Sequence Analysis of Protamine mRNA from the Rainbow Trout

DEPURINATION AND NEAREST NEIGHBOR ANALYSIS OF PROTAMINE cDNA*

(Received for publication, September 14, 1976)

PETER L. DAVIES,* L. NORMAN FERRIER, AND GORDON H. DIXON§

From the Department of Medical Biochemistry, University of Calgary, Calgary, Alberta, T2N 1N4, Canada

Protamine cDNA, which was a full length copy of protamine mRNA, was labeled during its synthesis by using deoxynucleoside [$\alpha-^{32}$P]triphosphates. Depurination analysis showed that there were 19 different pyrimidine oligonucleotides in protamine cDNA, some of which contained isomeric sequences. The stoichiometry of the pyrimidine oligonucleotides indicated that, while some sequences probably occur in each of the protamine mRNA components, other sequences are clearly absent from one or more of the components. Several of the pyrimidine oligonucleotides had sequences consistent with the amino acid sequences of the rainbow trout protamines. The longest oligopyrimidine tract, C,T,, showed that there were 19 different pyrimidine oligonucleotides to be determined. Information on the type of arginine codons used in protamine mRNA was deduced partly from nucleotide sequences and partly from nearest neighbor frequency analysis of protamine mRNA. This report presents sequences of oligopyrimidine tracts released from protamine cDNA after depurination. In addition, the fact that the $^{32}$P-labeled protamine cDNAs prepared with reverse transcriptase were full length copies of the mRNA enabled the stoichiometries of the pyrimidine nucleotides to be determined. Information on the type of arginine codons used in protamine mRNA was deduced partly from nucleotide sequences and partly from nearest neighbor frequency analysis.

EXPERIMENTAL PROCEDURES

Materials

Highly purified reverse transcriptase from avian myeloblastosis virus was kindly supplied by Dr. Joseph W. Beard, Life Sciences Research Laboratories, Gulfport Laboratory, St. Petersburg, Fla. Snake venom phosphodiesterase, spleen phosphodiesterase, micrococcal nuclease, and bacterial alkaline phosphatase were purchased from Worthington Biochemical Corp. The latter enzyme was dialyzed against distilled water before use to remove ammonium sulfate. Omnifluor and $\alpha-^{32}$P-labeled deoxynucleoside triphosphates of specific activity 135 Cl/mmol were obtained from New England Nuclear. Oligo(dT)$\beta_4$ was obtained from P-L Biochemicals Inc. and $\alpha-^{32}$P-labeled oligo(dT)$\beta_4$ gel markers were a gift from Dr. Hans van de Sande of this department. Calf thymus DNA was from Sigma Chemical Co. and yeast RNA for preparing homomixtures was from B. D.
Sequence Analysis of Protamine mRNA from Rainbow Trout

H. Chemicals Ltd. MN300 cellulose and DEAE-cellulose powders were from Macherey, Nagel and Co., cellulose acetate membrane strips were from Schleicher and Schuell, Inc., oligo(dT)-cellulose (type T1) was from Collaborative Research Inc., and DE81 ion exchange paper was from Whatman. Autoradiography was done on Kodak RP X-omat or Dupont Cronex 4-x-ray film. Trout testes for the preparation of protamine mRNA were collected from Dantrout, Brande, Denmark at the beginning of October 1974 and after freezing on dry ice were stored at -80°C.

Methods

Isolation of Protamine mRNA - Postmitochondrial supernatant was prepared from rainbow trout testis as described by Gedamu et al, with the exception that NaCl replaced KCl in all buffers used. Poly(A)-containing RNA was extracted from this supernatant using the method of Keftescheck et al. (14). The NaCl concentration of the postmitochondrial supernatant was adjusted to 0.5 M by the addition of 2 M NaCl. The temperature of the supernatant was quickly raised from 4°C to 60°C, and the filter was passed at a rate of 1 mL/min through an oligo(dT)-cellulose column (15-mL bed volume) which had been equilibrated with 0.5 M NaCl, 0.5% SDS, 10 mM Tris/HCl (pH 7.5) at 22°C. The column was then washed with the same buffer (150 mL) followed by buffer without SDS (150 mL). At 4°C a further 200 mL of buffer without SDS and 190 mL of 0.1 M NaCl, 10 mM Tris/HCl (pH 7.5) were passed through the column before the poly(A)-containing RNA was eluted with H2O. This RNA was precipitated overnight in 0.7 M ammonium acetate and -40°C by the addition of two volumes of ethanol. Protamine mRNA was isolated from this RNA fraction following a sucrose gradient sedimentation, a further oligo(dT)-cellulose chromatography step and a second sucrose gradient sedimentation, each carried out essentially as described by Gedamu and Dixon (9). The yield of protamine mRNA by this method was 687 pg from 400 g of frozen testis (1.67 mg/kg).

