A NEW LOOK AT OLD FRIENDS: THE MOLECULAR BIOLOGY OF THE PROTEIN MARKERS

G.F. Sensabaugh
Forensic Science Group, School of Public Health
University of California, Berkeley CA 94720 USA

Until the advent of DNA technology, protein markers occupied center stage in the armamentarium of forensic biology. Body fluid identification is based on the detection of tissue specific protein markers, e.g., hemoglobin for blood, prostatic acid phosphatase and prostatic antigen p30 for semen, amylase for saliva, and so on. More prominently, the core group of genetic markers used for individualization and paternity testing consisted primarily of polymorphic red cell enzymes and plasma proteins; these markers now have been eclipsed for the most part by DNA markers.

Before taking full leave of the protein markers, it is instructive to look at what has been learned about them using the tools of molecular biology. Many of the protein markers have been sequenced at the protein, cDNA, and/or genomic DNA levels. This sequence information provides insight into the basis of polymorphic and tissue specific variation; it can also indicate homologies among proteins and evolutionary relationships. The 3-D structure has been solved for some of the markers, providing additional insights into structure-function relationships. Tables 1 and 2 summarize the current level of knowledge about the tissue specific markers and polymorphic markers.

The discussion that follows focuses on insights provided by molecular biology to some of these markers with an emphasis on the markers of tissue origin. Other markers are discussed in accompanying papers by Hopkinson, Mayr, and Dissing.

Table 1
MOLECULAR KNOWLEDGE OF POLYMORPHIC MARKERS

Marker	Sequence	Gene Struct.	Polymorphism	Homology	3-D
Hp	P, cDNA	+	+	+	
Gc	cDNA	+	+	+	+
Ig-Km	P	+	+	+	+
Ig-Gm	P, cDNA	+	+	+	+
PGM	cDNA	+	+	+	
AK	P		+	+	+
ADA	cDNA	+	+		
EsD	cDNA	+[a]	+	+	
ACP1	P, cDNA	+[a]	+		
GLO	cDNA				
ABO	cDNA	+[a]	+		
MN	P, cDNA	+	+	+	+[b]
Rh	cDNA	+	+		
HLA I	P, cDNA	+	+	+	+
HLA II	cDNA	+	+	+	+

---

a - Detail not complete, b - Antigenic region structure modeled from NMR data

Table 2
MOLECULAR KNOWLEDGE OF TISSUE SPECIFIC PROTEIN MARKERS

Marker	Sequence	Gene Struct.	Regulation	Homology	3-D
Hemoglobin	P, cDNA	+	+	+	+
Pros. ACP	P, cDNA	+	+	+	(+)
Pros. p30	P, cDNA	+	+	+	(+)
Sal. Amylase	cDNA	+	+	+	

PROSTATIC ACID PHOSPHATASE (PAP)

Prostatic acid phosphatase has been used for nearly 50 years as a marker for semen; it also has a venerable history as a marker in clinical testing for prostatic cancer. It was established many years ago that the activity of PAP is inhibited by tartrate and that this could be used to distinguish PAP from red cell (cytoplasmic) acid phosphatase. Subsequent biochemical studies showed that PAP belongs to a group of acid phosphatases defined by a molecular weight of 100-110,000, referred to here as the 100K acid phosphatases. The 100K acid phosphatases are dimeric proteins with subunit molecular weights of about 50-55,000; they are glycosylated and tartrate inhibition is an additional defining characteristic. [For a recent review, see ref. 1] In addition to PAP, the 100K group includes lysosomal acid phosphatase (LAP) and, of particular forensic interest, the acid phosphatase present in vaginal fluids (VAP).

Genetic studies established that there exist at least two distinct genes for 100K acid phosphatases, ACP2 located on chromosome 11 and ACP3 with an unknown chromosomal location [2]. LAP has the same electrophoretic mobility as ACP2 and was presumed to be a product of the ACP2 locus. VAP has an electrophoretic mobility similar to the ACP3 gene product and PAP migrates somewhat anodal to both. PAP, VAP, and ACP3 gene product from placenta are immunologically cross-reactive suggesting a close molecular relationship [3]. The mobility of all three can be altered by desialidation and it is clear that variable glycosylation can alter electrophoretic migration patterns. Based on this evidence, I previously suggested that PAP and VAP might be products of the ACP3 locus [3]. The major forensic implication of this suggestion is that the specificity of the acid phosphatase test for semen resides in the 1000-fold excess of acid phosphatase activity in semen relative to other body fluids and tissues; in short, it was suggested that the specificity of the test is quantitative rather than qualitative.

