-exchange chromatography on DEAE–cellulose resulted in the further resolution of the B2 component into four subcomponents [12]. Alternatively, when column electrophoresis was applied, four fractions labelled I–IV were obtained, though inspection of the zone electrophoretic pattern showed only two poorly separated protein bands, labelled I and III [12]. Subsequent application of gel filtration on Sephadex G100 resulted in the further fractionation of the BI peak into BIA and BIB and the BIII peak into BIIIA and BIIIB, though it was later demonstrated that BIA was equivalent to component B1, and BIB to component B2 from the fractional precipitation approach [12]. This has resulted in a hybrid nomenclature system for the major HSP families in wool cortex, where only three are now recognised, specifically B2, BIIIA and BIIIB [10,11]. The HSPs are also included in the new nomenclature system (Table 1), though, as this system does not deal with the known polymorphic variants of the B2A family, it is still necessary to use the old nomenclature.

Amino acid sequencing has shown that the B2 family proteins range in size from 151 to 181 amino acids, are neutral–basic and contain on average 22 mol% cysteine [12]. The amino acids are highly conserved and differ in the number of tandem copies of the decapeptide repeat QTSCCQPT(I)SI; B2A having four decapeptide repeats [23,24], B2B three [23,25], B2C two [26] and B2D five [23]. Several genes encoding B2A, B2C and B2D have been located and sequenced, though strong homology within the 3′ non-coding region of the tandem B2A/B2D gene pair suggest that they may have arisen from a more recent duplication than the separate B2A and B2C genes [23]. However, subsequent studies have shown that the situation is more complicated than that, with at least three polymorphic variants of the B2A gene identified and six polymorphic variants of the B2C gene [27]. Moreover, the B2A gene exhibits length polymorphism with nucleotide sequences encoding three (B2Aγ), four (B2Aβ) or five (B2Aα) decapeptide repeats [27]. It is thought that the inserted/deleted nucleotides in the B2A gene have arisen through short gene conversion events whereby B2Aγ has been derived from the B2B gene and B2Aα from the B2D gene. Polymorphism is also thought to exist in other genes from the B2 family.

The application of a chromatographic approach utilising DEAE–cellulose led to the identification of the two further protein families, BIIIA and BIIIB [28]. Five subcomponents have been observed in the neutral–basic BIIIB family, for which sequences for BIIIB2 [29], BIIIB3 [30] and BIIIB4 [31] are available. Unlike the B2 family, it has no repetitive

cysteine-rich motifs. In contrast the higher molecular mass BIIIA family, for which 12 sequences are available [11,32], contains the repeat motif CCXPY, where X can be D, G, Q or R, and Y can be C, I, S, T or V.

Three families of UHSPs are known in wool, the cortical UHSPs, or the KAP4 family, and the cuticle UHSPs, or the KAP5 and KAP10 families (Table 1). In the case of the KAP5 family, five proteins are known to exist [33] and they have a distinctive amino acid composition with a strikingly ordered structure of glycine- and cysteine-rich repeats of approximately 10 and 20 amino acids, respectively [34]. In addition to the KAP5 proteins in the cuticle, a further family is recognised, that of the KAP10 family, for which the sequence of KAP10.1 is known [11]. This is the largest known KAP with 294 amino acids in which various short cysteine-rich motifs are arranged into higher order repeats of 28 and 42 amino acids. In the case of the cortical UHSPs, the full gene sequence is available for KAP4.2 [11], along with a partial gene sequence for the KAP4.1 protein [35]. From the one gene sequence that has been completed, it is apparent that the predicted wool protein KAP4.2 contains 211 residues, of which cysteine, serine, proline, arginine and threonine comprise 80% of the amino acid composition [11]. The protein pentapeptide repeat motifs containing these amino acids are arranged into two higher repeat orders of 48 amino acids [11].

The remaining family of matrix proteins is that of the HGTPs, so named because they contain between 35 and 60% of glycine and tyrosine amino acids. In the past, they were further subdivided into two further families, the Type I and Type II HGTPs, on the basis of amino acid content and solubility [10]. The main repeat motif in both families is either a YG pair or a GYG triplet. The Type I subfamily is considered to be quite heterogeneous [36], consisting of two major groups: a single protein in group F, for which both gene and amino acid sequences are available [37,38], and several components in group C. Both gene and amino acid sequences for component C2 are available, though amino acid sequence analysis for component C3 has so far shown it to be identical to C2 [36,38]. At least nine genes have been located for the Type II subfamily [23], for which one full gene sequence (KAP6.1) [39] and one partial amino acid sequence (KAP6.2) [40] are available.

4. Wool protein sequences available in web databases

There are three main web-based protein databases in which wool keratin sequences can be found: the SWISS-PROT database on the ExPASy Molecular Biology Server of the Swiss Institute of Bioinformatics in Geneva, Switzerland (http://www.expasy.ch/); the Protein Information Resource (PIR) of the National Biomedical Research Foundation at Georgetown University Medical Center (http://www-nbrf.georgetown.edu/pirwww/aboutpir.html); and the National Center for Biotechnology Information database (NCBInr) at the National Institutes of Health in Washington, DC, USA (http://www.ncbi.nlm.nih.gov/). All of the known wool keratin sequences in the literature are found in these databases, though there are some inconsistencies in the way the databases or sequences have been assembled, and some of the proteins classified.

The SWISS-PROT database has incomplete entries for wool keratins but the remainder can be found in the TrEMBL database, which contains sequences derived from the European Molecular Biology Laboratory (EMBL) nucleotide sequences. Both the SWISS-PROT and TrEMBL sequences are also found in the PIR database. However, two of the entries, S05408 and S29094, are confusingly referred to as cytoskeletal Type II keratins, even though the original literature does not refer to them as such [13,14]. This confusion in nomenclature may have arisen because of a tendency in the past to refer to the ‘‘hard α-keratins’’ as hair-type cytokeratins [41] when in fact the cytokeratins are now recognised as a group of IFPs found in epithelia, which are chemically and immunologically related to, but not identical with the microfibrillar α-keratins [42].