Preparation of .32P-labeled Protamine cDNA - DNA complementary to protamine mRNA was synthesized using reverse transcriptase from avian myeloblastosis virus according to the procedure of Iatrou and Dixon. The primer for transcription was oligo(dT),. Incubations were done in a volume of 500 M and 32P label was introduced into the cDNA by replacing one of the four unlabeled deoxynucleoside triphosphates at 0.6 mM with the corresponding a-32P-labeled nucleotide at 74 µM. After alkaline hydrolysis of the protamine mRNA template followed by neutralization, the cDNA products were purified by dialysis against H2O at 4°C. When necessary, the protamine cDNA was purified from partial transcription products by polyacrylamide gel electrophoresis as described by Iatrou and Dixon.

Polyacrylamide Gel Electrophoresis - Electrophoresis was carried out on a 15% polyacrylamide, 0.375% bisacrylamide gel (30 x 18 x 0.15 cm) without cooling using the buffer system described by Sanger and Coulson (15). The gel was pre-electrophoresed overnight at 200 V. Lyophilized samples were dissolved in 99% formamide (12 µL mixed with 50% sucrose (3 µL) containing bromphenol blue. Electrophoresis was at 400 V until the bromphenol blue was within 5 to 10 cm of the bottom of the gel.

Depurination - Oligopyrimidine tracts were liberated from 32P-labeled protamine cDNA according to the method of Burton (16). Calf thymus DNA (1 mg/ml) was used as a carrier.

Two-dimensional Separation of Oligonucleotides - The partial purine nucleoside digestion products were separated in two dimensions according to the procedure of Brownlee and Sanger (17). Homomixture c was hydrolyzed for 30 min and was diluted to 3% RNA. Lyophilized samples were dissolved in 7 µL of 0.1 M formic acid (50/50) containing 10 mM EDTA and 10 µL DDEPA for application to the cellulose acetate strip. 32P-labeled oligonucleotides were located on the DEAE-cellulose thin layers by autoradiography and were eluted with 30% triethylamine carbonate (18).

Partial Exonuclease Digestion - To facilitate sequence analysis of oligopyrimidine tract dephosphorylation products were phosphorylated with dialyzed bacterial alkaline phosphatase (19) at pH 9.0 prior to their separation in two dimensions. Partial exonuclease digestion of an oligonucleotide (5,000 to 30,000 Cerenkov cpm) with either snake venom or spleen phosphodiesterase in 12 to 15 µL was done essentially as described by Ling (19). Aliquots (2 to 3 µL) were removed from a snake venom phosphodiesterase digest of the deoxynucleoside triphosphates at 0, 10, 20, 30, 60, and 120 min to a tube at -80°C. Shorter oligonucleotides required a shorter time course. The combined aliquots were lyophilized directly. Aliquots from the spleen phosphodiesterase digest were removed at 1 h intervals up to 4 h and were treated as described above.

Alkaline Phosphatase Treatment of Oligopyrimidines - 32P-labeled oligopyrimidines (greater than 2,000 Cerenkov cpm) extracted from DEAE-cellulose thin layers were digested with dialyzed bacterial alkaline phosphatase (60 µg in 10 µL of 10 mM Tris/HCl (pH 9.5) at 37°C for 70 min. The labeled products, located by autoradiography, were cut out and counted in toluene/Omnifluor.