Recent molecular studies shed new light on the relationships among the 100K acid phosphatases and the tissue specificity of PAP. LAP and PAP have been sequenced as cDNA and the gene structure determined for both [4-6]. The gene for LAP has been determined to be on chromosome 11, affirming its probable identity with ACP2. The gene for PAP has been assigned to chromosome 3 at 3q21-ter. The secreted form of PAP has two identical subunits 354 amino acids in length. The LAP subunit is 393 residues long and consists of three parts: a 350 residue core enzyme followed at the C-terminal end by a 24 residue transmembrane segment and a 19 residue cytoplasmic segment. The sequences of PAP and the LAP core segment are about 50% homologous; critical features such as active site residues, cysteine positions, and glycosylation sites are conserved. The positioning of introns in the two genes is nearly identical; the LAP gene has an additional exon 3' to the core sequence exons which encodes the transmembrane and cytoplasmic segments. There is a significant difference in the structure of the two genes; despite the common placement of introns, the PAP gene spans about 50 Kb compared to about 10 Kb for LAP.

Nevertheless, the homologies in sequence and intron structure are evidence of a common evolutionary origin. Comparison of the human PAP and LAP sequences to rat secretory and lysosomal acid phosphatase sequences shows stronger homologies between enzyme types than between the enzymes from a species [7], suggesting that the PAP and LAP ancestral genes split prior to the split between primates and rodents some 50 million years ago.

The regulation of the PAP and LAP genes is markedly different. The promoter region of LAP is typical of "housekeeping" genes and the gene appears to be expressed in most tissues [5]. In contrast, high expression of PAP is known to be androgen dependent and PAP mRNA can be detected only in prostatic tissue among tissues surveyed [8]. Significantly, PAP mRNA was not detected in placental tissue, a tissue which does express ACP3; this provides evidence that PAP and ACP3 are encoded at separate genetic loci.

There have been no studies on acid phosphatase expression in vaginal tissue. Accordingly, it is not known whether PAP is expressed in this tissue or whether VAP is related to ACP3. The question of PAP's tissue specificity is thus nearly but not completely resolved. However, the tools are available to answer the question.

PROSTATIC ANTIGEN p30

The prominent semen protein of about 30,000 molecular weight has been "discovered" by at least four different groups [9]; it has been called variously gamma-semenoprotein, E1 protein, p30, and prostate specific antigen (PSA). Although there has been some argument that these proteins were not the same, sequence analyses and immunological studies have firmly established that they are in fact identical entities. Under the name PSA, the protein has become the marker of choice for early detection of prostatic cancer and for monitoring the progression of cancer in affected individuals. As p30, it has been used as a marker for semen in forensic testing.

The initial sequence analysis of PSA/p30 established its identity, theretofore unknown, as a serine protease [10]. Subsequent studies refined this identification and placed PSA/p30 in a specific family of serine proteases, the kallikreins [11]. There are two other members of the human kallikrein family, tissue kallikrein (KLK1) which is expressed in various tissues including kidney, pancreas, and salivary gland, and glandular kallikrein (KLK2) which like PSA/p30 appears to be expressed only in prostate tissue. PSA/p30 shows 60% sequence homology to KLK1 and 78% sequence homology to KLK2. The homology of PSA/p30 with KLK1 could account for the anecdotal reports of rare positive immunological reactions of anti-p30 with saliva; the two proteins could share a cross-reactive epitope recognized by some antisera.
PSA/p30, KLK1, and KLK2 have nearly identical gene structures, each with 5 exons and each spanning about 5 Kb, further affirming their common evolutionary origins. The three genes are located in a gene complex on chromosome 19 at 19q13. The gene order is KLK1 - 31 Kb - PSA/p30 - 12 Kb - KLK2.

Like PAP, both PSA/p30 and KLK2 mRNA expression is regulated by androgens; this is not surprising given the near identity of the promoter regions of the two genes [12,13]. Although KLK2 is expressed at the mRNA level, a secreted protein corresponding to KLK2 has not been identified; either the mRNA is not translated or tests with the necessary sensitivity and specificity have yet to be employed. The expression of PSA/p30 mRNA appears to be limited to the differentiated epithelial cells of prostatic tissue; this affirms previous findings on the tissue specific expression of PSA/p30 protein.

PSA/p30 has an amino acid substitution at its active site that precludes true kallikrein activity. The enzyme is present in an active form in semen and has been demonstrated to be involved in the liquefaction of the seminal clot by specific cleavage of a major seminal vesicular protein [14]. (In contrast, PSA/p30 leaked into blood plasma is over 90% complexed with the serine protease inhibitor alpha anti-chymotrypsin.) Although the 3-D structure of PSA/p30 has not been solved, the 3-D structure is known for a number of serine proteases; based on structural homologies, it is possible to model the physical structure of PSA/p30. This allows the use of rational drug design methods to develop specific synthetic substrates.

