The amino acid sequence of human sperm protamine P1

D. J. McKay, B. S. Renaux and G. H. Dixon

Department of Medical Biochemistry, Faculty of Medicine, University of Calgary, Health Sciences Centre, 3330 Hospital Drive N.W., Calgary, Alberta T2N 4N1, Canada

(Received 25 April 1985)

Human sperm protamines have been extracted from spermatozoa pooled from several donors, converted to their S-pyridylethylated derivatives and resolved into two major components, P1 and P2, by Bio-Rex 70 chromatography. Protamine P1 was further purified by Bio-Gel P-10 chromatography and sequenced directly on a gas phase protein sequencer for 43 residues. To complete the sequence, P1 was cleaved at methionine 36 and the C-terminal tetradecapeptide was purified by h.p.l.c. and sequenced completely. The 50 residue sequence is:

\[
\begin{array}{cccccccc}
10 & 20 & 30 & 40 \\
\text{ARYRC CRSQS RSRYY RQRQR SRRRR RRSCQ TRRRA MRCCR} \\
50 & \\
\text{PRYRP RCRRH.}
\end{array}
\]

This sequence has a calculated molecular weight of 6674 and is homologous with four other published mammalian protamine sequences.

As part of a programme to study the structure and organization of the sperm protamine genes through phylogeny, we wished to complete the sequence of a human protamine polypeptide in order to design a suitable oligonucleotide probe for the isolation of the human gene. Partial sequences for two human protamines prepared by the method of Kolk and Samuel (1975) had been determined by Gaastra et al. (1978). We were particularly interested in the M-Y-H-R sequence near the C-terminal end of component I for designing an oligonucleotide probe and we wished to confirm and extend this sequence. A complete sequence for the corresponding protamine from bull had been published (Coelingh et al., 1972), and during our work three more complete sequences, from boar (Tobita et al., 1983), ram (Sautiere et al., 1984) and mouse (as a cDNA; Kleene et al., 1985), became available. These sequences indicated that mammalian protamines were 47-50 residues in length and contained 52-54% arginine and 12-18% cysteine. This high cysteine content is a characteristic feature of mammalian protamines and distinguishes them from bird and fish protamines, which lack cysteine. Bloch (1969) classified mammalian protamines as 'stable protamines or basic
keratins', recognizing that the presence of the high cysteine content led, during the maturation of mammalian sperm, to an extensive system of inter-molecular disulfide bridges which impart great stability to the mammalian nucleoprotamine complex and hence to the entire sperm nucleus.

Only one protamine has been found in bull (Coelingh et al., 1972), ram (Sautiere et al., 1984) and boar sperm (Tobita et al., 1983), while two have been found in mouse (Kleene et al., 1983; Balhorn et al., 1977) and two (Kolk and Samuel, 1975; Gaastra et al., 1978) or more (Pongsawasdi and Svasti, 1976; Svasti and Talupphet, 1979) are present in human sperm. Our primary interest has been in the sequence of human protamine P1 but during this work we have also become interested in the sequence of the second major human protamine, P2, whose sequence is quite different from that of protamine P1 (Gaastra et al., 1978).

In this communication we present the complete amino acid sequence of human protamine P1. This sequence essentially confirms the N-terminal sequence of Gaastra et al. (1978) but differs markedly from their C-terminal sequence.