Nearest Neighbor Frequency Analysis - Nearest neighbor frequencies for each of the four protamine cDNAs individually labeled with one of the deoxynucleoside (a-32P)triphosphates (dTTP, dCTP, dGTP, dATP) were determined by the procedure of Kleppe et al. (20). The 3'-mononucleotides from protamine cDNA (1 to 6 x 104 Cerenkov cpm) were separated by electrophoresis, located by autoradiography, cut out, and quantitated by counting in toluene/Omnifluor.

RESULTS

Catalogue of Pyrimidine Tracts in Protamine cDNA - Reverse transcriptase is capable of synthesizing a full length single-stranded transcript from protamine mRNA primed with oligo(dT), as reported by Iatrou and Dixon. Fig. 1 shows that T-labeled4 protamine cDNA was equal in length to protamine mRNA. When shorter transcription products occurred in other preparations of protamine cDNA they were present in minor amounts and were readily removed by preparative polyacrylamide gel electrophoresis.

Depurination of full length C and T-labeled protamine cDNAs therefore gave a complete catalogue of the pyrimidine tracts in stoichiometric amounts. The base composition of each oligopyrimidine was deduced after two-dimensional separation shown in Fig. 2 from its position on the fingerprint in accordance with the observations of Ling (19). Orientation to the graticule was helped by the absence of pCp, pCCp, and pCCCp from the T-labeled series. The identities of these mono-, di-, and trinucleotides were confirmed by digestion with bacterial alkaline phosphatase. Release of free 32P at 100% from the mononucleotides, 50% from the dinucleotides, and 30 to 35% from the trinucleotides.

The stoichiometry of the oligopyrimidine tracts in the cDNAs was calculated for both the C-labeled and T-labeled series in Fig. 2 in the following way. For the C-labeled series, the number of times a particular oligopyrimidine appears in protamine cDNA

\[
\text{cpm in the oligopyrimidine} = \frac{\text{sum of the cpm in all the oligopyrimidines}}{\text{the number of C residues in protamine cDNA}} \times \text{the number of C residues in the oligopyrimidine}
\]

The corresponding calculation was done for the T-labeled series. However, it was necessary to correct the number of thyminylate residues in protamine cDNA for the oligo(dT) region complementary to the poly(A) tract. The values for the number of deoxyctydylate and thyminylate residues in protamine cDNA were taken to be 75 and 66, respectively.

1. L. Gedamu, G. H. Dixon, and P. L. Davies (1977) Biochemistry, in press.
2. The abbreviation used is: SDS, sodium dodecyl sulfate.
3. K. Iatrou and G. H. Dixon (1977) Cell, in press.
4. T-, C-, G-, and A-labeled protamine cDNAs refer to protamine cDNAs into which a-32P label was introduced by replacing unlabeled dGTP, dCTP, dATP, or dTTP, respectively, with the corresponding a-32P-labeled nucleotide.
Sequence Analysis of Protamine mRNA from Rainbow Trout

Fig. 1. Analysis of protamine cDNA transcript on a 15% acrylamide gel. The gel was prepared and run as described under "Methods." Slot A contained 32P-labeled oligo(dT)12-18 gel markers (50,000 Cerenkov cpn). Slot B contained a 1-μl aliquot from the preparation of protamine cDNA labeled with [α-32P]dATP. The aliquot was removed prior to alkaline hydrolysis of the template. Protamine mRNA (15 μg), prepared as described under "Methods," was run in Slot C. The protamine mRNA was made visible on staining with a 0.005% solution of Stains-all in 50% formamide (21). The gel markers and protamine cDNA were detected on autoradiographs which overlaid their respective slots (A and B).

