PROTAMINE GENES OF RAINBOW TROUT.

D.S.ANSON

Thesis submitted for degree of Ph.D., 1983

I declare that this work is solely my own.

D.S.Anson

Acknowledgements.

I would like to thank all those who have helped me in any way. Special thanks go to my supervisor,John Bishop,without whose help this thesis would not exist.

D.S.Anson

Index

List of figures and tables

Abstract

The protamines of rainbow trout are a family of small (30-33 amino acid residues) and extremely basic proteins that displace histones during spermatogenesis.

In this study two protamine gene sequences were cloned via a genomic DNA library constructed in Charon 4A. The gene sequences cloned correspond to only one of the six *Eco*RI fragments hybridizing to protamine cDNA in genomic Southern transfers.

The cloned sequences were analysed by restriction enzyme mapping, nuclease S1 and enoxuclease VII mapping and by nucleotide sequencing. These analysis show that the two protamine genes cloned do not contain introns. The clones were shown to be almost identical. One gene encodes a protamine sequence previously described both as an amino acid sequence and as a cDNA nucleotide sequence. The second gene encodes a novel protamine sequence.

Introduction

Sequence organization of the eukaryotic genome

DNA reassociation analysis of the eukaryotic genome has demonstrated that it contains three sequence classes of DNA (Britten and Kohne, 1968). These are defined by the degree of sequence repetition per haploid genome. The three classes are unique (sequences present once or a few times), moderately repetitive (sequences present up to 10^4 times) and highly repetitive (sequences present more than 10^4 times).

Most unique sequences are interspersed in the genome with moderately repetitive (Davidson and Britten, 1973; Davidson et al., 1973; Graham et al., 1974; Goldenberg et al., 1975; Meunier-Rotival et al., 1982) and some highly repetitive (Jelinek et al., 1980; Singer, 1982) sequences. In the sea urchin and many other eukaryotes the average sizes of the interspersed repetitive and unique sequences are 300 and 100 bp (base pairs) respectively (Graham et al., 1974). However, long interspersed repeats (approximately 5000 bp or longer) have also been found (Meuncier-Rotival et al., 1982; Singer, 1982). Both short and long repeated sequences have also been found interspersed within satellite DNA (Singer, 1982).

A large proportion of highly repetitive sequences are found in tande repeats as satellite DNA (Walker, 1971a, b; Corneo et al., 1971). Satellite DNA is located in the heterochromatic regions of centromeres and telomeres (Corneo et al., 1971; Jones, 1970; Rae, 1972).

The use of specific sequence probes in DNA reassociation analysis and the use of Rot analysis (RNA excess hybridization) has shown that most messenger RNA coding sequences are found in the unique sequence class

(Bishop et al., 1972; Harrison et al., 1972; Suzuki et al., 1972; Goldberg et al., 1973; Galau, 1974). However some coding sequences have been shown to be moderately repeated. The histone genes (Kedes and Birnsteil, 1971) and ribosomal RNA genes (Brown et al., 1972) are both examples of moderately repeated genes.

The interspersed unique sequences have been shown to include messenger RNA coding sequences (Davidson et al., 1975). While messenger RNA coding sequences are (usually) unique, transcription of repeated sequences does occur as they are found represented in heterogeneous nuclear RNA (Spradling et al., 1974; Lewin, 1975).

Cloning of eukaryotic genes

The use of DNA cloning techniques has made possible the analysis of individual genes at the nucleotide sequence level.

The first cloning systems, using plasmids of *E.coli* as vectors were relatively inefficient. The methods used for screening populations of clones for specific sequences were laborious, especially if large numbers of clones are involved. However, for genes such as the sea urchin histone genes and *Xenopus* ribosomal RNA genes that can be purified prior to cloning the limitations on cloning efficiency and screening were acceptable. These genes were therefore among the first to be cloned (Morrow et al., 1974; Kedes et al., 1975). However, the limitations of plasmids make them unsuitable for the construction of gene libraries (Clarke and Carbon, 1976) and they have been superceded by phage lambda and cosmids as primary cloning vectors (Blattner et al., 1977; Leder et al., 1977; Maniatis et al., 1978; Collins and Hohn, 1978). These vectors are used in conjunctio with highly efficient *in vitro* systems for the packaging and infection of lambda DNA (Hohn and Murray, 1977; Sternberg et al., 1977) and rapid,

high density screening methods (Benton and Davis, 1977; Hanahan and Meselson, 1980). These techniques make the isolation of any gene for which a specific sequence probe is available relatively easy.

A number of methods for making and identifying sequence probes have been developed. One of the most widely used of these is the construction of cDNA clones from messenger RNA (Maniatis et al., 1976). cDNA clones can be identified by methods such as hybrid arrest translation (Paterson et al., 1977) and hybrid selection directed translation (Stark and Williams, 1979). Synthetic oligonucleotide sequence probes have also been used in a number of ways (Comb et al., 1982; Suggs et al., 1981). The sequence to be synthesized is derived from amino acid sequence data and this approach makes possible the cloning of the gene for any protein for which even a very limited amount of amino acid sequence is known.

A variety of other methods have also been used including immunoadsorption of polyribosomes (Kraus and Rosenberg, 1982), complementation of yeast mutants with homologous (yeast) DNA (Petes, 1980) and non-homolegous (Drosophila) DNA (Henikoff et al., 1981) and isolation of DNA that will transform mutant cell lines (Lowy et al., 1980). Although the number of eukaryotic genes that have been cloned is quite large, and growing rapidly, it is extremely small when compared with the number of genes in the average eukaryotic genome. Man, for example, is thought to have of the order of 50,000 genes and only tens of these have been cloned either as genomic or cDNA sequences.

The structure of eukaryotic genes

The structural analysis of eukaryotic genes has revealed that most genes have non-contiguous messenger RNA coding sequences. Blocks

of messenger RNA coding sequence, termed exons, are separated by blocks of non-coding sequence termed introns. Introns are mainly found in the amino acid coding region of genes and have also been found in the 5' non-coding regions of genes. No gene has been shown to have an intron in the 3' non-coding region. Not all genes have introns. The known exceptions are the histone genes (Hentschel and Birnsteil, 1981) and interferon genes (Nagata et al., 1980). The ability of several yeast genes to complement *E.coli* mutants (Petes, 1980) suggests that these genes do not contain introns in their amino acid coding sequences. Sequence and nuclease S1 analysis of one such gene, the yeast iso-1-cytochrome C gene, confirms this view (Smith et al., 1979; Faye et al., 1981). However, it is possible that other such genes contain introns in their 5' or 3' non-coding regions.

The number of introns in a gene can be very large. The amount of gene sequence in introns may greatly exceed the amount in exons. The *Xenopus* vitellogenin genes contain at least 33 introns and the messenger RNA sequence of 6kb (kilobasepairs) is spread over 21kb of genomic DNA (Wahli et al., 1980). Another extreme example of intron/exon gene structure is the chicken pro α2 collagen gene. This gene has at least 50 introns and the mRNA sequence of 5kb is spread over about 40kb of genomic DNA (Wozney et al., 1981).

Evolution of eukaryotic genes

The discovery of the intron/exon structure of many eukaryotic genes has led to the proposal of theories about their origin and function (Gilbert, 1978; Blake, 1979; Reanney, 1979). The splitting of genes may allow more rapid evolution of complex and large proteins to occur by "exon shuffling". Such a process would allow new genes to be provided by the

combination of previously unjoined exons into a single transcription unit. Evolution of genes could also occur by the internal duplication of one exon of a gene. Support for such theories is given by the discovery that in some genes exons appear to encode separate functional domains of the protein chain. An example of this is the central domain of the globin genes which codes for the haem binding domain of the protein (Craik et al., 1980). Other examples include the immunoglobulin genes (Sakano et al., 1979; Brack et al., 1978) and the ovomucoid gene (Stein et al., 1980). To avoid loss of genes during exon shuffling duplication of a sequence prior to shuffling would have to occur. Evidence for evolution of a gene by internal duplication of a DNA sequence is the structure of the collagen gene. This suggests that it has evolved by repeated duplication of a sequence containing an exon of 54 base pairs (Yamada et al., 1980).

The analysis of gene families, such as the globin genes of man (Maniatis et al., 1980; Proudfoot et al., 1980) and mouse (Heder et al., 1980) and the histone genes of several organisms (Hentschel and Birnsteil, 1981) has shown the importance of gene duplication in evolution. The discovery of pseudogenes suggests that gene duplication may occur with a higher frequency than is suggested by the number of active genes (Leder et al., 1981; Hollis et al., 1982; Wilde et al., 1982; Steinmetz et al. 1981). However, several of the pseudogenes so far discovered lack introns (Nishioka et al., 1980; Hollis et al., 1982; Wilde et al., 1982), and may represent transposition of messenger RNA sequences by retroviruses rather than normal gene duplication (Newmark,1982).

The primary transcript, RNA processing

It is now widely accepted that interrupted structural genes are transcribed colinearly to produce a precursor RNA molecule containing intron sequences. The intron sequences are then presumably removed by

a series of splicing reactions. If a gene contains more than one intron this will involve a number of RNA molecules intermediate in size between the primary transcript and the mature mRNA. The existence of putative primary transcripts and processing intermediates has been demonstrated for globin genes (Leder et al., 1980; Proudfoot et al., 1980; Maniatis et al., 1980) and the chicken ovalbumin gene (Roop et al., 1978). Direct evidence for a precursor-product relationship between such molecules and the mature messenger RNA has been produced by pulse-chase experiments with the ovalbumin system (Tsai et al., 1980) and nuclease SI mapping of rabbit β-globin precursors (Grosveld et al., 1981). The latter data suggests that both the small and large intron sequences are removed by two splicing reactions each and that excision occurs in an ordered manner. These, and similar experiments with Ad2 suggest that the 5' end of the primary transcript is the same as that of the mature messenger (Ziff and Evans, 1978; Nevins and Darnell, 1978). More recent experiments using nucleotide triphosphate analogues lacking a hydrolyzable β- γ bond show that G and U uncapped transcripts from the Ad2 EIV and protein IX gene initiate at the cAP site in vitro (Bunick et al., 1982). These experiments also demonstrate that initiation shows an absolute requirement for a hydrolyzable β- γ bond in ATP whatever the initiating base. Therefore no uncapped A initiated transcripts could be demonstrated as transcription initiation is halted when the ATP analogue is used in the reaction.

The position of the 3' end of the primary transcript is less clear. It is clear that transcription can go through polyadenylylation sites, most notably in the transcription of the μ and δ, heavy chain immunoglobulin genes. In these genes transcription through polyadenylylation sites and differential splicing of transcripts occurs. This allows the simultaneous expression of IgM and IgD with the same antigenic specificity and also the simultaneous synthesis of the membrane bound and secreted forms of these molecules (Maki et al., 1981; Earlt et al., 1980; Chang

et al., 1982; Alt et al., 1982). The existance of four dihydrofolate reductase mRNAs in mouse cells which differ only in the length of their 3' untranslated regions also suggests that readthrough of polyadenylylation sites and utilisation of different polyadenylylation sites in the same gene can occur (Setzer et al., 1980). Studies of the transcripts of the mouse β major globin gene (Hofer and Darnell, 1981) and the two chicken α · globin genes (Weintraub et al., 1981) suggest that polyadenylylation occurs after endonucleolytic cleavage of the primary transcript. Analysis of nascent nuclear transcripts of the mouse β major globin gene suggest that 95% of the gene transcripts terminate about 1.4kb from the polyadenylylation site (Hofer et al., 1982). The rapidity of polyadenylylatic makes the analysis of *in vivo* termination sites extremely difficult and there are at present no *in vitro* systems that have been shown to support specific termination or polyadenylylation.

The mechanism of the RNA splicing reactions is unknown. However comparative sequence studies have led to the formulation of consensus sequences for intron/exon boundaries. The sequence at the 5' boundary of the intron is termed the donor site, the 3' boundary the acceptor site. The donor consensus sequence is 5'$^{A}_{C}$AGGT$^{A}_{G}$AGT3'. The corresponding acceptor consensus is 5'PyPyPyPyPyPyNCAGG$^{G}_{T}$3' (Breathnach and Chambon, 1981). These consensus sequences are derived from analysis of 90 donor and 85 acceptor sites. At most splice sites the sequences are redundant to a certain degree so that the exact site of the splice cannot be determined. However, in the ovomucoid gene several sites are exactly defined (Stein et al., 1980). In these cases the intron starts with a 5' GT and ends with a 3' AG. All the splice junction sequences so far elucidated show these two dinucleotides in positions at which they could define the ends of the intron sequence.

The importance of the consensus sequence in the splicing reactions has been demonstrated in a number of ways. Firstly the stepwise

removal of the small and large intron sequences from the rabbit β globin gene transcript (Grosveld et al., 1981) can be explained by the presence of acceptor and donor sites in the intron sequence. These can recombine with the end sites to reform the appropriate site to allow removal of the remainder of the intron sequence. A more direct demonstration of the importance of the splice point consensus sequences has come from analysis of the genotype of β^0 thalassemias. Sequence analysis has shown that several are caused by single base mutations in splice junction consensus sequences (Busslinger et al., 1981; Baird et al., 1981; Treisman et al., 1982) that prevent complete processing. Transcriptional analysis of one of these genes in a SV40 vector suggest that the removal of the two introns may be coupled (Treisman et al., 1982).

Initiation of transcription by RNA polymerase II

Comparison of the sequences around and 5' to the cAP (transcription initiation) site of eukaryotic structural genes has revealed several regions of homology. Consensus sequences for these regions have been formulated by sequence comparisons. The first of these consensus sequences is for the cAP site itself. This consensus is 5' PyAPyPyPyPy 3'. (Breathnach and Chambon, 1981). This consensus is derived from the comparison of 22 genes for which the start site is known unambiguously. The A residue represents the first transcribed nucleotide. A second consensus sequence is found approximately 30bp 5' to the cAP site. This sequence has been designated the Hogness or TATA box. The consensus sequence is 5' TATA$^{A}_{T}A^{A}_{T}$ 3' (Breathnach and Chambon, 1981). The consensus was derived from comparison of 60 gene sequences. The first T is found between nucleotides -34 to -26 (relative to the cAP site). In 85% of genes studied it occurs within

2bp of nucleotide -31. The TATA box and cAP site sequences appear to be ubiquitous in genes transcribed by RNA polymerase II. A third consensus sequence, 5'GG$^{C}_{T}$CAATCT 3', the CAAT box, is found in some genes (Efstratiadis et al., 1980).

A number of different transcription systems have now been developed that give accurate RNA polymerase II initiation and transcription on cloned gene sequences. These include two based on HeLa cell free extracts, one requiring the addition of exogenous RNA polymerase II (Weil et al., 1979), the other utilizing the endogenous enzyme (Manley et al., 1980). Several SV40 vector systems have also been developed to allow the introduction of cloned genes into mammalian cells for transcriptional analysis (Hamer et al., 1980; Mellon et al., 1981). Injection of cloned DNA into *Xenopus* oocytes has also been used (McKnight et al., 1981). The use of viral vectors such as the SV40 derived ones, has the additional advantage that these systems give proper processing and polyadenylylationof transcripts. The use of these transcription systems has allowed the importance of different sequences in the transcription process to be evaluated by transcriptional analysis of DNA that has had its structure altered *in vitro*.

Deletion mutants of the 5' regions of a number of genes have been constructed and assayed in this way. With the chicken ovalbumin (Tsai et al., 1980) and conalbumin genes (Corden et al., 1980), the rabbit β-globin (Grosveld et al., 1981) and the *Bombyx mori* silk fibroin genes (Tsujimoto et al., 1981) this approach has shown that the TATA box region is essential for efficient initiation. Deletions approaching this region from the 3' side in the rabbit β-globin and the chicken conalbumin suggest that the TATA box sequence may be sufficient to direct

initiation. Initiation appears always to occur approximately 30bp downstream from the TATA box, the exact site depending on the sequence around the initiation point. However the existence of "false" TATA box consensus sequences in the ovalbumin clone used by Tsai et al., (1980), including a TATATAT sequence identical to the genuine ovalbumin TATA box, that do not support transcription suggest that the consensus alone is not sufficient to cause initiation. In the conalbumin gene a T to G transversion in the TATA homology (at the second T) causes a 10 to 20 fold decrease in *in vitro* transcription (Wasylyk et al., 1980).

Similar deletion studies of the sea urchin H2A histone gene (Grosschedl and Birnsteil 1980), and the HSVI *tk* (thymidine kinase) gene (McKnight et al., 1981) show that the TATA box sequence in these genes can be deleted without abolishing transcription. In the case of the H2A gene deletion of the TATA consensus generates a number of new initiation sites of lower (than wild type) efficiency which map into the 5' non coding region of the (normal) mRNA sequence. Partial deletion (from the 3' side) of the TATA consensus in the *tk* gene produced heterogeneous initiation. However, total deletion appeared to restore accurate initiation. The accurate initiation of the *tk* gene was shown to be dependent on sequences between nucleotides -100 and -40 (from the cAP site).

The differences between these results may be due to the different transcription assays used. The sea urchin H2A gene and the HSVI *tk* gene were analysed by injection into *Xenopus* oocytes. The chicken ovalbumin and conalbumin genes, the rabbit β-globin and the *Bombyx mori* silk fibroin gene were all analysed in Manley HeLa systems. The effects of the TATA homology would therefore appear to be greater in the HeLa cell extracts than in *Xenopus* oocytes. However a more subtle

analysis of the HSVI *tk* gene in *Xenopus* oocytes would seem to refute this generalisation (McKnight and Kingsbury, 1982). In this study small regions were altered by recombining 5' and 3' *in vitro* deletions via *Bam*HI linkers. The only sequence changed is the sequence replaced by the *Bam*HI linker (10bp). All the other sequences are unchanged and in the normal spatial relationship with each other. Two mutants generated in this manner altered the TATA box sequence. The first was an A to C transversion at the last base of the wild type consensus (5' TATTAA 3') This produced a fifteen fold reduction in transcription. The second mutant introduced multiple base changes into the consensus and produced no detectable authentic *tk* mRNA. This analysis therefore suggests that the TATA box sequence is of importance in the *Xenopus* oocyte system when in the normal spatial relationship to other 5' sequences. The reported transcription of the *tk* gene with the TATA box deleted suggests that such spatial relationships may be important. In the deletion experiment the deleted sequences are replaced with pBR322 sequences. These may also affect transcription. Results from transcription experiments using deletion mutants should therefore always be interpreted with such factors in mind.

These deletion and mutation studies have also allowed the analysis of the role of the cAP site consensus. Most experiments suggest it defines the exact site of initiation but that it is not essential for initiation. In its absence initiation occurs at one or more new sites around 30bp from the TATA box. When it is not present initiation still appears to occur preferentially at A residues (Corden et al., 1980). Deletion of the cAP site has also been shown to reduce the efficiency of transcription of the histone H2A and conalbumin genes (Grosschedl and Birnsteil, 1980; Corden et al., 1980).

As well as the TATA box and cAP site sequences other, more upstream, 5' sequences have been shown to greatly influence transcription in a number of systems. The *Bam*HI linker mutants of the HSVI *tk* gene (McKnight and Kingsbury, 1982) described above reveal two further 5' sequence elements required for transcription. The first of these occurs between nucleotides -61 and -47. Mutations at this site predominantly affect a G rich sequence and reduce transcriptional efficiency ten fold. The second sequence is found between nucleotides -105 and -80. Quantitatively two types of mutants can be distinguished in this sequence. The first affect a C rich sequence (-105 to -97) and produce a twenty-fold decrease in transcription efficiency. The second produce multiple base changes in the sequence between nucleotides -97 and -80 and reduce transcriptional efficiency ten fold. Both the G rich sequence (-61 to -47) and the C rich sequence (-105 to 97) contain the same six bp inverted repeat. The authors suggest that this could function to form an intrastrand interaction producing a 42bp loop from both strands between the two elements.

A far upstream sequence that is important for efficient *in vitro* transcription of the histone H2A gene has also been found (Grosschedl and Burnsteil, 1981). This sequence occurs between nucleotides -139 and -111 and its removal reduces transcription five-fold. The effect of this sequence can be mimicked by free DNA ends or by specific pBR322 sequences. A similar sequence has also been found in the silk fibroin gene (Tsuda and Suzuki, 1981). This sequence maps upstream from nucleotide -74. Interestingly the enhancing effect of this sequence is detected only in homologous cell free transcription systems. The same sequence, again in homologous transcription systems, causes preferential transcription of the silk fibroin gene when alone or in competition with a mouse β-globin gene or the Ad 2 late promotor. This

effect is not seen in heterologous systems.

Another transcription enhancing element is the 72bp repeat sequence found 5' to the SV40 early transcription unit. This sequence has been shown to enhance transcription of the rabbit β-globin gene when in a cis position. When the rabbit β-globin gene is cloned into a vector lacking the 72bp sequence and introduced into cells its transcription is reduced 200 fold (Banerji et al., 1981). The position of the 72bp element relative to the initiation site of the β-globin gene appears to be unimportant. The enhancing effect occurs equally with the element 1400bp upstream or 3300bp downstream from the β-globin cAP site. However the human α1-globin gene does not appear to require the 72bp element as it is efficiently expressed in an SV40 vector lacking this sequence (Mellon et al., 1981).

Termination and polyadenylylation of RNA polymerase II transcripts

Polyadenylylation of RNA polymerase II transcripts is thought to be related to the consensus sequence AAUAAA. This sequence is found approximately 20 nucleotides from the 3' end (excluding poly A tail) of most eukaryotic mRNAs (Proudfoot and Brownlee, 1976). This sequence or a close derivative of it, has also been found in all genomic clones corresponding to polyadenylylated mRNAs, again approximately 20bp from the presumed or mapped polyadenylylation site. The polyadenylylation site usually corresponds to a number of A residues so the exact site of polyadenylylation cannot be defined (Benoist et al., 1980). The exception to this is yeast genes which do not contain the AAUAAA consensus.

Deletion of the AAUAAA consensus in SV40 has shown that it is required for polyadenylylation (Fitzgerald and Shenk, 1981). In

addition it was shown that polyadenylylation occurred at a fixed distance downstream from this sequence. The role of the consensus in directing polyadenylylation is supported by its absence in non yeast histone mRNAs which are not polyadenylylated and do not contain the consensus (Hentschel and Birnsteil, 1981). The existence of multiple transcripts from the DHFR gene also provides evidence for the role of the consensus sequence as the shortest mRNA has only a weak consensus sequence, AAUA, while the longest mRNA has a full consensus, AAUAAA (Setzer et al., 1980). Yeast genes lack the consensus. However all yeast RNA polymerase II transcripts are polyadenylylated (Hereford and Rosbash, 1977) including histone mRNAs (Fahrner et al., 1980) suggesting a non specific polyadenylylation mechanism.

The relationship between polyadenylylation and transcription termination, if any, is unknown. Polyedanylylation is extremely rapid and may occur before termination has occurred. The existence of stable non-polyadenylylated mRNAs, which do not appear to be the result of degradation or deadenylylation of polyadenylylated mRNAs (Milcarek et al., 1974; Nemer et al., 1974) suggests that termination can occur independantly of polyadenylylation.

Little is known about the sequence requirements for RNA polymerase II termination. It has not even been demonstrated that termination occurs specifically in most genes. Histone gene transcripts however do appear to terminate specifically as there is no evidence for precursor mRNAs or processing (Mauron et al., 1981; Seiler-Twyns and Birnsteil, 1981). A 23bp sequence has been shown to be necessary for termination of the sea urchin H2A histone gene in Xenopus oocytes (Birchmeier et al., 1982). This sequence is found at the 3' end of the mRNA and is conserved between histone genes. The sequence

includes a 16bp hyphenated inverted repeat and deletion of only 12bp out of the middle of the sequence is enough to abolish termination. However the 23bp sequence is in itself not sufficient to direct termination, sequences 3' to the termination site are also required.

Evidence has also been produced for a specific termination site in the mouse β-globin gene (Hofer et al., 1982). Analysis of labelled nascent RNA by hybridisation to DNA sequences immobilized on nitrocellulose show that 95% of transcripts appear to terminate approximately 1400bp downstream from the polyadenylylation site. These appear to be genuine globin transcripts as they are coordinately stimulated by stimulation of globin transcription.

An *in vivo* deletion mutant of the yeast CYC1 (iso-1-cytochrome C) gene that causes readthrough of the normal transcriptional termination site has also been described (Zaret and Skerman, 1982). This deletion is 38 base pairs long and has occurred between two 7 base pair direct repeats, leaving only one of the repeats. The deletion ends approximately 15 base pairs before the end of the wild type transcribed sequence. The deletion has two effects. It reduces the steady state level of CYC1 messenger RNA to between 5 and 10% of the wild type level and causes most CYC1 transcripts to be extended at their 3' ends by up to 1000 base pairs. The wild type and mutant transcripts are all polyadenylylated. However the CYC1 gene, in common with 11 out of 15 yeast gene sequences compared in this paper, does not contain the AAUAAA consensus sequence in its 3' non coding region. This sequence is therefore not required for either polyadenylylation or termination in yeast. Unlike higher eukaryotes yeast appears to polyadenylylate all RNA polymerase II transcripts (Hereford and Rosbash, 1977), including histone messenger RNA (Fahrner et al., 1980). Yeast may simply couple polyadenylylation to termination of transcription by RNA polymerase II.

Although the CYC1 deletion does not contain any obvious sequence structure a comparison of the deleted sequence with the 3' non coding regions of 14 other genes allows a consensus sequence to be formulated. This consensus sequence was not present in all the sequences looked at and its importance has not been demonstrated directly.

Steroid receptors and other induction sequences in DNA

Some genes are known to be induced by steroid hormones. These include the ecdysone induced genes of Drosophila, chicken egg oviduct genes and various liver specific genes.

The induction of gene activity by steroid hormones is thought to be mediated by the binding of hormone-receptor complexes to specific DNA sequences (Yamamoto and Alberts, 1976). The cloning of several steroid responsive genes has made it possible to search for hormone-receptor complex binding sites in these sequences. Sequences 5' to the chicken egg oviduct genes ovalbumin, conalbumin, ovomucoid, X and Y have a high affinity for purified chicken oviduct progesterone-receptor complex in a competition assay (Mulvihill et al., 1982). *In vitro* deletion mapping has localised the DNA sequence showing this high affinity to between 250 and 300bp upstream from the transcription initiation site. Comparison of this sequence with the same region in the conalbumin, ovumucoid, X and Y genes has revealed sequence homology. A 19bp consensus sequence has been formulated from these homologies. The importance of this sequence *in vivo* has not been tested.

The sequences required for glucocorticoid induction of the human growth hormone gene have been shown to be with 500bp of the cAP site by transformation of cells that retain glucocorticoid receptors with cloned DNA. The rat $\alpha_{2\mu}$ globulin gene and endogenous mouse mammary

tumour virus both retain hormone responsiveness when cloned and reintroduced into cells by transformation (Kurtz, 1981; Hynes et al., 1981; Buetti and Diggelman, 1981).

A fusion gene consisting of the Harvey MSV p21 transforming gene and the long terminal repeat of MMTV that is glucocorticoid induced (Huang et al., 1981) suggests that the LTR is responsible for the hormone responsiveness of MMTV. However some other MMTV sequences were present so this conclusion is not definitive. An *in vitro* binding assay of purified glucocorticoid receptors to MMTV DNA suggests that there is at least one receptor binding site outside the LTR (Payvar et al., 1981). This site is several kb downstream from the normal start site for MMTV transcription. However, no *in vitro* manipulations have yet been done to try and identify directly the sequences involved in hormone responsiveness.

The sequences required for cadmium regulation of the mouse metalliothionein I gene have been shown to be within 90bp of the cAP site by analysis of 5' deletion mutants of a metalliothionein/HSV I *tk* fusion plasmid. The plasmid was microinjected into mouse eggs and the effect of cadmium on the level of *tk* activity analysed (Brinster et al., 1982).

The *in vivo* structure of active genes.

In an attempt to understand the mechanisms of gene control the *in vivo* structure of genes has been analysed in a number of ways. Two features have been found that appear to be associated with gene activity. Firstly the DNA of the gene and immediately surrounding sequences tends to be undermethylated in the active gene when compared with the same sequences in the inactive gene. This can be assayed by digestion with

restriction enzymes sensitive to CpG methylation (Shen and Maniatis, 1980; Ploeg and Flavell, 1980; Mandel and Chambon, 1979). The second feature is the increased sensitivity of expressed genes to *in vivo* digestion with DNAase I or micrococcal nuclease (Weintraub and Groudine 1976; Wu et al., 1979a, b). The use of restriction enzymes and the Southern blotting technique have allowed the identification and mapping of DNAase hypersensitive sites around genes. This was first done in two of the heat shock genes of *Drosophila melanogaster* (Wu, 1980). This analysis showed that DNAase I hypersensitive sites were present at the 5' ends of the genes studied both in embryos and a cell line. The heat shock genes are readily inducible in both the embryos and the cell line. Upon heat shock the whole gene appears to become very sensitive to digestion (Wu et al., 1979b). DNAase I hypersensitive sites have now been mapped in a number of *Drosophila* genes. The *Drosophila* histone genes each have a single 5' DNAase I hypersensitive site (Samal et al., 1981) and the four small heat shock genes (hsp 22, 23, 26 and 28) each have two 5' hypersensitive sites (Keene et al., 1981). The *Drosophila sgs* 4 glue protein gene normally has five 5' DNAase I hypersensitive sites in tissue where it is being expressed. All these sites are absent in tissue where the gene is inactive (Sherman and Beckendorf, 1982). A number of mutants that do not synthesise the *sgs* 4 gene product, or synthesise it at a very low level, show deletions that affect the DNAase hypersensitive sites (Shermoen and Beckendorf, 1982; Muskavitch and Hogness, 1982). One of these mutations has the sequence corresponding to the third (from the gene outwards) site, normally at -330bp, deleted. The site is also deleted but the four other sites are found as normal. Another deletion removes the DNA sequences corresponding to the fourth and fifth sites (-405 and -480bp). However, all the hypersensitive sites are absent in this mutant. This suggests a hierarchical relationship between the sites. The -330 site is associated

with a partially self-complementary inverted repeat and around the -405 site there is a 14bp sequence showing homology to three other glue protein and hsp 70 gene 5' sequences (although these are found at different distances from the cAP site). Whether the sequences also correspond to DNAase I hypersensitive sites in these genes is unknown. The -480 site is associated with a three-fold direct repeat.

Tissue specific DNAase I sites have also been mapped 5' to the chicken α and β globin genes (Stalder et al., 1980; Weintraub et al., 1981). These DNAase I hypersensitive sites appear to accurately reflect the transcriptional switch from the embryonic β and α globin genes to the adult genes. The inactivation of these genes is also reflected by increased methylation of the DNA. Gene activity in this system is also reflected by the presence of high mobility group proteins (Weisbrod et al., 1980). Analysis of lines of chicken red blood cell precursors transformed by ts-AEV that can be induced to make globin upon temperature shift suggests that changes in DNA methylation and chromatin structure precede transcription and can be independently established (Weintraub et al., 1982). Another analysis by the same group shows the presence of SI nuclease sensitive sites 5' to active chicken globin genes. These sites are associated with DNAase I hypersensitive sites but do not map to exactly the same position (Larsen and Weintraub, 1982). Similar SI sensitive sites are found in supercoiled cloned chicken globin genes. However, the *in vivo* sites do not appear to be stress dependant as they are unaffected by pretreatment with ethidium bromide, restriction enzymes or nicking/closing enzyme.

As well as the association of undermethylation with gene activity *in vivo* by restriction enzyme analysis the importance of methylation in gene control has been more directly demonstrated in a number of ways.

Stein et al., (1982) contransformed *tk* and *aprt* (adenine phosphoribosyl transferase) genes into tk^- $aprt^-$ cells. The *tk* DNA was unmethylated the *aprt* DNA was either unmethylated or methylated *in vitro* with *Hpa*II methylase. Transformed cells were selected using the *tk* gene and the *aprt* genotype and phenotype then determined (Stein et al., 1982). The *aprt* phenotype was expressed only when the transforming DNA was unmethylated. Compere and Palmiter (1981) have shown that a cell line which is non-inducible for metallothionein I activity becomes inducible when exposes to 5 azacytidine (a cytidine analogue that cannot be methylated and appears to block methylation of cytidine residues). The *in vitro* methylation of cloned DNA has been shown to affect its subsequent transcription in *Xenopus* oocytes (Waechter and Baserga, 1982; Vardiman et al., 1982). These experiments suggest that the methylation of specific sites may be important rather than the general level of methylation.

Whether the relationship between the *in vivo* chromatin structure and DNA methylation is causal or merely a reflection of gene activity is unknown. However, one of the ts AEV transformed red cell precursors cell lines isolated by Weintraub et al., (1982) has globin gene specific DNAase I hypersensitive sites at the permissive temperature but only shows undermethylation of the respective gene sequences upon temperature shift. This suggests that changes in chromatin structure can precede gene expression and can occur independently of changes in DNA methylation.