**Materials and Methods**

**Purification**

Human semen was obtained from the Calgary Infertility Clinic by arrangement with Drs. P. Taylor and D. Mortimer, Department of Obstetrics and Gynaecology, Faculty of Medicine, University of Calgary, following positive male fertility tests, and stored at -20°C until use. The purification procedure is based on that of Kolk and Samuel (1975). The spermatozoa from 20 ml of semen were isolated by centrifugation, washed twice with 0.896 sodium chloride, dissolved in 10 ml of 6 M guanidine hydrochloride (GuHCl) plus 0.5 M Tris/HCl plus 0.1 M 2-mercaptoethanol, incubated at room temperature for 1 h and pyridylethylated with vinyl pyridine (0.25 M, 1.5 h, room temperature). This method of alkylation was chosen because the corresponding PTH-pyridylethyl cysteine is more easily identified by our h.p.l.c. method. The resulting solution was acidified with 2 ml of 5 M HCl (Coelingh et al., 1969) and dialyzed exhaustively in Spectrapore 3 tubing against several 4-liter changes of 2% HCl. The dialysate supernatant obtained following centrifugation was brought to 10% GuHCl and 0.1 M sodium phosphate, pH 7.5, in 20 ml and applied to a Bio-Rex 70 column (Bio-Rad, 0.6 x 60 cm) equilibrated with the same solution. The column was eluted successively with 50 ml of 10% GuHCl, 50 ml of 20% GuHCl, and a linear gradient of 100 ml of 20-60% GuHCl, all in phosphate buffer. Fractions under the four main 230 nm absorbance peaks were pooled, exhaustively dialyzed against water and concentrated on a vacuum centrifuge. An aliquot of each fraction was taken for amino acid analysis. The protamine P1-containing fraction was purified on a Bio-Gel P-10 column (Bio-Rad, 1.5 x 90 cm) equilibrated and eluted with 4 M GuHCl (Bellvé et al., 1975). The 80-drop fractions under the 230 nm absorbance peak containing protamine P1 were pooled, dialyzed against water, dried down in a vacuum centrifuge, redissolved in 2% acetic acid and stored at -20°C.
Amino acid analysis

Amino acid analysis was performed on either a Beckman 121 M or 6300 amino acid analyzer following HCl hydrolysis (6 N HCl plus 0.1% phenol plus 0.1% thioglycolic acid, 100°C, 24 h, in vacuo).

Protein sequencing

Protamine was sequenced on an Applied Biosystems 470A gas phase protein sequencer using methanolic HCl conversion and reagents from Applied Biosystems. Biobrene (30 µl; Applied Biosystems' polybrene) was precycled for three cycles (2 x 02 nfil degradation cycles plus 1 x 0.2 nvac degradation cycle) prior to each run. (0.2 nfil and 0.2 nvac are standard programme cycles supplied by Applied Biosystems for polybrene pre-conditioning and protein degradation, respectively.) A modified 02 nvac degradation cycle in which the TFA (trifluoroacetic acid) delivery time was increased from 15 to 25 min was used where Pro-Arg bonds, which are incompletely cleaved with the standard 02 nvac cycle, were indicated. The PTH-amino acids from the sequencer were identified by h.p.l.c. using a Varian 5040 LC equipped with a Varian 401 data system, a Waters WISP 710B autosampler and a Waters 440 two-channel u.v. detector (254 nm and 313 nm). A Beckman Ultrasphere ODS column (5 micron, 0.46 x 25 cm) with a Whatman Co:Pell ODS guard column (0.21 x 5.0 cm) was eluted with a 25-min sodium acetate (pH 5.2, 15-50 cm)/acetonitrile programme at 50°C and a flow rate of 1.0 ml/min. All PTH-amino acids except the Lys/Phe pair were well resolved and PTH/pyridylethylCys eluted in front of PTH-Met. PTH-norleucine was used as an internal standard in each sequencer vial.

Peptide h.p.l.c.

Peptides from digests of protamine P1 were purified by h.p.l.c. on a Varian 5060 LC equipped with a Varian 401 data system, a Waters WISP 710B autosampler and a Waters 441 two-channel u.v. detector (214 nm and 280 nm). Samples were applied in 0.1% aqueous TFA and eluted with gradients of 0.09% TFA in acetonitrile at 30°C on a Vydac C18 column (5 micron, 0.46 x 25 cm) equipped with a Whatman Co:Pell guard column (0.21 x 5.0 cm).