Additional sequence information can be obtained by determining the base adjacent to the 3' end of the oligopyrimidines in the following manner. The oligopyrimidines released by depurination in the presence of diphenylamine have phosphate groups at both ends. The phosphate group at the 5' end originates from the α position of the pyrimidine nucleotide at the 5' end. However, the phosphate group at the 3' end originates from the α position of the purine nucleotide which was adjacent to this end but which was removed by depurination. The identity of this base, whether adenine or guanine, was determined by separately depurinating and fingerprinting A-labeled and G-labeled protamine cDNAs as shown in Fig. 5. Oligopyrimidines showed up on the autoradiograph of the A-labeled series if adenine was the base adjacent to the 3' end or on the autoradiograph of the G-labeled series if guanine was the base adjacent to the 3' end. The shorter sequences which tend to occur several times in protamine cDNA have both adenine and guanine nearest neighbors to their 3' ends, whereas the longer sequences tend to have either adenine or guanine. The oligopyrimidines C1T3, C1T4, C2T4, C3T2, and C4T2 appear in the A-labeled series but not in the G-labeled series. Conversely C1T3, C2T4, C3T2, and C4T2 appear in the G-labeled series but not in the A-labeled series. T3, which has a stoichiometry of 1.52 in protamine cDNA, appears mainly in the A-labeled series but also to a lesser extent in the G-labeled series. This indicates that the sequence ATTT is the major one and that GTTT occurs less frequently, perhaps in only one of the protamine mRNA components. In accordance with this observation the sequence ATTT has been seen in an endonuclease IV fragment of protamine cDNA which is believed to occur in each of the protamine mRNA components.

The oligopyrimidine sequences extended by nearest neighbor considerations are shown in Table III together with their complementary mRNA sequences and possible coding assignments. The isomeric sequences pCpT and pTcP were shown to be present in equal amounts. Digestion with alkaline phosphatase of these pyrimidine dinucleotides from either C- or T-labeled protamine cDNA liberated 50% of the total cpm as 32P. Several of the RNA sequences tabulated, such as AGAG, have been seen in pancreatic RNase fragments of protamine mRNA sequenced after iodination of the mRNA or following postlabeling of the fragments with polynucleotide kinase and [γ-32P]ATP (22). Other sequences have

---

L. N. Ferrier, et al., manuscript in preparation.

Unpublished observations.
Sequence Analysis of Protamine mRNA from Rainbow Trout

Number of nucleotides in the nonadenylated portion of protamine mRNA (270) and dividing this figure by 1000. Thus ApG occurs 88 × (270/1000) = 24 times and CpG occurs 46 × (270/1000) = 12 times.

DISCUSSION

Depurination of highly labeled DNA complementary to protamine mRNA released short oligopyrimidine tracts which were suitable for nucleotide sequence analysis. In addition, the availability of full length protamine cDNA enabled some information to be deduced which could not come from partial transcripts. Specifically, this information concerned the stoichiometry of the oligopyrimidine tracts in protamine cDNA and the nearest neighbor frequencies of the bases.

The same 19 oligopyrimidines were consistently released by depurination from different preparations of C- and T-labeled protamine cDNAs in a reproducible ratio (Table I). This vouches for the fidelity of transcription. An average stoichiometry of close to 1/DNA copy was obtained for the longer oligopyrimidines, C7T9, C10T7, and C13T4, which were subsequently shown to be unique sequences (Table II). However, the stoichiometry of other oligopyrimidines, such as C9T5 and C7T4, falls clearly below 1. Furthermore, two of these oligopyrimidines (C7T5 and C7T4) are each a mixture of at least two isomeric sequences (Table II). A stoichiometry for these sequences of less than unity is consistent with protamine mRNA being a mixture of mRNA components. Heterogeneity in protamine mRNA is to be expected since protamine itself, as isolated from the rainbow trout (10) or as translated from protamine mRNA in a cell-free assay (9), is made up of at least three polypeptides. Although these polypeptides are closely related in both length (32 to 33 residues) and amino acid sequence, coding assignments dictate that they must in part have different nucleotide sequences. Further heterogeneity in

A number of the RNA sequences shown in Table III must come from the noncoding region of protamine mRNA since in any reading frame they are inconsistent with the known amino acid sequences of the three major protamine components from the rainbow trout (10). By the same criterion other RNA sequences are consistent with a location in the coding region, although this does not prove that they occur there rather than in the noncoding region. The longest oligopyrimidine tract from protamine cDNA, CT13, is complementary to AGGAGAGAGAGG which in turn can code for Arg-Arg-Gly-Gly. This is an amino acid sequence which occurs towards the COOH terminus in each of the protamine components. Nearest neighbor considerations place a cytidine residue at the 5' end of the oligoribonucleotide. Since in each of protamine components an arginine residue precedes the sequence Arg-Arg-Gly-Gly it follows from a knowledge of the arginine codons that CG must precede the dodecanucleotide listed in Table III to give a total sequence CGCAGGAGAGAGG.