The evolution of the rainbow trout

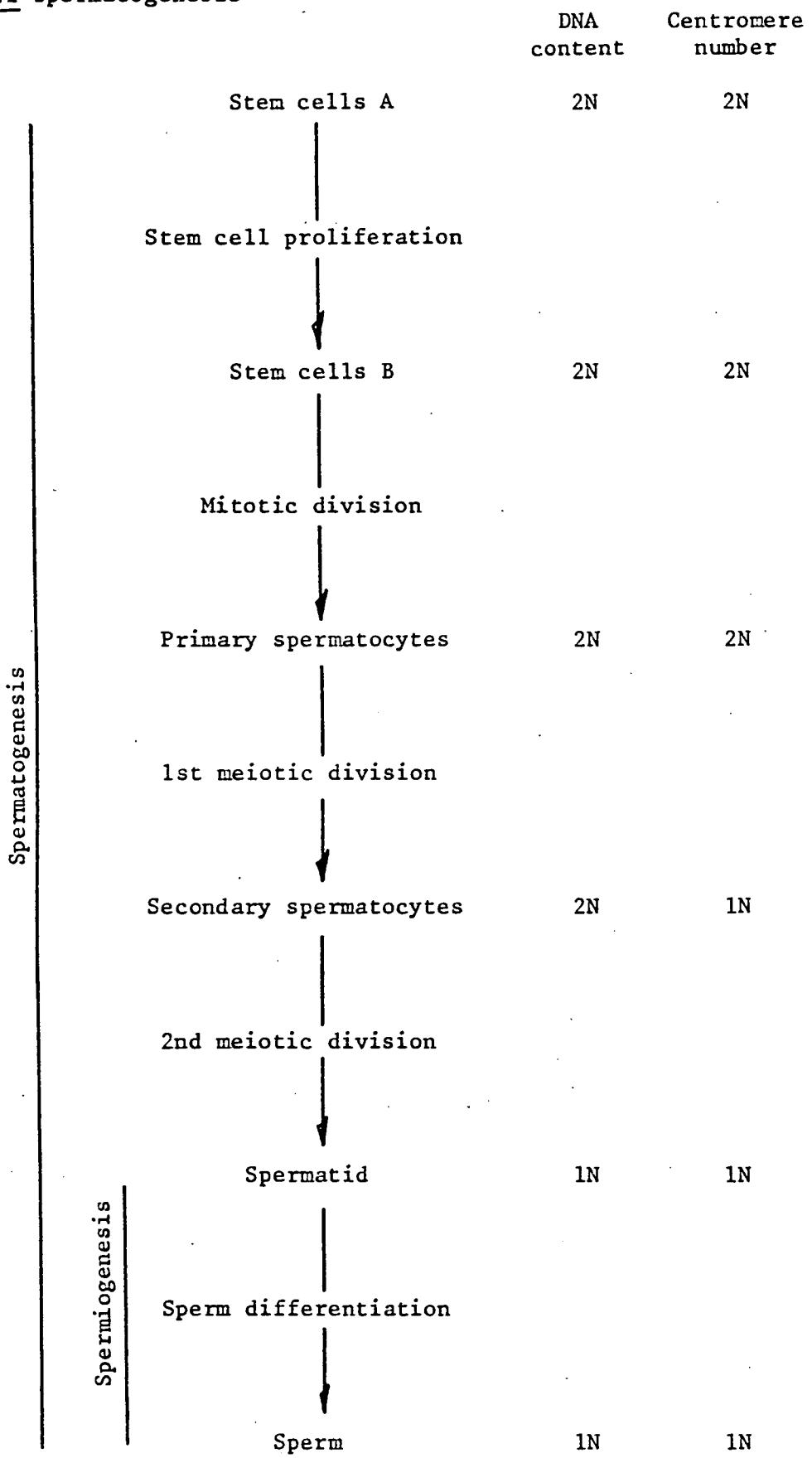
The rainbow trout, *Salmo gairdnerii* is a member of the family *Salmonidae* of the order *Isospondyli*. The *Salmonidae* are closely related to the *Clupeidae* which include the herring of the same order.

It has been proposed that the fishes evolved via a number of polyploidisations (Ohno et al., 1968). Evidence for this includes the results from counts of chromosome number and analyses of the cellular DNA content of different fishes (Ohno and Atkin, 1966; Ohno et al., 1968). Included in this evolutionary scheme is the evolution of the *Salmonidae* from the *Clupeidae* by a tetraploidisation event. The evidence for this tetraploidisation is found in the chromosome counts and DNA content analyses mentioned. Isozyme studies of the tetrameric enzymes malate dehydrogenase (Bailey et al., 1969) and lactate dehydrogenase (Massaro and Markert, 1968; Bailey and Wilson, 1968) also suggest that a tetraploidisation event occurred during the evolution of the *Salmonidae*. Isozymes of both enzymes are due to the existence of two monomer types α and β, which can combine in different numerical combinations. As the *Clupeidae* are thought to represent a tetraploid form of the original vertebraete genome (Ohno et al., 1968) the *Salmonidae* are, in an evolutionary sense, octaploid. The genetic study of the *Salmonidae* especially in comparison with other fish families, is therefore interesting from an evolutionary viewpoint.

Spermatogenesis.

Spermatogenesis is the developmental process by which sperm are produced. The process involves several stages (Fig. 1). The first step is the repeated division of diploid stem cells. The stem cells then undergo a further mitotic division to produce primary spermatocytes. The primary spermatocytes then undergo meiotic division. The first division of meiosis gives rise to secondary spermatocytes which although being diploid in DNA content have only a haploid number of centromeres (as the chromatids remain paired). The second division of meiosis gives rise to spermatids. Spermatids are haploid in

Fig.1 Spermatogenesis



both DNA content and centromere number and are the direct cellular precursors of sperm. The process of differentiation of spermatids to sperm is termed spermiogenesis. Spermiogenesis involves many biochemical and morphological changes. One of these changes is the condensation of the nucleus which occurs during the latter half of the process. In many species the nuclear condensation involves replacement of the histones normally found bound to DNA with novel, sperm specific proteins called protamines. This gives rise to highly condensed nucleoprotamine which is transcriptionally inactive. The appearance of protamines and their binding of DNA is one of the more easily characterised events of spermatogenesis. Protamines from a wide range of organisms have been described and characterised to a greater or lesser extent (Coelingh et al., 1969, 1972; Nakano et al., 1970; Kistler et al., 1973, 1974; Subirana et al., 1973; Monfoort et al., 1973; McMaster-Kaye and Kaye 1976; Bellve et al., 1975; Bellve and Carraway, 1978; Bols et al., 1980). The complete amino acid sequences of several fish protamines have been described. These are the rainbow trout protamines (iridines) (Ando and Watanabe, 1969; Gredamu et al., 1981), the tuna fish thynnins (Bretzel, 1972a, b, 1973), the herring clupeines, (Ando and Suzuki, 1966, 1967) and one component of salmon salmine (Ando and Watanabe, 1969).

Several models for the binding of protamine to DNA have been proposed. The fact that the volume of sperm nuclei is very close to the volume of the DNA contained in the nucleus favours a model where the protamine is almost entirely contained in one of the grooves of the DNA helix. One model has the protamine binding in the large groove via polyarginine and helical domains (Warrant and Kim, 1978). A second proposes that protamine binds in the minor groove in an extended conformation (Balhorn, 1982). X ray diffraction studies and infra red data support the latter model.

Protamines, however, are not found in the sperm of all species. The sperm of crabs appears to lack basic protein completely (Vaughn and Hinsch, 1972) while sea urchin sperm contain a novel histone but no protamines (Ozaki, 1971). One fish, the goldfish, also appears to lack protamines. In the goldfish sperm chromatin appears indistinguishible from somatic chromatin. No novel sperm specific chromatin proteins of any type are detectable (Munoz-Guerra et al., 1982).

Spermatogenesis in the rainbow trout

In contrast to mammals, where spermatogenesis is continuous, spermatogenesis in rainbow trout is seasonal and hence discontinuous. Spermatogenesis occurs naturally in rainbow trout during the months August to January but can be induced in immature fish by pituitary extracts from sexually mature salmon (Robertson and Rinfret, 1957). The seasonal, discontinuous nature of spermatogenesis in rainbow trout results in the process being semi-synchronous. This means that at any given time one cell type will be predominate in the testis. However all other cell types will also be present in lower numbers. During spermatogenesis the weight of the testis increases approximately a thousand times, from around 10 milligrams to between 5 and 10 grams.

In a rainbow trout, as in all other organisms, transcriptional activity decreases during spermatogenesis. This is reflected in a decrease both in the total amount, and in the sequence complexity of messenger RNA (Ando and Hashimoto, 1958; Levy and Dixon, 1977a) and also in a reduction in the ability of isolated chromatin to support RNA synthesis (Marushige and Dixon, 1969). This latter effect occurs in two distinct stages. A marked decrease is first seen at an early stage and is associated with an increase in the histone content and a decrease in

the non histone protein content of chromatin. A second large decrease, resulting in complete cessation of the ability to support RNA synthesis, occurs during the displacement of histones from chromatin by protamines.

The semi-synchronous nature of spermatogenesis in rainbow trout and the ability to induce spermatogenesis in sexually immature fish make the rainbow trout an ideal species for studying the processes of spermatogenesis.

The protamines of rainbow trout

The rainbow trout protamines are small and extremely basic. Fractionation of total protamine on alumina (Ando and Watanabe, 1969) or carboxymethyl cellulose (Ling et al., 1971), yields three components. The amino acid composition of the three components reveals that they are heterogeneous (Ling et al., 1971). Amino acid sequencing confirms this (Ando and Watanabe, 1969; Gedemu et al., 1981a) and suggests that there are 5 protamine species. These data give the length of the protamines as 30 to 32 amino acid residues. The extreme basicity of the protamines is due to their high molar arginine content of 66-71%.

The time course of protamine synthesis and DNA binding during spermatogenesis can be followed by the use of unit gravity gradients to separate the different cell types found in the testis (Louie and Dixon, 1972). The protamines first appear in the middle spermatids and are rapidly synthesised in middle and late spermatids. Synthesis occurs on characteristic disomes. The polysome profile of the testis reflects the sexual maturation of the fish through the number of disomes present (Ling and Dixon, 1970). At first synthesis of protamine is actinomycin

D sensitive but synthesis later becomes insensitive to actinomycin D (Ling and Dixon, 1970). This suggests that protamine messenger RNA is synthesised prior to maximum protamine synthesis and is metabolically stable. The presence of protamine messenger RNA in cytoplasmic ribonucleoprotein particles, both associated with ribosomes and in the post ribosomal cytoplasmic supernatant, suggest that such particles may represent a stable store of protamine messenger RNA (Gedamu et al., 1977a). The kinetics of the synthesis of the three protamine components, CI, CII, CIII (Ling et al., 1971) differ suggesting that the three components may play different specific roles in nuclear condensation. It is not known whether the kinetics of synthesis are controlled at the transcriptional or translational level.

Shortly after synthesis the protamines become phosphorylated (Ingles and Dixon, 1967; Sander and Dixon, 1972). Phosphorylation occurs on the serine residues. It appears that all the serine residues can become phosphorylated as the number of different phosphorylated species that are seen corresponds to the number of serine residues in the protamine molecule. The protamines subsequently become dephosphorylated before the sperm are completely mature (Ingles and Dixon, 1967). The dephosphorylation of protamine appears to be correlated with the formation of nucleoprotamine (Louie and Dixon, 1972). An enzyme capable of phosphorylating protamine has been partially purified from rainbow trout testes (Jergil and Dixon, 1970).

Modification of histones is also observed in the developing testis. Histones are extensively modified by both phosphorylation of serine residues and acetylation of lysine residues during all stages of spermatogenesis up to the middle spermatid when protamine replacement occurs (Dixon, 1972).

The modifications of protamines by phosphorylation, and of histones by phosphorylation and acetylation, reduces their net positive charge and hence will decrease their binding to DNA. This, and the time at which these modifications occur, suggest that this may help facilitate the ordered displacement of the histones by the protamines. The existence of a highly heterogeneous series of basic protein fragments during nucleoprotamine formation suggests that histone proteolysis may also be important (Marushige and Dixon, 1971).

The structure and synthesis of protamine messenger RNA

As might be expected from the small size of protamine, protamine messenger RNA (mRNA) is extremely short. Most protamine messenger is also polyadenylated and so can be readily purified by oligo dT-cellulose chromatography, sucrose gradient centrifugation and preparative polyacrylamide gel electrophoresis (Gedamu and Dixon, 1976).

On high resolution polyacrylamide gel systems, purified polyadenylated protamine mRNA can be resolved into 4 components with lengths of 270, 290, 310 and 330 bp (Iatrou and Dixon, 1977). Translation of the four separated mRNA components in wheat germ and rabbit reticulocyte cell free translation systems reveals that each codes for all three protamine components (Gedamu et al., 1979). Fractionation of deadenylated protamine mRNA using the same system also reveals 4 components (Gedamu et al., 1977b). Translation assays again reveal that each component is heterogeneous in its coding for protamine species (Gedamu and Dixon, 1979). Naturally occurring poly(A)$^-$ protamine mRNA is also found in rainbow trout testis. This poly(A)$^-$ mRNA is biologically active and appears to have a different cellular distribution from poly(A)$^+$ protamine mRNA during the early stages of protamine synthesis. At this

stage poly(A)$^-$ protamine mRNA is found almost entirely in polysomes while poly(A)$^+$ protamine mRNA is almost equally distributed between polysomes and post ribosomal supernatant (Iatrou and Dixon, 1977). However the low level of poly(A)$^-$ protamine mRNA as a percentage of the total amount of protamine mRNA (<4%) make its biological significance questionable.

Some primary structure of protamine mRNA has been determined by sequence analysis of radioactive single strand cDNA (Davies et al., 1977; Ferrier et al., 1977) and by direct sequence analysis of T1 ribonuclease fragments of protamine mRNA (Davies et al., 1979). These experiments have revealed the presence of the putative polyadenylation signal AAUAAA in an untranslated region (Ferrier et al., 1977). The T1 ribonuclease digestion suggests that considerable secondary structure exists in the protamine mRNA molecule because large T1 oligonucleotides containing internal G residues arise from the 3' non coding region. The existence of secondary structure in protamine mRNA is also suggested by the melting profiles of the purified mRNA (Davies et al., 1976).

Evidence suggests that protamine mRNA is synthesised long before it is translated (Iatrou et al., 1978). Analysis of testis cells separated on unit gravity gradients show that protamine mRNA is present in large quantities in primary spermatocytes while protamine synthesis occurs only in spermatids. In primary and secondary spermatocytes protamine mRNA appears only in the post ribosomal cytoplasmic supernatant and is not found on ribosomes until the spermatid stage. This suggests that the protamine mRNA is stored in the cytoplasm in an untranslatable form. This may be the ribonucleoprotein particles found in the cytoplasm. These contain polyadenylated RNA that, *in vitro*, codes only for protamine

(Gedamu et al., 1977). The synthesis of protamine mRNA in primary spermatocytes also raises the possibility that transcription only occurs in truly diploid cells and before meiosis begins (primary spermatocytes are, before meiosis begins, diploid in both DNA content and centromere number being tetraploid in DNA content during the division process. Secondary spermatocytes are diploid in DNA content and haploid in centromere number).

Number and repetition of protamine genes

Synthesis of radioactively labelled cDNA, using highly purified protamine mRNA as a template, allows a specific protamine sequence probe to be prepared. Two groups have used such a cDNA probe in reassociation experiments designed to estimate the number, and repetition of, the protamine genes.

In the first of these studies cDNA was reassociated with a vast excess of total genomic DNA at a single temperature (Levy and Dixon, 1977b). The results of this experiment suggest the reiteration frequency of the protamine genes is between 0.4 and 4 per haploid genome. No estimate of the number of different protamine genes is made.

In a second study cDNA was again reassociated with a vast excess of total genomic DNA. However the experiment was repeated at two different temperatures (Sakai et al., 1978). The results of reassociation at the higher temperature suggests that the protamine genes are unique. However at the lower temperature the results indicate that a large percentage of the sequences present in the cDNA are repeated about six times per haploid genome. This suggests that there are about six unique protamine genes that share considerable sequence homology. This would agree with the amino acid sequence data that shows that the protamines differ

little in primary structure. That this homology is retained at the nucleotide level suggests either recent duplication and divergence or correction mechanisms to prevent nucleotide sequence drift.

Cloning and sequence analysis of protamine cDNA

Several different groups have cloned and sequenced protamine cDNA (Jenkins et al., 1979; Jenkins, 1979; Sakai et al., 1981; Gedamu et al., 1981a). These sequences are shown in figure 2.

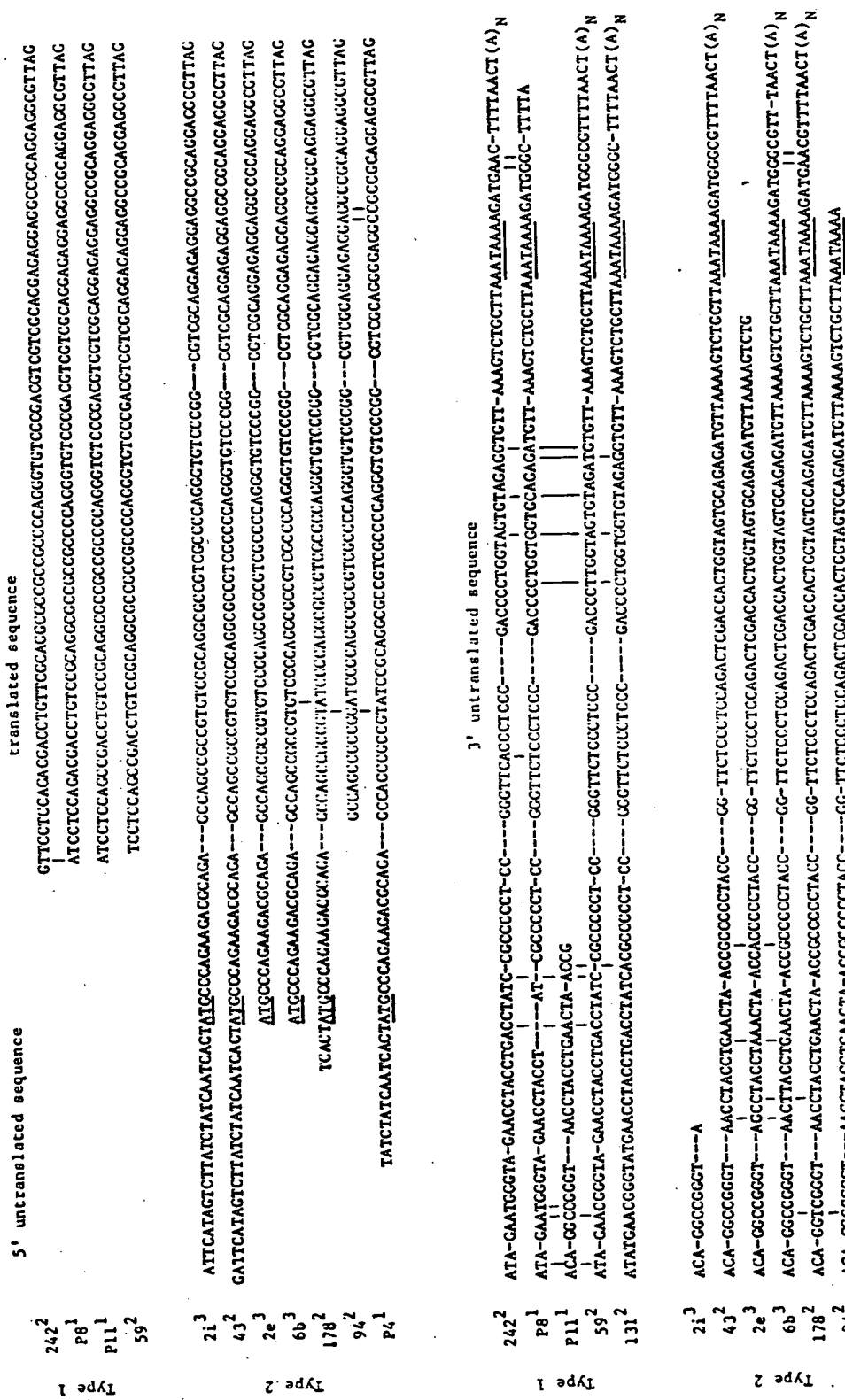
All the protamine cDNA sequences with a long 3' non translated sequence show the consensus sequence AAUAAA starting between 85 to 100bp from the termination codon. This suggests the length of the 3' non-coding region in protamine mRNA is about 100 to 120bp.

Over half of the cDNA's end at the same point in the mRNA, 71bp upstream from the termination codon. This suggests that there is a very tight secondary structure at this point in the mRNA which causes premature termination by reverse transcriptase.

Examination of the codon usage in the protamines shows that there is preferantial usage of certain arginine and serine codons (Table 1). Approximately 30% of the arginine codons are CGC and 27% AGG. The codons AGA and CGT are used approximately equally and account for most of the remaining arginine codons. The two CGPu codons are used very rarely, accounting for only 8% of the arginine codons between them. The serine codon TCC accounts for 73% of the serine codons, the codon AGC the remainder. The four other serine codons TCT, TCA, TCG and AGT are not used at all.

Fig.2 Nucleotide sequences of protamine cDNA clones

Sequences were aligned for maximum homology



1. Jenkins,(1979) 2. Gedamu et al.,(1981a) 3. Sakai et al.,(1981)

Table 1 Codon usage in protamine cDNA sequences

TTT Phe 0	TCT Ser 0	TAT Tyr 0	TGT Cys 0
TTC Phe 0	TCC Ser 16	TAC Tyr 0	TGC Cys 0
TTA Leu 0	TCA Ser 0	TAA Term 0	TGA Term 0
TTG Leu 0	TCG Ser 0	TAG Term 8	TGG Trp 0
CTT Leu 0	CCT Pro 4	CAT His 0	CGT Arg 27
CTC Leu 0	CCC Pro 11	CAC His 0	CGC Arg 52
CTA Leu 0	CCA Pro 0	CAA Gln 0	CGA Arg 8
CTG Leu 0	CCG Pro 0	CAG Gln 0	CGG Arg 5
ATT Ile 0	ACT Thr 0	AAT Asn 0	AGT Ser 0
ATC Ile 3	ACC Thr 0	AAC Asn 0	AGC Ser 6
ATA Met 0	ACA Thr 0	AAA Lys 0	AGA Arg 18
ATG Met 3	ACG Thr 0	AAG Lys 0	AGG Arg 40
GTT Val 1	GCT Ala 0	GAT Asp 0	GGT Gly 0
GTC Val 4	GCC Ala 4	GAC Asp 0	GGC Gly 8
GTA Val 0	GCA Ala 0	GAA Glu 0	GGA Gly 8
GTG Val 8	GCG Ala 0	GAG Glu 0	GGG Gly 0

Taken from Gedamu et. al. (1981)

Preferential codon usage has been reported for a large number of genes and several theories suggested to explain it. These include the proposals that it is an evolutionary strategy to reduce the number of mutations with drastic effects (Modiano et al., 1981), that it reflects dinucleotide frequency preference (Nussinov 1981) and that codon usage reflects the genus or type of organism, (Grantham et al., 1980), this perhaps being due to tRNA availability. Another theory also suggests that tRNA availability and codon anticodon binding energies may be reflected in third position preferences in codons (Wilson et al., 1980). Obviously constraints can operate at a number of different places. The theory that preferential codon usage may be a strategy to reduce the number of mutations with drastic effects (i.e. mutations to give stop codons or nonpolar hydrophobic/hydrophilic amino acid substitutions) goes some way to explaining the preferential codon usage in the protamine genes. There is preferential use of the four CG arginine codons. This would allow single bp mutations to occur in the wobble (third) position of the codon without producing an amino acid change. There is also preferential usage of the CGU and C codons in this group. These two codons are two changes away from any termination codon while the CGA and CGG need only a single change to give a stop codon. This explanation also fits the preferential use of the arginine AGG codon over the AGA codon. The preferentially used serine codon, TCC, also allows third position change without changing the amino acid. It also needs two changes to give a stop codon, however so do all the TC serine codons. It is unlikely that preferential codon usage reflects a single factor but is probably due to a number of constraints operating simultaneously. At present a full explanation is therefore impossible.

Comparison of the coding sequences of the cDNA clones (Fig. 2) shows that they are remarkably conserved at the nucleotide level as well as at the amino acid level. Most of the sequence variation is found between codons 6 and 11. This corresponds to the major site of serine phosphorylation. The sequence variation found in this region can be used as a basis for dividing the cDNA sequences into two groups. The first group is characterised by having two or three serine residues and a single proline residue in the variable region. All the cDNAs represent ~~a time~~ alone of this group terminate prematurely, only 27 codons of the coding sequence are present. Comparison of the four sequences that fall into this group show that two amino acid and three nucleotide sequences are represented. The amino acid change is at the third codon represented, this being either arginine (AGA) or serine (AGC). The third nucleotide sequence is due to a change in the third, or wobble, base of a valine codon. This change may represent allelic variation as the comparable cDNAs showing this change were constructed by different groups.

The second group of sequences is characterised by having a single serine residue and no proline residue in the variable region. Five out of six of the cDNA clones representative of this group have the whole of the coding region of the mRNA represented. The five full-length clones all have the ubiquitous ATG initiation codon at the beginning of the coding region immediately followed by a proline residue. Comparison of the coding sequences shows that there are three amino acid sequences and four nucleotide sequences represented. One protein sequence difference is at residue ten (valine or isoleucine), the second at residues twenty four to twenty six (Arginine, glycine, glycine, with either valine or isoleucine at residue ten, or glycine, glycine, arginine with

isoleucine at residue ten). The fourth nucleotide sequence is caused by a change in the third, or wobble, base of an arginine codon at residue nine (CGT in five of the sequences, CGG in one). Again this could represent allelic variation.

Comparison of the 3' non coding sequences represented in the cDNAs show that they are almost as well conserved, within each of the two groups, as the coding sequences. However the two groups have distinctly different 3' non coding sequences. The one exception to this phenomenon is the clone pTPII (Jenkins, 1979), which contains a group 1 type coding sequence but a group 2 type 3' non-coding sequence.

This analysis of cDNA sequences suggests that the protamine genes may have evolved by repeated duplication of the genome, the most recent of these being the duplication during the evolution of the *Salmonidae* from the *Clupeidae*. However, it is impossible to present a detailed evolutionary scheme without knowing if nucleotide sequence variation is allelic or not. Cloning of genomic sequences should make this possible.

Materials and Methods

Plating bacteriophage lambda

Charon 4A and recombinant derivatives were plated on *E. coli* ED8654 (*sup* E, *sup* F, *hsd* R^- M^+ S^+, *met*$^-$, *trp* R. Murray 1977). A static culture, grown in LB, was diluted 1:50 into fresh LB and grown, with shaking, to an OD_{650} of 0.5.Cells were then pelleted by centrifugation and resuspended in an equal volume of 10mM $MgSO_4$. Cells in $MgSO_4$ could be kept for several days at 4°C. Phage were then absorbed by addition of 100µl of phage (in phage buffer) to 500µl of plating bacteria, vortexing, and incubating at 37°C for 20 to 30 minutes. 2.5mls of LB top agar (42°C) made 10mM in $MgSO_4$, was then added and mixed by rolling. The mixture was then poured onto a fresh, thick, 9cm diameter LA plate. The plates were then inverted and incubated at 37°C overnight. All volumes were scaled up if plates larger than 9cm diameter were being used.

Phage plate lysates

Phage plate lysates were prepared by plating phage at a density of 8 to 10 x 10^3 pfu/cm^2 (if possible using phage picked from a single plaque). After lysis became confluent the phage were harvested by flooding the plate with phage buffer and then scraping off the top layer, with the phage buffer, into a suitable container. A small volume of chloroform was then added and the mixture vortexed to ensure complete lysis. The phage suspension was then clarified by centrifugation (Sorvall, SS34 4°C, 10k rpm , 10 minutes) and kept at 4°C.

Preparation of DNA from phage plate lysates

DNA was prepared from phage plate lysates essentially as described by Cameron et al., (1977) except that the phage suspension recovered from the plate lysate was first treated with 10µg/ml DNAase and RNAase (37°C, 30min). The DNA was also routinely phenol/chloroform extracted after recovery. The DNA was then stored in 10mM TrisHCl pH8.0, 1mM EDTA rather than 100mM TrisHCl pH7.5, 1mM EDTA as described.

DNA prepared by this method was restricted in 4mM spermidine as described (see restriction digests).

Phage liquid lysates

A static culture of cells was diluted 1:20 LB (made 10mM in $MgSO_4$) and grown at 37°C, with vigorous aeration to an OD_{650} of 0.45-0.6 (approximately 2-3 x 10^8 cells/ml). Vigorous aeration was ensured by using LB at 10% of flask volume and using a shaking incubator at full speed. When the OD_{650} reached the appropriate range phage were innoculated at a multiplicity of infection of one and incubation continued as before. The growth and subsequent lysis of the cells can be followed by following the rise, and then the fall to a minimum level, of the OD_{650}. When lysis was complete (i.e. when the OD_{650} reached a minimum) chloroform was added to 0.2% and shaking continued for 10 minutes. The lysate was the clarified by centrifugation (Sorvall G5A, 10Krpm, 4°C, 10 minutes). Contaminating E. coli DNA and RNA were removed by treatment with 10µg/ml DNAase and RNAase at 37°C for 30 minutes. Phage were then purified by PEG precipitation and centrifugation through CsCl step gradients.

Polyethylene glycol (PEG) precipitation of phage

To PEG precipitate phage from liquid lysates NaCl was first added

to 4% w/v. Solid PEG 6000 was then added to 10% w/v and dissolved by gentle swirling. The lysate was then left at 4°C for at least an hour. The PEG precipitate was then pelleted by centrifugation (Sorvall GSA, 10K rpm, 4°C, 10 minutes). The supernatant was discarded and the pellet resuspended in a small volume (2-5% starting volume) of phage buffer. The resulting phage suspension was then partially clarified by centrifugation (Sorvall HB4, 5K rpm, 4°C , 10 minutes). Phage were then purified by two cycles of centrifugation through CsCl step gradients.

Caesium chloride step gradients

CsCl step gradients were run in a Beckman SW40 rotor using clear (nitrocellulose) tubes. Step densities were 1.3, 1.5 and 1.7g/ml (31%, 45%, 56% w/v CsCl at 20°C) in phage buffer. The step volumes were 1 to 1.5ml. The steps were introduced by pump into the bottom of the tube (lightest first) and the phage suspension was then carefully layered on top. Centrifugation was at 35K rpm, 20°C for 1.5 to 2 hours. Phage, showing as a clean white band approximately 1/3 of the way through the steps, were collected by side puncture. CsCl was removed by dialysis against two changes of 10 TrisHClpH8,1mM EDTA.

Preparation of phage DNA

Phage DNA was prepared from purified phage by three phenol/ chloroform extractions followed by two chloroform extractions. Phage DNA was then either dialysed into 10mM TrisHClpH8.0, 0.1mM EDTA (using four to six buffer changes) or ethanol precipitated and redissolved in the same buffer. Phage DNA's were stored at -20°C.

Preparation of *Eco*RI methylase

*Eco*RI methylase was prepared from frozen *E.coli* RY13 cells essentially as described by Greene et al., (1974, 1975). 200g of cells were thawed in EB buffer (10mM K-P pH7.0, 7mM β MSH, 1mM EDTA) and disrupted by sonication. Cell debris was removed by centrifugation (Sorvall SS34, 19K rpm, 45min) and the resulting supernatant streptomycin precipitated. The precipitate was removed by centrifugation (Sorvall GSA, 8 Krpm, 30min) and the supernatant fractionated with 50% v/v saturated ($4^{o}C$) ammonium sulphate. The ammonium sulphate precipitate was recovered by centrifugation (Sorvall GSA, 8K rpm, 30 minutes) and redissolved in EB + 0.2M NaCl. The enzyme was then fractionated on a 240ml phosphocellulose (Whatman P11) column, concentrated on HAP and refractionated on a 40ml carboxymethyl cellulose (Whatman CM52) and a 50cm sephadex G100 column exactly as described by Greene et al., (1975). Between the carboxymethyl cellulose and sephadex G100 columns the enzyme was concentrated using a Millipore immersible CX ultra filter. The active fractions from the sephadex G100 column were pooled and dialysed into 5mM K-P pH7.0, 3.5μM β MSH, 0.5μM EDTA, 0.1% NP40, 0.1M NaCl, 50% v/v glycerol and stored at $-20^{o}C$. The purified enzyme was then assayed for contaminating endonuclease on supercoiled plasmid DNA and for 5' and 3' exonuclease using a religation assays on restriction enzyme (*Eco*RI and *Pst*I) sticky ends. In methylase reaction buffer (0.1M TrisHClpH8.0, 1mM EDTA) both these tests were negative. One unit of *Eco*RI methylase is defined as the amount of enzyme that will incorporate 1 pmole of methyl groups into DNA in one minute under standard reaction conditions (0.1M TrisHClpH8.0,10mM EDTA,6μM SAM).

Cloning in lambda Charon 4A

The lambda cloning vector Charon 4A (Fig. 3) was used to construct a genomic library essentially as described by Kemp et al., (1979). The original rationale of the system is best described by Maniatis et al., (1978).

(1) Preparation of Charon 4A *Eco*RI arm fragments

Charon 4A DNA was restricted to completion with an excess of *Eco*RI and recovered by phenol/chloroform extraction and ethanol precipitation. The DNA was redissolved at 50-100μg/ml in 100mM TrisHCl pH8, 10mM $MgCl_2$ and the lambda cohesive ends annealed at 42°C for two hours. The solution was then cooled on ice and made 10mM in EDTA. The DNA fragments were then size fractionated on 10-40% sucrose gradients (25-50μg/gradient).