Results

The purification procedure for human protamine P1 was developed from the procedure of Kolk and Samuel (1975) but contains two major modifications. Pyridylethylolation by vinyl pyridine was used in place of aminooethylation by ethylenimine because PTH-pyridylethylCys is more easily identified on our h.p.l.c. PTH-amino acid analysis system. Secondly, since retention of any three-dimensional structure was unimportant for sequencing, an acid precipitation step was added after the derivatization to remove considerable amounts of protein contaminants and to improve the resolution during the subsequent Bio-Rex 70 chromatography. The elution profile from the Bio-Rex 70 column is shown in Fig. 1. Amino acid analysis was used to monitor the purity of the P1 after the last two stages of the purification because it was found to be more sensitive to protein contaminants than the more traditional polyacrylamide gel electrophoresis. As shown by its
amino acid composition in Table 1, the P1-containing fraction from Bio-Rex 70 was still contaminated with other proteins which were largely removed by gel filtration on a Bio-Gel P10 column (elution profile not shown). The amino acid composition shown in Table 1 for the P-fraction after Bio-Gel P-10 chromatography indicates the highest purity of P1 obtained from eleven preparations. In this preparation only 20 nmol (approx. 133 µg) of pure P1 was obtained from 20 ml of semen.

When 4 nmol of P1 was subjected to Edman degradation on the gas phase protein sequencer, 44 of the first 45 residues were unequivocally identified and methionine was definitely established at position 36 as shown in Fig. 3. The strategy then was to isolate and sequence the small C-terminal peptide obtained after cyanogen bromide (CNBr) cleavage. In the first attempt 8 nmol of P1 was digested with CNBr (70% formic acid, 5% CNBr, 20 h room temperature, glass tube) and the fragments were purified by h.p.l.c. to yield 3 nmol of the C-terminal peptide, CB2. All of CB2 was subjected to 14-02 n voc degradation cycles on the gas phase protein sequencer yielding an unequivocal 13 residue sequence despite heavy carryover following the two Pro-Arg bonds; however, no histidine could be identified at position 14 of this peptide. Also, the tyrosine at position 7 was completely converted to an unidentified derivative. In a second CNBr cleavage reaction with the remaining 8 nmol of P1 a few crystals of phenol were added to the reaction mixture to protect, at least partially, the tyrosine residue, and a plastic tube was substituted for the glass tube to try to improve the yield of CB2. Following h.p.l.c. of the digest (see Fig. 2) 4 nmol of CB2 was obtained, for a 50% yield. The sequencing of CB2 was repeated as before using all 4 nmol
Table 1. Amino acid compositions (residues/mol)
Cysteine was determined as pyridylethylcysteine in this work and as aminoethylcysteine by Kolk and Samuel (1975). Tryptophan was not determined. The residue numbers in brackets were found by sequencing.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Bio-Rex 70</th>
<th>Bio-Gel P-10</th>
<th>CB2 Kolk &amp; Samuel (1975)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>P1</td>
<td>P1</td>
<td>P1</td>
</tr>
<tr>
<td>Asp</td>
<td>2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>1.8</td>
<td>1.0 (1)</td>
<td>1.06</td>
</tr>
<tr>
<td>Ser</td>
<td>3.9</td>
<td>4.1 (5)</td>
<td>4.95</td>
</tr>
<tr>
<td>Glu</td>
<td>5.2</td>
<td>4.2 (4)</td>
<td>4.18</td>
</tr>
<tr>
<td>Pro</td>
<td>2.6</td>
<td>2.1 (2)</td>
<td>2.0 (2)</td>
</tr>
<tr>
<td>Gly</td>
<td>2.1</td>
<td>0.7 (0)</td>
<td></td>
</tr>
<tr>
<td>Ala</td>
<td>2.7</td>
<td>2.0 (2)</td>
<td>1.99</td>
</tr>
<tr>
<td>Cys</td>
<td>7.1</td>
<td>5.6 (6)</td>
<td>2.7 (3)</td>
</tr>
<tr>
<td>Val</td>
<td>1.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>0.9</td>
<td>0.6 (1)</td>
<td>0.73</td>
</tr>
<tr>
<td>Ile</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>2.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>2.5</td>
<td>3.4 (4)</td>
<td>0.9 (1)</td>
</tr>
<tr>
<td>Phe</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>1.4</td>
<td>1.0 (1)</td>
<td>0.9 (1)</td>
</tr>
<tr>
<td>Lys</td>
<td>1.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>16.1</td>
<td>22.9 (24)</td>
<td>7.2 (7)</td>
</tr>
<tr>
<td>Basis</td>
<td>MW=6670</td>
<td>His=1.0</td>
<td>Pro=2.0</td>
</tr>
</tbody>
</table>