One of the more confusing aspects of all of these databases is that they contain a mixture of nomenclature systems, ranging from the one in use in the early 1970s to the more currently accepted system. In the case of the NCBInr database all the entries in it have

been drawn from both the SWISS-PROT and PIR databases without any effort to examine any individual entries, resulting in a duplication of every protein sequence: the SWISS-PROT entry for a particular protein and its PIR equivalent both being present. The SWISS-PROT database also introduces its own nomenclature. Thus, the Type I IFP component 8c-1 is also known as K1M1; the Type II IFP component 7c as K2M2; the Type II IFP component 5 as K2M3; while an unspecified Type I IFP, for which only the gene sequence is given, is known as K1M2 and the HSPs, UHSPs and HGTPs are also allocated unique names.

5. Identification of wool proteins

More recently wool proteins, alkylated with iodoacetic acid (IAA), have been separated and classified on the basis of their electrophoretic mobility into their broad classes by one-dimensional electrophoresis (1DE) on polyacrylamide gels (Fig. 2) [43–45]. The application of non-equilibrium (NE) two-dimensional electrophoresis (2DE), whereby proteins are separated by means of either acidic or alkaline PAGE in the first dimension and sodium dodecyl sulfate (SDS)–PAGE in the second dimension, resulted in the resolution of some of the bands on the 1DE gels into several protein spots on the NE-2DE gels [46] and these were initially separated into the three main groups of proteins, based on their amino acid composition [47]. While the location of the UHSPs was determined by running UHSP-enriched fractions separately on the NE-2DE system, it would appear that the location of some of the other proteins, such as BIIIA (Fig. 3) [47], on the NE-2DE map were determined by running translation products of cloned genes on NE-2DE gels [48].

When isoelectric focusing (IEF) was used in the first dimension of 2DE (IEF–2DE) instead of acidic or alkaline gel electrophoresis to separate IAA-alkylated wool proteins, the protein patterns appeared similar to those seen on NE-2DE gels, though some of the Type II IFP spots appeared to have been stretched into short streaks [49]. However, when IEF was performed in immobilized pH gradient strips, a significant increase in the number of wool proteins on the gels was observed [50]. The protein patterns

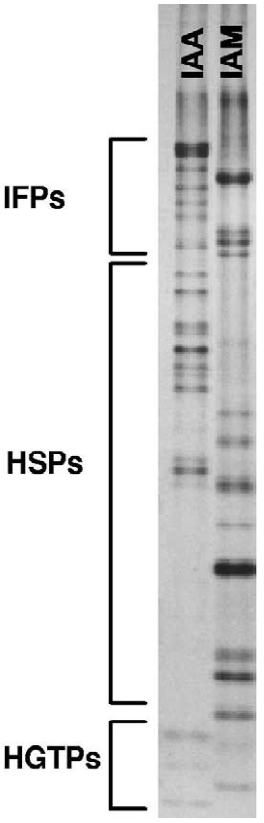

Fig. 2. A 1D autoradiograph of wool proteins extracted from a Romney sheep showing the protein patterns when the wool was alkylated with IAA or IAM. The labels on the left of the gel refer to the IAA lane only (unpublished results, Plowman and Flanagan).

observed were the same regardless of whether the proteins were alkylated with iodoacetamide (IAM) or were in their native state [51]. Unfortunately, these IEF–2DE gels are not strictly comparable with the NE-2DE gels because wool proteins alkylated with IAA have a different isoelectric point (p*I*) values and a different electrophoretic mobility in SDS–PAGE gels to proteins alkylated with IAM (Fig. 2). The IAA-alkylated HSPs have been observed to exhibit anomalous electrophoretic mobility in SDS gels, as in the case of BIIIA1 and BIIIA8, which were estimated as having masses of 28 and 35 kDa, respectively, though they differed in only seven out of 131 amino acids [52]. One notable difference between these proteins is that BIIIA8 contains three more cysteines than BIIIA1 and hence the higher

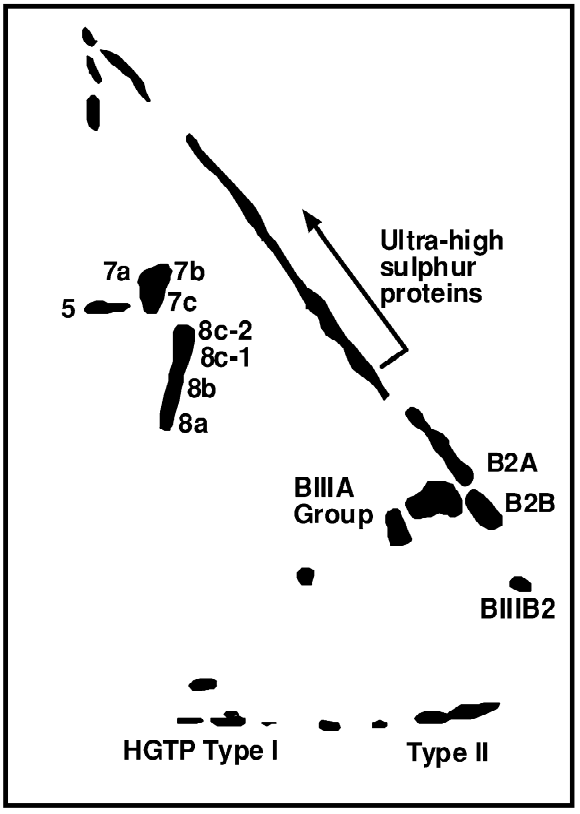

Fig. 3. A NE-2DE map of IAA-alkylated wool proteins adapted from Powell and Rogers [47].