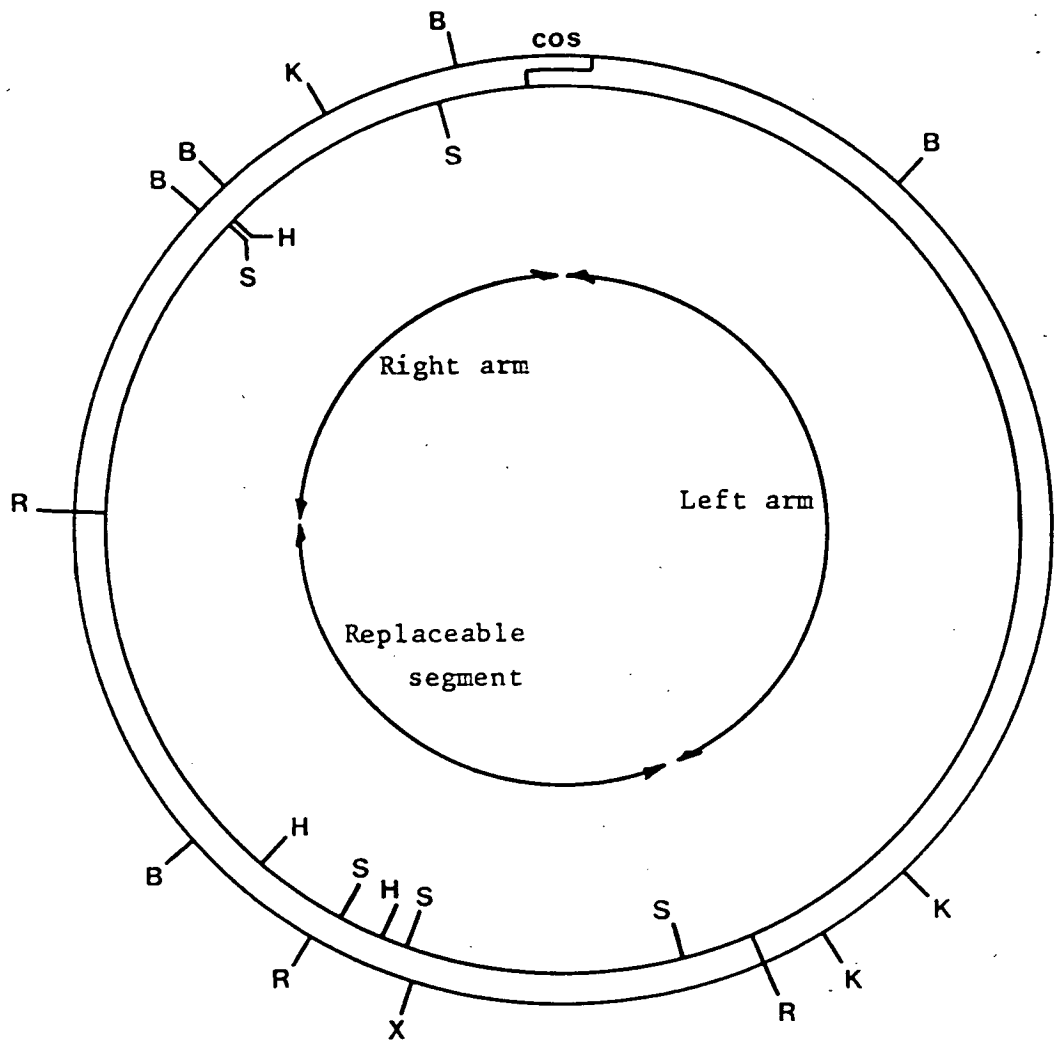
(2) Preparation of trout genomic DNA *Eco*RI* fragments

Trout testis DNA was first methylated to completion with *Eco*RI methylase in 100mM TrisHClpH8.0, 10mM EDTA, 6μM SAM with 1 unit of *Eco*RI methylase/10μg DNA at 37°C for 2hrs. The degree of methylation was determined by removing a portion of the reaction immediately after addition of the enzyme, adding a small amount of lambda (C1857S7) DNA (0.2-0.5μg) and incubating in parallel with the main reaction. The lambda DNA was then tested for restriction with *Eco*RI

After recovery of the methylated DNA by phenol/chloroform extraction, dialysis (to remove EDTA) and ethanol precipitation the DNA was restricted with *Eco*RI* activity. The *Eco*RI* activity of *Eco*RI is promoted by restriction in low salt buffer (Polisky et al., 1975).

Fig.3 Charon 4A restriction map

Total size 45.41kb



B *Bam*HI
H *Hind*III
K *Kpn*I
R *Eco*RI
S *Sst*I
X *Xba*I
cos lambda cohesive ends

From de Wet et al.,(1980).

The reaction was in 25mM TrisHClpH8.9(at 37°C) 2mM $MgCl_2$, at 37°C for 8 hours using several different enzyme concentrations (all producing only partial digestion). DNA was recovered by phenol/chloroform extraction and ethanol precipitation before being size fractionated on 10-40% sucrose gradients (100-200µg/gradient).

(3) Size fractionation of DNA

DNA was size fractionated on 10-40% sucrose gradients exactly as described by Maniatis et al., (1978). Fractions of 0.5ml were collected via an ISCO ultraviolet flow analyzer. Appropriate fractions were then analysed on agarose gels.

(4) Ligation and packaging of DNA

Charon 4A *Eco*RI arms and size fractionated (12-20kb) *Eco*RI* genomic fragments were ligated either using the DNA concentrations of Maniatis et al., (1978) or Kemp et al., (1979). Charon 4A arms were first annealed in 100mM TrisHClpH7.6, 10mM Mg Cl_2 at 42°C for two hours. The solution was then cooled on ice, genomic *Eco*RI* fragments added and the buffer adjusted to 66mM ~~TMSCHl~~ TrisHCl pH7.6, 10mM Mg Cl_2, 1mM EDTA, 40mM NaCl, ~~2mM NaCl~~, 2µM DTT, 0.1µM ATP, 125µg/ml BSA. T4 DNA ligase was then added and the ligation incubated at 12°C overnight.

The ligation mix was then placed on ice and packaged exactly as described by Grosveld et al., (1981). The packaging extracts were kindly supplied by Melville Richardson. After packaging the recombinant phage were diluted into 0.5ml of phage buffer and kept at 4°C.

Amplification of recombinant Charon 4A library

The recombinant phage were amplified on LA plates by plating at

a density of 50pfu/cm^2. This low plating density was to minimize selective amplification. Phage were harvested as described. After clarification the phage suspension was PEG precipitated and loaded onto CsCl step gradients. After centrifugation was complete the entire volume below the protein band was collected and stored at 4^{o}C. The phage titre in this stock was determined as described.

Assay of non-recombinant phage

The percentage of non-recombinant phage in the Charon 4A library was determined by plating on BB2 agar containing 40µg/ml 5 chloro 4 bromo 3 indolyl-β-D-galactoside (x gal) using an *E. coli lac z* deletion strain C344 (*thr*$^-$, *leu*, B_1, *sup* E, *ton* A, *hsd* R^-M^-, *lac z*). Recombinant phage produce colourless plaques, non recombinant blue plaques.

Screening recombinant plaques for specific DNA sequences

Recombinants were screened using nick translated probes essentially as described by Benton and Davis (1977). Phage were plated at a density of 70pfu/cm^2 on 22cm x 22cm plates of LA agar using top agarose (0.7%) instead of top agar. Phage were grown overnight and then cooled to 4^{o}C before blotting. Replica filters were absorbed sequentially for 1 minute and 5 minutes. After denaturation and neutralization filters were blotted dry between Whatman 3MM paper and baked for 1.5 to 2 hours at 80^{o}C in a vacuum oven. Filters were hybridized as described except that sonicated *E. coli* DNA (50µg/ml) was included as carrier (instead of salmon sperm DNA).

Recombinant phage were purified by two further cycles (after the original screen) of screening at low plating density. Plaques were picked, using a sterile toothpick, into 0.5ml phage buffer containing

a drop of chloroform and stored at $4^{o}C$.

Use of the lambda vector EMBLI

EMBLI is a spi type lambda cloning vector which is a close derivative of λ1059 (Karn et al., 1980). It is a BamHI vector with a cloning capacity of 6.3 to 24.4 kb. The restriction map of EMBLI is shown in Fig. 4. The bacterial strains used in conjunction with EMBLI were Q358 (r_k^-, m_k^+, SU_{II}^+, 80^R) and Q359 (r_k^-, m_k^+, SU_{II}^+, 80^R, P2).

Genomic fragments for cloning in EMBLI were prepared by partial *Sau*3A or *Bam*HI digestion. The DNA was then size fractionated on 0.6% agarose gels and fragments of the appropriate MW recovered by electroelution.

EMBLI DNA was restricted with a two fold excess of BamHI and recovered by phenol/chloroform extraction and ethanol precipitation. The BamHI restricted EMBLI DNA was then used directly.

Ligation of vector and insert DNA was carried out as described by Maniatis et al., (1978) and DNA was packaged as described by Grosveld et al., (1981). The phage from the packaging reaction were then plated on Q359.

The bacterial media and methodology were the same as for propogation and use of charon 4A.

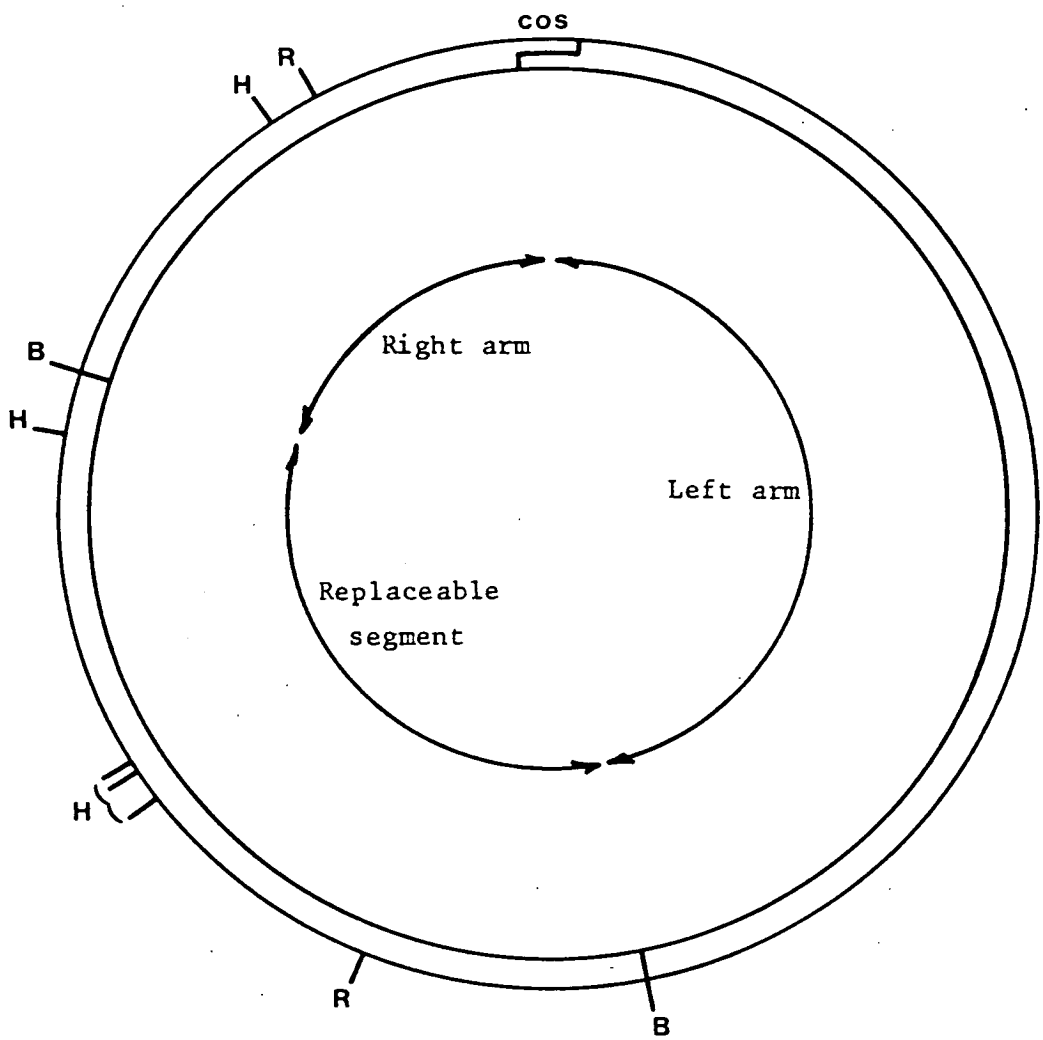
Sub cloning in pAT153

(1) Preparation of recombinant DNA

Sub clones of charon 4A recombinants were constructed using the plasmid vector pAT153 (Twigg and Sherratt, 1980; Fig. 5). Plasmid

Fig.4 EMBL 1 restriction map

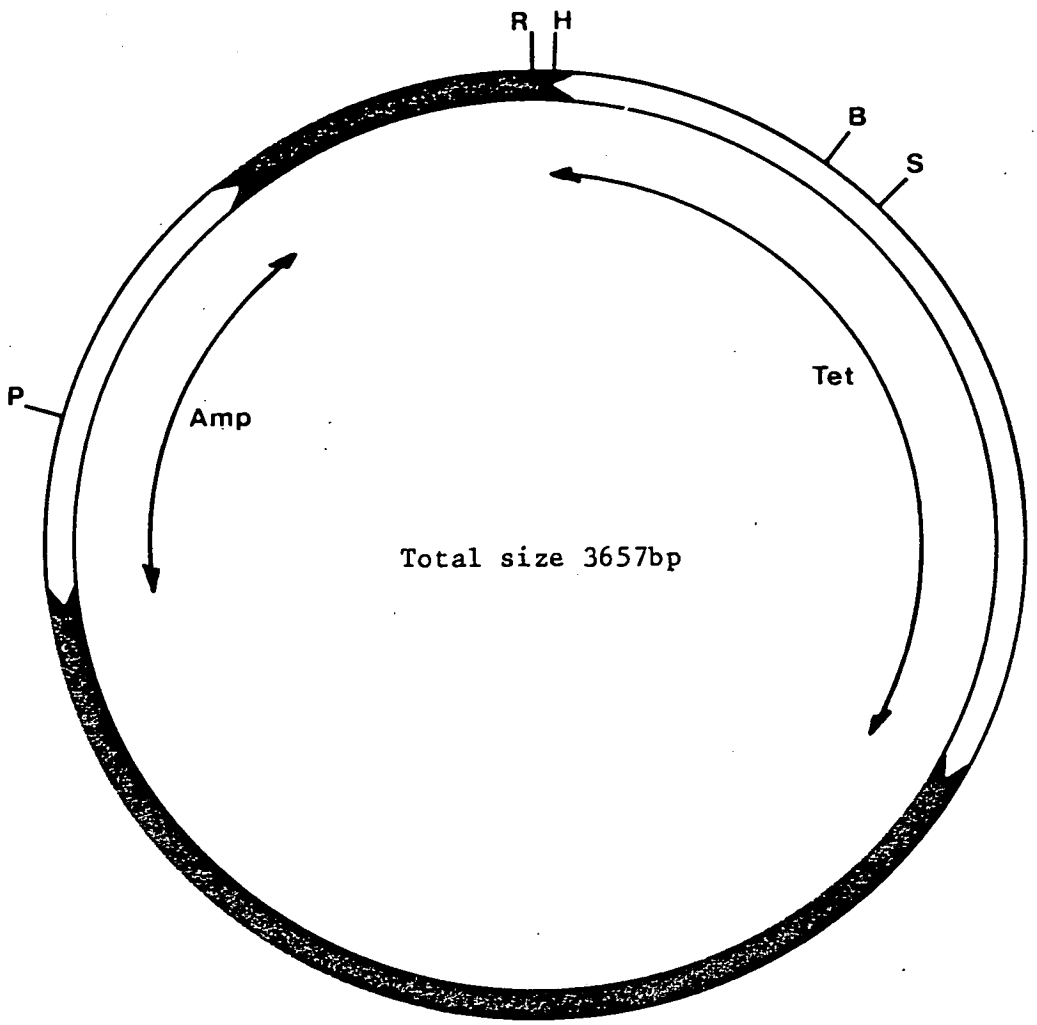
Total size 46kb



B *Bam*HI
H *Hind*III
R *Eco*RI
cos lambda cohesive ends

From Karn et al.,(1980).

Fig.5 pAT153 restriction map



B *Bam*HI
H *Hind*III
R *Eco*RI
S *Sal*I
P *Pst*I
Amp Ampicillin resistance marker
Tet Tetracyclin resistance marker

From Sutcliffe, (1978) and Twigg and Sherrat, (1980)

DNA was restricted and recovered by phenol/chloroform extraction and ethanol precipitation. Small fragments released by double enzyme restriction were removed by sepharose CL2B chromatography in 0.3M NaCl 0.1M NaOAc pH5. Phage DNA was restricted using the same enzyme(s) and recovered by phenol/chloroform extraction and ethanol precipitation. 0.5µg of plasmid and 0.25µg of phage DNA were coligated in 50µl of 66mM TrisHCl pH7.6, 1mM EDTA, 10mM $MgCl_2$, 40mM NaCl, 2mM DTT, 0.1mM ATP, 125µg/ml BSA using T4 DNA ligase. Ligation was at 10°C for 2 hours and then 0°C for 12 to 16 hours. The ligation was then diluted by addition of 0.3ml of ice cold TMC (10mM TrisHCl pH7.5, 10mM $MgCl_2$, 10mM $CaCl_2$), and transfected into competent HB101 (Boyer and Roulland-Dusseix, 1969) using a modification of the method of Mandel and Higa (1970).

(2) Transfection into HB101

To prepare competent cells a static culture of HB101 was diluted 1:50 in LB and grown, with shaking, at 37°C, to an OD_{650} of 0.5. The culture was cooled on ice for 15 minutes and the cells pelleted by centrifugation and resuspended in ½ volume of ice cold 50mM $CaCl_2$. After standing on ice for a further 15 minutes the cells were again pelleted and resuspended in 1/10 volume of ice cold 50mM $CaCl_2$. Transfection was carried out by the addition of 50µl of DNA/TMC to 100µl of component cells. After vortexing the mix was then incubated on ice for 15 minutes. The transfection mix was then warmed at 37°C for 2 minutes and 1ml of LB, at 37°C, added and incubation at 37°C continued for a further 30 minutes. The cells were then plated on LB plates in 2.5ml of BBL top agar. The LB plates contained antibiotics to select for transformants, antibiotics at 50% concentration were also included in the top agar. Tetracycline was used at 10µg/ml, ampicillin at 100µg/ml.

(3) Sizing of recombinants

Recombinants were identified by replica plating on suitable antibiotic plates and plasmids sized using a quick lysis procedure modified from that of Barnes (1977). Colonies or patches of recombinant plasmids were picked into 100µl of 1 x TBE buffer (see agarose gels) and suspended by vortexing. 20µl of 5% SDS, 100mM EDTApH7.0 were added and mixed by vortexing. The lysis mixture was then heated at 70°C for 10 minutes, vortexed again and cooled. The cell lysate was then electrophoresed on a 0.6% agarose/TBE gel until the BPB marker was at the bottom of the gel.

Growth of plasmids and preparation of plasmid DNA

A single colony of transfected HB101 was innoculated into 25ml of LB containing a plasmid selecting antibiotic. The culture was grown overnight with shaking, at 37°C and then innoculated into 200ml of LB/antibiotic. Incubation was continued at 37°C, again with shaking, until the OD_{650} reached 1:0. Replication of chromosomal DNA was then blocked by the addition of a suitable antibiotic (for pAT153 chloramphenicol, at 150µg/ml, was used), and incubation continued overnight.

The culture was then chilled on ice and cells collected by centrifugation (Sorvall GSA, 10K rpm, 4°C, 15 minutes). Cells were resuspended in 40mls of ice cold 10mM TrisHCl pH7.4, 1mM EDTA and repelleted (as before, 10 minutes). The cells were then resuspended in 4mls of ice cold 25% sucrose, 50mM TrisHCl pH8.1, 40mM EDTA by homogenisation and 1.2mls of 10mg/ml lysozyme (in the same buffer) added. The mixture was then left to stand on ice for 10 minutes (with occasional mixing) 1.2mls of ice cold 0.5M EDTApH8.1, were added and the mixture was left on ic

for 5 minutes. Finally 10.8mls of ice cold 0.1% triton X-100, 62.5 mM EDTA, 50mM TrisHClpH8.1 were added. After standing on ice for 10 minutes the lysed cells were centrifuged at 20K rpm (Sorvall, SS34), $4^{o}C$ for 1 hour and the resulting supernatant collected. Plasmid DNA was then purified by two cycles of caesium chloride/ethidium bromide density gradient centrifugation.

Agarose gels

(1) DNA gels

Gels of 0.6% to 2% agarose were run vertically. The gels were 16cm x 20cm (wide) x 0.8cm. Agarose was first dissolved in distilled water by refluxing, cooled to $50^{o}C$ and made to 1 x TBE (90mM Tris, 90mM boric acid, 2mM EDTA, pH8.3) by the addition of 10 x TBE and then cast. Samples were made 6% w/v Ficoll, 0.01% BPB, 20mM EDTA ph7.0 by the addition of 5 x stock. Samples containing bacteriophage lambda DNA were heated to $70^{o}C$ for 5 minutes and then cooled on ice before loading to melt the cohesive ends of lambda. Electrophoresis was at 30 to 60v overnight. Gels were stained in 1µg/ml ethidium bromide, 1 x TBE for 45 to 60 minutes and photographed using a short wavelength UV transilluminator and a polaroid camera with a red filter.

Gels of below 0.6% agarose were run horizontally. Gels were 26cm x 20cm (wide) x 0.5cm. Gels were made and cast as before except that ~~this~~ tris acetate buffer (50mM Tris, 20mM NaOAc, 2mM EDTA, 10mM NaCl pH'ed to 7.9 with glacial acetic acid) was used and ethidium bromide at 1µg/ml, was included in the gel. The gel was connected to buffer reservoirs using wicks of Whatman 3MM paper soaked in gel buffer. Electrophoresis was at 50 to 75v overnight. The gel was covered with a thin sheet of plastic during electrophoresis.

(2) RNA gels

RNA was run on agarose formaldehyde gels essentially as described by Rave et al., (1979) except that gel buffer was 1 x MOPS (20mM MOPS, 1mM EDTA, 5mM NaOAc, pH7.0). Vertical slab gels were prepared as described for DNA gels. 10 x MOPS and 50% v/v formaldehyde (to 6% v/v) being added after refluxing and cooling of the agarose in distilled water. Samples were made 50% v/v formamide, 6% v/v formaldehyde, 1 x MOPS, heated at 60°C for 5 minutes and quenched on ice. Samples were then made 3% w/v ficoll, 0.005% BPB, 10mM EDTA pH7.0 by the addition of 10 x stock before loading. Electrophoresis was at 30-35 v overnight. For visualisation of RNA the gel was stained in 50µg/ml acridine orange/1 x MOPS for 30 to 90 minutes and then destained in two changes of 1 x MOPS (30 minutes each).

Electroelution of DNA from agarose gels

To elute DNA from a gel a slice of agarose was first removed from directly in front of the DNA to be eluted. A piece of sterile dialysis membrane was then placed in the slot and under the band (the gel being placed horizontally). The trough was then filled with buffer and the DNA electrophoresed onto the membrane. When the DNA had all moved onto the membrane it was rapidly removed into a small volume of distilled water and washed thoroughly. The DNA was then recovered by ethanol precipitation.

Native acrylamide gels

Acrylamide gels 20cm x 15cm (wide) x 1.5mm were used. Gel buffer was 1 x TBE. The acrylamide : bisacrylamide ratio was 30:1.

The gel mix was degassed thoroughly before polymerization. Polymerization was catalysed by the addition of ammonium persulphate to 0.03% and a predetermined volume of TEMED. Gels were aged overnight after polymerization and preelectrophoresed for 30 minutes before use. Electrophoresis was at 100v.

Southern transfers

Southern transfers were carried out essentially as described by Southern (1975) using the modifications described by Wahl et al., (1979). Transfer was in 20 x SSC and after transfer the nitrocellulose membrane was washed in 2 x SSC, blotted dry using Whatman 3MM paper and baked at 80°C in a vacuum oven for 1½ to 2 hours.

Northern transfers

Northern transfers of RNA from agarose/formaldehyde gels to nitrocellulose were carried out essentially as described by Thomas (1980). Transfer was usually overnight. The nitrocellulose membrane was baked, without rinsing in 2 x SSC as described for Southern transfers.

Hybridisation of nitrocellulose membranes

Nitrocellulose membranes were hybridised to labelled probes exactly as described by Maniatis et al., (1978) except that 10% w/v dextran sulphate was included in the hybridisation (Wahl et al., 1979) Hybridisations with G/ C tailed cDNA plasmids as probes contained 10µg/ml poly I to prevent non-specific hybridisation to G/C rich sequences. Occasionally the stringency of washing was increased by adding a final wash in 0.1 to 0.5 x SET.

Preparation of trout testis DNA

Trout testis DNA was prepared from nuclei isolated as described by Marushige and Dixon (1971). Isolated nuclei were washed four times by resuspension and centrifugation in saline/EDTA as described. The nuclei were then resuspended in NTE (100mM NaCl, 100mM TrisHCl pH8.0, 100mM EDTA). Proteinase K was added to 100µg/ml and mixed thoroughly by gentle homogenization. SDS was then added to 1% w/v and the mixture incubated at 37°C for 3 hrs. The resulting DNA solution was then phenol/chloroform extracted twice, chloroform extracted twice and then extensively dialysed against 20mM TrisHCl pH8.0. DNA was then purified by two cycles of caesium chloride/ethidium bromide density equilibrium centrifugation.

Caesium chloride/ethidium bromide density equilibrium centrifugation

DNA solutions were made up to a density of 1.55g/ml, by the addition of solid CsCl, and 1mg/ml in ethidium bromide. The DNA was then banded by centrifugation (Beckman A4 or MSE 10 x 10ml, 35K rpm, 20°C for 60 hours). DNA was visualised by UV illumination and collected by side puncture with a syringe. Ethidium bromide was removed by repeated extraction with isopropanol saturated with water and CsCl. The DNAs were finally dialysed against 10mM TrisHCl pH8.0, 0.1mM EDTA and stored at -20°C.

Preparation of trout testis RNA

(1) Preparation of total cellular RNA

Total cellular RNA was prepared using the methodology of

Chirgwin et al., (1979). Testis tissue was homegenized in 4M guanidium thiocyanate,0.5% w/v sodium N-laurosyl sarcosine, 25mM sodium citrate, 0.1% w/v Sigma antifoam A and 0.1M βMSH in a Sorvall omnimix. The resulting solution was then centrifuged at 8Krpm, 10°C for 10 minutes (Sorvall HB4). The resulting pellet was discarded and the supernatant remixed. The supernatent was then layered onto 1.2ml cushions of 5.7M CsCl, 25mM NaOAc pH5 and centrifuged at 36K rpm, 20°C for 12 hrs (Beckman SW50). The pelleted RNA was then recovered by first carefully removing most of the overlaying solution with a pipette draining off the remainder and redissolving the RNA in 7.5M guanidine hydrochloride, 25mM NaCitrate pH7.0, 5mM DTT. The RNA was then ethanol precipitated by the addition of 0.025 volumes of acetic acid and 0.5 volumes of ethanol.

(2) Preparation of cytoplasmic polysomal RNA

Total cytoplasmic polysomal RNA was prepared exactly as described by Jenkins et al., (1979).

(3) Fractionation of poly $(A)^+$ RNA

Poly $(A)^+$ RNA was fractionated by two cycles of oligo dT cellulose chromatography essentially as described by Aviv and Leder (1972). The loading buffer was 0.5M NaCl,20mM TrisHClpH7.0,1mM EDTA,0.1% w/v SDS. Elution was in the same buffer without NaCl. RNA was recovered by ethanol precipitation.

Nuclease SI and Exonuclease VII mapping

(1) Labelling of DNA

Total plasmid DNA was labelled by nick translation as described

After labelling for 60 minutes ATP was added to 1mM along with T4 DNA ligase. Incubation was then continued at 37°C for a further 60 minutes. DNA was then recovered and unincorporated nucleotides removed by phenol/chloroform extraction, sephadex G50 chromatography and ethanol precipitation. The labelled DNA was then restricted with suitable restriction enzymes and again recovered by phenol/chloroform extraction and ethanol precipitation.

(2) Hybridisation of DNA with mRNA

The DNA and mRNA were mixed and ethanol precipitated. The coprecipitated nucleic acid was then redissolved directly in hybridisation mix (40mM PIPES 7.7, 400mM NaCl, 1mM EDTA, 80% v/v formamide). The hybridisation mix was then taken up into a siliconised glass capillary which was then flame sealed. The nucleic acids were then denatured by incubating the capillary at 90°C for 5 minutes. Hybridisation was then at 52°C for 4 hours.

(3) Digestion with nuclease SI and exonuclease VII

Nuclease SI digestion was in 0.28M NaCl, 0.03M NaOAc pH4.5, 5mM $ZnCl_2$, 5% glycerol, 10µg/ml denatured salmon sperm DNA, 100 units/ml nuclease SI (Berk and Sharp, 1978; Favalora et al., 1980). The hybridisation mix was expelled directly into the digestion buffer containing the enzyme, at 37°C. Digestion was continued for 30 minutes. The reaction was stopped by addition of EDTA to 5mM. The reaction was then ethanol precipitated after the addition of 5µg carrier RNA.

Exonuclease VII digestion was in 30mM KCl, 10mM TrisHCl pH7.4, 10mM EDTA, 0.25 units/ml exonuclease VII (Berk and Sharp, 1978).

As with the SI digestion the hybridisation mix was expelled directly into prewarmed (45°C) digestion buffer. Digestion was continued for 60 minutes. The reaction was then made 200mM in NaOAc pH7.0 and ethanol precipitated after the addition of 2.5µg of carrier RNA.

(4) Gel electrophoresis

The products of the nuclease S1 and exonuclease VII digestion were dissolved in 15µl of 100% formamide, denatured by boiling for 3 minutes and cooled in ice/water. Electrophoresis was on 20cm x 15cm x 1.5mm gels of 8% w/v acrylamide (acrylamide/bis acrylamide ratio of 5.67:1) gels containing 98% v/v formamide in 20mM $NaPO_4$, pH7.5 (Maniatis et al., 1975) at 100v. Marker dyes were run in adjacent slots. The gels were preelectrophoresed for 30 minutes before loading and the buffer constantly recirculated. The gel was fixed in 3 changes (20 minutes each) of 40% methanol, 10% acetic acid and dried under vacuum at 70°C.

Labelling DNA by nick translation

Nick translation of DNA was carried out by a procedure modified from that of Rigby et al., (1977). DNA was first nicked using DNAase by giving approximately one nick per 0.5 to 1.0kb. The reaction was in 66mM Tris HCl pH7.5, 6mM Mg Cl_2. DNAase was diluted in 50mM Tris 7.4, 100µg/ml BSA and added to 20% v/v. The reaction was at 20°C for 2-15 minutes. DNA was recovered by phenol/chloroform extraction and ethanol precipitation.

Nicked DNA was then labelled using E. coli DNA polymerase (Klenow fragment) in 66mM Tris HCl pH7.5, 6mM $MgCl_2$, 2.5mM DTT, 30µM

cold dNTPS (dGTP, dATP, TTP) and 50μCi α^{32} P dCTP per 0.5μg of DNA. The reaction was at 20-30°C and the degree of incorporation followed by TCA precipitation of 1μl samples. The reaction was terminated by addition of EDTA to 10mM and deprotenised by phenol/chloroform extraction. Unincorporated nucleotides were removed by Sephadex G50 chromatography.

End labelling with *E.coli* DNA polymerase 1 (Klenow fragment)

E.coli DNA polymerase 1 (Klenow fragment) was used to label restriction enzyme 3' recessed ends. The reaction conditions were the same as those for nick translation except that the reaction was at 5-20°C and cold and α^{32}P dNTPs were chosen to give specific labelling.

End labelling DNA with T4 DNA polymerase

T4 DNA polymerase can be used in either to fill in restriction enzyme 3' recessed sticky ends or to replace the 3' adjacent nucleotide at the end of double stranded DNA. The reaction buffer was 66mM Tris HCl pH8.0, 30mM KCl, 6mM Mg Cl_2, 1mM DTT, 100μg/ml BSA. Cold dNTPS were added to 30μm when necessary and α^{32} P dNTPS were added to give an approximate two fold molar excess over sites to be labelled. The reaction temperature was at 20°C. The reaction was terminated by addition of four volumes of 2.5M NH_4 OAc and DNA recovered by ethanol precipitation.

End labelling using AMV reverse transcriptase

AMV reverse transcription was used to label 3' recessed restriction enzyme sticky ends. The reaction was in 50mM Tris HCl pH7.8, 10mM Mg Cl_2, 50mM NaCl 6mM β MSH. Cold dNTPS were added to 6μM as necessary and α^{32} P dNTPS added to give a two fold molar excess over sites

to be labelled.

DNA Sequencing

DNA sequencing was carried out as described by Maxam and Gilbert (1980). DNA was end labelled with either AMV reverse transcriptase or T4 DNA polymerase and cleaved using the G, G+A, T+C and C specific reactions. Sequencing gels were 40cm x 20cm x 0.35mm and were prepared and run as described by Sanger and Coulson (1978). 6%, 8% and 20% gels were used. For autoradiography the 6% and 8% gels were dried onto Whatman 3MM paper under vacuum and at 80°C. 20% gels were exposed wet. Exposure was if possible at room temperature to maximise resolution.