SALIVARY AMYLASE

Human saliva contains very high levels of amylase activity; this activity is higher by several orders of magnitude than amylase activity in other human body fluids and tissues [15,16]. Salivary amylase is also found in other primates, in rodents, and in lagomorphs but not at appreciable levels in other mammals [17]. Amylase is ubiquitously expressed in mammalian pancreatic tissue and as a result is found in blood plasma, urine, and feces. The human salivary and pancreatic amylases can be distinguished electrophoretically and have been determined to be the products of two distinct loci, AMY1 and AMY2, respectively [18]. Forensic efforts to develop enzymatic and immunological tests to distinguish the salivary and pancreatic amylases have not been highly successful.

Molecular biological studies have provided considerable insight into our understanding of amylase expression in humans. Sequence comparison at the protein and DNA levels shows that the salivary and pancreatic amylases are highly homologous; the protein sequences are 97% identical [19]. The two human amylases are more similar to each other than either is to the corresponding mouse amylases, indicating that the human AMP1 and AMY2 genes diverged after the split of the two species. The high sequence homology between the salivary and pancreatic amylases no doubt accounts for the relative failure of enzymatic and immunological tests to qualitatively distinguish the two.

The AMY1 and AMY2 genes are located in a gene complex on chromosome 1 at 1p21. The normal human genome in fact contains three AMY1 genes (AMY1A, AMY1B, and AMY1C) and two AMY2 genes (AMY2A and AMY2B); variant genotypes containing more or fewer AMY1 genes have been reported [20]. The AMY1 genes differ from the AMY2 genes in having two inserted elements, an actin pseudogene and a retroviral-like segment, located 5' to the coding sequence. A 1 Kb segment immediately 5' to the coding sequence, consisting entirely of sequence from the inserted elements, has been demonstrated in reporter gene constructs to be responsible for tissue specific expression in parotid gland; the inserts appear to have combined to form a tissue specific promoter for AMY1 expression [21]. The identification of this unique tissue specific promoter accounts for the expression of amylase in human saliva.

GROUP SPECIFIC COMPONENT

Group specific component (Gc) was one of the first genetically polymorphic proteins to be identified. At present, three common alleles are recognized, Gc*1F, Gc*1S, and Gc*2, and over 100 rare variants are known [22]; it is one of the most polymorphic of the plasma proteins. The protein possesses strong binding affinity for several biologically significant ligands including vitamin D, G-actin, complement factor 5a, fatty acids, and lymphocyte immunoglobulin receptor; its primary function is believed to be the plasma transporter for vitamin D but obviously other functions cannot be discounted [23].

The first protein and cDNA sequence analyses of Gc indicated homology to two other plasma proteins, albumin and alpha-fetoprotein; Gc is a somewhat truncated form of the other two [24]. This homology allows a 3-D structure for Gc to be modeled on the 3-D structure of albumin; since albumin is also a ligand binding protein, the ligand binding domains of Gc can also be posited. The vitamin D and G-actin binding regions appear to be situated at the opposite ends of the protein, in segments spanning residues 14-58 and 350-403 respectively [23]; these correspond to binding domain IA and III in albumin.

The homology of Gc with albumin and alpha-fetoprotein allowed primers to be designed for sequencing Gc exons without interruption by introns. The sequence substitutions for the common *1F, *1S, *2, and *1A1 variants were located in exon 11 at codons 416, 420, and 429. The substitutions distinguishing the *1F, *1S, and *2 alleles are at 416 and 420 and are both transversions, T->G at 416 and C->A at 420; a G->A transition at 429 defines the 1A1 allele [25]. Interestingly, the 1A1 polymorphism appears identical in Australian aborigines and American blacks. The corresponding amino acid substitutions are indicated in the table below and account for the observed electrophoretic variation among the types.

	416	420	429	net charge pH7
Gc*1F	asp	thr	arg	0
Gc*1S	glu	thr	arg	0
Gc*2	asp	lys	arg	+1
Gc*1A1	asp	thr	his	-1

This pattern of substitution suggests that Gc*1F is the ancestral form of the gene. None of these residues are located within the identified ligand binding regions although they are found in the same domain as the G-actin binding region. Their biological significance, if any, remains to be determined.

HAPTOGLOBIN

Haptoglobin (Hp), like Gc, was one of the first protein polymorphisms to be identified in humans. It is unique among the common genetic polymorphisms in possessing allelic forms that differ in size as well as charge. Three common alleles, Hp*1F, Hp*1S, and Hp*2, are found in all populations. Each encodes a single polypeptide chain which undergoes posttranslational cleavage yielding an alpha-chain and a beta-chain; these combine as alpha-beta dimers in the Hp 1 types and as extended multimers in the Hp 2-1 and 2 types [24].