Fig. 2. H.p.l.c. separation of a CNBr digest of S-pyridylethylated human protamine P1. The fraction size was 0.5 ml. (For experimental details see Materials and Methods).
except that a modified degradation cycle was used for the two Pro-Arg bonds (see Materials and Methods). This time the tyrosine residue was only 50% derivatized, the carryover was reduced to 20% at the 14th cycle and histidine was clearly identified as the 14th and C-terminal residue in approx. 15% yield. The sequence of CB2 matched its amino acid composition and completed the sequence of PI. Thus, human protamine PI is 50 residues long and has a mol. wt. of 6674. Fig. 3 shows the complete amino acid sequence of PI together with the N-terminal and C-terminal sequences of Gaastra et al. (1978) for comparison.

Discussion

Despite its small size and unusually basic nature, human protamine PI proved to be exceedingly difficult to purify and sequence. The results presented here were obtained with the last and most satisfactory of 11 preparations. Perhaps because of our choice of a different alkylation method both the quality and the quantity of PI obtained was quite variable in contrast to the results of Kolk and Samuel (1975). As judged by amino acid analysis, P2, the second major protamine, was the most common contaminant but other proteins were also often present. Purification of the protamine from bull (Coelingh et al., 1972), boar (Tobita et al., 1983) and ram (Sautiere et al., 1984) seems not to have been a major problem but these species each have only one protamine. Although mouse has two protamines the sequence of component 1 was obtained by DNA sequencing methods (Kleene et al., 1985) so that purification of the two proteins was not necessary. In purified preparations of PI, methionine typically gave values in the 0.5-0.7 residue range while other amino acids, most notably glycine, which was subsequently found not to be in the sequence, frequently gave higher values. This result when taken together with the failure during two initial attempts to

Fig. 3. The complete amino acid sequence of human protamine PI. Each residue determined by a gas phase protein sequencer is indicated by '->'. The N-terminal and C-terminal sequences of Gaastra et al. (1978) are shown underneath.
cleave P1 with CNBr suggested that methionine might not actually be present in the sequence (or that it might have been oxidized to the sulfoxide during the protamine fractionation). The early sequence work was done with a Beckman 890C liquid phase sequencer which required much more material than the gas phase sequencer. In three different attempts on 15-20 nmol each, the sequence became uninterpretable beyond residue 34 because of extreme carryover, the serine at position 21 was missed and residues 28-32 were mis-assigned. Two digests each with chymotrypsin, CNBr and N-bromosuccinimide on material less pure than presented here failed to yield any useful fragments on h.p.l.c. The completion of the P1 sequence became possible because the gas phase sequencer produced a 43-residue N-terminal sequence on only 4 nmol of protein and at the same time clearly established that methionine was present in the sequence at position 36.

The N-terminal sequence of Gaastra et al. (1978) agrees at 14 of 19 positions with the sequence presented here. These authors experienced difficulty in identifying AminoEthylCys at positions 5 and 6 and Arg at positions 13, 16 and 18. Their C-terminal sequence differs markedly from our sequence particularly with respect to the positions of Met, Tyr and His, all of which show low codon redundancy and potentially would have been useful in the design of an oligonucleotide probe for the P1 cDNA but which are well separated in our sequence.