Scintillation counting

Dry samples (e.g. GFC filters) were counted in 5mls of PPO/POPOP toluene counting fluid in plastic minivials. Liquid samples were counted in NE260 (Nuclear Enterprise) using 9 volumes of NE260 to 1 volume of aqueous sample. All samples were counted in a Packard 3320 Tri-Carb scintillation counter.

Autoradiography

Autoradiography was carried out using Kodak X-Omat S film and Kodak X-Omatic regular intensifying screens. Exposure was normally at -70°C using preflashed film (Laskey and Mills, 1975). For sequencing gels, where resolution is of primary importance autoradiography was without intensifying procedures if sufficient counts were present.

Restriction digests

Restrictions using *Eco*RI, *Bam*HI, *Hind*III, *Pst*I and *Sal*I were in *Eco*RI buffer (10mM TrisHCl pH7.5, 10mM $MgCl_2$, 100mM NaCl, 10mM β MSH). All other enzymes were used in the buffers recommended by the suppliers. For gel electrophoresis digestion was termination by addition of ¼ volume of FDE (30% ficoll, 0.05% BPB, 100mM EDTA pH7.0). Restrictions containing bacteriophage lambda DNA were heated to 70°C for 5' and quenched on ice, after addition of FDE, to melt the lambda cohesive ends. Alternatively DNA could be recovered by phenol/chloroform extraction and ethanol precipitation. Trout testis DNA for southern transfers were always recovered in this way, and redissolved in 10mM TrisHCl pH8.0 before elctrophoresis.

Phenol/chloroform extraction of nucleic acid

Nucleic acids were deproteinized by phenol/chloroform extraction. Phenol was added to 50% v/v and mixed thoroughly for 5 to 15 minutes. A similar volume of chloroform was then added and mixed thoroughly. The aqueous and organic phases were then separated by low speed centrifugation. The aqueous layer was then removed and extracted twice and equal volumes of chloroform. The organic phases were routinely back extracted with aqueous buffer to minimise losses.

Ethanol precipitation of nucleic acid

DNA was precipitated by addition of NaCl or NaOAc pH7.0, to a final concentration of 0.3M and 2½ volumes of ethanol. The ethanol was mixed in thoroughly and the DNA precipitated at -70°C for 30 to 45 minutes or at -20°C overnight. The DNA was then pelleted by centrifugation (Sorvall HB4, 10K rpm, 30 minutes, Eppendorf minifuge,

full speed, 15 minutes) at 4°C.

RNA was precipitated in a similar manner by the addition of NaOAc pH5.0 and 3 volumes of ethanol.

TCA precipitation of DNA

DNA samples for scintillation counting were added to 1ml of ice cold 0.2M Na_4PPi, 150µg/ml BSA and mixed by vortexing. 0.3ml of ice cold 50% w/v TCA was then added and mixed by vortexing. After standing on ice for 15 minutes the precipitated DNA was collected on GFC filters by vacuum filtration. The filters were washed thoroughly with 5% TCA before being dried and counted in PPO/POPOP scintillation fluid.

Molecular weight markers

DNA molecular weight markers were lambda (C1857S7) cut with *Eco*RI, *Hind*III, *Bgl*II, *Sal*I, *Kpn*I or *Eco*RI + *Hind*III, pBR322 cut with *Alu*I, *Hpa*II and *Hae*III.

The sizes of fragments produced by these digests is shown in Table 2.

RNA size markers were ribosomal RNA. The sizes of 25S, 18S and 7S + RNA were taken as 5kb, 1.97kb and 121b respectively.

Unknown molecular weights were determined from a plot of $\log_{10}$ MW versus mobility drawn using the known molecular weights of the markers.

Table 2

(a) Lambda C1857S7 molecular weight markers

*Eco*RI	*Hind*III	*Bgl*II	*Bam*HI	*Eco*RI / *Hind*III
21.24kb	23.15kb	22.01kb	16.84kb	21.24kb
7.24	9.42	13.29	7.23	5.14
5.81	6.56	9.7	6.785	4.965
5.65	4.38	2.39	6.53	4.28
4.88	2.32	0.65	5.62	3.54
3.54	2.02	0.435	5.53	2.02
	0.56	0.06		1.91
	0.125			1.595
				1.37
				0.95
				0.845
				0.560
				0.125

*Kpn*I	*Sal*I
29.96kb	32.76kb
17.07	15.27
1.5	0.5

(b) pBR322 molecular weight markers

*Alu*I		*Hpa*II		*Hae*III	
910bp	46	622bp	110	587bp	89
659	19	527	90	540	80
655	15	404	76	504	64
521	11	309	67	458	57
403		242	34 x2	434	51
281		238	26 x2	267	21
2572		217	15	234	18
226		201	9 x2	213	11
100		190		192	7
90		180		184	
63		160 x2		124	
57		147 x2		123	
49		122		104	

Recrystallization of formamide

Formamide was recrystallised by stirring slowly at 0°C for 3hrs and then standing at 0°C to allow recrystallization. The liquid remaining after recrystallization was discarded and the formamide remelted. This process was then repeated a further two times. Formamide was deionized immediately before use by stirring with Bio-Rad AG501 mixed bed resin.

Recrystallization of urea

A saturated solution of urea was prepared at room temperature and filtered through nitrocellulose (0.45μM). The urea was then recrystallized by standing the solution at 4°C overnight. The urea was then collected on a glass scintre under vaccuum and dried.

Recrystallization of acrylamide

Acrylamide was recrystallized by dissolving at 70g per litre in chloroform at 50°C, filtering through Whatman number 1 paper (while still hot) and standing the filtrate at -20°C overnight. The recrystallized acrylamide was then collected on a glass scintre under vacuum and dried.

Recrystallization of bis-acrylamide

Bis-acrylamide was dissolved at 10g per litre in acetone at 45 to 50°C and then filtered and recrystallized as described for recrystallisation of acrylamide.

Media and solutions not specified in text

L Broth (LB), 1% Difco Bacto tryptone, 0.5% Difco Bacto yeast extract, 1% NaCl.

L Agar (LA), LB plus 1.2% Difco agar.

L Agar top, LB plus 0.7% Difco agar.

BBL agar top, 1% Baltimore Biological Laboratories trypticase, 0.5% NaCl, 0.7% Difco agar.

Phage buffer, 0.3% KH_2PO_4, 0.7% Na_2HPO_4 (anhydrous), 0.5% NaCl, 10mM $MgSO_4$, 1mM $CaCl_2$, 0.001% gelatin.

X gal indicator plates. BBL agar plates containing 40µg/ml X Gal. The X gal is first dissolved in dimethyl formamide at 2mg/ml and then added to molten agar cooled to 50°C immediately before the plates are poured.

Antibiotics, antibiotics were added to agar and broth from stock solutions. Agar was first cooled before antibiotics were added. The concentrations of stock solutions and the concentrations antibiotic were used at are as follows:-

Tetracyline, stock 10µg/ml in methanol, used at 10µg/ml.
Ampicillin, stock 100mg/ml in 0.35 N NaOH, used at 100µg/ml.
Chloroamphenicol, stock 20mg/ml in methanol used at 150mg/ml.

The concentration of antibiotics in top agar was generally half that used in agar or broth.

Results

Southern transfer mapping of trout testis DNA

To determine the number and size of restriction fragments in the trout genome that contain protamine gene sequences Southern transfers of total genomic DNA were made and hybridised with nick translated protamine cDNA plasmid clones.

DNA prepared from a single testis was restricted with *Eco*RI, *Bam*HI, *Hind*III, *Kpn*I, *Sst*I and *Pst*I. Each digest contained 10µg of trout testis DNA and a two fold excess of enzyme. The *Bam*HI digest was electrophoresed on a 0.8% agarose gel, the other digests on a 1.0% agarose gel. Lambda DNA restricted with *Bgl*II and *Eco*RI plus *Hind*III was used as molecular weight markers. Electrophoresis was continued until the BPB dye marker was two thirds down the gel. The gel was then stained, photographed and transferred. The transfer was then hybridised with an equal mixture of nick translated pTP4,8 and 11 (Jenkins, 1979). The hybridisation mixture contained 5×10^6 cpm/ml of this probe. Hybridisation was for 16 hours after which the transfer was washed to a final stringency of 0.1 x SET, 68°C. Hybridising fragments were then identified by autoradiography (Fig. 6). The molecular weights of the hybridising fragments were estimated by comparison with the lambda marker digests and are shown in Table 3.

In all the digests two strongly hybridising bands are seen (asterisked both in fig.6 and table 3). In addition additional bands with varying signal strengths are seen. The *Pst*I digest shows a single moderately hybridising band. The *Bam*Hl and *Sst*I digests show two such bands. All the digests also contain several weekly hybridising bands. The *Eco*RI and *Hind*III digests both contain four or five such bands (the possible fifth band being in an area of high background).

Fig.6 Southern transfers of genomic trout DNA.

(a) 1.0% agarose gel transfer.

S *Sst*I

P *Pst*I

K *Kpn*I

H *Hind*III

R *Eco*RI

M molecular weight scale

* strongly hybridising fragments

(b) 0.8% agarose gel transfer.

B *Bam*HI

M molecular weight scale

* strongly hybridising fragments

Each digest contained 10µg trout DNA.Molecular weight markers were 0.1µg of λC1857S7 restricted with *Bgl*II or *Eco*RI plus *Hind*III and coelectrophoresed with 10µg of *Eco*RI restricted trout DNA.λ fragments were identified by transfer and hybridisation with nick translated λC1857S7 DNA.

Fig.6

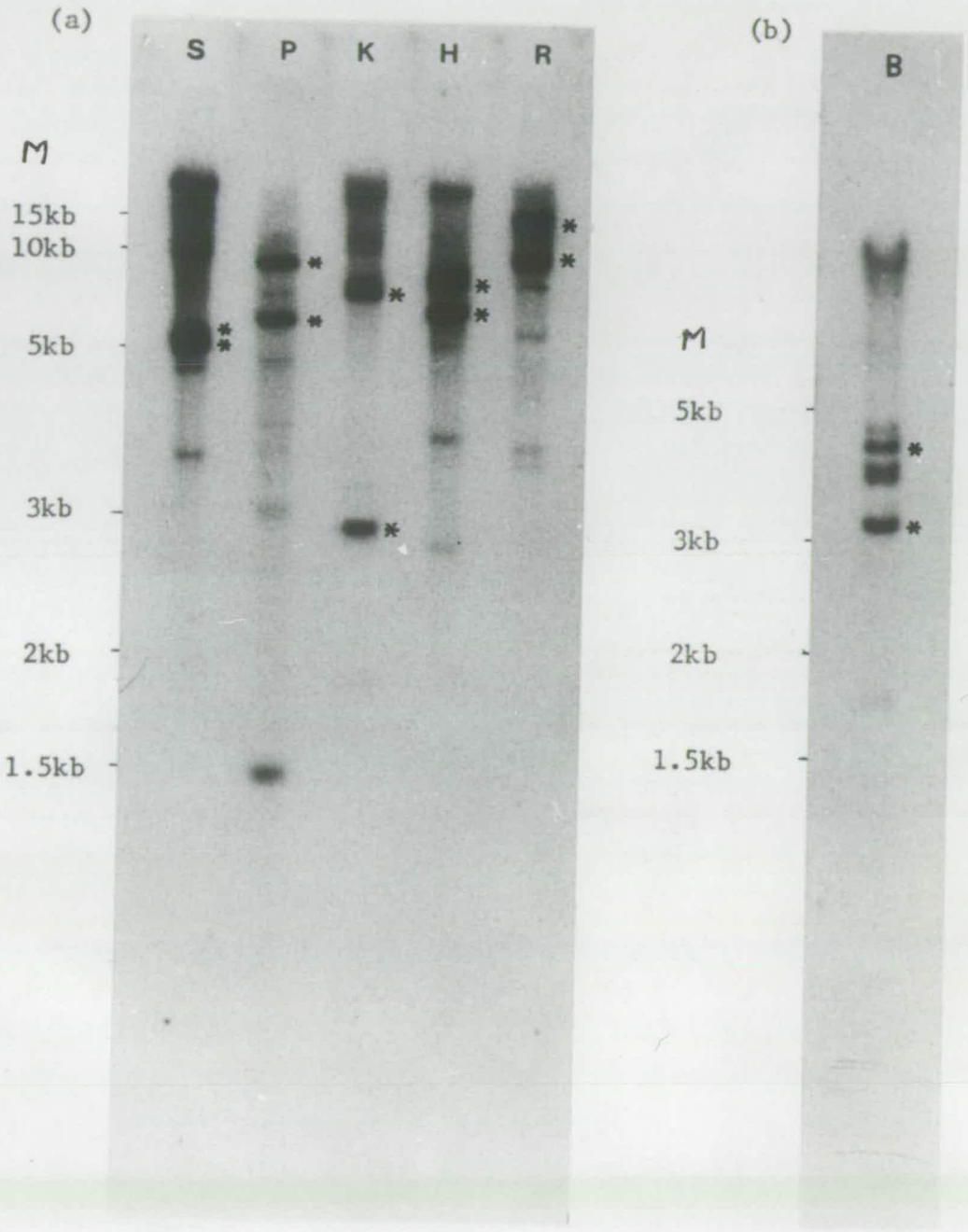


Table 3. Size of genomic fragments hybridizing to protamine cDNA.

EcoRI	*HindIII*	*KpnI*
16.6kb*	(8.4kb)	10.2kb
(10.8)	7.8*	9.5
9.1*	5.8*	7.6
7.8	5.3	7.3*
5.1	5.0	2.7*
3.6	3.6	
3.4	2.6	

PstI	*SstI*	*BamHI*
9.2kb*	∿20kb+	4.7kb
7.7	10.2+	4.3*
5.8*	5.3*	3.9+
4.8	5.0*	3.8+
3.0	4.5	3.2*
1.5+	3.5	1.8

* Bands of strong intensity

\+ Bands of intermediate intensity

( ) Putative bands either obscured by high background and/or adjacent high intensity bands

The *Pst*I digest contains four weakly hybridising bands. The *Kpn*I digest contains at least three but because of the high background in the high molecular weight region of the gel where these are found it is possible that additional bands are not being resolved. In addition the *Kpn*I digest appears to contain two very weakly hybridising bands. The *Bam*HI and *Sst*I digests both contain two weakly hybridising bands. At a lower stringency of post hybridisation washing (1 x SET, 68°C) the hybridisation signal from the weaker bands increases in proportion to the strongly hybridising bands. Some of the difference in signal would therefore appear to be due to the degree of homology between the cDNA probe and the genomic fragments hybridising (1 x SET, 68°C, allows hybridisation of sequences with approximately 22% mismatch). However, even at this low stringency the difference in signal is approximately four fold. In addition hybridisation with either type of cDNA clone (pTP4 or pTP8, Jenkins, 1979) separately showed the same pattern of hybridisation even after a high stringency post hybridisation wash (0.1 x SET, 68°C).

The differences in hybridisation signal of the various bands could be explained in a number of ways. Firstly they could be repeated or contain a number of protamine genes. Cloning of the largest (and strongly hybridising) *Eco*RI fragment shows that it contains only a single protamine gene sequence. Alternatively the weakly hybridizing bands could represent pseudogenes or short cross hybridising sequences. A third possibility is that the strongly hybridising bands represent protamine genes with contiguous coding sequences while the weakly hybridising bands represent intron containing genes. Again cloning and analysis of the longest *Eco*RI fragment shows that the gene contained in this fragment does not contain intervening sequences. It is not known whether the weakly hybridising bands contain intron containing protamine

genes. The presence of three introns in the human γ interferon gene (Taya et al, 1982), in contrast with the intronless α and β interferon genes, demonstrates that closely related genes may have very different gene structures. All three introns in the γ interferon gene split the amino acid coding part of the gene, the introns being found between amino acid residues 18/19, 41/42 and 102/103 (the coding frame is 166 residues in length). As the coding sequence in the protamine genes is extremely short, even the presence of a single intron in the coding sequence could be expected to diminish the signal obtained in a Southern transfer. This would be true even if the non-contiguous parts of the gene were contained in the same restriction fragment.

Another possible explanation is that there are two families of repeated protamine genes, these corresponding to the two major bands in Southern transfers. The minor bands could then represent (unique) polymorphic variants. If this explanation is true the extent of the homology must be much greater than the size of the gene itself (subsequent clonal analysis is showing the protamine gene to be less than 0.5kb in length). Unfortunately no copy number estimate was included in this experiment (such as genome equivalents of CH4A/TP3A) making an estimate of the copy number of the fragment represented in the strongly hybridising bands impossible.

The genomic Southern transfers are therefore not straightforward in that they do not give a clear picture of the number or repetition frequency of the protamine genes. Cloning of all hybridising fragments followed by copy number determination would be required to determine the relationships of the various hybridising fragments seen in Southern transfers.

Construction of the trout genomic library in Charon 4A

Charon 4A *Eco*RI arms and trout genomic 12-20kb *Eco*RI* fragments were made and purified as detailed in materials and methods. The trout genomic DNA used was from the same preparation as that used

for the genomic Southern transfers. After methylation with *Eco*RI methylase the DNA was digested with 2,4,6,8 and 12 units/μg of *Eco*RI under *Eco*RI* conditions. Digestion was for 8hrs at 37°C. The fragments produce were then sized as described in materials and methods. Representative sucrose gradient profiles and sample gels of charon 4A/*Eco*RI and trout/*Eco*RI* digests are shown in Figs. 7 and 8 respectively. The purified Charon 4A arms and trout *Eco*RI* fragments were combined in two ligations. The first ligation contained the two DNAs at the concentration specified by Maniatis et al., (1978). The second contained the two DNAs at the concentrations specified by Kemp et al., (1979). The ligations contained 11μg and 10μg of arms and 4μg and 8μg of genomic *Eco*RI* fragments respectively. After packaging the two ligations produced 2.55×10^5 and 0.95×10^4 pfu respectively. The first ligation was therefore much more efficient producing 6.4×10^4 pfu/μg of eukaryotic DNA. The second produced only 0.95×10^4 pfu/μg of eukaryotic DNA. Maniatis et al., (1978) report efficiencies ranging from 3.8×10^4 to 5.6×10^5 pfu/μg using Charon 4A to clone eukaryotic DNA via *Eco*RI linker molecules. Kemp et al., (1979), using the *Eco*RI* methodology reported an efficiency of 1.3×10^5 pfu/μg of eukaryotic DNA. Both ligations were therefore somewhat less efficient than might have been expected.

The total number of recombinant phage obtained was 3.5×10^5. This is approximately 1.75 genome equivalents (assuming a size of $3.\times 10^6$ kb for the trout genome and an average insert size of 15kb). Assuming a random distribution of sequences within the library this gives an 82% probability of the library containing any single copy sequence (Clarke and Carbon, 1976).

The recombinant phage were divided into seven pools (5×10^4 pfu

Fig.7 Representative Charon 4A/*Eco*RI sucrose gradient.

(a) OD_{260} profile.
Gradients were pumped out from bottom to top via an ISCO ultraviolet flow analyser.Fractions were approxiamately 0.5ml.

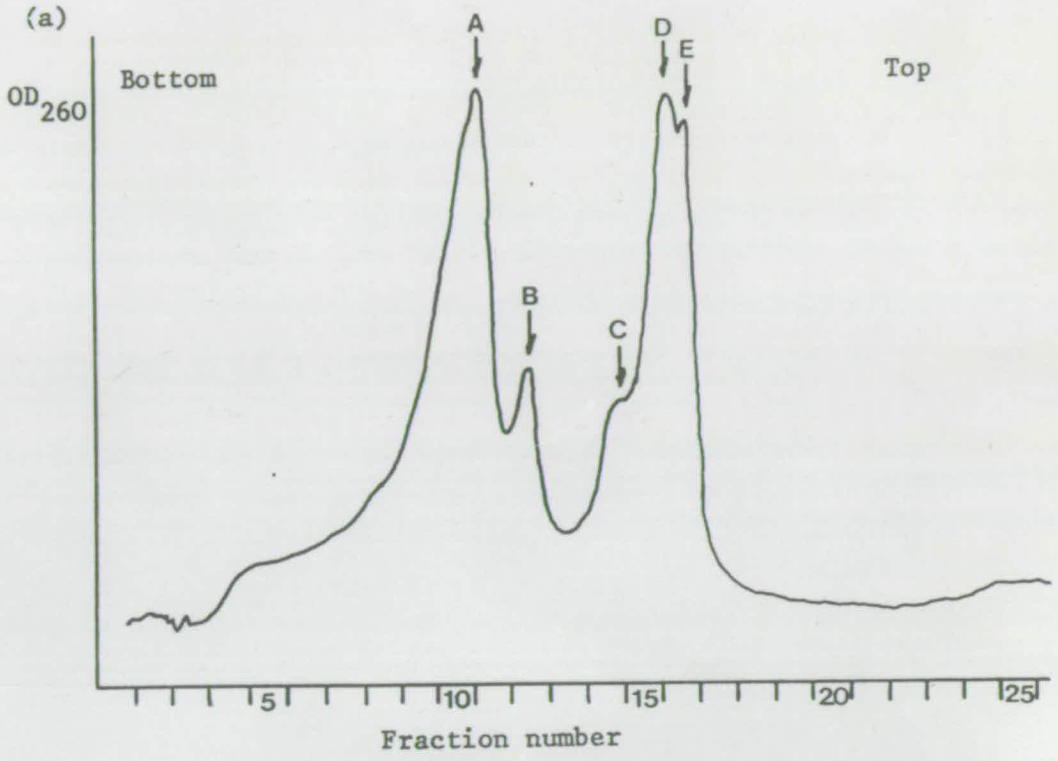
A Annealed Charon 4A *Eco*RI arms (30.7kb).
B Charon 4A left arm (19.8kb).
C Charon 4A right arm (10.89kb).
D Charon 4A 7.8kb internal fragment.
E Charon 4A 6.9kb internal fragment.

(b) 0.4% agarose gel of fraction samples.Tracks, from left to right,are,

1. Fraction number 7
2. " " 8
3. " " 9
4. " " 10
5. " " 11
6. " " 12
7. " " 13
8. " " 14
9. " " 15
10. " " 16

Samples were 15 to 25µl in size.
Fractions containing arm pieces and no visible internal fragments were pooled and DNA recovered by ethanol precipitation.Fractions 9 to 13 inclusive were precipitated from this gradient.

Fig.7



(b)

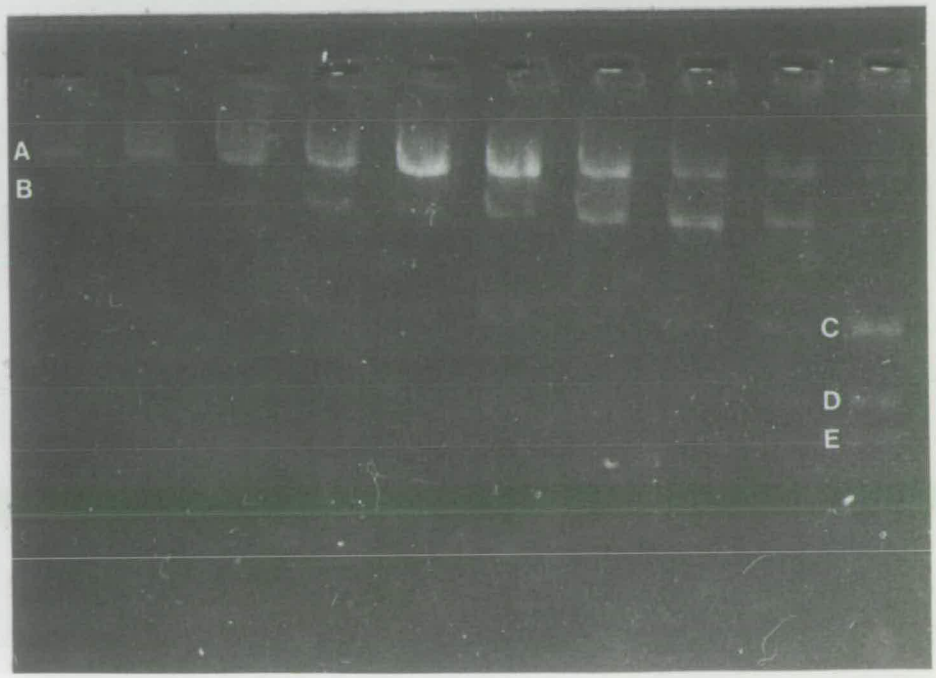


Fig.8 Representative trout genomic DNA/*Eco*RI* sucrose gradient.

(a) OD_{260} profile.

Gradients were pumped out from bottom to top via an ISCO ultraviolet flow analyser.Fractions were of about 0.5ml.

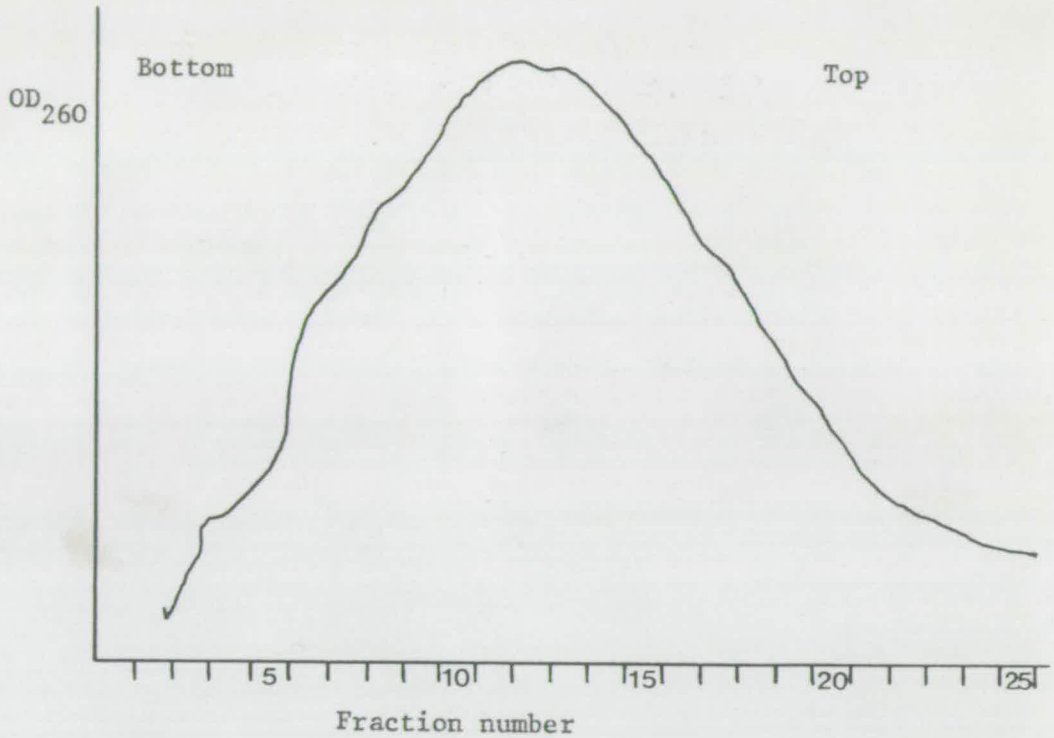
(b) 0.4% agarose gel of fraction samples.Tracks,from left to right,are,

1. Fraction number 9
2. " " 10
3. " " 11
4. " " 12
5. " " 13
6. " " 14
7 " " 15
8. " " 16
9. " " 17
10. λC1857S7/*Bgl*II

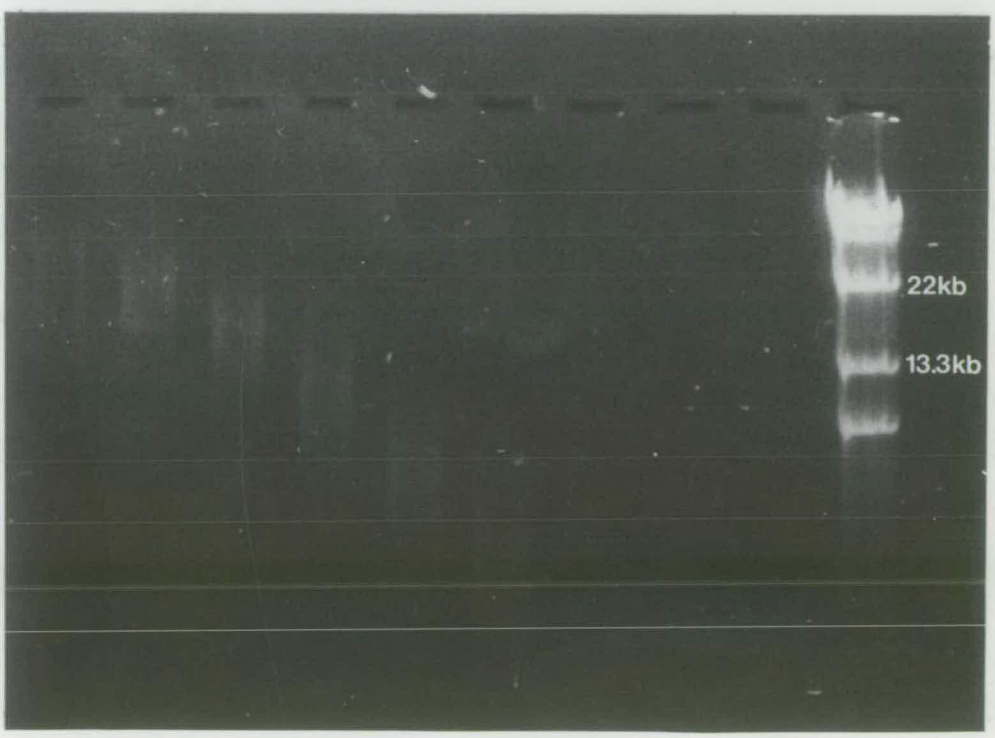
Samples were 15 to 25µl per fraction.
Fractions containing primarily fragments between 14 and 20kb were pooled and DNA recovered by ethanol precipitation(ie. fractions 10 and 11).
The top two bands in the λ/*Bgl*II digest represent incomplete digestion products.

Fig.8

(a)



(b)



each) and amplified. The degree of amplification was about 10^4. After amplification a representative assay from two pools, each representing one of the two ligations, was carried out on X gal indicator plates. This revealed a background of 4% non recombinants.

Screening the trout genomic library for protamine gene sequences

The genomic library was screened by plating 35,000 pfu from each pool at a density of 60pfu/cm^2 (i.e. one 20 x 20cm plate per pool). Replica filters were prepared and hybridised with an equal mixture of the protamine cDNA clones pTP4, 8 and 11 (Jenkins, 1979). Two strongly hybridising and two weakly hybridising positives were identified and purified by two further low density plaque screens. The weakly hybridising phage were later shown by hybridisation to testis mRNA, not to contain protamine gene sequences. Neither of these phage corresponded to the weakly hybridising bands seen in genomic Southern transfer experiments. Addition of polyI (10µg/ml) to prehybridisation and hybridisation mixes abolished hybridisation between the nick translated protamine cDNA plasmids and these phage. This suggests that the hybridisation originally observed was due to non specific binding between the dG/dC tails used to insert the cDNA sequences into pBR322 (Jenkins et al., 1979) and dG/dC rich sequences in these phage. Subsequent to this discovery polyI (10µg/ml) was included in all hybridisations containing the protamine cDNA clones. A second library screen, in which polyI was included during the prehybridisation and hybridisation, revealed only the same two strongly hybridising phage found in the first screen. It was concluded that these were the only two phage that contained protamine gene sequences in the library. The two phage were found in different pools, this signifies they represent independent cloning events. Both were grown up and DNA prepared

They were designated CH4A/TP3A and CH4A/TP4A as they originated from pools 3 and 4 of the library respectively.

The finding of only two protamine gene clones in the library was surprising. If the library was completely random it should contain 10-11 different protamine clones (1.75 genome equivalents of library x six genes) representing 4 or 5 different protamine genes (6 genes x 80% probability of finding any single copy sequence). The presence of only two protamine gene clones therefore suggests that the library is not random. This means some sort of selection for certain sequences has occurred during the construction of the library. Selection is most likely to occur during the making of size fractionated DNA (by non-random digestion) or during amplification of the library (by unequal growth of recombinant phage).