The Hp*2 alpha-chain is nearly twice as large as the Hp*1F and *1S alpha chains, ca. 16,000 vs. 8,000. It was originally postulated that the Hp*2 allele arose through an end-to-end partial duplication. Gene sequence analysis has demonstrated that this is in fact the case [26]. The junction point for the duplication connects a point in intron 4 of the *1F gene sequence with a point in intron 2 of the *1S gene sequence, forming a FS hybrid; the appropriate splice sites were in frame, allowing the formation of a functional extended gene.

Gene sequence analysis has also revealed the basis for the Hp 2-1mod phenotype found in Blacks [27]. This phenotype is determined by a mutation in the promoter region of an Hp*2 allele; this mutation results in reduced expression of the Hp*2 allele product, possibly through altered interaction with an IL-6 response element. There is evidence that other phenotypic variation in Hp expression might be the result of mutation in the Hp promoter region.

ADENYLATE KINASE

Due to its small size and important role in energy metabolism, adenylate kinase (AK) has been extensively studied [28]; protein sequences are known for several species and the 3-D structure has been solved for the pig enzyme. This allows interpretation of the sequence data generated for the human allelic products.

Human AK is a monomeric enzyme, 194 amino acids in length. The AK1*1 and AK1*2 allelic proteins differ in a glu->gln substitution at residue position 123. This substitution accounts for the different electrophoretic mobilities of the allelic products. Residue position 123 sits at the end of an alpha helix which arches over the top of the catalytic site and which contains two args which interact with substrate. Glu at this position forms an ion pair with arg-127 further on in the helix, contributing to the stability of the helix and the catalytic site. Substitution of a gln at this position potentially destabilizes both. Thus, the substitution can account also for allelic differences in activity and stability.

CONCLUSION

This excursion into the molecular biology of the classical markers provides some insights into their behavior. Although these markers are fading from use, we are now in a position to understand at a fundamental level the nature of our experience with their past use.

REFERENCES

1. Vincent JB, Crowder MW, and Averill BA (1992) Trends Biochem. Sci. 17: 105.
2. Harris H and Hopkinson DA (1976) Handbook of Enzyme Electrophoresis in Human Genetics (Elsevier Publ.) sec. 3.1.3.2.
3. Sensabaugh GF (1982) In Isozymes: Current Topics in Biological and Medical Research 6: 247.
4. Peters C, et al. (1989) Biol. Chem. Hoppe-Seyler 370: 177.
5. Geier C, Figura K, and Pohlmann R (1989) Eur. J. Biochem. 183: 611.
6. Sharief FS and Li SSL (1992) Biochem. Biophys. Res. Comm. 184: 1468.
7. Roiko K, Janne OA, and Vihko P (1990) Gene 89: 223.
8. Solin T et al. (1990) Biochem. Biophys. Acta 1048: 72.
9. Sensabaugh GF and Blake ET (1990) J. Urol. 144: 1523.
10. Watt KWK, et al. (1986) Proc. Nat. Acad. Sci. USA 83: 3166.
11. Riegman PHJ, et al (1992) Genomics 14: 6.
12. Schedlich LJ, Bennetts BH, and Morris BJ (1987) DNA 6: 429.
13. Riegman PHJ, et al (1989) Biochem. Biophys. Res. Comm. 59: 95.

14. Lilja H (1985) J. Clin Invest. 76: 1899.

15. Willott G (1974) J. Forensic Sci. Soc. 14: 341.

16. Kipps AE and Whitehead PH (1975) Forensic Sci. 6: 137.

17. Meisler MH and Gumucio DL (1986) In Molecular and Cellular Basis of Digestion (P Desnulle, H Sjostrom, and O Noren, eds., Elsevier Press, Amsterdam) p.457.

18. Merritt AD, et al (1987) Amer. J. Human Genet. 25: 510.

19. Horii A, et al (1987) Gene 60: 57.

20. Bank RA, et al (1992) Human Genet. 89: 213.

21. Ting CN, et al (1992) Genes and Develop. 6: 1457.

22. Cleve H and Constans J (1988) Vox Sang 54: 215.

23. Haddad JG, et al (1992) Biochem. 31: 7174.

24. Bowman BH and Yang F (1987) In The Plasma Proteins, 2nd Ed. Vol V. (FW Putnam, ed., Academic Press) p.1.

25. Reynolds R and Sensabaugh GF (1990) Adv. Forensic Haemogenet. 3: 158.

26. Maeda N, et al. (1984) Nature 309: 131.

27. Maeda N (1991) Amer. J. Human Genet. 49: 158.

28. Schultz GE, et al. (1986) Eur. J. Biochem. 161: 127.

29. Luz CM, et al. (1990) Biochem. Biophys. Acta 1038: 80.

In the interests of brevity, the references listed above tend to be survey articles; a complete reference list would have been much longer. I apologize to any that may not have been included.