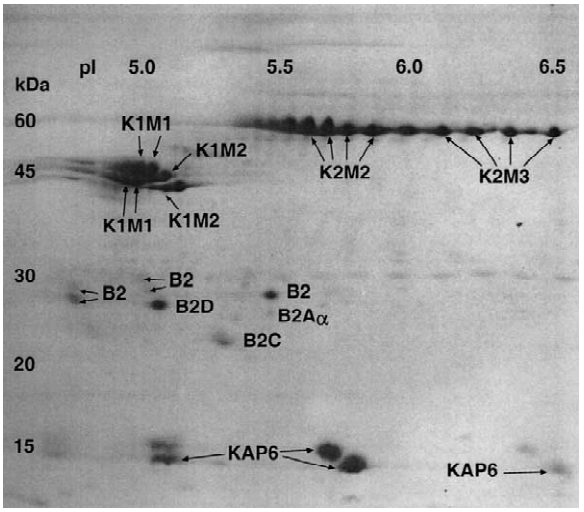

Fig. 4. An IEF–2DE map of a Lincoln-like staple of wool from Chimera 120 sheep showing the location of proteins currently identified. A carbamylated protein marker string is also visible at 30 kDa. HSPs not uniquely identified are indicated as B2. The possible location of the HGTPs is indicated.

negative charge of its IAA-alkylated form could result in a larger apparent denatured volume and shape. Nevertheless, while the proteins were not identified in the IEF–2DE gels at this stage, an attempt was made to determine the locations of the basic family groups on the gels on the basis of their electrophoretic mobility in NE-2DE gels and their known molecular masses. From this it appeared that the Type II IFPs were spread out in a long train of proteins, running from acidic to neutral pH, whereas the Type I IFPs separated over a much narrower range at acidic pH and into four distinct trains of spots (Fig. 4) [51]. Treatment of the cysteine thiol modified keratin proteins with alkaline phosphatase led, in some experiments, to a reduction of the number of spots in the Type II IFP train, suggesting that this variability could be partly explained by differences in the degree of phosphorylation [51]. Studies of glandular keratin IF phosphorylation have shown that these long trains can arise from the post-translational modification of these proteins by glycosylation, phosphorylation or both [53]. The HSPs were considered to be confined to a region below the Type I IFPs, between 20 and 30 kDa and pH 4.8 to 5.6 with the major HGTPs below them around 15 kDa [54].

Relatively few papers have been published on the identification of wool proteins using modern proteomic methods. Herbert et al. [55] employed amino acid compositional matching to identify seven Type II IFP spots from the wool map and found they could not distinguish between K2M2 (IFP Type II component 7c) and K2M3 (IFP Type II component 5) because of the 77% sequence homology between the two. One further limitation of amino acid profiling was that there were at least two other Type II IFPs for which only incomplete sequence information was available, so these partially sequenced proteins could not be included in the matching process. However, it was possible to distinguish between the Type I and Type II subfamilies of IFPs with amino acid profiling, because there is only 27% sequence homology between them. Reversed-phase chromatographic separation was also applied to

separate tryptic peptides obtained from protein spots from the Type I IFP region [55]. Four peptides from at least one protein spot were sequenced. All four tryptic peptides were found to have 100% sequence identity to both known Type I IFPs, K1M1 (IFP Type I component 8c-1) and K1M2. Thus, it was not possible to distinguish between these two proteins on the basis of the peptides selected and it was considered that this approach would have limited success because of the high sequence homology between them (92%). Phosphoserine was also detected in all seven Type II IFPs, confirming that these proteins were post-translationally modified.

More recently the MALDI-TOF-MS peptide mapping approach has been applied to wool keratin proteins [54]. With trypsin digestion generating numerous peptides covering 67% of the sequence, the MS peptide mapping approach had no difficulty uniquely identifying an IFP as the Type I K1M2, and was also able to uniquely identify a HSP as B2Aα by utilising the masses of only four peptide peaks covering only 39% of the sequence [54]. Using either the program MoverZ, available from the PROWL website (http://prowl.rockefeller.edu/cgi-bin/ProFound) to search for protein matches on the NCBInr database, or Data Explorer 3.5.0.0 (Applied Biosystems) to search for protein matches on either the NCBInr or SWISS-PROT databases, this study was extended to identify a total of 14 IFPs and three HSPs (Fig. 4). An additional five spots were classified as being from the B2A family and four spots were classified as being members of the Type II HGTP family [56].

6. Application of peptide mass fingerprinting to wool protein identification

Successful identification of proteins by MALDI-TOF-MS peptide fingerprinting is dependent on a number of factors, the most important of which is the number of peptides generated: the greater the number of peptides generated the better the fit, with a minimum of 30% of the measured masses belonging to genuine peptides from a protein being required for a successful identification of a protein [57]. The size of protein is also important, large proteins generally produce many more protease fragments and hence the chance of identifying such a protein by chance is much higher than identifying a smaller protein that generally only produces a few protease fragments [58]. It is also dependent on the availability of relevant protein sequences in the various web-based databases.

IFPs both have high molecular masses and generate a large number of peptides on tryptic hydrolysis and on this basis ought to be suitable for identification using the MS fingerprinting approach. In the case of wool, only two sequences for the Type I IFPs (component 8c-1 and a gene sequence) are available in these databases despite the fact that four Type I and four Type IIs are postulated to exist [11], while for the Type II IFPs three complete (components 5 and 7c and a gene sequence for a cortical follicular Type II IFP) and three partial sequences exist. While Type I IFP component 8a has also been partially sequenced, the others have not and this may be because of difficulties in purifying those particular components. Components 7b and 8b were poorly resolved by ion-exchange chromatography, while components 7a and 8c-2 were not evident at all [12].