Subsequent restriction analysis of the phage show that the inserts in both are defined by *Eco*RI sites at the junction with the Charon 4A arms. If cleavage of the genomic DNA occurred only at *Eco*RI* sites and not at any full *Eco*RI sites only one in four of the Charon 4A/insert junction sites should be full *Eco*RI sites. This suggests that methylation of the genomic DNA was not complete allowing selection to occur by the preferential cleavage of unmethylated *Eco*RI sites during *Eco*RI* digestion. The fact that the size of the insert in one of the clones, CH4A/TP3A, is approximately the same size as a genomic *Eco*RI fragment reinforces this view. The sizes are 15.41kb and 16.6 kb respectively. The difference in the size is well within the accuracy of measurement for fragments of this size, especially as the genomic blot was from a 1% agarose gel. Another possibility is that the *Eco*RI* digestion is not as random as once thought. Polisky et al., (1975) suggest that the *Eco*RI* recognition site is 5' NAATTN 3' and

such sequences are cleaved to give 5' tetranucleotide extensions (AATT). Nearest neighbour analysis showed that the 3' N is most likely to be dC (59.2%) or T (26.1%) and less likely to be dA (14.2%) and very rarely dG (0.55%). This suggests an effective specificity of more than 4bp, however the digestion should still be random enough to produce a good library if a wide range of digestion conditions are used. However a second analysis of the *Eco*RI* activity (Woodbury et al., 1980a), raises doubts about the premise of the 5'AATT3' recognition specificity. This analysis suggests that after full *Eco*RI sites the preferred cleavage site is 5'GGATTT3'. This sequence would not give a AATT 5' extension and therefore would be unlikely to clone with a high efficiency in an *Eco*RI site (i.e. in Charon 4A). The T_M of *Eco*RI sticky ends in ligation buffer is about 5-6^{o}C (Dugaiczyke et al., 1975). A one base pair mismatch would lower this considerably. Although one half of the ligation site would be normally paired the GA mismatch at the other half would presumably hinder the ligation process. It is possible that the site would be half ligated and repaired upon introduction into *E. coli*. The GGATTT sequence would occur, on average, every 4,629 bp in 60% TA DNA. Cleavage of this sequence could therefore rapidly make a large proportion of DNA either unclonable or only clonable at low efficiency. The same analysis of the *Eco*RI* recognition site suggests that the hierarchy of cleavage after the 5'GGATTT3' sequence is 5'AAATTT3', 5'GAATT$^{A}_{T}$3' and 5'NAATTN'3'. If N is dC or N' dG the last sequence is cleaved only very slowly. With these data in mind it is clear that *Eco*RI* digestion is not an ideal way to produce a "random" digest of DNA for cloning, even after protection of full *Eco*RI sites by *Eco*RI methylase.

A possible way to improve the *Eco*RI* cloning methodology would be to use partial *Eco*RI* methylation to protect the most rapidly cleaved *Eco*RI* sites

Under low ionic strength conditions the *Eco*RI methylase also shows a reduced specificity and protects DNA from *Eco*RI* digestion (Woodbury et al., 1980b). Assuming the hierarchy of sequence recognition is similar for both the *Eco*RI endonuclease and methylase under *Eco*RI* conditions a combination of partial *Eco*RI* methylation and digestion should produce a much more random collection of clonable fragments. However with the recent development of *Bam*HI lambda vectors such as Charon 28 and 30 (Rimm et al., 1980) and λ1059 (Karn et al., 1980). the easiest way to produce random digests of DNA for preparing genomic libraries is by partial digestion with *Sau*3A. This enzyme recognises the sequence 5'GATC3' to produce the same tetranucleotide extension as *Bam*HI allowing easy cloning of such fragments into *Bam*HI sites. *Sau*3A cleavage appears to be unaffected by the nucleotides adjacent to the recognition sequence. Another advantage is that as the recognition sequence contains all four bases in equal proportion the frequency of occurrence of the site is affected only slightly by variation in the GC content of the DNA.

Analysis of recombinant phage by restriction mapping and Southern transfer

The recombinant phage CH4A/TP3A and 4A were mapped with the restriction enzymes *Eco*RI, *Bam*HI and *Hind*III. Initially 1.0μg of phage DNA was restricted using restrictions enzymes singly and in combination. The DNA was then electrophoresed and sized against lambda molecular weight markers. Generally 0.8μg of the digest was electrophoresed on a 1.0% agarose gel and the remaining 0.2μg on a 0.4% agarose gel. Southern transfer and hybridisation with nick translated cDNA plasmids was used to identify fragments containing protamine gene sequences. Fig. 9 shows the pattern of bands produced by restriction of

Fig.9 Restriction digest analysis of CH4A/TP3A and CH4A/TP4A

(a) 1.0% agarose gel.Tracks,from left to right are,

1. CH4A/TP3A *Eco*RI
2. " *Bam*HI
3. " *Hind*III
4. " *Eco*RI/*Bam*HI
5. " *Eco*RI/*Hind*III
6. " *Bam*HI/*Hind*III
7. λC1857S7 *Eco*RI/*Hind*III
8. λC1857S7 *Bgl*II

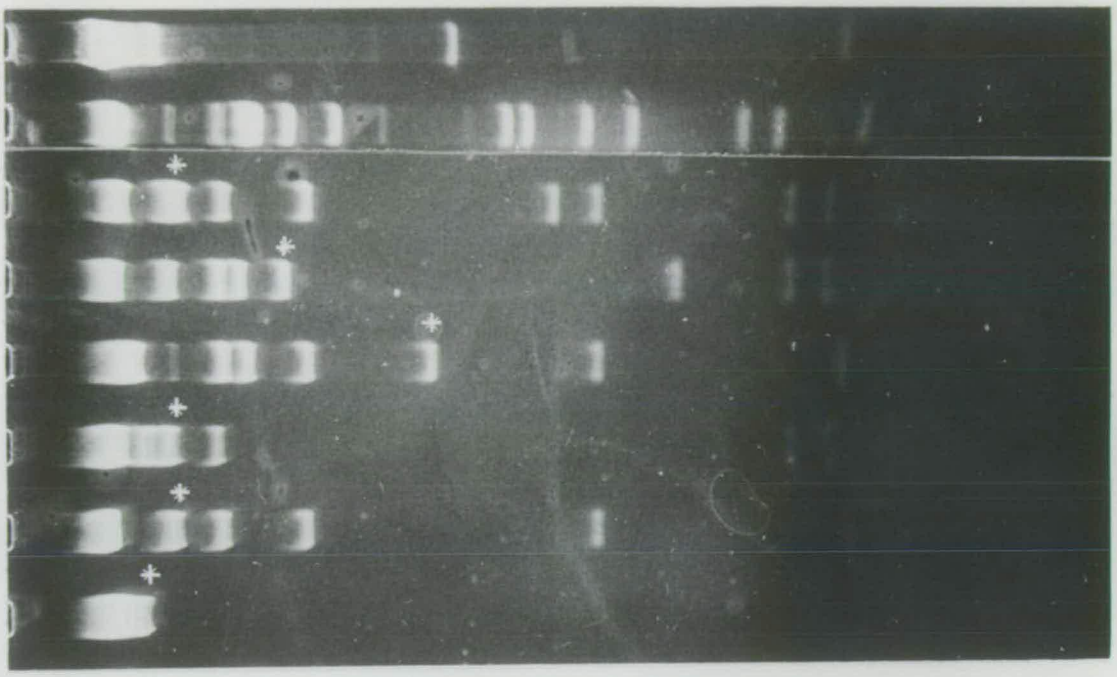
(b) 1.0% agarose gel.Tracks from left to right are,

1. CH4A/TP4A *Eco*RI
2. " *Bam*HI
3. " *Hind*III
4. " *Eco*RI/*Bam*HI
5. " *Eco*RI/*Hind*III
6. " *Bam*HI/*Hind*III
7. pCM2 *Bam*HI/*Eco*RI

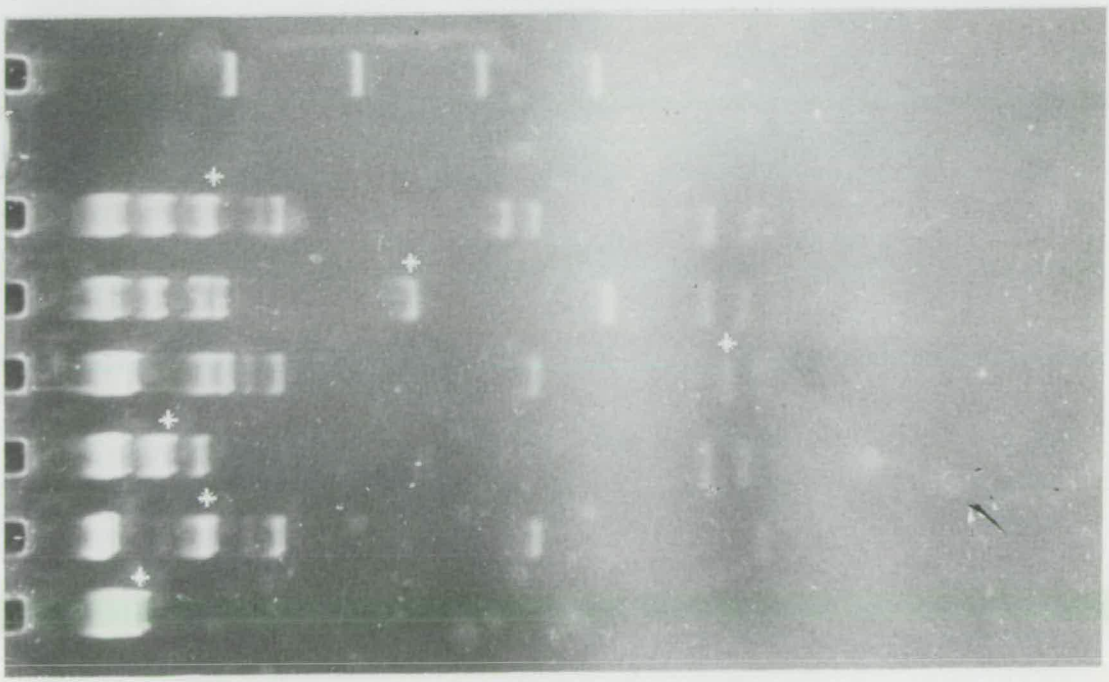
pCM2 *Bam*HI/*Eco*RI fragment sizes are 4.8kb,2.5kb,1.6kb and 1.1kb.

* Asterisk denotes fragments hybridising to protamine cDNA clones.

Fig.9



(b)



(a)

CH4A/TP3A and 4A with *Eco*RI, *Bam*HI and *Hind*III after electrophoresis on a 1.0% agarose gel. Fragments containing protamine sequences are asterisked.

Restriction mapping was extended to include the enzymes *Xba*I, *Sst*I and *Kpn*I using the same methods. Both CH4A/TP3A and 4A were seen to contain two small *Hind*III fragments. Further experiments show that the larger of these contains a *Kpn*I site. To map these fragments they were first sized on a 5% acrylamide gel (Fig. 10). This analyses shows that both phage actually contain four small *Hind*III fragments. To order these fragments both phage were restricted to completion with *Eco*RI (cutting out the phage insert which has no internal *Eco*RI sites) and then partially digested with *Hind*III. Half of each digestion was then electrophoresed on a 0.5% agarose gel and the second half on a 0.9% agarose gel. Both gels were transferred onto nitrocellulose. The transfer of the 0.5% gel was then hybridised with the protamine cDNA probe to identify *Hind*III partials extending from the end of the insert next to the right arm of Charon 4A. The 0.9% gel transfer was hybridised with a subclone (from CH4A/TP 3A) of the 1.2kb *Eco*RI/*Hind*III fragment from the opposite end of the insert (fragment a, Figs. 12 and 13) to identify *Hind*III partials extending from this end of the insert. The resulting autoradiograms and *Hind*III restriction maps are shown in Fig. 11.

The restriction maps of CH4A/TP3A and 4A (Figs. 12 and 13) show that the two clones are very similar. Extending from the *Eco*RI site at the junction with the left arm of Charon 4A to the *Eco*RI site in CH4A/TP4A at the junction with the right arm of Charon 4A both phage are identical except for the presence of a second *Xba*I site in CH4A/TP3A. In addition CH4A/TP3A extends for an extra 2.1kb. The size of

Fig.10 *Hind*III restriction fragments from CH4A/TP3A and CH4A/TP4A.

Digests contained 1.5µg of phage DNA and were sized on a 5.0% polyacrylamide/TBE gel. Tracks, from left to right, are,

1. CH4A/TP3A *Hind*III/*Kpn*I
2. CH4A/TP4A *Hind*III/*Kpn*I
3. pBR322 *Alu*I
4. CH4A/TP3A *Hind*III
5. CH4A/TP4A *Hind*III

Fragment sizes are,

Track 1. 820bp, 617bp, 302bp and 263bp.

Track 2. " " " " "

Track 4. 661bp, 617bp, 302bp, 263bp and 155bp.

Track 5. " " " " " "

Fig.10

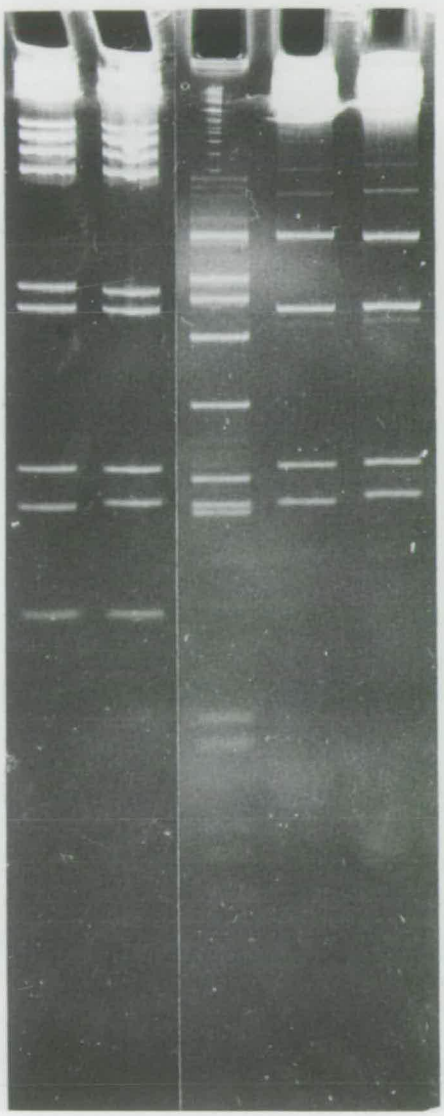


Fig.11 Mapping of *Hind*III restriction sites in CH4A/TP3A and CH4A/TP4A.

Phage were restricted to completion with *Eco*RI and then partially with *Hind*III.

(a) Southern transfer of phage electrophoresed on a 0.9% agarose gel and hybridised with a subclone of *Eco*RI/*Hind*III fragment a (fig.12) from CH4A/TP3A.
Track 1. CH4A/TP4A
Track 2. CH4A/TP3A
The sizes of the two smallest fragments in both tracks are 1.1kb and 1.6kb.
Molecular weight markers were λC1857S7 digested with *Bgl*II,*Kpn*I and *Eco*RI/*Hind*III.

(b) Southern transfer of phage electrophoresed on a 0.6% agarose gel and hybridised with protamine cDNA clones.
Track 3. CH4A/TP3A
Track 4. CH4A/TP4A
The sizes of the four smallest fragments in each digest are,
Track 3. 4.2kb,5.0kb,5.3kb and 6.0kb
Track 4. 2.1kb,2.95kb,3.2kb and 3.9kb
Molecular weight markers were λC1857S7 digested with *Bam*HI,*Eco*RI,*Bgl*II and *Eco*RI/*Hind*III.

(c) The resulting *Hind*III restriction map from the measured *Hind*III fragment sizes(fig.10) and the above results.
RA Charon4A right arm.
LA Charon4A left arm.
H *Hind*III
R *Eco*RI

Fig.11

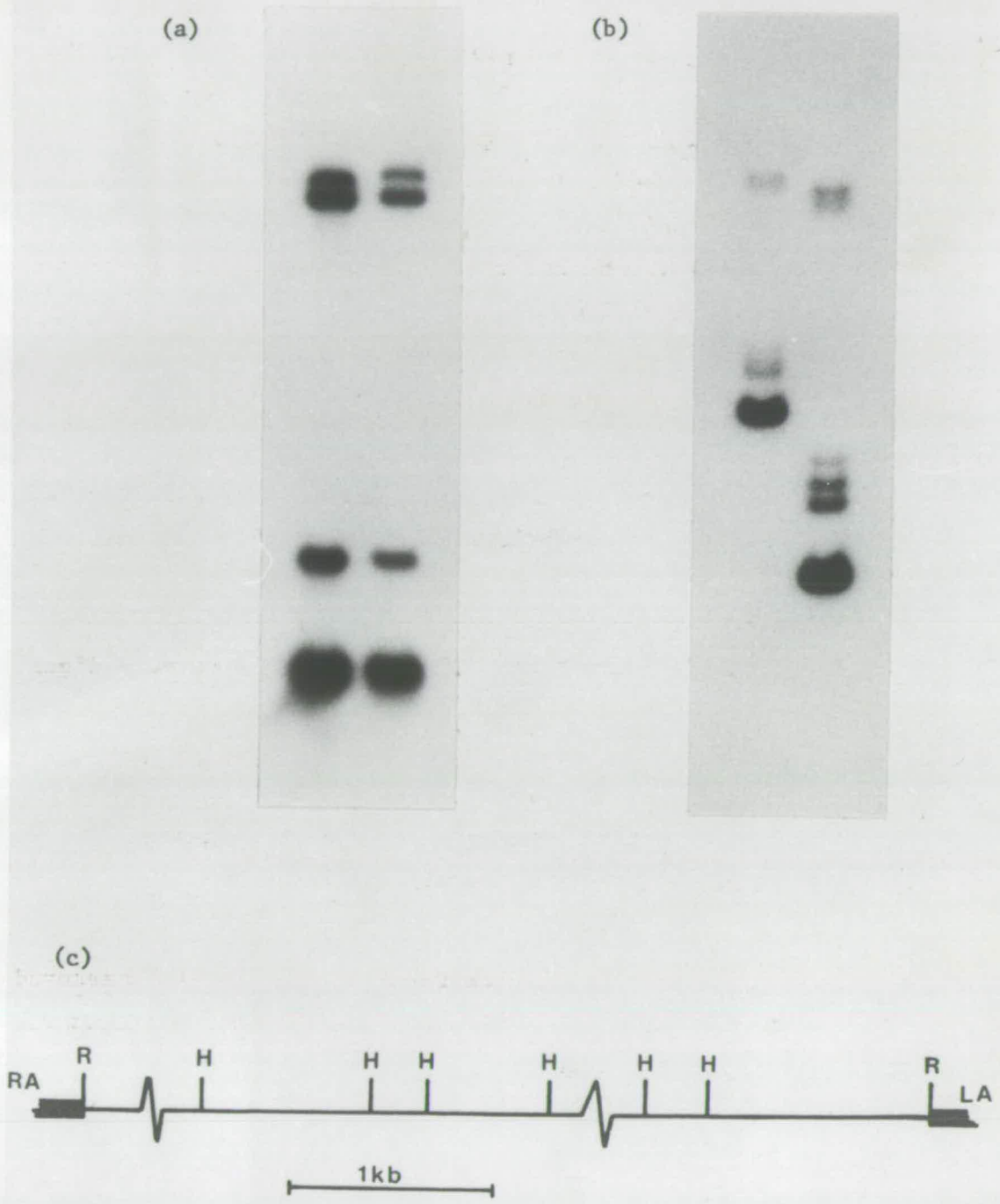


Fig.12 Restriction map CH4A/TP3A and pTP3A

B *Bam*HI
Bg *Bgl*II
H *Hind*III
Hp *Hpa*I
K *Kpn*I
Ps *Pst*I
Pv *Pvu*II
R *Eco*RI
X *Xba*I

——— trout genomic sequences

vector DNA(charon 4A or pAT153)

trout genomic fragments hybridizing to protamine cDNA

pTP3A contains no sites for *Sst*II,*Xho*I,*Ava*I,*Sma*I.

Fig.12 Restriction map CH4A/TP3A and pTP3A

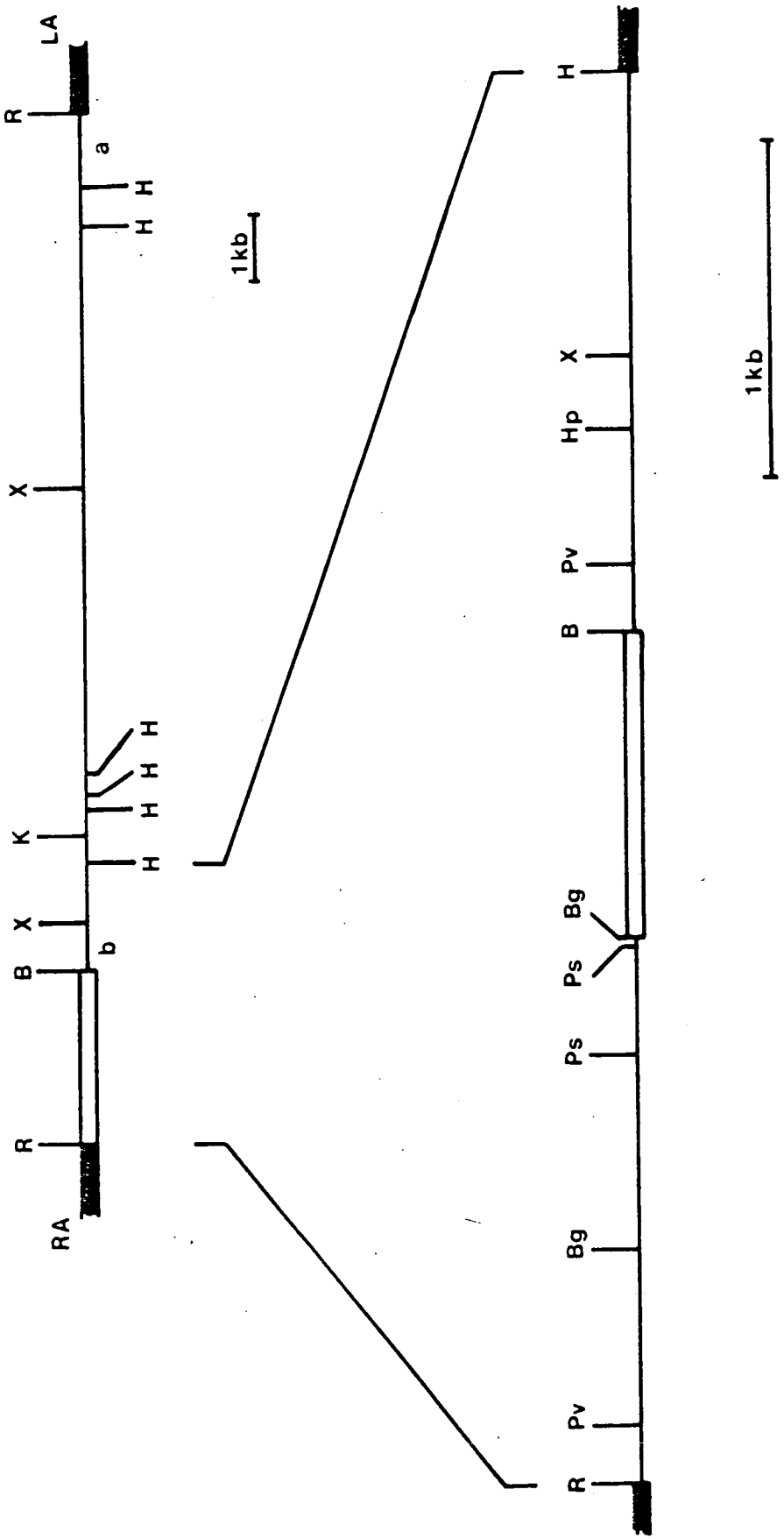


Fig.13 Restriction map CH4A/TP4A and pTP4A

A *Ava*I
B *Bam*HI
Bg *Bgl*II
H *Hind*III
Hp *Hpa*I
K *Kpn*I
Ps *Pst*I
Pv *Pvu*II
R *Eco*RI
X *Xba*I

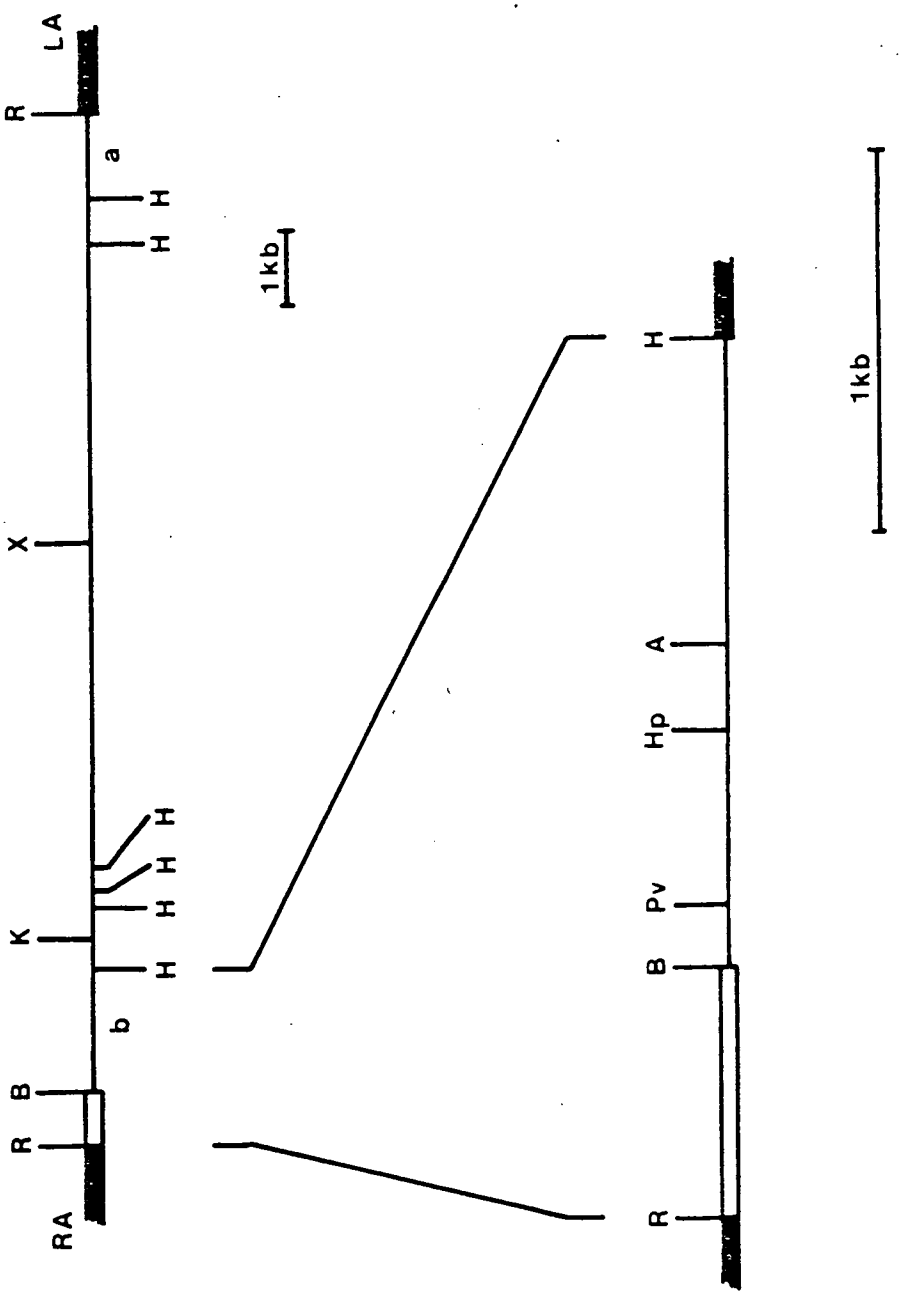
——— trout genomic sequences

▬▬ vector DNA(charon 4A or pAT153)

▭ trout genomic fragments hybridizing to protamine cDNA

pTP4A contains no sites for *Sst*II,*Xho*I,*Sma*I.

Fig.13 Restriction map CH4A/TP4A and pTP4A



the inserts (13.5 and 15.6 kb) means that they could only represent one of the *Eco*RI genomic fragments that hybridise in Southern transfer experiments, this being the 16.6 kb fragment. CH4A/TP3A probably contains the whole of this fragment (the difference being due to measuring innacuracies) and CH4A/TP4A a *Eco*RI/*Eco*RI* piece of this fragment as it is too long (13.5kb) to represent the second longest genomic *Eco*RI fragment (9.1 kb) even after allowing for measuring inaccuracies. Subsequent sequence analysis has shown that CH4A/TP3A contains an *Eco*RI* site in a homologous position to the *Eco*RI site at the end of the insert in CH4A/TP4A.

The mapping of the *Hind*III sites at the left-end of both phage inserts by hybridisation with an *Eco*RI/*Hind*III subclone from CH4A/TP3A showed that this fragment hybridised equally well with both phage. This proves that the phage retain sequence homology as well as restriction site homology throughout their extire length. That they are not clones of the same piece of DNA is shown by the extra *Xba*I site in CH4A/TP3A. As the library was made with DNA from a single testis these differences do not represent variation between individual fish.

To determine whether the difference in restriction pattern (a single *Xba*I site) is statistically significant the data was analysed according to the methodology of Nei and Li (1979) and Brown et al., (1979).

First the proportion of unchanged ancestral sites, was estimated using the equation:

(1) $$\hat{S} = \frac{\Sigma n_{xy}}{\Sigma n_x + \Sigma n_y - \Sigma n_{xy}}$$ (Brown et al., equation 1)

where $\hat{S}$ = estimated s

n_{xy} = number of restriction sites shared by both sequences

n_x = number of restriction sites in sequence X

n_y = number of restriction sites in sequence Y

This estimate is used when summing data for different enzymes (with the same number of bp in their recognition sequence).

Sequence X = CH4A/TP3A

Y = CH4A/TP4A

$$\hat{S} = \frac{9}{10 + 9 - 9} = 0.9$$

The mean number of nucleotide substitutions per nucleotide site, δ, was then estimated using the equation:

(2) $$\hat{\delta} = -(\tfrac{3}{2}) \ln [(4S^{\frac{1}{2}r} - 1)/3]$$ (Nei and Li, equation 9).

Where $\hat{\delta}$ = estimated δ

r = number of bp in enzyme recognition sequence

Substitution of $\hat{S} = 0.9$ in this equation gives:

$\hat{\delta} = 0.0176$

The approximate variance of $\hat{\delta}$ is given by:

(3) $V(\hat{\delta}) = [81\ \hat{S}^{1/r}\ V(\hat{S})]/[(4\ \hat{S}^{\frac{1}{2}r} - 1)r\ \hat{S}]^2$ (Nei and Li, equation 14).

where

(4) $V(\hat{S}) = [\hat{S}(1-\hat{S}) - \hat{S}^2(1-\hat{S}^{\frac{1}{2}})^2]/\bar{n}$ (Nei and Li, equation 12)

(5) Where $\bar{n} = (n_x + n_y)/2$

Substitution of

$\bar{n} = (10 + 9)/2 = 9.5$

and $\hat{S} = 0.9$ into equation (4) gives

$V(\hat{S}) = 0.835$

Substitution of this value into equation (5) gives

$V(\hat{\delta}) = 0.259$

The estimated nucleotide substitution rate is therefore not significant. This is due mainly to the small sample size. However the results confirms the subjective impressions obtained from the restriction maps of CH4A/TP3A and 4A.

Subcloning CH4A/TP3A and CH4A/TP4A

To facilitate more detailed analysis of the protamine gene sequences in CH4A/TP3A and 4A *Eco*RI/*Hind*III fragments of these phage were subcloned into the plasmid vector pAT153 (Fig. 5, Twigg and

Sherratt, 1980). The small (29bp) fragment released from pAT153 by *Eco*RI/*Hind*III digestion was removed by Sepharose CL-2B chromatography to minimise plasmid reformation. Both CH4A/TP3A and 4A contain only three *Eco*RI/*Hind*III fragments. One of these is the 5.72kb fragment from the right (short) arm of Charon 4A. In addition to this fragment both phage clones contain a 1.1kb *Eco*RI/*Hind*III fragment adjacent to the left (long) arm of Charon 4A (fragment a in Figures 12 and 13). The third fragment contains the protamine gene sequence (fragment b in Figures 12 and 13). In CH4A/TP3A this fragment is 4.2kb, in CH4A/TP4A it is only 2.1kb. Because of the small number, and differences in size, of the *Eco*RI/*Hind*III fragments from both Charon 4A clones it was possible to identify the resulting subclones solely by molecular weight. Therefore recombinant plasmids were first identified by replica plating on ampicillin and tetracyclin (recombinants being Amp^R, Tet^S) and then analysed on agarose gels using a quick lysis procedure. pAT153 transformed colonies were used as markers for the quick lysis gels. The subclones containing fragments a and b from the Charon 4A clones were identified and grown up. The two subclones containing the protamine gene sequences (fragment b) from CH4A/TP3A and 4A were denoted pTP3A and pTP4A respectively.

Restriction mapping of subclones pTP3A and pTP4A

The restriction maps of the two subclones containing protamine gene sequences, pTP3A and pTP4A, were extended to include the enzymes *Hpa*I, *Sst*II, *BgL*II, *Pst*I, *Pvu*II, *Sma*I, *Cla*I, *Xho*I and *Ava*I. The cleavage sites for these enzymes were mapped relative to the *Eco*RI and *Hind*III sites in the plasmids (defining the ends of the insert) by double digestions with these two enzymes. Each digest generally contained 0.5µg of DNA. The digest was then electrophoresed on a 1% agarose gel

with suitable molecular weight markers.

The restriction fragments containing protamine gene sequences were localised by Southern transfer of CH4A/TP3A and 4A restricted with the same enzymes. The transfers were hybridised with the mixed cDNA probe (pTP4, 8, 11, Jenkins et al., 1979) to identify the size of the coding fragments. The Charon 4A clones were used for these Southern transfer experiments to avoid the plasmid/plasmid hybridisation that would have occurred using the cDNA clones to probe subclone digest transfers.

The resulting subclone restriction maps are shown in Figures 12 and 13. The maps confirm that the two clones are very similar, pTP4A being approximately 2.1kb shorter at the *Eco*RI end than pTP3A The restriction maps of the sequences common to both clones (i.e. up to 2.1kb from the *Hind*III site) are identical except for the apparent replacement of an *Xba*I site in pTP3A by an *Ava*I site in pTP4A. However there is no relationship between the recognition specificities of these two enzymes (*Xba*I recognises 5'TCTAGA3', *Ava*I recognises 5'CPyCGPuG3'). Therefore, rather than a change in site this difference must represent simultaneous loss and gain of neighbouring sites. This again shows that the clones do not represent identical gene sequences but closely related sequences.