Nearest Neighbor Frequency Analysis – The availability of full length A-, C-, G- and T-labeled protamine cDNAs enabled the nearest neighbor frequencies to be calculated for all of the 16 doublets as shown in Table IV. The values for ApT, CpT, GpT, and TpT obtained from T-labeled protamine cDNA were distorted by the large amount of TpT in this sample which arose from extension of the oligo(dT)18 primer in a region complementary to the poly(A) tract. Therefore, each of the values in this set were calculated from the three other sets. Thus, for example, ApT = 100% – (ApA + ApC + ApG). The number of times each doublet occurs in the nonadenylated portion of protamine mRNA can be obtained from the figures in Table IV by multiplying parts per 1000 by the

been confirmed in endonuclease IV fragments of protamine cDNA.b

A number of the RNA sequences shown in Table III must come from the noncoding region of protamine mRNA since in any reading frame they are inconsistent with the known amino acid sequences of the three major protamine components from the rainbow trout (10). By the same criterion other RNA sequences are consistent with a location in the coding region, although this does not prove that they occur there rather than in the noncoding region. The longest oligopyrimidine tract from protamine cDNA, CT13, is complementary to AGGAGAGAGAGG which in turn can code for Arg-Arg-Gly-Gly. This is an amino acid sequence which occurs towards the COOH terminus in each of the protamine components. Nearest neighbor considerations place a cytidine residue at the 5' end of the oligoribonucleotide. Since in each of protamine components an arginine residue precedes the sequence Arg-Arg-Gly-Gly it follows from a knowledge of the arginine codons that CG must precede the dodecanucleotide listed in Table III to give a total sequence CGCAGGAGAGAGG.

Nearest Neighbor Frequency Analysis – The availability of full length A-, C-, G- and T-labeled protamine cDNAs enabled the nearest neighbor frequencies to be calculated for all of the 16 doublets as shown in Table IV. The values for ApT, CpT, GpT, and TpT obtained from T-labeled protamine cDNA were distorted by the large amount of TpT in this sample which arose from extension of the oligo(dT)18 primer in a region complementary to the poly(A) tract. Therefore, each of the values in this set were calculated from the three other sets. Thus, for example, ApT = 100% – (ApA + ApC + ApG). The number of times each doublet occurs in the nonadenylated portion of protamine mRNA can be obtained from the figures in Table IV by multiplying parts per 1000 by the
TABLE I

Stoichiometry of oligopyrimidine tracts in protamine cDNA

Oligopyrimidines from C-labeled and T-labeled protamine cDNAs were prepared as described in the legend to Fig. 2. The 32P-labeled oligopyrimidines were recovered from the DEAE-cellulose thin layers as described under "Methods" and the radioactivity of each was measured by Cerenkov counting. The stoichiometry of the oligopyrimidine tracts was calculated from this data as described under "Results." Duplicate values are presented for both the C-labeled and T-labeled series. The sum of Cerenkov counts in the spots from each plate was 257,540 and 262,870 for Experiments 1 and 2 of the C-labeled series, and 162,970 and 557,234 for Experiments 1 and 2 of the T-labeled series.

| Oligopyrimidine | T-labeled series | C-labeled series | Average stoichiometry |
|-----------------|------------------|------------------|----------------------|
|                 | Experiment 1     | Experiment 2     |                      |
| T,              | 10.08            | 8.88             | 9.48                 |
| T,              | 4.21             | 4.62             | 4.41                 |
| T,              | 1.62             | 1.43             | 1.52                 |
| CT              | 5.74             | 5.25             | 5.48                 |
| CT,             | 3.52             | 3.74             | 3.61                 |
| CT,             | 1.52             | 1.91             | 1.72                 |
| CT              | 0.68             | 0.78             | 0.73                 |
| CT,             | 0.91             | 1.38             | 1.15                 |
| CT,             | 2.55             | 1.80             | 2.18                 |
| CT,             | 0.78             | 0.50             | 0.64                 |
| CT              | 0.60             | 0.68             | 0.64                 |
| CT,             | 1.13             | 1.15             | 1.14                 |
| CT              | 1.40             | 1.35             | 1.38                 |
| CT,             | 0.64             | 0.54             | 0.59                 |
| CT,             | 1.10             | 0.60             | 0.80                 |
| CT              | 0.76             | 0.82             | 0.80                 |
| C,              | 2.96             | 3.35             | 3.15                 |
| C,              | 3.19             | 3.39             | 3.29                 |
| C,              | 20.74            | 18.53            | 19.64                |