In the case of the Type I IFPs, for instance, it was possible to identify six spots from the four trains observed in the IEF–2DE map as either being K1M1 or K1M2, despite a sequence homology of 92% between them [56]. However, a seventh spot could not be identified because the sequence coverage was insufficient. Likewise, it was possible to identify eight Type II IFP spots as either K2M2 or K2M3 using this approach, though if MoverZ was used in conjunction with ProFound to explore the NCBInr database then there was a tendency for it to rank the ''cytoskeletal'' Type II IFPs, originating from the PIR database, higher than the ''microfibrillar'' Type II IFPs from the SWISS-PROT database. As noted earlier, this situation would cause confusion if the researcher was not aware of the duplication problem for these proteins in this web-based database. The sequence for K2M3 in the SWISS-PROT database also contains an error in which the serine after S446 has been deleted, which could lead a researcher to believe that they were dealing with a different, but strongly homologous protein.

While the NE-2DE wool protein map shows the location of the four major Type I and II IFPs [47], only two major Type I and II IFPs were identified in

the IEF–2DE gels. Despite there being three complete and three partial sequences in SWISS-PROT and TrEMBL databases available for the Type II IFPs, only two Type II IFPs were picked when the PeptIdent (http://au.expasy.org/tools/peptident.html) was used to search for matches. Examination of the search results revealed that the unannotated sequences present in TrEMBL were picked, though always with a lower score and sequence coverage than the prime candidate, thus giving reasonable confidence that these identifications were correct.

In contrast to the IFPs, the MS peptide fingerprinting approach is less successful when applied to the HSPs. In the case of the B2 family, the mass spectrum is dominated by two peaks at 1049.2 and 1274.4 Da (Fig. 5), these peaks corresponding to two peptides common to all family members; W^{111}–R^{117} and P^{149}–R^{158} (B2D sequence numbering), respectively. Thus, the presence of these two peaks in the MALDI-TOF-MS of eight HSP spots in the wool protein map has meant that these proteins have been identified as belonging to this family (Fig. 4) [56]. However, identification of proteins in this family is more difficult because of their lower molecular mass, high sequence homology (96% between B2Aα and B2D) and relatively few acidic or basic residues. Inspection of Fig. 6 reveals that, with the exception of B2C, there are no acidic or basic residues in the N-terminal half of the molecule and only four arginine residues or three acidic residues in the second half, largely as a result of the presence of the decapeptide repeats. This results in the generation of relatively few peptides from a tryptic digest and thus poor matches when database searching is undertaken.

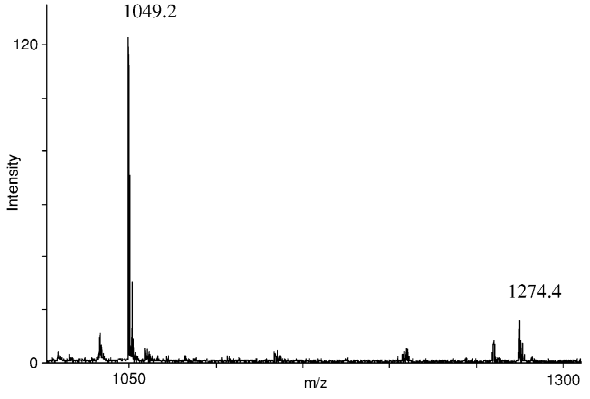

Fig. 5. A MALDI-TOF-MS peptide mass fingerprint of a tryptic digest of a B2Aα HSP spot from the Merino-like staple from Chimera 120 showing the peaks that are diagnostic for the B2 family.

```
B2A    1 ACCSTSFCGFPICSTGGTCGSSPCQPTCCQTSCCQPTSI----------Q
         ||||||||||||||  ||||||  ||||  ||||||||||
B2B    1 ACCSTSFCGFPICSSVGTCGSSCCQPTCSQTSCCQPTSI-----------
         ||||||||||||||  |||||||    ||||||||||||
B2C    1 ACCSTSFCGFPICSTAGTCGSSCCRSTCSQTSCCQPTSI-----------
         ||||||||||| ||| |||||  |  || ||||||||||
B2D    1 ACCSTSFCGFPTCSTGGTCGSNFCQPTCCQTSCCQPTSIQTSCCQPTSIQ

B2A   41 TSCCQPISIQTSCCQPTSIQTSCCQPTCLQTSGCETGCGIGGSIGYGQVG
                 |||||||  |||||||||||||||||||||||||||||||  |||
B2B   41 ---------QTSCCQPISIQTSCCQPTCLQTSGCETGCGIGGSIGYDQVG
                           ||||||||||||||||||||||||  ||  |||
B2C   31 --------------------QTSCCQPTCLQTSGCETGCGIGGSTGYGQVG
                           ||||||||||||||||||||||||  ||||||
B2D   51 TSCCQPTSIQTSCCQPISIQTSCCQPTCLQTSGCETGCGIGGSIGYGQVG

B2A   91 SSGAVSSRTRWCRPDCRVEGTSLPPCCVVSCTPPSCCQLYYAQASCCRPS
         |||||||||||||||||||||||||||||||| ||||||||||||||||
B2B   81 SSGAVSSRTRWCRPDCRVEGTSLPPCCVVSCTSPSCCQLYYAQASCCRPS
         |||||||||||||||||||||||||||||||||||||||||||||||||
B2C   71 SSGAVSSRTRWCRPDCRVEGTSLPPCCVVSCTSPSCCQLYYAQASCCRPS
         |||||||| |||||||||||||||||||||||||||||||||||||||||
B2D  101 SSGAVSSRTKWCRPDCRVEGTSLPPCCVVSCTSPSCCQLYYAQASCCRPS

B2A  141 YCGQSCCRPACCCQPTCIEPICEPSCCEPTC
         |||||||||||||||| ||| |
B2B  131 YCGQSCCRPACCCQPTCIEPVCEPTC
         |||||||||||||||| |||||||
B2C  121 YCGQSCCRPACCCQPTCTEPVCEPTCSQPIC
         |||||||||||||||| ||||||| | |
B2D  151 YCGQSCCRPACCCQPTCIEPVCEPTCCEPTC
```

Fig. 6. Amino acid sequences of the sheep HSPs; B2A [23,24], B2B [23,25], B2C [26] and B2D [23]. The sequences corresponding to the peptide masses 1049.2 and 1274.4 Da are indicated in bold and underlined, the common sequences for B2B, B2C and B2D are double underlined and the diagnostic N-terminal sequence of B2C is dotted underlined.