Northern blotting of trout testis RNA

Polysomal and total cellular RNA were prepared from frozen mature trout testis. Half of the RNA in each preparation was fractionated into poly(A)$^-$ and poly(A)$^+$ RNA by two cycles of oligo dT cellulose chromatography. Samples of unfractionated poly(A)$^-$ and poly(A)$^+$

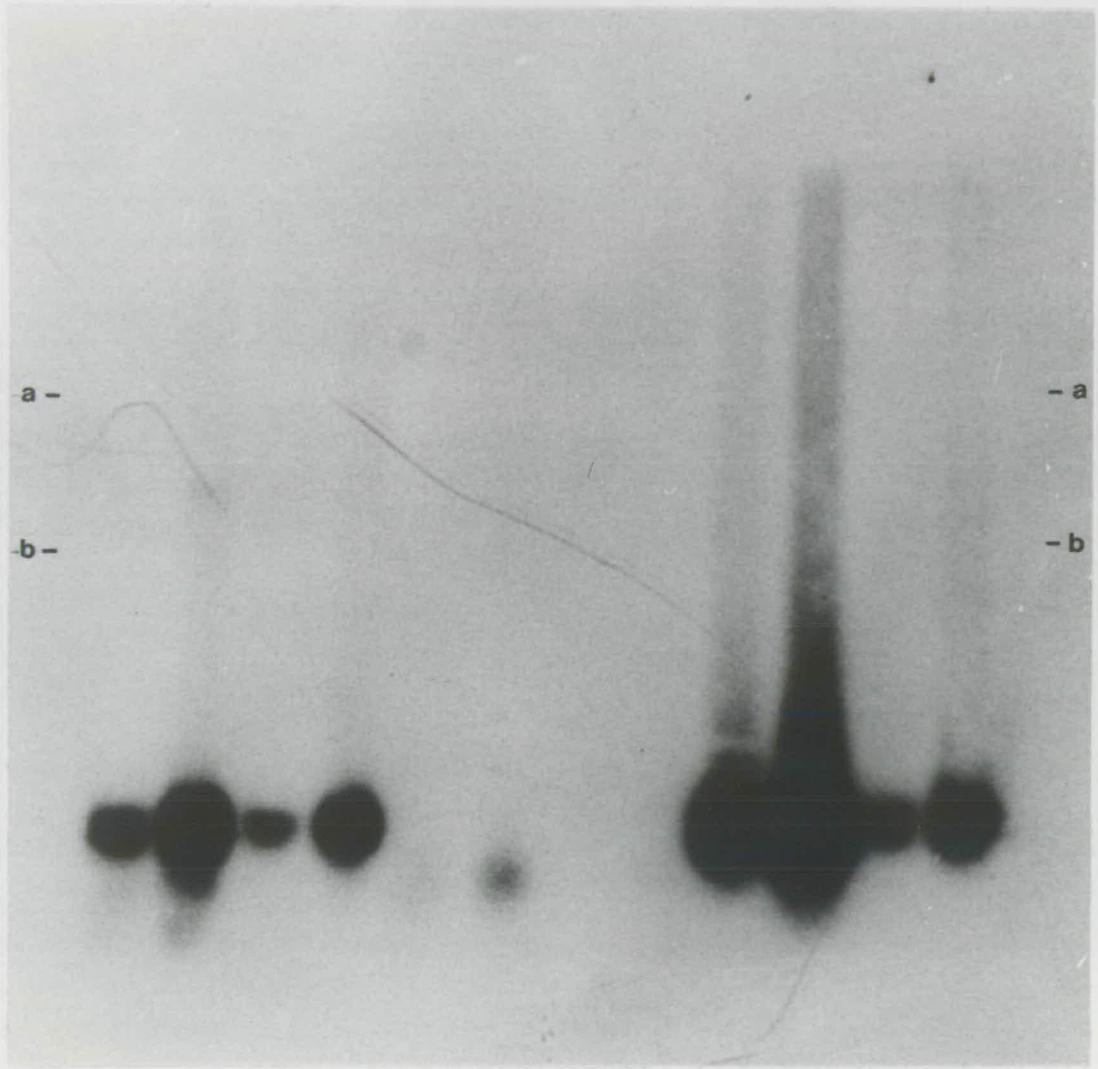
Fig.14 Northern transfer of trout testes mRNA hybridised with protamine cDNA clones.

1.4% agarose/formaldehyde gel, tracks, from left to right are,

1. Unfractionated polysomal RNA, 5µg
2. " " " , 25µg
3. " total cellular RNA, 5µg
4. " " " " , 25µg
5. Poly(A)$^-$ polysomal RNA, 5µg
6. " " " , 25µg
7. " total cellular RNA, 5µg
8. " " " " , 25µg
9. Poly(A)$^+$ polysomal RNA, 0.5µg
10. " " " , 2.5µg
11. " total cellular RNA, 0.5µg
12. " " " " , 2.5µg

Molecular weight markers, a 28s RNA(5.0kb).
b 18s RNA(1.97kb).

Fig.14



RNA from both the polysomal and total cellular RNA preparations were electrophoresed on a 1.4% agarose/formaldehyde gel and subsequently transferred to a nitrocellulose filter. The resulting transfer was then hybridised with an equal mixture of nick translated pTP4, 8 and 11 (Jenkins, 1979) and subsequently washed to a final stringency of 0.1 x SET, 68°C. A transfer of unfractionated polysomal RNA from the same gel was hybridised to a nick translated Xenopus ribosomal DNA probe, pCM4305 (Bishop, 1979) to provide molecular weight markers.

The autoradiograms resulting from the cDNA hybridised filter is shown in Figure 14. Unfortunately it is impossible to calculate the size of the protamine message from this gel as the smallest ribosomal molecular weight marker (7S, 121bp) does not appear on the marker transfer. Presumably this molecule is too small to bind effectively to nitrocellulose.

The first four tracks show that the percentage of protamine mRNA is slightly higher in unfractionated polysomal RNA than in unfractionated total cellular RNA. This is at variance with the results of Iatrou and Dixon (1977, 1978) who suggest that at the early protamine stage of spermatogenesis the amount of protamine mRNA in the polysomes and cell sap is about equal but as a percentage of total RNA, the amount of protamine mRNA is about 9 times higher in the cell sap than in the polysomes. This difference may be caused by the fact that the testis used in this study are at a later stage of maturation than those used by Iatrou and Dixon. During maturation protamine mRNA is first stored in ribonucleoprotein particles and is subsequently translated on disomes after mRNA synthesis is essentially complete. (Iatrou et al., 1978).

The second set of four tracks show that poly(A)$^-$ protamine

mRNA does occur at low concentrations in the cell as reported by Gedamu et al., (1977). The results show that most of the poly(A)$^-$ protamine mRNA is associated with polysomes. Several reports suggest that poly(A)$^-$ mRNAs are often simply a degradation production of poly(A)$^+$ mRNAs (Sheiness and Darnell, 1973; Gorski et al., 1974; Gorski et al., 1975). That this may be the case for poly(A)$^-$ protamine mRNA is suggested by the fact that this is found mainly in polysomal RNA (Iatrou and Dixon, 1977) and that it may in fact have an extremely short (2-5 nucleotides) poly(A) tail (Gadamu et al., 1977).

The third set of four tracks represent polysomal and total cellular poly(A)$^+$ RNA. Again it can be seen that the percentage of protamine mRNA is higher in the polysomes than in the total cell. The second two tracks of these four demonstrate that no nuclear protamine mRNA precursors can be seen. Any precursors would be expected to be poly(A)$^+$ as polyadenylylation occurs very rapidly after transcription (Nevins and Darnell, 1978). However it is possible that poly(A)$^-$ precursors occur in this system as reported by Iatrou and Dixon (1978). However no such precursors can be seen in either the unfractionated or poly(A)$^-$ total cellular RNA. A repeat experiment, loading 3 times as much RNA on the gel, and overexposing the resulting hybridised filter, also failed to detect any precursors.

Mapping the coding regions of pTP3A and pTP4A with restriction enzymes known to cleave protamine cDNA sequences

To further define the protamine gene sequences in pTP3A and pTP4A restriction fragments known to contain the gene sequences (by Southern transfer experiments) were restriction mapped with the enzymes *Hae*II, *Hae*III and *Hpa*II. These three enzymes cut all known cDNA

sequences in an easily recognisable pattern. In addition, a fourth enzyme, *Alu*I, which does not cut any of the known protamine cDNA sequences, was used. The restriction mapping was carried out using the rapid method described by Smith and Birnsteil (1976).

The 0.94kb *Bam*HI/*Pst*I fragment from pTP3A and the 0.65kb *Bam*HI/*Eco*RI fragment from pTP4A were purified by electroelution from agarose gels. Both fragments were then specifically labelled at the *Bam*HI site using E. coli DNA polymerase I, dATP, dGTP, TTP and α^{32} PdCTP to fill in the 5' tetranucleotide extension (5'GATC3'). Neither the *Pst*I site or the *Eco*RI site will be labelled as the former has a 3' tetranucleotide extension and the latter a 5' tetranucleotide extension that contains only A and T residues. The labelled DNA fragments were then digested to completion with an excess of the *Hae*II and electrophoresed on a 5% acrylamide/TBE gel. The gel was dried and autoradiographed. The resulting film showed only a single band (smaller than the original fragment) for each digest. This demonstrates that no internal labelling had taken place.

The labelled DNA was then partially restricted with *Hae*II, *Hae*III, *Hpa*II and *Alu*I. Cold carrier DNA was added (0.5µg) and then the DNA was incubated with enough enzyme (0.5 standard units) to give complete digestion in 60 minutes. Samples were taken at 2, 5, 10, 20, 40 and 60 minutes and pooled. The DNA was then electrophoresed on a 5% acrylamide/TBE gel. Radiolabelled molecular weight markers (pBR322 *Hpa*II, λC1857S7 *Eco*RI/*Hind*III) were electrophoresed in parallel slots. The gel was then dried down and the labelled fragment visualised by autoradiography. The resulting autoradiogram is shown in Figure 15. The fragment sizes were measured and used to construct restriction maps. As the *Bam*HI/*Pst*I fragment from pTP3A is relatively large it was decided to check the map obtained by mapping in the opposite direction. To do this

Fig.15 Smith and Birnsteil restriction mapping of pTP3A *Bam*HI/*Pst*I and pTP4A *Bam*HI/*Eco*RI fragments.

Samples were elecrophoresed on a 5.0% polyacrylamide TBE gel.Tracks,from left to right,are

M. Molecular weight scale.
1. pTP3A *Bam*HI/*Pst*I fragment restricted with *Alu*I.
2. pTP4A *Bam*HI/*Eco*RI " " " "
3. pTP3A *Bam*HI/*Pst*I " " " *Hae*II
4. pTP4A *Bam*HI/*Eco*RI " " " "
5. pTP3A *Bam*HI/*Pst*I " " " *Hae*III
6. pTP4A *Bam*HI/*Eco*RI " " " "
7. pTP3A *Bam*HI/*Pst*I " " " *Hpa*II
8. pTP4A *Bam*HI/*Eco*RI " " " "

Molecular weights were calculated from pBR322/*Hpa*II and λC1857S7/*Eco*RI+*Hind*III markers.

Fragment sizes are,

Track 1. 952bp and 550bp.
Track 3. 952bp and 447bp.
Track 5. 952bp,851bp and 407bp.
Track 7. 952bp,781bp and 361bp.
Track 2. 645bp.
Track 4. 645bp and 473bp.
Track 6. 645bp and 439bp.
Track 8. 645bp and 381bp.

The *Hpa*II site represented by the 781bp fragment in the pTP3A *Bam*HI/*Pst*I fragment(asterisked in the fig.,track 7) was shown to be two *Hpa*II sites seperated by 20bp when using the pTP3A *Bgl*II/*Pvu*II fragment,labelled at the *Bgl*II site,to confirm the above results by mapping in the opposite direction.

Fig.15

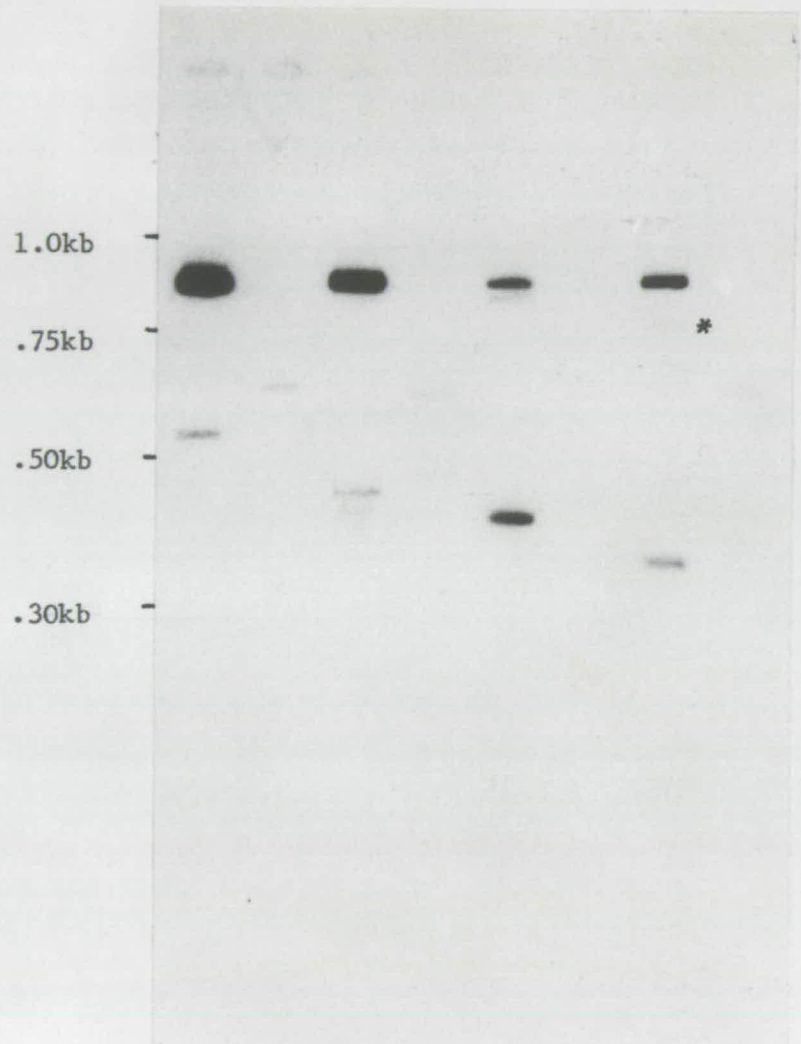
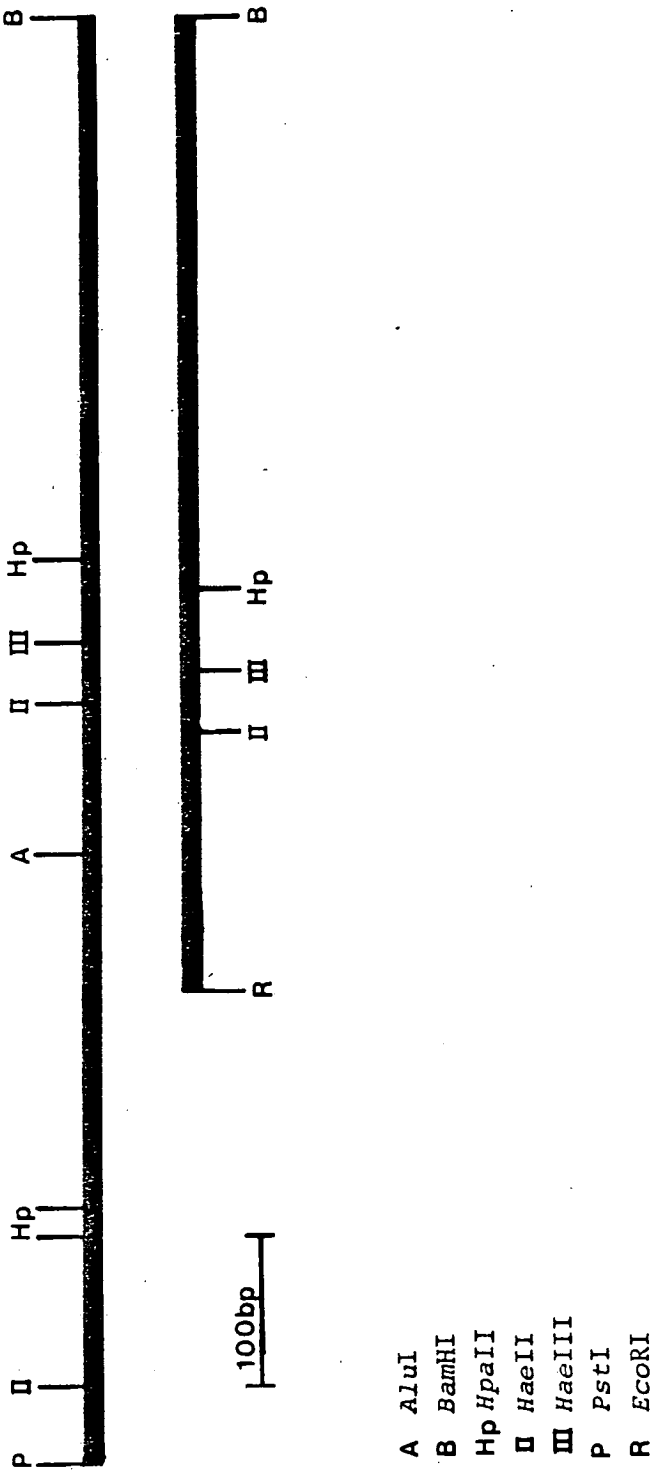


Fig.16 Restriction maps of pTP3A *Bam*HI/*Pst*I and pTP4A *Bam*HI/*Eco*RI protamine gene containing restriction fragments.



the *Pvu*II/*Bgl*II fragment was purified, labelled and digested in an analogous manner. The results obtained confirmed the map obtained but revealed that the first *Hpa*II site seen in the *Bam*HI labelled frequent (i.e. the largest sub fragment in Figure 15, marked with an asterisk) is in fact a doublet representing two *Hpa*II sites separated by 20bp.

The resulting restriction maps, aligned by the *Bam*HI sites in the insert in pTP3A and pTP4A, are shown in Figure 16. The restriction maps show that both fragments contain the characteristic restriction pattern of the protamine cDNA sequences. This is a *Hpa*II site, a *Hae*III site and a *Hae*II site separated by 51bp and 42bp respectively. These restriction sites also give the orientation of the coding sequence, the *Hae*II site being towards the 5' end of the coding sequence and the *Hpa*II site being in the 3' non coding region of the mRNA. The spacing of the sites suggests that there are no introns in the region between the *Hpa*II and *Hae*II sites. Within the accuracy of measurement these sites are at the same spacing in the gene as they are in the cDNA.

The restriction maps also show that the distance between the *Bam*HI site and the gene is about 20bp longer in pTP4A than in pTP3A. A further difference between the clones is found in the presence of an *Alu*I site immediately 5' to the gene (70bp from the *Hae*II site) in pTP3A which is absent in pTP4A. Comparison with cDNA sequences suggest that this site is well outside the coding region.

Nuclease SI and exonuclease VII mapping of the protamine coding sequences in pTP3A

To delineate the protamine gene in pTP3A more accurately and to determine if the gene contains introns the coding region in this clone was mapped using nuclease SI and exonuclease VII.pTP3A was used for

this rather than pTP4A because its larger size make it more likely to contain a complete protamine gene.

The labelling of pTP3A, the DNA/mRNA hybridisations and the SI and exo VII digestions were carried out as described in methods. The hybridisation contained approximately 50ng of DNA and 1µg of trout testis RNA. Assuming protamine mRNA is 5-10% of total trout testis poly(A)$^+$ RNA and that the (probable) six genes are transcribed with equal efficiency this is an approximate 10 to 20 fold excess of each protamine mRNA over the respective gene sequence. Therefore the hybridisation will essentially be controlled by the RNA concentration and most of the mRNA homologous DNA sequences should be hybridised. This relatively large excess of RNA over DNA should also help to stop partially homologous RNA sequences (i.e. other protamine mRNAs) hybridising to the protamine sequence in pTP3A.

Hybridisations were done using plasmid DNA either linearized with *Eco*RI, restricted with *Hae*II or restricted with *Hae*III. Three hybridisations were carried out using *Eco*RI linearized plasmid, two of these were digested with S1, one with exonoclease VII. Hybridisations containing *Hae*II or *Hae*III digested plasmid were digested with S1. A control hybridisation containing *Hae*II or *Hae*III cut plasmid but no RNA was digested with SI or exoVII to check for DNA/DNA reassociation. S1 digestion/hybridisations containing *Hae*II, or *Hae*III and *Eco*RI linearized plasmid were electrophoresed on a 8% acrylamide/formamide gel along with radiolabelled pBR322/*Hae*III size markers. The exonuclease VII digestion/hybridisation of *Eco*RI linearised plasmid was electrophoresed on an identical gel in paraellel with an S1 digestion of an identical hybridisation. Controls (digestions of hybridisations lacking RNA) were included on both gels. The gels were dried down and bands visualised by autoradiography. The resulting autoradiograms are shown

Fig.17 Nuclease S1 and exonuclease VII mapping of the protamine gene in pTP3A.

(a) Nuclease S1 mapping.

8% acrylamide,98% formamide gel.Tracks,from left to right are,

1. pBR322/*Hae*II molecular weight markers
2. S1 digest of *Hae*II cut pTP3A/trout testes mRNA hybridisation
3. S1 digest of *Hae*III cut pTP3A/trout testes mRNA hybridisation
4. S1 digest of *Eco*RI cut pTP3A/trout testes mRNA hybridisation
5. S1 digest of *Eco*RI cut pTP3A control(- mRNA) hybridisation

Fragment sizes are,

Track 2. 130bp,123bp and 100bp

Track 3. 176bp,167bp and 58bp

Track 4. 232bp and 224bp

(b) Exonuclease VII mapping.

8% acrylamide,98% formamide gel.Tracks from left to right are,

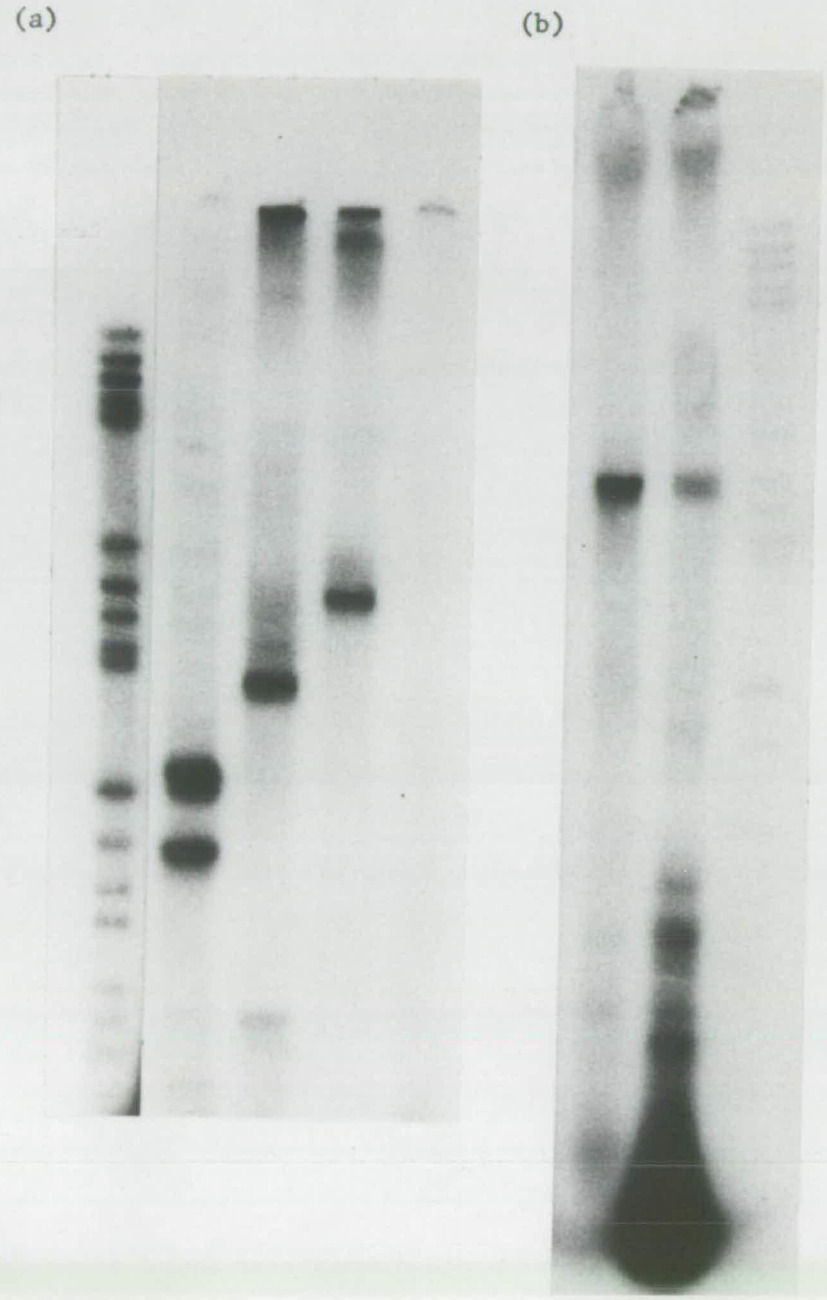
1. S1 digest of *Eco*RI cut pTP3A/trout testes mRNA hybridisation.
2. Exo VII digest of *Eco*RI cut pTP3A/trout testes mRNA hybridisation
3. pBR322/*Hae*II molecular weight markers

Fragment sizes are,

Track 1. 232bp and 224bp

Track 2. 235bp and 229bp

Fig.17



in Figure 17.

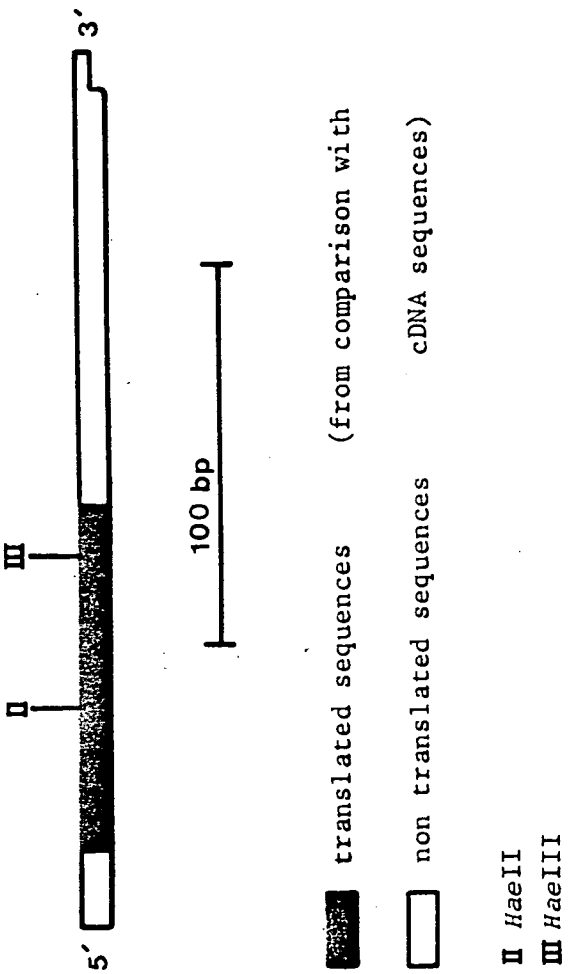
The S1 results allow the mapping of the 5' and 3' ends of the protamine gene in pTP3A and also suggest that the gene contains no introns. The *Eco*RI digested DNA track shows two bands of 224 and 232bp respectively. This suggests length heterogeneity in the hybridising mRNA rather than two exons as this is approximately the size expected, deadenylated protamine mRNA being 215-275 nucleotides long (Gedamu and Dixon, 1979). The 232bp band is much darker than the shorter band. The *Hae*II digested DNA track allows the mapping of the 5' and 3' ends of the gene as cDNA sequences show that the *Hae*II site is closer to the 5' end of the mRNA than to the 3' end. The shorter band seen, 58bp, therefore gives the distance to the 5' end of the gene. This result suggests a short 5' non translated sequence of about 16 bases in the mRNA. The 3' boundary of the gene is given by the longer *Hae*II track bands. Two bands are seen, these being 167 and 176 bp. These therefore show the same spacing as the "*Eco*RI" bands. Again the longer band is much heavier. This suggests the length heterogeneity of the hybridising mRNA is at its 3' end rather than it's 5' end. This result also gives the length of the 3' non translated region of the mRNA as 117 to 126bp. Full length 3' cDNA sequences show a 3' non translated sequence of between 113 and 120bp. The results obtained with the *Hae*III cut DNA confirm these findings, the 5' and 3' end points matching to with 2bp. The exoVII result confirms the absence of introns in the coding sequence in pTP3A. The exoVII digestion gives two bands of 229 and 235bp compared with S1 which gives bands of 224 and 232bp. The 3-5bp difference is explained by the fact that exoVII leaves 2-3bp extensions beyond the end of ds nucleic acid while S1 will digest to give blunt ds ends (Ghangas and Wu, 1975). One surprising difference is that the two exoVII bands are

of approximately equal intensity while, as before, the large S1 band is much darker than the smaller one. The reason for this is unknown.

In conclusion these results allow the mapping of the 5' and 3' boundaries of the protamine gene in pTP3A (Figure 18) and show that the gene, in common with histone (Hentschel and Birnsteil, 1981) and interferon (Nagata et al., 1980) genes contain no introns. The 3' length heterogeneity most probably represents hybridisation of the poly A tail to oligo dT sequences followed by nibbling of the AT (and hence relatively weak) hybrid. Polyadenylylation normally occurs at a stretch of coding strand A residues making the exact site impossible to define by either S1 mapping or sequencing. It is in fact likely that polyadeylylation does not occur at a single nucleotide. Alternatively it may represent a cross hybridising transcript from a second protamine gene. This is less likely however as all cDNA sequences so far published show enough differences to make protection of the whole gene in these experiments unlikely. The map of the gene in pTP3A (Figure 18) also suggests that pTP4A, although shorter that pTP3A, is long enough to contain the whole gene and 50bp of 5' sequence, assuming the pTP4A contains a gene of identical or similar structure.

The absence of an intron in the protamine gene in pTP3A is interesting in that the two other higher eukaryotic genes that do not contain introns, histones and interferon, share the feature of being relatively short genes. Interferons are 166 amino acids long, the gene includes a 21 to 23 amino acid signal peptide coding region. The total length of the reading frame is therefore 187 to 189 codons. Histones range in size from 102 to 207 amino acids in length, all but one (Histone HI) being between 102 and 135 amino acids. That these small genes lack introns can be explained by the exon equals functional

Fig.18 Nuclease S1 map of the protamine gene sequences in pTP3A



structure theory (Gilbert, 1978; Blake, 1979) by assuming they consist of only a single functional unit or domain. Alternatively they may represent ancestral polypeptides that evolved separately instead of by recombination of several exon defined ancestral coding units.

The DNA sequence of the protamine gene and surrounding sequences in pTP4A

The nucleotide sequence of the protamine gene in pTP4A was determined by sequencing the 650bp *Eco*RI/*Bam*HI fragment. Southern blotting and restriction analysis had already shown that the gene was contained within this fragment. The sequencing strategy in shown in Fig. 19. To label the *Bam*HI site pTP4A was restricted with *Bam*HI and the *Bam*HI site 3' labelled with α^{32}P CTP and AMV reverse transcriptase. The plasmid was then restricted with *Eco*RI and the uniquely labelled 650bp *Bam*HI/*Eco*RI fragment recovered by preparative polyacrylamide gel electrophoresis on a neutral 5% gel. The *Eco*RI site was 3' labelled in an analogous manner using α^{32}P GTP and T4 DNA polymerase. Again the 650bp *Eco*RI/*Bam*HI was recovered. To sequence from the internal (to the gene) *Hpa*II site the 1.36 *Alu*I fragment was recovered from a preparative 1.5% agarose gel. The recovered fragment was then restricted with *Hpa*II. The *Hpa*II site was then 3' labelled with T4 DNA polymerase and α^{32}P CTP or α^{32}P GTP. The labelled DNA was then restricted with *Taq*I and the uniquely labelled 0.48kb *Alu*I/*Hpa*II and the 0.59kb *Hpa*II/*Taq*I fragments recovered as described above. The labelled DNA was then cleaved using the G, G+A, T+C and C specific cleavage described by Maxam and Gilbert (1980). The reaction times used were 5, 60, 7 and 5' minutes respectively. It was found that to maintain the specificity of cleavages the first ethanol precipitation had to be carried out immediately after stopping the reaction. A representative sequencing gel is shown in Fig. 20.