In Fig. 4 the human P1 sequence is compared to the corresponding protamine sequences from bull (Coelhingh et al., 1972), ram (Sautiere et al., 1984), boar (Tobita et al., 1983) and mouse (Kleene et al., 1985). It is necessary to insert a three residue gap into the bull sequence in order to maximize its homology with the ram sequence and to bring its length up to 50 residues as in the other four completed sequences. The published evidence for the peptide bond between Cys38 and Tyr39 in the bull sequence seems rather weak.

Fig. 4. A sequence comparison of 5 completed mammalian protamine amino acid sequences in 1 letter code (IUPAC (1968)). For clarity 'r' has been used for Arg. Each sequence is split into 3 regions with the number of Arg residues in each region shown underneath.
so that a C-R-R tripeptide may actually be present between them as in the ram and boar sequences. The five sequences shown in Fig. 4 have been divided into three regions of 14, 22 and 14 residues, respectively, to highlight the striking concentration of arginine residues in the central region of the sequence as noted previously by Coelingh et al. (1972) for the bull sequence. A similar concentration of arginine residues in the central region is also observed in the ram, boar and mouse sequences; however, in the human P1 sequence this concentration is not so marked since the arginine concentration of 54% in the central region is only slightly above the value of 48% arginine for the whole sequence. The run of six arginines in the middle of the central region is conserved in all five sequences. In the N-terminal 14-residue region, the N-terminal hexapeptide sequence A-R-Y-R-C-C is totally conserved as are Ser10, Ser12 and Arg13, for a total of nine identical residues in all five sequences. At position 8, a hydroxyamino acid is always present with the serine in mouse, boar and man replaced by a threonine in bull and ram. The C-terminal 14-residue region has only residues Arg46 and Cys47 identical in all five sequences; however, if the C-R-R tripeptide is inserted into the bull sequence then the homology among the bull, ram and boar sequences matches that of their N-terminal regions at 11 of 14 identical residues and all five sequences have a C-C sequence starting at position 38 (36 in the mouse), an Arg at position 40 and a Tyr at position 42 or 43. Now that more sequence data are available, it is clear the C-terminal region is not as variable in length or sequence as was previously suggested (Monfoort et al., 1973).

In some recent comparative studies of protamine gene sequences in salmonid fishes by dot-blot hybridization (Dixon et al., 1985; Moir and Dixon, in preparation), we have observed that the divergence of the protamine gene nucleotide sequence during the evolution of these species has been much more rapid than that of the corresponding histone genes. By inspection of the mammalian protamine sequences in Fig. 4, a similar phenomenon is seen at the amino acid level. In comparing the human P1 sequence with that of, respectively, mouse, boar, ram and bull (assuming C-R-R inserted at 39-41) there are 21/50, 19/50, 23/50 and 24/50 amino acid changes. Even some of the cysteine residues which are involved in disulfide bond formation and stabilize the nucleoprotamine complex (Bloch, 1969) and which are among the amino acids least likely to mutate (Dayhoff et al., 1972) have changed at positions 14, 21 or 22 and 29-30. According to Dayhoff et al. (1972) the probability of cysteine changing to tyrosine as at position 14 in human P1 is essentially zero. Thus, it appears that mammalian protamines have been quite free to evolve provided that the N-terminal sequence and certain of the cysteine residues are conserved and the arginine content remains around 50%.

Acknowledgements

This work was supported by the Medical Research Council of Canada through an operating grant to G.H.D. and the Alberta Heritage Foundation for Medical Research which provided the Applied Biosystems Gas Phase Sequencer, the Beckman 6300 Amino Acid Analyzer and the Varian h.p.l.c. system for the Faculty of Medicine Protein Analysis Laboratory.
References