Thus, alternative approaches have to be found to identify HSPs separated on gels. In the case of the B2 HSPs, manual examination of the mass spectrum utilising theoretical masses of the likely peptides generated by PAWS (Protein Analysis Work Sheet) has been found to be the best approach [56]. From an examination of Table 2 it is apparent that the C-terminal tryptic peptides differ sufficiently from each other to enable unique identification of the protein spot if masses from these peptides are detected in the MALDI-TOF spectrum. It was on this basis that B2Aα (Figs. 5 and 7) [54] and B2D were identified [56]. The HSP B2C is also unique in having an arginine residue at position 25 in what is normally a long sequence, unbroken by basic residues, thus enabling this protein spot to be identified in the wool protein map [56]. The peak arising from the sequence V^{118}–R^{148} also offers a way to distinguish most of the B2A gene variants from the rest of the

Table 2
Masses of the main tryptic peptides in the B2 family of HSPs

B2D sequence numbering	B2A	B2B	B2C	B2D
N-Terminus	10 805.0	9614.6	2894.2 5600.1	11 991.3
111–117	**1049.2**	**1049.2**	**1049.2**	**1049.2**
118–148	3666.5	3656.5	3656.5	3656.5
149–158	**1274.4**	**1274.4**	**1274.4**	**1274.4**
C-Terminus	2899.0	2237.3	2810.9	2885.0

The bold and underlined masses relate to the sequences common to all members of this family. Those masses with double underlines are common to B2B, B2C and B2D, while the dotted underlined mass relates to the N-terminal tryptic peptide of B2C. These sequences are shown in Fig. 5.

B2 family, as the proline at position 134 of B2A, B2Aα and B2Aγ is replaced by a serine in B2Aβ, B2B, B2C and B2D. Thus in the wool protein map, the spot above B2Aα at p*I* 5.5 and 27 kDa (Fig. 4) has a peptide with a mass of 3656.5 Da which is indicative of it being either B2Aβ, B2B, B2C or B2D (Table 2).

No members of the BIIIB family have yet been located on the IEF–2DE wool protein map and it is apparent from inspection of their sequences (Fig. 8) and theoretical tryptic peptide masses (Table 3) that distinguishing between different members of the family in the 2DE map will be difficult using the MS peptide fingerprinting approach. As for the B2 family, identification of BIIIB proteins is dependent on a limited number of peptides, some of which are high in molecular mass, as determined by PAWS. BIIIB2 differs significantly from the other two known members of this family, particularly in the N-terminal region from residues 1–26. However, the sequences of BIIIB3 and BIIIB4 are identical in this region. While there is strong sequence homology between BIIIB3 and BIIIB4 in the remaining two thirds of the sequence, the presence of an arginine residue at position 51, in both, could potentially lead to the generation of two peptides around 3000 and 5000 Da. The detection of either or both would result in the unique identification of either protein.

Only a single strong mass of 1700.3 Da (Fig. 9) was observed in the MALDI-TOF-MS spectrum of the spots around 15 kDa in the IEF–2DE map (Fig. 4) and this was insufficient to obtain a match using the standard database search routines. However, it was apparent from a determination of the masses likely to be generated using PAWS and an examination of the known HGTP sequences (Fig. 10) that this mass corresponds to the mass of the sequence R^{35}–R^{49} from KAP6.1 or KAP6.2, suggesting a possible match to the Type II HGTP family [56]. The absence of other peaks in the spectrum meant that a unique identification was not possible for these

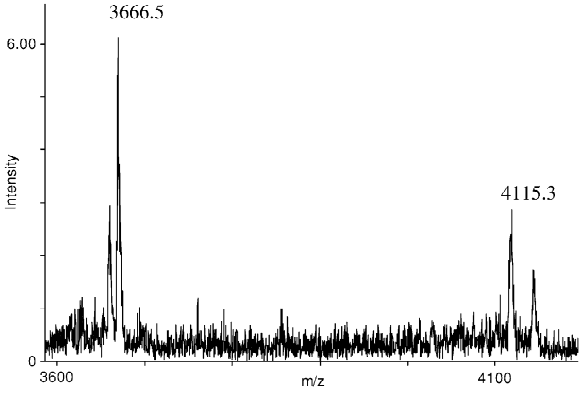

Fig. 7. A MALDI-TOF-MS peptide mass fingerprint of a tryptic digest of a B2Aα HSP spot from the Merino-like staple from Chimera 120 showing the peaks unique to B2Aα.

```
BIIIB2   1 ACCAPRCCSVRTGPATTICSSDKFCRCGVCLPSTCPHNISLLQPTCC-DN
           ||||| |||| | ||||||||||||||||||||||||   ||||||| ||
BIIIB3   1 ACCARLCCSVPTSPATTICSSDKFCRCGVCLPSTCPHTVWLLQPTCCCDN
           ||||||||||||||||||||||||||||||||||||||||||| |||||||||
BIIIB4   1 ACCARLCCSVPTSPATTICSSDKFCRCGVCLPSTCPHTVWFLQPTCCCDN

BIIIB2  50 SPVPCVYPDTYVPTCFLLNSSHPTPGLSGINLTTFIQPGCENVCEPRC
             | |   |  |||||||||| |||||  |||||  |  ||      |
BIIIB3  51 RPPPYHVPQPSVPTCPLLNSSQPTPGLESTNLTTYTQSSCEPCIPSCC
           ||||  | ||||||||||||||||||||||||||||||||| ||||||||||
BIIIB4  51 RPPPCHIPQPSVPTCPLLNSSQPTPGLESTNLTTYTQPSCEPCIPSCC
```

Fig. 8. Amino acid sequences of the sheep HSPs; BIIIB2 [29], BIIIB3 [30] and BIIIB4 [31]. The sequence corresponding to the peptide mass 481.6 Da is bold and underlined, while the common sequences for BIIIB3 and BIIIB4 are bold.