Fig.19 Sequencing strategy for protamine gene region in pTP4A

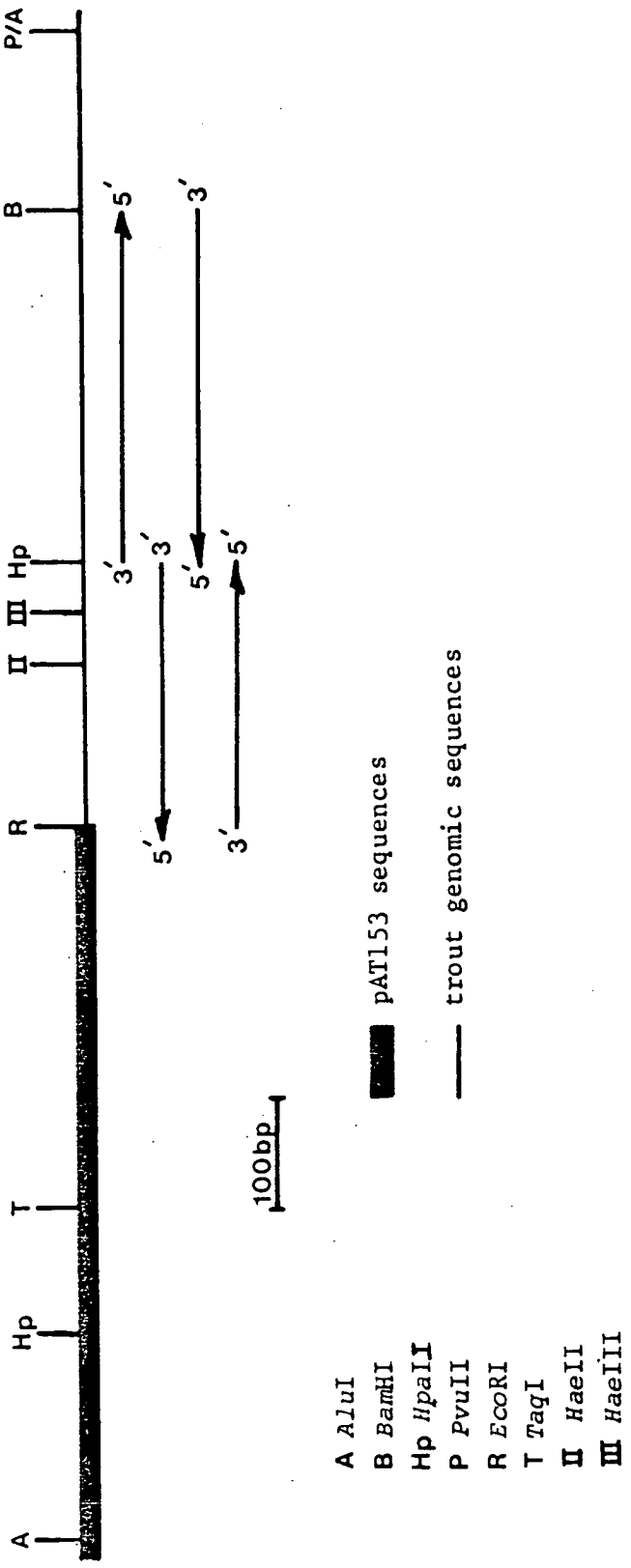
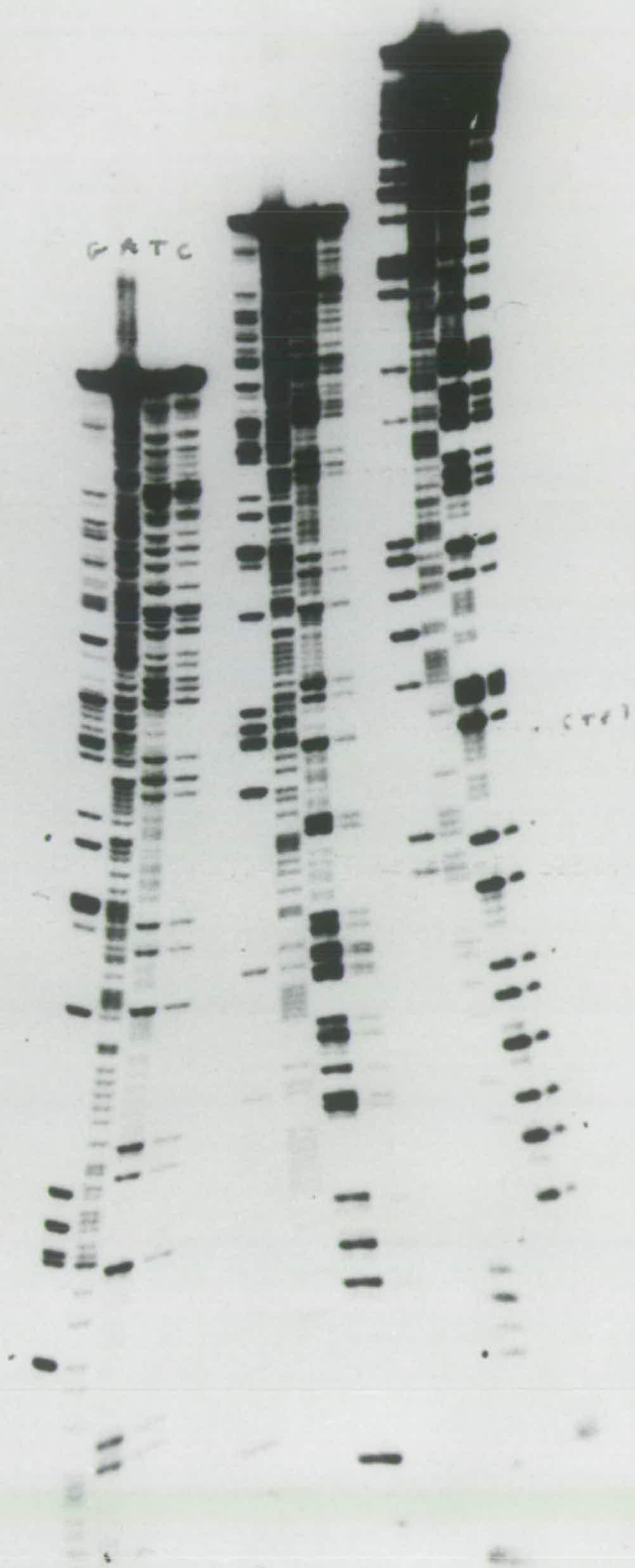


Fig.20 Representative sequencing gel.

Gel shown is an 8%,20cm x 40cm x 0.4mm,sequencing gel of the *Hpa*II/*Alu*I fragment giving the sequence from the *Hpa*II site towards the *Bam*HI site 3' to the gene (see fig.19).The gel was run at a constant current of 25mA.After the gel had been preelectrophoresed for 30 minutes the first set of four tracks was loaded.The second set of four tracks was loaded when the BPB dye marker in the first tracks was approxiamately 1cm from the bottom of the gel.The third set of tracks was loaded when the BPB dye marker in the second had reached the xylene cyanol dye in the first.Electrophoresis was continued until the BPB dye marker in the last loading was approximiately 5cm from the bottom of the gel.The ~~four~~ three loadings(left to right in the fig.)represent electrophoresis times of about 90, 190 and 290 minutes.The ~~four~~ [illegible] tracks in each loading represent,from left to right,the G,G+A,T+C and C specific cleavages.The sequence is read from bottom to top.

Fig.20



This sequencing strategy gave the complete nucleotide sequence of the 650bp *Eco*RI/*Bam*HI fragment. Nearly all the sequence was confirmed by having the sequence of both strands. The only unconfirmed sequence was 33bp on the *Eco*RI side of the *Hpa*II site and 19bp immediately preceeding the *Bam*HI site. The sequence is shown in Figure 21 along with the sequence of pTP3A (determined by S.P. Gregory, N.O. Dillon and P.H.W. Butterworth) and the homologous cDNA sequence pRTP59 (Gedamu et al., 1981). A comparison of the consensus sequences found in the protamine gene region of pTP3A and pTP4A is shown in Fig. 22.

The first feature of interest is that the sequences of pTP3A and 4A are very similar, even in the regions most distant from the coding sequence. This is to be expected from the restriction mapping data which show the two clones to be virtually identical. The differences found between the two sequences are four transitions, five transversions, four 1bp deletion/insertions, one 2bp deletion/insertion and one 16bp deletion/insertion.

The pTP3A sequence contains an *Eco*RI* site in a position analogous to the *Eco*RI site delineating the boundary of the insert in pTP4A (and CH4A/TP4A). The sequence of this *Eco*RI* site in pTP3A is 5'TAATTC3' (opposite strand 5'GAATTA3'). According to Woodbury et al., (1980), this is one of the most rapidly cleaved *Eco*RI* sites (excluding the full *Eco*RI site) than can produce 5'AATT overhangs and hence be clonable into *Eco*RI sites. The half of this site corresponding to the end of the insert in pTP4A would also reconstitute a full *Eco*RI site upon cloning into an *Eco*RI site. This suggests that pTP4A is derived from the genomic *Eco*RI fragment represented in CH4A/TP3A by *Eco*RI* cleavage at this site. Interestingly a second *Eco*RI* site, 5'CAATTT3' is found 17bp from the position of the first, in both pTP3A and pTP4A.

Fig.21 Nucleotide sequence of protamine gene region in pTP3A and pTP4A and comparison with a homologous cDNA sequence

Details of differences

pTP3A	pTP4A	cDNA[1]	Nucleotide
A	T	-	-121
C	T	-	-91
C	A	-	-67
T	Δ	-	-47
C	A	-	-6
C	G	C	+66
C	C	T	+176
G	T	T	+189
Δ	Δ	G	+221
Δ	T	-	+309
T	C	-	+319
Δ	T	-	+335
Δ	C	-	+337
Δ	TA	-	+347/348
T	C	-	+385
G	A	-	+411
Δ	16bp[2]	-	+412 to 427

1 pRTP59

2 Deleted sequence is TATTGGTATTGAAAAC

Total transversions,5

Total transitions,4

1. Gedamu et al.,(1981a)

The pTP3A sequence was determined by S.P.Gregory,N.O.Dillon and P.Butterworth, department of Biochemistry,University College,London.

EcoRI * site,TATA box,cAP site,ATG initiation and TAG termination codons are underlined

According to Woodbury et al., (1980) this is one of the most slowly cleaved *Eco*RI* sites.

The sequence of pTP4A reveals that the sequence from the *Eco*RI site to the ATG initiation codon is 60% TA rich. The (amino acid) coding sequence is 70% GC rich and the 3' non coding region of the gene is about 50% GC rich. The 3' region beyond the gene however is 70% TA rich. The GC rich protamine gene therefore appears surrounded by TA rich sequences. The sequence of pTP3A shows that the 5' TA rich sequence extends for at least another 327bp beyond the end of the sequence in pTP4A.

As pTP4A extends 120bp 5' to the putative transcription start site (by SI mapping of pTP3A) both pTP3A and 4A are long enough to include all known 5' consensus sequences. The first of these, 5' GG$^{C}_{T}$CAATC3', occurs about 70 to 80bp upstream from the transcription start site in many eukaryotic genes (Efstratiadis et al., 1980). However, it is not ubiquitous and neither of the protamine clones have a comparable sequence. The closest sequence to the consensus is 5'TCTCAATT3' which occurs between 105 and -112. In pTP4A a C to A transversion produces the sequence 5'ATCCAATAA3'. This sequence occurs between residues -70 and -62.

The second 5' consensus is the TATA box homology. The consensus for this sequence is 5' G-GTATA$^{A}_{T}$A$^{A}_{T}$-G--G3' (Breathnach and Chambon, 1981) with the first T occurring between residues -34 and -26. The pTP3A/4A sequences show the sequence 5'CCCTATAAAAGGGAC3' with the first T at residue -33. This sequence shows an exact 11bp homology (Fig.22) with the TATA box sequences of the adenovirus 2 major late promoter (Ziff and Evans, 1978) and the conalbumin gene (Cochet et al., 1979

Fig.22 Consensus sequences in the pTP3A and pTP4A protamine gene region.

TATA box homology

Consensus[1]	G-GTATA$^{A}_{T}A^{A}_{T}$-G--G	
pTP3A	AGCCCTATAAAAGGGAC	T -33
pTP4A	AGCCCTATAAAAGGGAC	T -33
Ad2[2]	GGGGCTATAAAAGGGGG	T -30
Conalbumin[3]	TCCTCTATAAAAGGGGA	T -30

cAP site homology

Consensus[1]	Y---AYYY	Y pyrimidine
pTP3A	TCTAAACATTTTATCAA	
pTP4A	TCTAAAAATTTTATCAA	

Polyadenylylation signal consensus

Consensus[3]	AATAAA
pTP3A	TAAATAAAAG
pTP4A	TAAATAAAAG
cDNAs[4,5,6]	TAAATAAAAG

1. Breathnach and Chambon,(1981).
2. Major late promoter,Ziff and Evans,(1978).
3. Cochet et al.,(1979).
4. Jenkins,(1979).
5. Gedamu et al.,(1981a).
6. Sakai et al.,(1981).

Both the conalbumin gene and Ad2 major late promoter are transcribed with high and equal efficiency in vitro (Corden et al., 1980). Preliminary results (S.P. Gregory, personal communication) suggest that the protamine gene in pTP3A is transcribed with comparable efficiency to the Ad 2 major late promoter in the Manley HeLa cell free transcription system (Manley et al., 1980). The extended homology of the TATA box sequence therefore appears to be reflected in the similar transcription efficiencies of these gene *in vitro*.

The third 5' consensus sequence is the one for the cAP or transcription initiation site. The consensus sequence for the cAP site is 5'Py---PyAPyPyPyPyPy3'. Two such sequences are found in the correct spatial relationship to the coding sequence (as defined by nuclease S1 mapping). These sequences in fact overlap. The first is 5'ATTTTATCCAT3' which has the consensus A residue 14bp from the initiation codon. The second is 5'TAAACATTTTA3' which has the consensus A residue 19bp from the initiation codon. This second consensus is found only in pTP3A. In pTP4A a C to A transversion at the C residue preceding the consensus A destroys the consensus. The nuclease S1 mapping data is not accurate enough to absolutely distinguish between the two sites. However nuclease S1 mapping to a sequence ladder shows that it is the first consensus that is utilised (S.P. Gregory, personal communication) and shows that transcription appears to start at the consensus A residue. This is consistent with the finding that 90% of protamine mRNA begins (after the cap nucleotide) with an A residue (Gedamu et al., 1981b). This A residue was therefore designated nucleotide +1. Nucleotide positions 5' to this base were given a minus sign prefix.

Surprisingly, considering the evidence that suggests that pTP3A and pTP4A may be allelic a single base substitution is found in the coding region of the two gene sequences. Even more surprising is that this

change causes a change in the amino acid sequence. The change is a C to G transversion that changes amino acid residue 17 in the protein sequence from proline (CCC) in pTP3A to alanine (GCC) in pTP4A. None of the known protamine amino acid sequences or cDNA sequences show such an amino acid change. However, it has been sequenced on both strands in pTP4A and has also been confirmed by Smith and Birnsteil restriction analysis. This is possible because the change creates on extra *Hha*I (5'GCGC3') site in pTP4A (Fig.23). The possibility exists that this change represent a cloning artifact. This, however is unlikely as cloning artifacts are usually deletions. If it represents a genuine change the combination of the evidence for allelism between pTP3A and 4A and the fact that such a change has not been seen before suggest that it may represent an *in vivo* mutation inherited by the fish from which the gene was cloned. Presumably, due to the time of expression of the gene, any deleterious effect would not be seen in the fish in which such a mutation occurs. Alternatively this change could be interpreted as evidence for the protamine genes being a repeated gene family containing a limited number of polymorphisms.

The 3' non translated part of the protamine gene in pTP3A and 4A contains a typical 5'AAATAAAA3' consensus sequence 13bp from an 8bp sequence containing 6 A residues. S1 mapping data suggest that this is ~~were~~ where polyadenylylation occurs. The sequence is 5'AACTAAAA3'. Comparison with cDNA sequences showing homology to pTP3A and 4A (Fig. 2) sug that polyadenylylation occurs in the 4A residues after the CT dinucleotide. This gives a total length for the 3' non translated region of 119bp (including the 4A residues). This agrees well with the longer of the 2 nuclease S1 bands (which corresponds to a 3' non translated region of 120bp). The second nuclease S1 band, which maps the 3' end of the gene 7bp shorter, probably represent nibbling of nuclease

Fig.23 Confirmation of the C to G transversion between the coding sequences of pTP3A and pTP4A by Smith and Birnsteil restriction mapping.

The two sequences are,

Codon	13 14 15 16 17 18
	ArgArgArgArgProArg
pTP3A	AGGCGCCGCCGCCCCAGG
pTP4A	AGGCGCCGCCGCGCCAGG
	ArgArgArgArgAlaArg

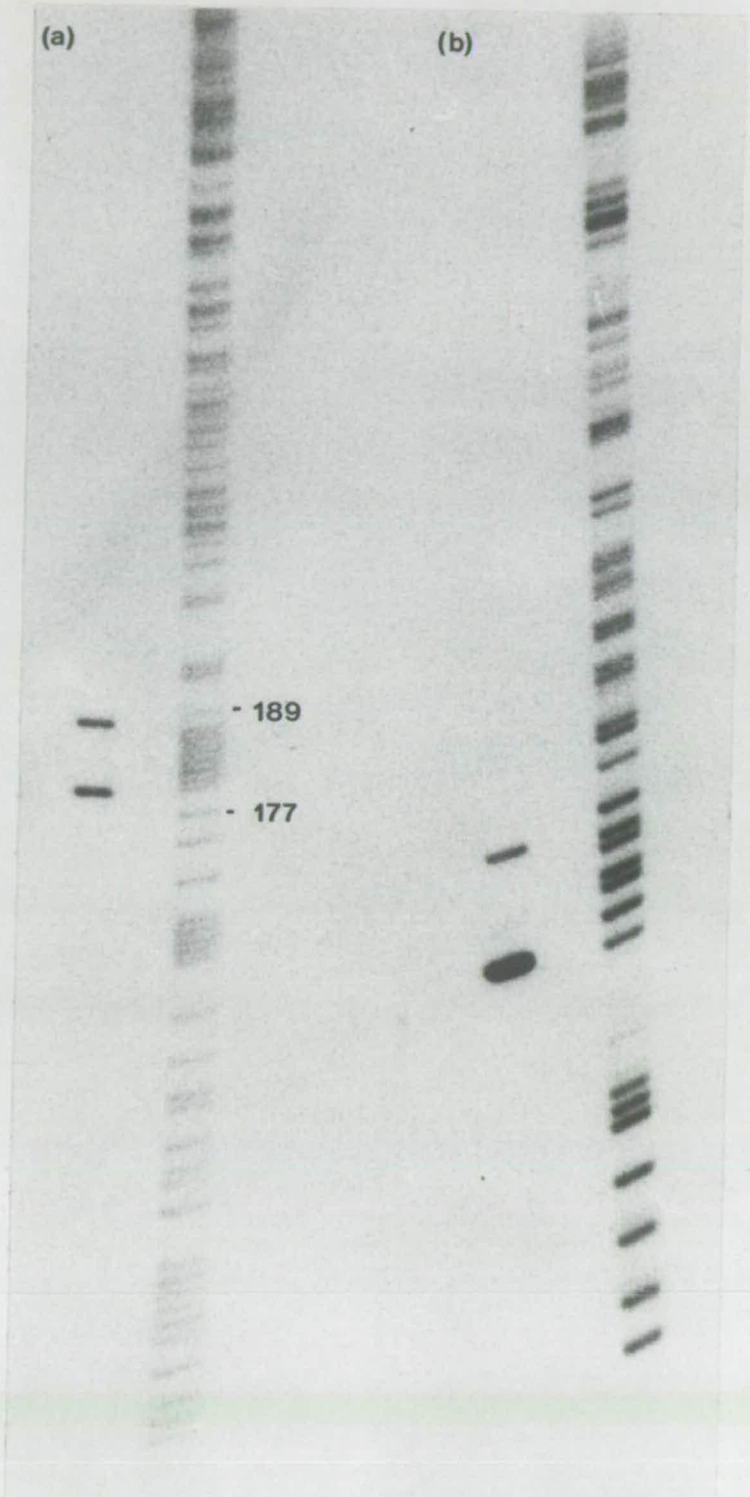
*Hha*I sites are underlined.The C to G transversion creates a second *Hha*I site in the pTP4A sequence. This sequence should therefore have two *Hha*I sites 8 bp apart.

(a) pTP4A *Eco*RI/*Bam*HI fragment,labelled at the *Eco*RI site,partially restricted with *Hha*I.The sequence ladder in the adjacent track is the G+A cleavage ladder of the *Hpa*II/*Taq*I fragment(fig.19).This gives the lengths of the two fragments as 179bp and 187bp.

(b) pTP4A *Hpa*II/*Taq*I fragment(fig.19)partially restricted with *Hha*I and run alongside a T+C sequence ladder of the same fragment.The two *Hha*I cleavage fragments align with the known sequence given above.

These results therefore confirm the pTP4A sequence as a previously undescribed protamine gene sequence.

Fig23



S1 through the CT dinucleotide to the two A residue preceding it. Such a relationship between the consensus sequence and the site of polyadenylylation is found in most other polyadenylylated mRNA sequences. There is evidence to suggest the consensus is a signal for polyadenylylation (Fitzgerald and Shenk, 1981).

In the region 3' to the gene the most striking feature is an exact 16bp inverted repeat sequence. The centre of this repeat is 130bp downstream from the polyadenylylation site. The 3' region also reveals a 16bp deletion in pTP3A (or insertion in pTP4A) 176bp from the polyadenylylation site. This deletion, and three 1bp deletions and one 2bp deletion in pTP3A confirm the S1 mapping data which suggested a 20bp increase in the distance from the *Bam*HI site to the *Hpa*II site (in the gene) in pTP4A when compared with pTP3A. The sequence shows the difference to be 21bp (16 + 3 x 1 + 1 x 2).

The approximately equal number of transversions (five) and transitions (four) is comparable with data obtained from sequence comparisons of duplicated and diverged genes (Smithies et al., 1981, Gojobori, et al., 1982). The approximately equal occurrence of transversions and transitions is not compatible with the assumption that any base change is equally likely or with analysis based on chemical considerations of the mutagenic process. The former would give twice as many transversions as transitions and the latter fifteen times as many transitions as transversions (Topal and Fresco, 1976). It is possible that transversions may obliterate the evidence of previous transitions and hence give an artifically low transition rate on analysis However the sequences in pTP3A and 4A are so closely related that it is reasonable to assume all changes represent single mutational events making this argument invalid. The observed mutation rates must represent the

interaction of both the mutational process (involving the chemical basis of base mismatch and the mechanisms of DNA replication) and of presumed repair mechanisms,

The total number of transversions (five) plus transitions (four) gives a substitution rate of 0.014 substitutions per nucleotide pair. The same calculation for the non transcribed sequence only gives a rate of 0.036 substitutions per nucleotide pair. This compares with the value of 0.018 calculated from the restriction site data of the charon clones. The relatively high mutation rate in the sequences immediately surrounding the transcribed sequence may be partly explained by their high AT content. Positive correlations between the rate of nucleotide substitution and AT content have been found in comparison studies of recently diverged genes (Smithies et al., 1981) and of active and pseudogenes (Gojoberi et al., 1982). The data of Smithies et al., (1982) suggest that an increase in the AT content from 40% to 60% (the region 5' to the gene in pTP4A has an AT content of 60% ,the region 3' an AT content of 70%) increases the nucleotide sustitution rate at least eight times. Restriction enzyme polymorphisms in the human β-globin gene cluster (Jeffreys,1979; Antonarakis et al.,1982) suggest that on average 1 base out of every 80-100 will be polymorphic (equal to substitution rates of 0.010 to 0.0125). The sequence heterogeneity between pTP3A and 4A is therefore not inconsistent with their being alleles or repeated genes given the high AT content of the regions containing the majority of the substitutions.

Attempted cloning in EMBL1

To attempt to construct a second, more random, trout genomic

library a new approach was used to try and take advantage of recent advances in cloning technology.

The first such advance is the use of *Bam*HI lambda cloning vectors. The use of *Bam*HI vectors allows the nearly random fragmentation of genomic DNA by partial digestion with *Sau*3A. *Sau*3A recognises the tetranucleotide sequence 5'GATC3', cleaving to produce 5' tetranucleotide extensions with the same sequence. *Sau*3A DNA fragments can therefore be readily ligated into *Bam*HI sites as *Bam*HI digestion also produces this 5' tetranucleotide extension. The use of *Sau*3A also has another advantage. As the recognition sequence contains all four nucleotides in equivalent amounts its frequency of occurrence (every 256bp in 50% GC random sequence DNA) will not vary appreciably with changes in base composition of DNA.

The second recent development is the construction of lambda cloning vectors that allow the direct selection of recombinants. These are based on the *spi*$^+$ phenotype which prevents growth of lambda on strains of *E. coli* lysogenic for phage P2 (Lindahl et al., 1970) The *spi*$^+$ phenotype is due to the lambda *red* and *gamma* functions as well as a P2 coded function. In the *spi* type lambda vectors the *red* and *gamma* functions are situated on the central replaceable lambda fragment and hence are deleted in recombinant phage being replaced by the cloned DNA fragment. The recombinant phage are therefore *spi*$^-$ and can be selected by plating on a strain of *E. coli* lysogenic for P2. The use of this selection system means that it is not necessary to purify phage arm pieces to reduce the number of non recombinants.

The vector chosen to exploit these new developments was a derivative of λ1059 (Karn et al., 1980) called EMBL1 (Fig. 4). For all functional purposes this phage is the same as λ1059. It does,

however, lack one of the *Eco*RI sites (at 67 map units) found in λ1059. EMBL1 is a *Bam*HI *spi*$^+$ replacement vector with a cloning capacity of 6.3 to 24.4kb. Size selection of genomic DNA is therefore essential to prevent cloning of multiple fragments.

To prepare genomic fragments suitable for cloning trout testis DNA was digested with 0.005, 0.01, 0.025, 0.05 and 0.1 units of *Sau*3A per μg of DNA. Digestion was for 60 minutes at 37°C. The digests were pooled and electrophoresed on a preparative 0.6% agarose gel. Molecular weight markers of λ(C185757) digested with *Bgl*II were included in adjacent slots. Genomic fragments of between 13.3 and 22kb were purified by electroelution and ethanol precipitation (using isopropanal to remove ethidium bromide). The recovered DNA was quantified by taking a small sample and electrophoresing it into a 0.8% tube gel containing 1μg/ml ethidium bromide. The intensity of flourescence under short wave UV illumination was then compared with that of known amounts of λ DNA under the same conditions.

EMBL1 DNA was restricted with a two fold excess of *Bam*HI, phenol/chloroform extracted and recovered by ethanol precipitation. A small aliquot was electrophoresed on a 0.6% agarose gel to check that restriction was complete.

0.5μg of EMBL1/*Bam*HI DNA was then ligated in a 5μl volume and 2.0μg of EMBL1/*Bam*HI DNA and 1.0μg of trout/*Sau*3A fragments were coligated in a 20μl volume. The two ligations were then packaged (as described for the charon 4A cloning) and assayed on Q358 and/or Q359, the former being the non-selecting strain and the latter the selecting P2 lysogenic strain. Religation of EMBL1/*Bam*HI DNA gave 10^7 pfu per μg on Q358. Coligation of EMBL1/*Bam*HI DNA and trout/*Sau*3A DNA gave 1.25×10^6 pfu/μg of DNA on Q358 and 5×10^4 on Q359.

To determine whether the phage selected by plating on Q359 were in fact recombinants the phage from the EMBL1/trout coligation/packaging were first amplified by plate lysis using Q359. The phage stock obtained in this way was then used to make a plate lysate. on Q358. DNA was then prepared from this plate lysate as detailed in methods. The DNA obtained was then restricted with *Eco*RI, *Bam*HI and *Eco*RI and *Bam*HI together. The restricted DNA was then electrophoresed on a 0.5% agarose gel along with λ weight markers (Fig. 24). The first feature to be noticed is that each digest gives distinct bands. Digestion of a collection of random genomic clones would not produce distinct bands in this way. First of all *Bam*HI digestion should produce a smear because ligation of *Sau*3A sites to *Bam*HI sites will only reconstitute full *Bam*HI sites in one in four cases. One quarter of recombinant phage will therefore have one of the *Bam*HI sites reconstituted giving one of the *Bam*HI arms on restriction with *Bam*HI. However most of the phage *Bam*HI sites will not be reconstituted. A collection of recombinants will therefore produce a smear due to random *Bam*HI cleavage in the inserted genomic fragments. The same sort of argument holds for *Eco*RI restriction, only the small *Eco*RI fragment wholly contained in the right arm should be seen. In fact one of the fragments produced by the *Eco*RI digestion corresponds to the small right arm fragment. The measured sizes of the restriction fragments found in given in Fig. 24. In each case the largest fragment corresponds to the two end arm fragments annealed together and can be ignored. A tentative restriction map, made by comparison with the EMBL1 restriction map is shown in Fig. 25 along with the EMBL1 map. Although the restriction map of the supposedly *spi*⁻ phage is only tentative it suggests that the vector has undergone rearrangement around the middle EMBL1 *Eco*RI site. The *red* and *gamma* functions map around this *Eco*RI site. It would

Fig.24 Restriction analysis of EMBL 1 *spi*$^-$ derivative.

Phage DNA was prepared from a plate lysate,restricted and electrophoresed on a 0.5% agarose gel.Tracks,from left to right,are

1. λC1857S7 *Bam*HI
2. λC1857S7 *Eco*RI/*Hind*III
3. EMBL 1 *spi*$^-$ phage *Eco*RI
4. EMBL 1 *spi*$^-$ phage *Eco*RI/*Bam*HI
5. EMBL 1 *spi*$^-$ phage *Bam*HI
6. λC1857S7 *Bgl*II

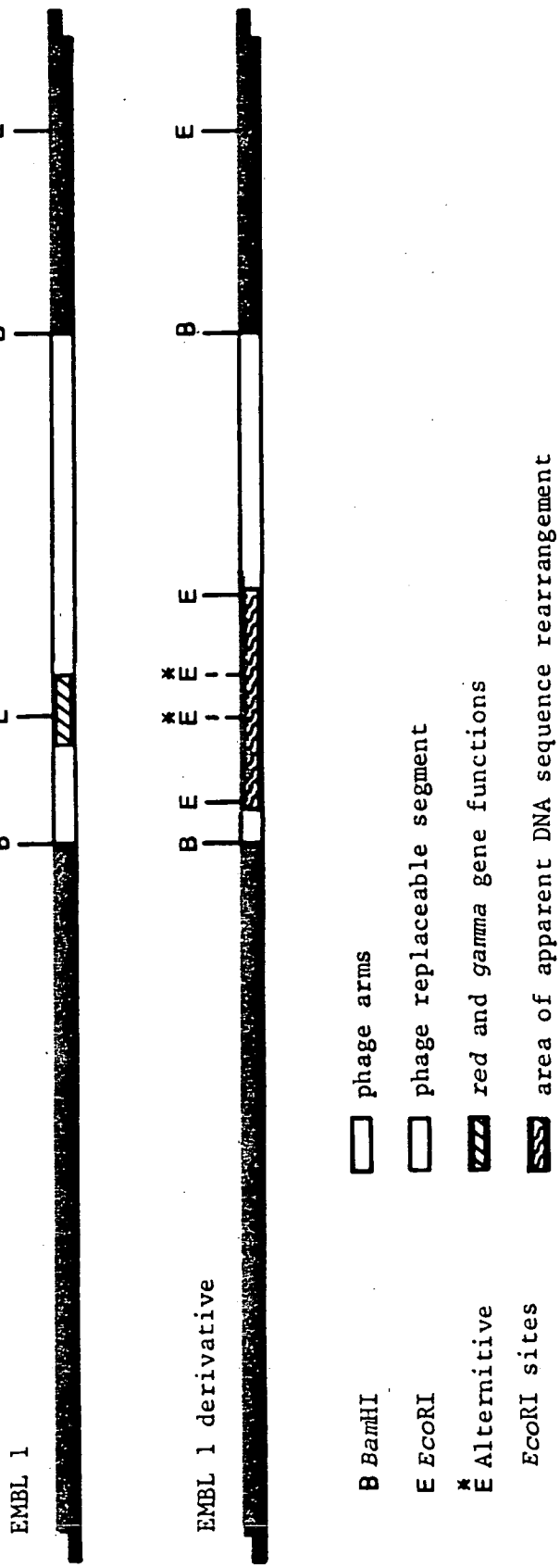
EMBL 1 *spi*$^-$ phage fragment sizes are

Track 3. 20.5kb,13.2kb,3.4kb,3.1kb,2.8kb
4. 23kb,19.5kb,7.5kb,5.7kb,3.4kb,3.1kb,2.8kb
5. 20kb,14.5kb,9.2kb

In tracks 3,4 and 5 the largest fragment in each digest represent the combination of the two restriction fragments with the lambda cohesive ends.The sizes of these double fragments were not calculated.Phage DNA prepared from plate lysates is usually badly nicked and so cannot be heated to 70°C after restriction to melt the lambda cohesive ends.Faint bands were shown to be partial digestion products or undigested DNA on further analysis.

Fig. 24

Fig.25 Restriction maps of EMBL 1 and EMBL 1 *Spi*⁻ derivative



seem that the *spi* selection system has selected for a variant, which presumably has had the *red* and *gamma* functions affected by the DNA rearrangement that has occurred. The fact that no other variants or recombinants are present suggest that the cloning procedure, at some stage, did not work. Whether the variant found was in the DNA stock used at a low concentration or arose during the cloning process is unclear. Unfortunately due to lack of time it was impossible to pursue the investigation of this cloning system further. To attempt to analyse the problem it would first be necessary to regrow phage stocks, checking the *spi* phenotype at each step. It would then be best to attempt to clone a suitable pure *Bam*HI fragment. It would allow easy analysis of the ligation step and also of protamine recombinants. The presence of recombinants could most easily be determined by Benton and Davis screening of packaged phage plate directly onto Q359 without amplification. Phage could be screened with both a central EMBL1 restriction fragment and also with the fragment being cloned.