protamine mRNA, aside from that introduced by the variable length of the poly(A) tract, could occur in the length and nucleotide sequence of the noncoding region. Recent observations indicate that protamine mRNA can, in fact, be subfractionated into four components on the basis of length after prolonged electrophoresis in denaturing 6% polyacrylamide gels.

While sequence heterogeneity must exist between the protamine mRNA components, sequence homology can also be expected to occur, both in the coding region, because of the similar amino acid sequences of the rainbow trout protamines (10), and in the noncoding region, where substantial sequence homology has previously been observed between eukaryotic mRNAs (3, 23). For these reasons the unique oligopyrimidines such as C,CT, CT, and CT, which have a stoichiometry of close to 1 are likely to represent homologous regions in each of the protamine mRNA components. In the case of Ct, the complementary RNA sequence fits the amino acid sequence Arg-Arg-Gly-Gly which occurs in each of the three major protamine components. Although the same RNA sequence could occur twice in one protamine mRNA component, once in the coding region and once in the noncoding region, and be absent from another component to give an average stoichiometry of 0.87, this is the less likely of the two explanations. Based on this rationale, CCTCTCTCCT is presently being synthesized chemically for use as a specific primer to initiate reverse transcription within the coding region of each protamine mRNA component.

The protamines are unusual polypeptides in that two-thirds of their amino acid residues are arginine and these arginine residues occur in blocks of up to 6 residues long. The arginine codons along with those for serine and leucine are the most degenerate in the genetic code. Of the six codons which specify arginine four are of the CGX series and the other two are AGA and AGG. It is of interest to know which, if any, of these codons are used predominantly in the protamine genes. The
information contained in this report indicates that neither the CGX series nor the AG ε codons predominate and that both AGA and AGG codons are used.

The oligopyrimidine tracts released from protamine cDNA after depurination are complementary to the purine tracts in the protamine mRNA. A sequence of AGA and AGG codons would appear as an unbroken oligopyrimidine tract in the cDNA. Exclusive use of either AGA or AGG codons, or both, for arginine would give rise to numerous, long oligopyrimidine tracts in protamine cDNA up to at least 18 nucleotides in length, corresponding to the arginine tracts in the protamines. This is not the case since protamine cDNA contains only one oligopyrimidine longer than a hexanucleotide. The sequences determined for the oligopyrimidine tracts (Table III) show no more than two AGA or AGG codons are used in series. This is the case in the sequences AGAGAGA, AGGAGG, and AGGAGA. The latter sequence is a portion of the undecanucleotide coding for Arg-Arg-Gly-Gly. Conversely, the CGX codons are clearly not used exclusively for arginine since some AGG and AGA codons occur, as in the undecanucleotide.

It is difficult to estimate from the sequence data alone the proportions of the CGX codons and AG ε codons used for arginine. Many of the mRNA sequences deduced (Table III) could represent AG ε codons for arginine but it is not possible to establish the reading frame with certainty nor can these sequences be unequivocally assigned to the coding region. However, the upper limit to the frequency of the CGX series can be obtained from nearest neighbor frequency analysis of protamine cDNA. Every CGX codon for arginine contains one CpG doublet. Given a length of 270 residues for the nonadenylated portion of protamine mRNA and the information in Table IV, there must be 12 CpG doublets. Hence, a maximum of 12 out of the 21 to 22 arginine codons in protamine mRNA can be of the CGX series. This value of 12 CGX codons is based on the unlikely premises that all the CpG doublets occur in the coding region, although the noncoding region of protamine mRNA is more than one and one-half times as long (160 to 170 residues) as the coding region (100 residues), and that the reading frame of CGX comprises the first and second bases in a codon rather than second and third bases, or third base in one codon and first base in the next. For these reasons the number of CGX codons for arginine is probably somewhat less than 12. A lower limit of seven codons of the CGX series can be deduced from a knowledge of the lengths of the arginine tracts in protamine (10) given that no more than two AG ε codons can occur in series based on the recovery of oligopyrimidine tracts in the cDNA (Table III).