Table 3
Masses of main tryptic peptides in the BIIIB family of HSPs

BIIIB2 sequence numbering	BIIIB2	BIIIB3&4 sequence numbering	BIIIB3	BIIIB4
1–6	775.8	1–5	**678.7**	**678.7**
7–11	680.7	6–23	**1984.2**	**1984.2**
12–23	1237. 3			
24–26	**481.5**	24–26	**481.5**	**481.5**
27–97	8140.0	27–51	3164.5	3198.5
		52–98	5294.0	5315.0

The bold and underlined masses relate to the sequences common to all members of this family, while those in bold are common to both BIIIB3 and BIIIB4. These sequences are shown in Fig. 7.

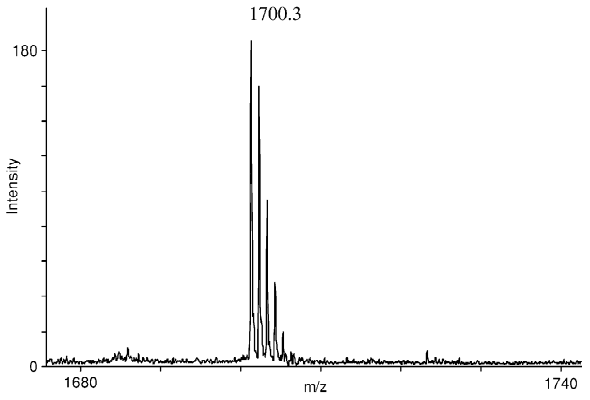

Fig. 9. A MALDI-TOF-MS peptide mass fingerprint of a tryptic digest of a HGTP spot from the Lincoln-like staple from Chimera 120.

proteins. Again, the small size of the protein, coupled with the high proportion of glycine and tyrosine residues and the presence of few basic (and no acidic) residues means that very few peptides are likely to be generated. Unfortunately, the use of alternative enzymes to digest the proteins in this family is also limited as most other commercial available enzymes tend to be endopeptidases.

```
KAP6.1      1  CGYYGNYYGGLGCGSYSYGG-LGCGYGSCYGSGFR RLG--CGYGC
KAP6.2      1        GGGYLGCGSYGYGG-LGCGYGSCYGNYFR RLG--CGYGC
Type I C2   1  TRFTCCGSYFPG-YP-SYGTNFIRTFRATPLNCVVPLGSPLGYGCN
Type I F    1  SYCLSSTV-FPGCYWGSYGYPLGYSVGCGYGSTYSPVGYGFGYGYD

KAP6.1     43 GYGYGSRSLCGSGYGYGSRSLCGSGYGCGSGYGSGFGYY-Y
KAP6.2     37 GYGYGSRSLCGSGYGCGSRPLYGCGYGCGSGYGSGFGYY
Type I C2  45 GYSSLGYG-FGGSSFSNLGCCYGGSFYRPYGSGSGFGYSTY
Type I F   46 GGSAFGCRRFWPFALY
```

Fig. 10. Amino acid sequences of the four known sheep HGTPs; KAP6.1 [39], KAP6.2 [40], Type I C2 [36,38] and Type I F [37,38]. The sequence corresponding to the peptide mass 1700.3 Da is indicated in bold.

7. Application of ESI-MS–MS to wool protein identification

Because of their high mass accuracy, resolution and sensitivity, Q-TOF mass analysers offer a considerable advantage over other mass spectrometers. An Applied Biosystems Q-STAR Pulsar i was applied to the identification of wool proteins (Plowman et al., unpublished results) and these results were compared with some preliminary investigations performed using a Micromass Q-Tof (Coffey, personal communication). Both instruments were able to distinguish between Type I and II IFPs and between the various members of each family when a peptide mapping approach was used for identification. Confirming these identities using de novo sequencing was less successful. In the case of the Type I IFPs this is partly because the high sequence homology within each family means that tryptic peptides common to both known members make up over half the sequence length of their respective proteins. Therefore, the sequences of a large number of peptides have to be determined before the protein identity can be established by this means with any degree of confidence.

The Micromass automatic sequencing routine, employing a double Bayesian approach [59,60], generated a number of sequences that were common to both proteins, but also found for each protein a

number of peptides that were unique to that protein. In the case of a Type I IFP spot it could not distinguish between K1M1 or K1M2, both proteins being ranked equally, because it located a similar number of peptides that were unique to K1M1 as well as to K1M2. Unfortunately, insufficient numbers of peptides were sequenced by the Applied Biosystems Q-STAR to be able to uniquely identify the Type I or II IFPs. Both instruments had difficulty in identifying matrix proteins, which is largely due the limited number of peptides generated by trypsin. However, in the case of one HSP spot examined the only peptides identified were the ones that were diagnostic to the B2 family.

8. Conclusions

Despite the advent of modern proteomic techniques, identification of wool keratin proteins is still not a straightforward matter. While it is possible to identify the higher molecular mass IFPs by the MALDI-TOF- or Q-TOF-MS peptide mapping approach, the high sequence homology observed in these proteins means that good sequence coverage is necessary to have a reasonable degree of confidence that the protein has been uniquely identified. Another limitation of this approach is the need for sequences in web-based databases against which unknown proteins can be matched. Currently this is a problem for Type I and II IFPs, where four of each are known though less have actually been fully sequenced.

Identification of the matrix proteins is even more difficult than it is for the IFPs. Their lower molecular mass, coupled with the presence of decapeptide and higher repeats, results in the production of a small number of peptides, some of which have very high molecular masses. Their high sequence homology also means that relatively few regions in these proteins have a unique sequence that can be said to be diagnostic for that protein. Hence the MS mapping approach is less successful in identifying these proteins as there are only a limited number of diagnostic peptides in the proteins in question.

There are also problems with the automatic MS–MS matching routines used by some Q-TOF instruments when applied to IFPs. The use of a double Bayesian probabilistic method to deduce the most plausible sequence or sequences from the ESI-MS–MS data does not work very well when the proteins have as high a sequence homology as the Type I IFPs. As with the MALDI-TOF, Q-TOF instruments cannot easily uniquely identify matrix proteins because there are only a limited number of diagnostic peptides in these proteins. To be absolutely certain of a keratin protein's identity, de novo sequencing by ESI-MS–MS may be the only practical solution for many wool proteins.

Despite these difficulties it has been possible to identify some of the major wool keratin protein spots separated and visualised in an IEF–2DE gel. From the map it is evident that two proteins are predominant among the neutral–basic Type II IFPs, K2M2 and K2M3, with the former being found at lower p*I* than the latter, this is despite the fact that three sequences for Type II IFPs are known. Likewise, among the more acidic Type I IFPs two trains of spots identified as K1M1 appear to be found at lower p*I* than the two trains of spots identified as K1M2. These results are in contrast to the NE-2DE protein map where the positions of four different Type I and II IFPs are shown. Unfortunately, as it is not known how these proteins were identified it is not possible to resolve these differences at this stage, however, it does suggest that while four gene sequences for each IFP family exist, only two of them are expressed in significant amounts. If this is true then the appearance of four spots for each IFP family in the NE-2DE gel may have arisen through post-translational modification of the two main expressed Type I and II IFPs. Unfortunately, it may not be possible to resolve this issue until sequences for more IFPs are available or until these protein spots from the gel can be subject to full de novo sequencing and the location and nature of their post-translational modifications determined by Q-TOF-MS.

The HSPs are concentrated between 20 and 30 kDa and p*I* 4.8 and 5.5 and appear to be predominantly from the B2 HSP family [56]. This is in marked contrast to results from NE-2DE, where spots from the BIIIA and BIIIB families have also been marked on the wool protein map [47]. The HGTPs appear to be located in two trains of spots at lower molecular mass, around 15 kDa, and appear to be entirely from the Type II HGTP family. Thus, despite the number of wool proteins sequenced,

results to date suggest that a limited number of them predominate in the IEF–2DE wool protein map.

Acknowledgements

I am grateful to Dr. Warren G. Bryson, Dr. Geoffrey Aitken and Joy Woods for their help throughout the preparation of this review, Dr. T. William Jordan of Victoria University for the MALDI-TOF-MS analysis of the wool proteins and for NZ Government funding via the Foundation for Research, Science and Technology, contract No. WROX0004.