It follows from these figures that the minimum number of AG ε codons in protamine mRNA is 9 to 10. This value is easily exceeded by the number of potential AG ε codons contained in the RNA sequences in Table III, and by the frequency of the doublet ApG (=24). However, sequences complementary to AG, AGA, and AGG have all been seen in endonuclease IV fragments which have originated from the noncoding region of protamine cDNA, whereas CG has not been seen in any of these fragments.5

Given this balance between the codons of the CGX and
Sequence Analysis of Protamine mRNA from Rainbow Trout

FIG. 5. Determination of the purine neighbor at the 3' end of the oligopyrimidine tracts in protamine cDNA. Protamine cDNAs labeled with \([\alpha-32P]dATP(A-cDNA)\) or with \([\alpha-32P]dGTP(G-cDNA)\) comprising 10^6 Cerenkov cpm were each depurinated and separated as described under "Methods."

TABLE IV

| Nearest neighbor frequencies of protamine mRNA |  |
|---------------------------------------------|---|
| Nearest neighbor frequency analysis was performed as described under "Methods." Doublet frequencies are expressed as parts per thousand for protamine mRNA corrected for the poly(A) tract. |   |
| ApA | UpU | 56 | 55 |
| CpA | UpG | 66 | 44 |
| GpA | UpC | 73 | 45 |
| CpU | ApG | 79 | 88 |
| GpU | ApC | 73 | 58 |
| GpG | CpG | 71 | 88 |
| UpA | 53 |
| ApU | 43 |
| CpG | 46 |
| GpC | 62 |

AG\(_2\) series it appears also that the two types are distributed heterogeneously within the sequences coding for arginine blocks since no more than two AGA or AGG codons are used in series. The sequence of the undecanucleotide, C\(_1\)T\(_n\), extended by nearest neighbor considerations, reveals three different arginine codons, CGG, AGG, and AGA used in succession.

There are numerous reports that the doublet CpG is remarkably rare in the DNA of vertebrates (24, 25). This scarcity carries over into RNA transcripts, both heterologous nuclear RNA and mRNA (26). In view of the unusual amino acid composition of the protamines, in which 2 out of every 3 amino acid residues are arginine, it might be expected that the mRNAs for these polypeptides would have a high CpG content since the sequence CpG occurs as first and second bases in four out of the six arginine codons. Moreover CpG occurs as second and third bases in the codons UCG, CCG, GCG, and ACG which code for the amino acids serine, proline, alanine, and threonine, respectively. Serine and proline are, after arginine, the most plentiful amino acids in protamine, and alanine occurs once in one of the three major components. When CpG occurs as third base in one codon and first base in the next codon there are 256 possible two-codon combinations (27), some of which would occur in protamine mRNA.

Despite the opportunity for the CpG doublet to appear in protamine mRNA it is instead quite rare, comprising 46 parts/thousand. This is despite the fact that the base composition for the nonadenylated protamine mRNA is 52% G + C (28). The selective pressures which have kept down the level of the sequence CpG in the DNAs of vertebrates have apparently been acting on the protamine genes also.

The greater abundance of the doublet ApG compared to CpG in protamine mRNA lends further support to the proposal that the archtypal protamine gene may have been a repetitive DNA of the type d(AGA)n\_1d(TCT)n (29).

REFERENCES
1. Marotta, C. A., Forget, B. G., Weissman, S. M., Verma, I. M., McCaffrey, R. P., and Baltimore, D. (1974) Proc. Natl. Acad. Sci. U. S. A. 71, 2300-2304
2. Forget, B. G., Marutia, C. A., Weissman, S. M., and Cohen-Solal, M. (1975) *Proc. Natl. Acad. Sci. U. S. A.* 72, 3614-3618
3. Proudfoot, N. J., and Brownlee, G. G. (1976) *Nature*, 263, 211-214
4. Proudfoot, N. J., and Brownlee, G. G. (1974) *Nature* 252, 359-362
5. Poon, R., Paddock, G. V., Heindell, H., Whitcombe, F., Salser, W., Kacian, D., Bank, A., Gambino, R., and Ramires, F. (1974) *Proc. Natl. Acad. Sci. U. S. A.* 71, 3502-3506
6. Salser, W., Brown, S., Browne, D., El Adli, F., Federoff, N., Fry, K., Heindell, H., Paddock, G. V., Poon, R., Wallace, B., and Whitcombe, R. (1975) *Fed. Proc.* 35, 23-35
7. Milstein, C., Brownlee, G. G., Cartwright, E. M., Jarvis, J. M., and Proudfoot, N. J. (1974) *Nature* 252, 354-359
8. Cowan, N. J., Secher, D. S., and Milstein, C. (1976) *Eur. J. Biochem.* 61, 355-368
9. Gedamu, L., and Dixon, G. H. (1976) *J. Biol. Chem.* 251, 1455-1463
10. Ando, T., Yamasaki, M., and Suzuki, K. (1973) in Proteins—Isolation, Characterization, Structure and Function, Springer-Verlag, New York
11. Lahee, F. (1969) *Nature* 231, 1217-1222
12. Temin, H. M., and Baltimore, D. (1972) *Adv. Virus Res.* 17, 129-186
13. Wu, R., Bambara, R., and Jay, E. (1974) *CRC Crit. Rev. Biochem.* 2, 455-512
14. Krystosek, A., Cawthon, M. L., and Kabat, D. (1975) *J. Biol. Chem.* 250, 6077-6084
15. Sanger, F., and Coulson, A. R. (1975) *J. Mol. Biol.* 94, 441-448
16. Burton, K. (1967) *Methods Enzymol.* 12A, 222-224
17. Brownlee, G. G., and Sanger, F. (1969) *Eur. J. Biochem.* 11, 395-398
18. Barrell, B. G. (1971) in *Procedures in Nucleic Acid Research* (Cantoni, G. L., and Davies, D. R., eds) Vol. 2, Harper and Row, New York
19. Ling, V. (1972) *J. Mol. Biol.* 64, 97-102
20. Kleppe, K., Ohtsuka, E., Kleppe, R., Molineux, I., and Khorana, H. G. (1971) *J. Mol. Biol.* 56, 341-361
21. Dahlberg, A. E., Dingman, C. W., and Peacock, A. C. (1969) *J. Mol. Biol.* 41, 139-147
22. Dixon, G. H., Davies, P. L., Ferrier, L. N., Gedamu, L., and Iatrou, K. (1976) in *The Proceedings of the Symposium on Molecular Biology of the Mammalian Genetic Apparatus* (Teo, P. O. P., ed) Vol. A, pp. 355-379, Associated Scientific Publishers, Elsevier-Excerpta Medica-North Holland
23. Proudfoot, N. J., and Brownlee, G. G. (1976) *Brit. Med. Bull.*, in press
24. Swartz, M. N., Trautner, T. A., and Kornberg, A. (1962) *J. Biol. Chem.* 237, 1961-1967
25. McCarthy, B. J. (1969) in *The Handbook of Molecular Cytology* (Lima-de-Faria, A., ed) North Holland Publishing Co., Amsterdam
26. Fraser, N. W., Burdon, R. H., and Elton, R. A. (1975) *Nucleic Acids Res.* 2, 2131-2146
27. Subak-Sharpe, J. H. (1969) in *The Handbook of Molecular Cytology* (Lima-de-Faria, A., ed) North Holland Publishing Co., Amsterdam
28. Ferrier, L. N. (1976) Masters thesis, University of Calgary, Canada
29. Davies, P. L., Dixon, G. H., Ferrier, L. N., Gedamu, L., and Iatrou, K. (1976) *Progr. Nucleic Acid Res. Mol. Biol.* 19, 135-155