References

[1] D.R. Goddard, L. Michaelis, J. Biol. Chem. 112 (1935) 361.
[2] R.C. Marshall, D.F.G. Orwin, J.M. Gillespie, Electron Microsc. Rev. 4 (1991) 47.
[3] K. Weber, N. Geisler, EMBO J. 1 (1982) 1155.
[4] D.A.D. Parry, P.M. Steinert, Intermediate Filament Structure, Springer-Verlag, New York, 1995.
[5] L. Jones, N.R. Watts, N. Cheng, D.A.D. Parry, A.C. Stevens, J. Invest. Dermatol. Symp. Proc. 4 (1999) 353.
[6] S. Heins, U. Aebi, Curr. Opin. Cell Biol. 6 (1994) 25.
[7] D.R. Goddard, L. Michaelis, J. Biol. Chem. 106 (1934) 605.
[8] W.G. Crewther, L.M. Dowling, K.H. Gough, A.S. Inglis, N.M. McKern, L.G. Sparrow, E.F. Woods, in: Proceedings of the Xth International Wool Textile Research Conference, Aachen, Vol. II, August 1975, p. 233.
[9] W.G. Crewther, L.M. Dowling, K.H. Gough, R.C. Marshall, L.G. Sparrow, in: D.A.D. Parry, L.K. Creamer (Eds.), Fibrous Proteins: Scientific, Industrial and Medical Aspects, Vol. 2, Academic Press, London, 1980, p. 151.
[10] B.C. Powell, Wool Technol. Sheep Breed. 44 (1996) 100.
[11] B.C. Powell, G.E. Rogers, in: P. Jollès, H. Zahn, H. Höcker (Eds.), Formation and Structure of Human Hair, Birkäuser Verlag, Basel, 1997, p. 59.
[12] W.G. Crewther, in: Proceedings of the Xth International Wool Textile Research Conference, Aachen, Vol. I, August 1975, p. 1.
[13] L.G. Sparrow, C.P. Robinson, J. Caine, D.T.W. McMahon, P.M. Strike, Biochem. J. 282 (1992) 291.
[14] L.G. Sparrow, C.P. Robinson, D.T.W. McMahon, M.R. Rubira, Biochem. J. 261 (1989) 1015.
[15] L.M. Dowling, W.G. Crewther, A.S. Inglis, Biochem. J. 236 (1986) 695.
[16] W.G. Crewther, L.M. Dowling, A.S. Inglis, L.G. Sparrow, P.M. Strike, E.F. Woods, in: Proceedings of the Xth International Wool Textile Research Conference, Tokyo, Vol. I, August 1985, p. 85.
[17] B.W. Wilson, K.J. Edwards, M.J. Sleigh, C.R. Bryne, K.A. Ward, Gene 73 (1988) 21.
[18] B. Powell, L. Crocker, G. Rogers, Development 114 (1992) 417.
[19] B.C. Powell, J.S. Beltrame, J. Invest. Dermatol. 102 (1994) 171.
[20] W.G. Crewther, A.S. Inglis, N.M. McKern, Biochem. J. 17 (1978) 365.
[21] B.C. Powell, L.A. Crocker, G.E. Rogers, J. DNA Seq. Map. 3 (1993) 401.
[22] B.C. Powell, G.R. Cam, M.J. Fietz, G.E. Rogers, Proc. Natl. Acad. Sci. USA 83 (1986) 5048.
[23] B.C. Powell, M.J. Sleigh, K.A. Ward, G.E. Rogers, Nucleic Acids Res. 11 (1983) 5327.
[24] T.C. Elleman, Biochem. J. 130 (1972) 833.
[25] T.C. Elleman, T.A. Dopheide, J. Biol. Chem. 247 (1972) 3900.
[26] T.C. Elleman, Biochem. J. 128 (1972) 1229.
[27] G.R. Rogers, J.G.H. Hickford, R. Bickerstaffe, Anim. Genet. 25 (1994) 407.
[28] T. Haylet, L.S. Swart, D. Parris, F.J. Joubert, Appl. Polym. Symp. 18 (1971) 37.
[29] T. Haylet, L.S. Swart, Text. Res. J. 39 (1969) 917.
[30] T. Haylet, L.S. Swart, D. Parris, Biochem. J. 123 (1971) 191.
[31] L.S. Swart, T. Haylet, Biochem. J. 123 (1971) 201.
[32] L.S. Swart, F.J. Joubert, D. Parris, in: Proceedings of the Xth International Wool Textile Research Conference, Aachen, Vol. II, August 1975, p. 254.
[33] B.J. Jenkins, B.C. Powell, J. Invest. Dermatol. 103 (1994) 310.
[34] P.J. McKinnon, B.C. Powell, G.E. Rogers, J. Cell Biol. 111 (1990) 2587.
[35] A. Fratini, B.C. Powell, P.I. Hynd, R.A. Keough, G.E. Rogers, J. Invest. Dermatol. 102 (1994) 178.
[36] R.C. Marshall, J.M. Gillespie, A.S. Inglis, M.J. Frenkel, in: Proceedings of the Xth International Wool Textile Research Conference, Pretoria, Vol. II, August 1980, p. 147.
[37] T.A.A. Dopheide, Eur. J. Biochem. 34 (1973) 120.
[38] E.S. Kuczek, G.E. Rogers, Eur. J. Biochem. 166 (1987) 79.
[39] A. Fratini, B.C. Powell, G.E. Rogers, J. Biol. Chem. 268 (1993) 4511.
[40] J.M. Gillespie, in: R.D. Goldman, P.M. Steinert (Eds.), Cellular and Molecular Biology of Intermediate Filaments, Plenum Press, New York, 1990, p. 95.
[41] H.W. Heid, I. Moll, W.W. Franke, Differentiation 37 (1988) 137.
[42] H.W. Heid, E. Werner, W.W. Franke, Differentiation 32 (1986) 101.
[43] I.J. O'Donnell, E.O.P. Thompson, Aust. J. Biol. Sci. 17 (1964) 973.
[44] R.D.B. Fraser, T.P. McRae, G.E. Rogers, Keratins. Their Composition, Structure and Biosynthesis, Charles C. Thomas, Springfield, IL, 1972.
[45] J.L. Woods, D.F.G. Orwin, Aust. J. Biol. Sci. 40 (1987) 1.
[46] R.C. Marshall, Text. Res. J. 51 (1981) 106.
[47] B.C. Powell, G.E. Rogers, in: J. Bereiter-Hahn, A.G. Matoltsy, K. Sylvia-Richards (Eds.), Biology of the Integument 2: Vertebrates, Springer-Verlag, Berlin, 1986, p. 695.

[48] P.J. MacKinnon, Ph.D. Thesis, University of Adelaide, Adelaide, 1989.
[49] R.C. Marshall, R.J. Blagrove, J. Chromatogr. 172 (1979) 351.
[50] B.R. Herbert, J.L. Woods, Electrophoresis 15 (1994) 972.
[51] B.R. Herbert, A.L.P. Chapman, D.A. Rankin, Electrophoresis 17 (1996) 239.
[52] R.C. Marshall, J. Invest. Dermatol. 80 (1983) 519.
[53] J. Liao, N.-O. Ku, M.B. Omary, Electrophoresis 17 (1996) 1671.
[54] J.E. Plowman, W.G. Bryson, T.W. Jordan, Electrophoresis 21 (2000) 1999.
[55] B.R. Herbert, M.P. Molloy, J.X. Yan, A.A. Gooley, W.G. Bryson, K.L. Williams, Electrophoresis 18 (1997) 568.
[56] J.E. Plowman, W.G. Bryson, L.M. Flanagan, T.W. Jordan, Anal. Biochem. 300 (2002) 221.
[57] L. Guang, M. Waltham, N.L. Anderson, E. Unsworth, A. Treston, J.N. Weinstein, Electrophoresis 18 (1997) 391.
[58] P. Berndt, U. Hobohm, H. Langen, Electrophoresis 20 (1999) 3521.
[59] D. Gostick, J. Langridge, R. Rachubinski, R. Wozniak, J. Smith, J. Cottrell, J. Hoyes, E. Kapp, J. Skilling, R. Bordoli, in: 25th Annual Lorne Conference on Protein Structure, Australia, February 2000, p. A107.
[60] D. Gostick, J. Brown, E. Kapp, J. Langridge, R. Bordoli, in: 25th Annual Lorne Conference on Protein Structure, Australia, February 2000, p. A105.