Discussion

The number and repetition frequency of the protamine genes of rainbow trout

The number and repetition of the protamine genes has previously been estimated by DNA reassociation studies. Sakai et al., (1978) compared the rate of reassociation of the protamine genes at two different temperatures and using two assay systems, nuclease SI and HAP chromatography. At 70°C a single transition was seen with both nuclease SI and HAP assays. Comparison with an integral unique sequence standard showed that the protamine genes were also unique. At 60°C two transitions were seen with nuclease SI assay of hybrids. One of these was unique, the other representing a protamine fraction reassociating about six times faster. This faster reassociating fraction was shown to have a higher degree of mismatch than the uniquely reassociating (60°C and 70°C) fractions. When HAP was used to analyse the 60°C reassociation experiment only the faster reassociating fraction was seen. This experiment therefore suggests that the rainbow trout contains six different protamine genes that are unique but contain closely related sequences. Sequence analysis of cDNA clones shows this to be essentially correct. The cDNA sequences from three different groups is shown in Fig. 2. In total these represent five protein sequences. However nucleotide sequence comparison suggests that the number of genes represented could be either six or seven. Because of the very close sequence relationship between the sequences it is impossible to be sure what represents gene difference and what represents sequence polymorphism. For example two of the clones sequenced by Gedamu et al., (1981), numbers 94 and 178, differ in the coding region by a single wobble position substitution and by a single substitution in the 3' non coding region. On these differences alone it is impossible to unambiguously assign these

sequences to different genes rather than to gene polymorphism.

The results of the genomic Southern transfer experiments in this study do not resolve this question. With all six of the restriction enzymes used two strongly hybridizing bands are seen. In addition the *Pst*I digest contains a single band of intermediate intensity and the *Sst*I and *Bam*HI digest two such bands. All the digests also contain a number of weakly hybridizing bands. Unfortunately no copy number estimate of the strongly hybridizing bands can be made as no genome equivalent standards were included in the gel. This severely limits the analysis as such an estimate could help distinguish between the several explanations of the data. Firstly, if the strongly hybridizing bands were single copy this would suggest that the weakly hybridizing bands also represent protamine genes (assuming that estimates of gene number from DNA reassociation kinetics and cDNA sequence analysis is essentially correct). Several factors argue against this assumption, firstly hybridization with either the cDNA clone pTP4 or pTP8 (Jenkins, 1979) separately gives the same pattern of hybridization, even with a high stringency post hybridization wash (0.1 x SET, 68^{o}C). Washing to a higher stringency (0.05 x SET, 68^{o}C) causes loss of both probes from all hybridizing fragments. Secondly, a low stringency post hybridization wash (1 SET, 68^{o}C, allowing approximately 22% mismatch) does not substantially increase the signal from the weakly hybridizing bands in relation to the signal from the more strongly hybridizing bands. These two points argue against the different in signal strength being due to differences in degree of genomic fragment/probe homology. Reinforcing this view is that cloning and sequence analysis of fragments corresponding to the largest (16.6 kb), strongly hybridizing band in Southern transfers (i.e. CH4A/TP3A and 4A) show that the gene sequence in these two clones is not exactly homologous to either

of the two cDNA probes used. However possible explanations still exist. For example the strongly hybridizing bands could represent intronless protamine genes while the weaker bands represent protamine genes containing introns. The presence of an intron in the gene sequences would reduce the length of hybridizing sequence (already short) and hence reduce the resulting signal. Alternatively the strongly hybridizing bands could represent sequences present several times in the trout genome. If this is the case the two strongly hybridizing bands could represent repeated gene families corresponding to the two types of cDNA sequence. The intermediate strength bands could then be due to restriction site polymorphisms and would therefore probably represent single genes. This puts the repetition frequency of the strongly hybridizing fragments at approximately 3 to 5. The weakly hybridizing bands could then possibly correspond to crosshybridising sequences or pseudogenes. However such proposals cannot be distinguished without clonal analysis and determination of sequence copy number.

Construction and use of genomic libraries

The relative ease of constructing and screening fully representative libraries of genomic DNA sequences in lambda vectors made this approach an obvious choice for cloning the protamine genes. Although single genes, contained in restriction fragments of known size, may be more easily cloned by judicious choice of restriction endonuclease and cloning vector the cloning of a number of genes is more easily done via a random library. The construction of a random library also allows easy isolation of adjacent and overlapping sequences.

The first method described for the construction of random genomic libraries was that of Maniatis et al., (1978). Random DNA fragments were produced by partial digestion of genomic DNA with *Alu*I

and *Hae*III. Both these restriction enzymes have a tetranucleotide specificity and cleave to produce blunt ends. These fragments were first size fractionated, then *Eco*RI methylated and finally cloned into the vector Charon 4A (Blattner et al., 1977) via synthetic *Eco*RI linkers. This approach has the advantage that the cleavage of DNA by enzymes with tetranucleotide specificity is relatively random. This ensures that there is a high and approximately equal chance of any fragment being represented in the library. The disadvantage of this approach is the large number of manipulations involved in preparing the DNA for cloning. The use of the *Eco*RI* specificity of the *Eco*RI endonuclease to prepare random genomic fragments for cloning, as described by Kemp et al., (1979), appeared to circumvent this disadvantage. Polisky et al., (1975) reported that the *Eco*RI* activity of the *Eco*RI endonuclease recognised the sequence 5'NAATTN'3'. The cleavage of full *Eco*RI sites by the *Eco*RI* activity is extremely rapid. For this reason Kemp et al., (1979) use *Eco*RI methylase to protect *Eco*RI sites prior to *Eco*RI* digestion. Although the nature of N and N' in the *Eco*RI* site appears to affect the cleavage of the site the reduced specificity should still ensure nearly random digestion of DNA. For example, if the effect of N and N' on cleavage is enough to make the specificity of cleavage approximate a 5bp recognition sequence such cleavages should still occur about every 1kb. As fragments of between 13 and 19kb are being cloned this site frequency is acceptable.

The results obtained using this method in this study suggest that for a number of reasons this approach was not as successfull as expected. A total of 350,000 recombinants were made and screened. This represents 1.75 trout genome equivalents. Using the formula $P=1-(1-f)^N$ where P is the probability of a given unique sequence being present in a collection of N clones each containing a fraction, f, of the genome cloned

(Clarke and Carbon, 1976) this number of clones gives P=0.82. This assumes 15kb per clone and a genome size of 3×10^6 kb for the rainbow trout (Louie and Dixon, 1972). Assuming six protamine sequences per genome a library screen should therefore give 10-11 positive clones (6 x 1.75) representing 4 or 5 of the protamine genes (6 x 0.82). As only two clones were isolated the library is obviously not random. Part of the reason for this is likely to be the recognition of sites other than 5' NAATTN'3' by the *Eco*RI* activity (Woodbury et al., 1980a). Preferential recognition of the sequence 5'GGATTT3' may render a large proportion of the DNA unclonable. That three of the four *Eco*RI sites at the ends of the inserts in CH4A/TP3A and 4A probably represent genomic *Eco*RI sites suggests that a second problem was the incomplete *Eco*RI methylation of the genomic DNA prior to *Eco*RI* digestion. Analysis shows that undermethylation can have a large effect on the degree of randomness of fragments produced by subsequent *Eco*RI* digestion. For example, if methylation is 90% complete an unmethylated *Eco*RI site will occur on average, every 31kb in 40% GC DNA (an *Eco*RI site will occur every 3086bp, 90% methylation will leave 1×10^{-1} unmethylated). 80% methylation would give an unmethylated *Eco*RI site every 15.5kb. Assuming unmethylated *Eco*RI sites are cleaved most rapidly upon *Eco*RI* digestion it is clear that undermethylated DNA will rapidly be reduced in size. Subsequent cleavage of *Eco*RI* sites will therefore produce a collection of fragments of lower than expected randomness.

The subsequent development of *Bam*HI lambda cloning vectors makes the use of Charon 4A unnecessary. Several restriction enzymes with tetranucleotide specificity that produce the same 5' extensions as *Bam*HI have been discovered and then can be used to produce random collectio of genomic DNA. *Bam*HI vectors include several charon derivatives (Rimm et al., 1980) as well as lambda derivatives that allow direct selecti

of recombinants through the *spi* phenotype (Karn et al., 1980; Loenen and Brammer, 1980). The attempted use of one such phage EMBL 1 was however unsuccessful. The reason for this is not known. Selection for the spi^- phenotype revealed a homogeneous population of phage. These were clearly related to EMBL 1 but showed sequence rearrangements in the sequences corresponding to the *red* and *gamma* genes responsible for the *spi* phenotype. This result suggests that no spi^- phage were produced by insertion of trout genomic cDNA indicating that something made this DNA unsuitable for cloning. As the spi^- phage were homogeneous it is possible that they represent a variant present in the phage DNA preparation used for cloning. To investigate this possibility the DNA used could be packaged without restriction/religation and any spi^- phage selected, grow up and restricted. To investigate the reasons why the trout genomic DNA was not cloning successfully control experiments with a defined *Bam*HI fragment, purified from a recombinant phage from another source would be ideal. Unfortunately the time was not available to carry out such experiments.

The structure of the protamine gene in pTP3A

The structure of the protamine gene in pTP3A was analysed by restriction mapping, nuclease SI and exonuclease VII mapping and sequence analysis. This showed that the gene sequence contained no intervening sequences and fixed the 5' and 3' ends of the mRNA homologous sequence. The assignment of the ends of the gene sequence are confirmed by comparison with T1 oligoribonucleotides (5' end) and with homologous cDNA sequences (3' end). The gene is 238 nucleotides long. The length of the 5' non translated sequence is 196bp. The 3' non translated sequence is 112bp. The translated sequence is identical to the sequence found in two cDNA clones, PII (Jenkins, 1979) and 59 (Gedamu

et al., 1981). However only 59 has a homologous 3' non translated sequence. The 3' sequence of 59 shows two base substitutions in comparison with pTP3A and a single base insertion.

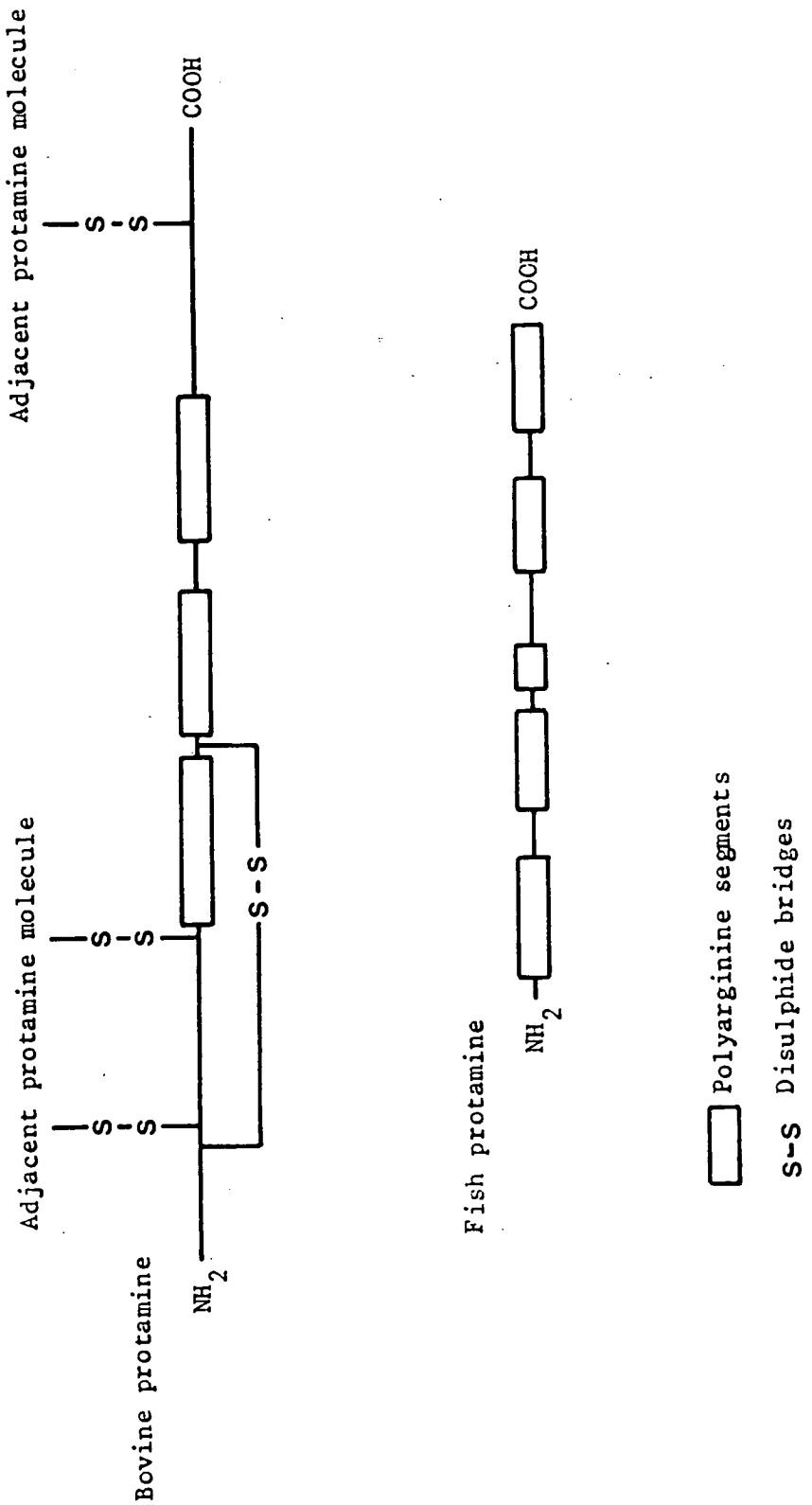
The intronless structure of the protamine gene in pTP3A is interesting in that the two other higher eukaryotic structural genes that have been shown not to contain introns (histones and interferon) have their small size in common with the protamine gene. Gilbert (1978) has suggested that the exon/intron structure of genes allows more rapid evolution of genes by recombination in introns producing exon "shuffling" or through novel splicing pathways. This proposal has been extended by the apparent relationship between exons and protein functional units or domains (Blake, 1979). Another proposal is that the intron/exon structure of genes and the RNA splicing reactions used to produce a mature mRNA may be due to the ancestral use of RNA as the primary genetic material (Reanney, 1979). RNA splicing would then represent the recombination mechanism of such an ancestral genetic system. All of these theoretical proposals about the evolution and function of the intron/exon structure of genes give essentially the same analysis of an intronless gene. The intronless structure either represents a simple, ancestral, polypeptide coding unit or it represents a gene that has lost its introns. The discovery of intronless pseudogenes (Nishioka et al., 1980; Wilde et al., 1982; Hollis, et al., 1982) suggests that eukaryotic cells contain mechanisms that allow the removal of introns from genes. The most likely mechanism for this is the reintegration of a cDNA molecule, possibly via a retrovirus. Retroviral oncogenes appear to be intronless counterparts of normal, intron containing cellular genes (Bishop, 1981). The retrovirus is therefore able to make intronless genes from intron containing cellular genes and express them. However, apart from such oncogenes, no active intronless counterpart of

a normal intron containing, cellular gene has been found. Structural analysis of actin genes from a variety of organisms suggests that introns can be removed from normal genes. Comparison of the nucleotide sequence of the rat skeletal muscle actin gene with actin genes from a number of other organisms (Zakut et al., 1982) suggests several evolutionary schemes. Firstly, if the theory that genes have originally evolved by assembly and rearrangement of several exons is correct the primordial actin gene was split in at least all those positions in which introns are found in the various present day actin genes. The variety in intron position and number is then due to varying deletion of exons. However, it is also possible that insertion of introns has occurred, perhaps by integration of transposable sequences. A chicken histone H3 gene containing 2 introns has been discovered (Engel et al., 1982). The gene has normal intron/exon splice junctions and the coding sequence contains no stop codons. The encoded amino acid sequence suggests the gene could specify a minor H3 histone found in low abundance in somatic tissues. However, the evolutionary relationships and transcriptional activity of this gene are unknown. No such histone gene has been discovered in any other organism.

The second possibility is that the protamine, interferon and histone genes have never contained introns. This would suggest that these sequences evolved without the recombination of different ancestral coding sequences. Because of their small size this may be quite likely to occur. This is especially true in the case of the protamine gene, not only because of its extremely small size, approximately 1/6 of the size of the interferon gene (187-189 amino acid residues) and 1/3 to 1/6 the size of the various histone genes (102 to 207 amino acid residues) but also due to its simple structure.

The protamine molecule is not only relatively simple at the primary sequence level but also appears to adopt a simple tertiary structure when bound to DNA. The primary amino acid sequence contains only five or six different amino acids and is basically formed of blocks of arginine residues separated by differing numbers of non arginine residues. Comparison of the known amino acid sequences of fish protamine (Fig. 31) shows that only a single non arginine residue appears to be absolutely conserved, this being the serine residue at position 8 in trout protamine CI. The protamine molecule is thought to bind in an extended manner to DNA (Balhorn , 1982), therefore there appears to be no vital tertiary structure considerations. It is therefore reasonable to assume such a simple molecule evolving without the combination of several different exons. In fact the protamine molecule is at least as simple as many of the exons on multi-exon genes.
In the context of the intron/exon structure of genes and gene evolution it would be interesting to determine the structure of the bovine protamine gene. The bovine protamine is 47 amino acid residues in length (Coelingh et al., 1972). The amino acid sequence shows that the bovine protamine has three central polyarginine segments and amino and carboxy terminal segments of 14 and 11 residues respectively. Comparison with the structure of fish protamine in a schematic diagram (Fig. 26) suggests that the amino terminal extension represents an addition to the presumably more primitive fish protamine. This extra sequence also has a novel function not seen in fish protamine. This is to provide both intra and intermolecular disulphide bridges via the three cysteine residues in the sequence (Balhorn, 1982). It would therefore be interesting to see if this extra sequence and function is represented by a second exon, separate to that containing the polyarginine segments (equivalent to the fish sequence). The carboxy terminal sequence

Fig.26 Schematic diagram of bovine and fish protamines.



of the bovine protamine is less obviously an extension to the fish protamine. However this sequence also participates in formation of intermolecular disulphide bridges and may also be considered an added sequence. This sequence may then be encoded in a third exon.

The sequences of the protamine gene regions in pTP3A and pTP4A

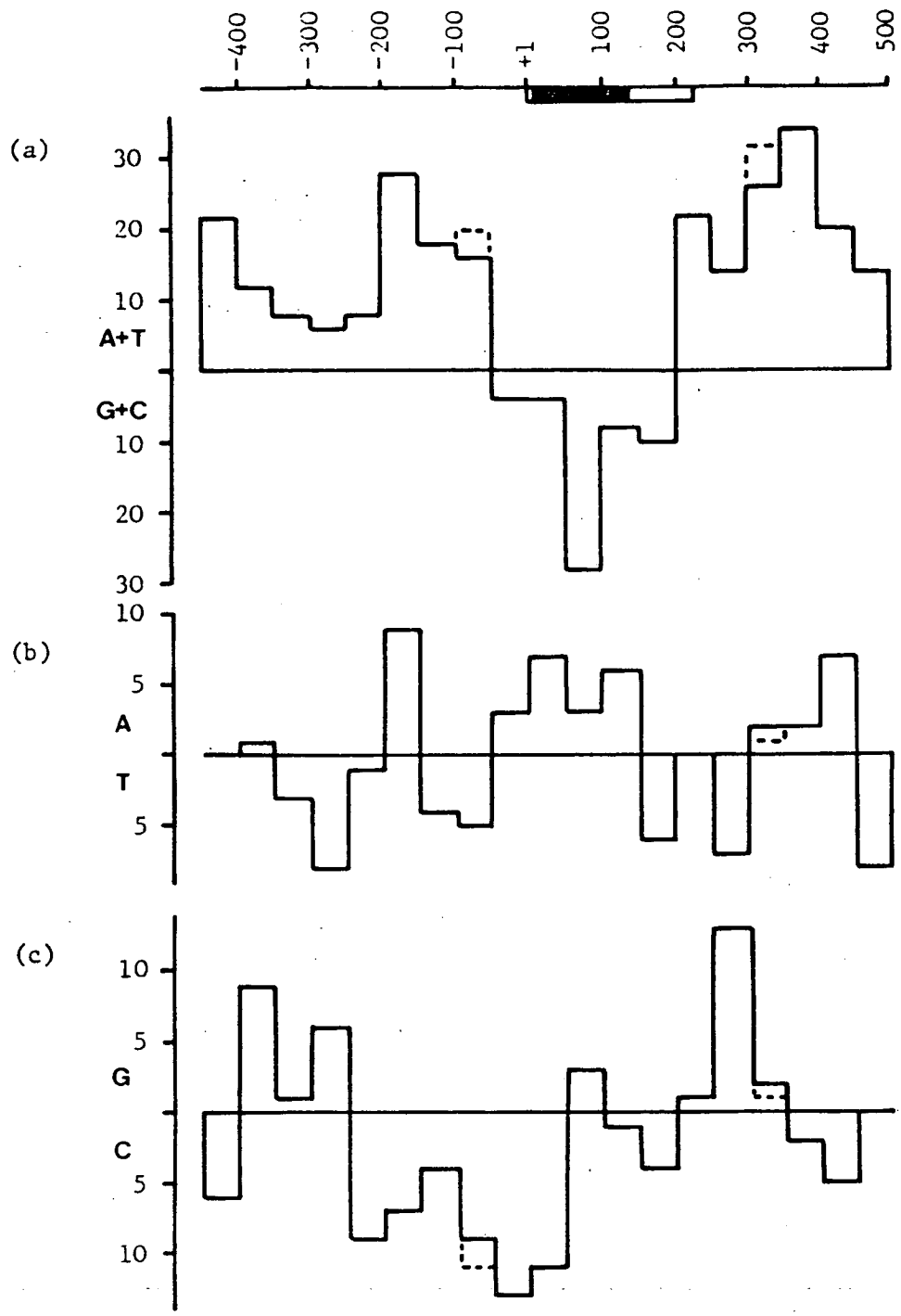
The sequenced regions of pTP3A and 4A are very similar. In the total of 648 base pairs common to both sequences there are 9 base substitutions and three 1bp, one 2bp and one 16bp deletion/insertions. The two sequences are shown in Fig. 22. The most surprising nucleotide substitution is a C (in pTP3A) to G (in pTP4A) transversion in the first base of the codon for amino acid residue 17. The codon change is CCC/proline to GCC/alanine. The pTP3A sequence, with a proline at position 17 has previously been described as a protamine amino acid sequence and as a cDNA nucleotide sequence. The protein sequence is that of the CII fraction of protamine (Gedamu et al., 1981a; Fig.31) and the nucleotide sequence has been described by Jenkins (1979), in pTP11, and by Gedamu et al., (1981a), in pRTP59. The pTP4A sequence has neither been described as an amino acid or cDNA sequence. However a very similar protein sequence with an alanine residue at the analogous amino acid residue, has been described in purified protamine from *Salmo irideus*. This is the iridine II sequence (Fig. 31). *Salmo irideus* is considered taxonomically indistinct from *Salmo gairdnerii*, the species used in this study. However, it is thought they represent distinct, genetically isolated populations. The sequence may then represent a gene polymorphism found at a very low frequency in *S. gairdnerii* but at a high frequency in *S. irideus* explaining its previous description only in the ~~later~~ latter species. Alternatively the two sequences may represent different genes. The previous description of a pTP4A type amino acid

sequence only in *S.irideus* may then be due to a higher level of expression of this gene in this species than in *S. gairdnerii*. The change could also represent a cloning artifact but this is highly unlikely.

Analysis of the sequenced regions by nucleotide content shows several interesting features. The sequences were analysed by AT and GC content and base excess (A over T and G over C) in the coding strand (Fig. 27). For analysis the sequence was divided into blocks of 50bp from the cAP site. The analysis shows that the generally GC rich gene sequence is surrounded by AT rich domains that extend to the end of the sequenced region. The GC richness of the gene is partially explained by the number of arginine residues and the preferential use of CGC and AGG arginine codons (Table 1). However the GC richness extends into the 3' non translated region as well. The reason for the surrounding AT rich sequences is unknown. The AT content of the surrounding regions approximately compensates for the GC content of the gene region. The region from nucleotide - 200 to +500, containing the gene and surrounding AT rich sequences, has an average AT content 58.7%, very close to the normal value for a higher eukaryote genome. Analysis of the coding strand for base excess (base asymmetry) reveals two domains, an A rich domain corresponding to nucleotides -50 to +150 and a C rich domain corresponding to nucleotides -250 to +50. The significance of the domains is unknown. Similar analysis has revealed similar domains in other sequences (Smithies et al., 1981). The part of the A rich domain corresponding to the coding sequence is mainly due to the preferential use of AGG and AGA arginine codons over the only T containing arginine codon, AGT.

The nucleotide sequence of the 5' non translated region and the 5' flanking region contains a number of repeated sequences. One sequence of 39-46bp is repeated three times (Fig.28(a)). The first example of this sequence is found between nucleotide positions -215 and

Fig.27 Nucleotide distribution in the protamine gene region



(a) Percentage AT or GC excess over fifty percent

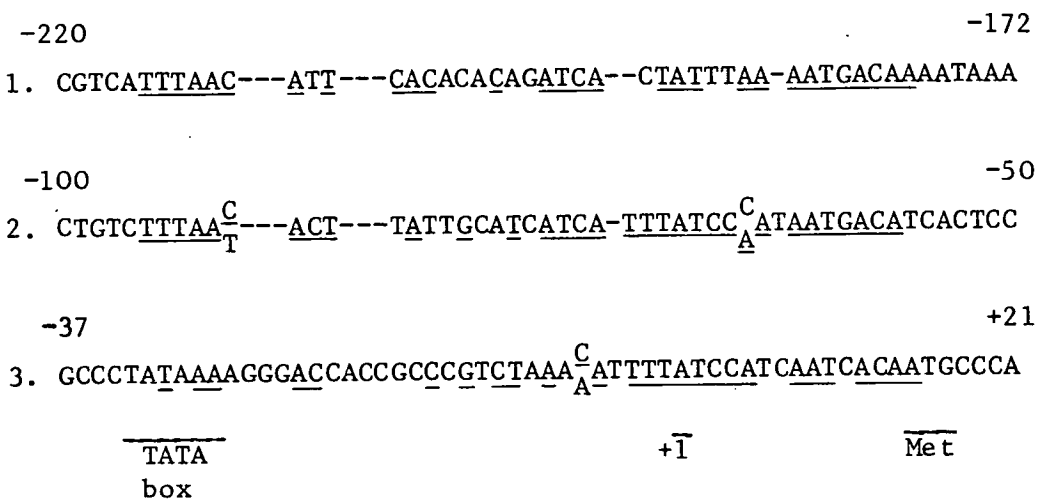
(b),(c) Nucleotide excess in coding strand as number excess per 50bp.(ie. A-T or G-C)

--- pTP4A if different from pTP3A

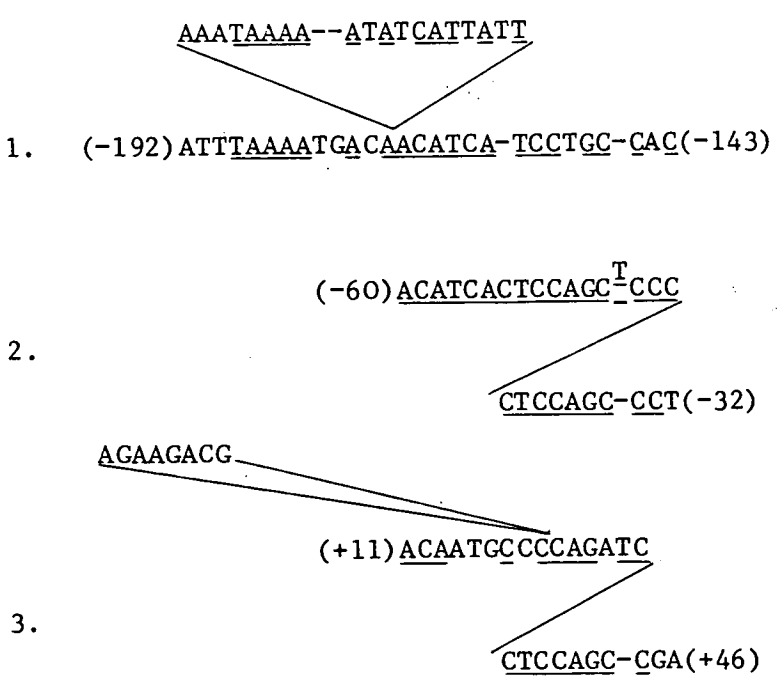
Nucleotides 401 to 450 taken from pTP4A only as pTP3A contains a 16bp deletion in this region.

Fig.28 Repeated sequence elements in the protamine gene region.

(a)



(b)



$NN^{X}_{Y}NN$ X pTP3A sequence, Y pTP4A sequence where different

-178. The second occurs between nucleotides -95 and -56 and the third between nucleotides -32 and +15. The latter therefore contains most of the TATA box sequence, the cAP site consensus and ends at the ATG initiation codon. The degree of homology between the elements varies. The first and second show 61% homology, the second and third 55% and the first and third 45%. Assuming that the protamine gene in pTP3A and pTP4A is extremely ancient the third sequence, which includes part of the gene structure, must be the original sequence. The first duplication event then presumably formed the first element as this shows least homology to the third element. The degree of sequence divergence between these two sequences (55%) gives an estimated duplication time of between 33 to 66 million years ago (assuming sequences are not selected for and a nucleotide substitution rate of 1% per 0.6-1.3 million years). A second duplication event, of the first element, to form the second, then presumably occurred. The first and second elements show 39% divergence corresponding to an estimated duplication time of between 22 to 44 million years ago.

The second group of sequences, shown in Fig.28(b) , show a more complex relationship. Two of the sequence elements, the second and third (represented by nucleotides -59 to -34 and +12 to +43 respectively) share a homologous sequence of 25bp, the third element containing a 8bp insertion. Each element also shows an internal direct repeat of a 9bp sequence at the 3' end. These two sequences can therefore be represented as abb'. The first element is more complex. A sequence of 16bp shows homology to the first 16bp of the second and third elements and can therefore be represented as ab. However the sequence contains a 19bp insertion that in part shows homology to sequences 5' to the ab sequence and also to the 5' half of the ab sequence. This structure ca- therefore be represented by $\overset{xa}{\underset{xab}{\nabla}}$

Consensus sequences in the protamine gene region of pTP3A and pTP4A

(1) Far upstream sequences, the CAAT consensus

The first putative consensus sequence in the protamine gene region in pTP3A is the sequence 5'TCTCAATTT3'. The 3' C in this sequence is nucleotide -103. The consensus sequence this partially matches is 5'GG$^{C}_{T}$CAATCT3' (Efstratiadis et al., 1980). The corresponding C residue is found between nucleotides -93 to -70 in the gene sequences from which the consensus was derived. All these genes also show conservation of at least one of the two G residues of the consensus. The sequence in pTP3A is therefore neither a good fit to the consensus sequence nor in the correct position being at least 10bp further upstream than normal.

In pTP4A a second putative fit to this consensus is found. A C to A transversion at nucleotide -66 produces the sequence 5'ATCCAATAA3'. The C residue of the CAAT sequence is at nucleotide -67. This sequence is not a good fit to the consensus either as again neither of the two 5' G residues of the consensus are conserved. This sequence is also too close to the gene to fit within the normal spatial relationship found between this consensus and gene sequences. However it is less (3bp) outside the normal range of the consensus/cAP site spacing than the first sequence. It may therefore be a coincidental rather than a conserved sequence. The absence of this consensus sequence in pTP3A would not be surprising as it is not ubiquitous and in the rabbit β-globin (Grosveld et al., 1981), conalbumin (Corden et al., 1980) and HSVI *tk* (McKnight and Kingsbury, 1982) genes, where it does occur, it does not appear to be necessary for efficient transcription.

(2) The TATA box and cAP site and the control of transcription

Both the pTP3A and 4A sequences contain a cAP site consensus

and a TATA box homology which is in the correct spatial relationship to the cAP site.

The cAP site utilised in pTP3A has been defined by nuclease S1 mapping (Fig. 22). Although the pTP3A sequence contains two potential cAP sites only one appears to be utilized *in vivo*. The cAP site not used in pTP3A is destroyed in pTP4A by a C to A transversion at nucleotide -6. The TATA box sequence shares a 12bp homology with the analogous sequences in the conalbumin gene and the Ad2 major late promoter (Fig.22). Both the conalbumin and Ad2 major late promoters are strong *in vitro* and *in vivo* promoters (Corden et al., 1980). Similarly the protamine gene promoter appears to be of equal efficiency to the Ad2 major late promoter in the Manley cell free system (S.P. Gregory, personal communication). However it is difficult to assess whether this reflects the *in vivo* activity of the gene as no direct measurement has been made. The protamine sequence encoded in pTP3A (a CII type polypeptide) is the most abundant *in vivo* (Ling et al., 1971). However as the time scale of spermatogenesis is long (several months) and protamine mRNA appears to be synthesised and stored in cytoplasmic ribonucleoprotein particles (Iatrou et al., 1978; Gedamu et al., 1977a) this may not reflect a high gene transcription rate. It is also possible that there are several genes coding for the same polypeptide sequence.

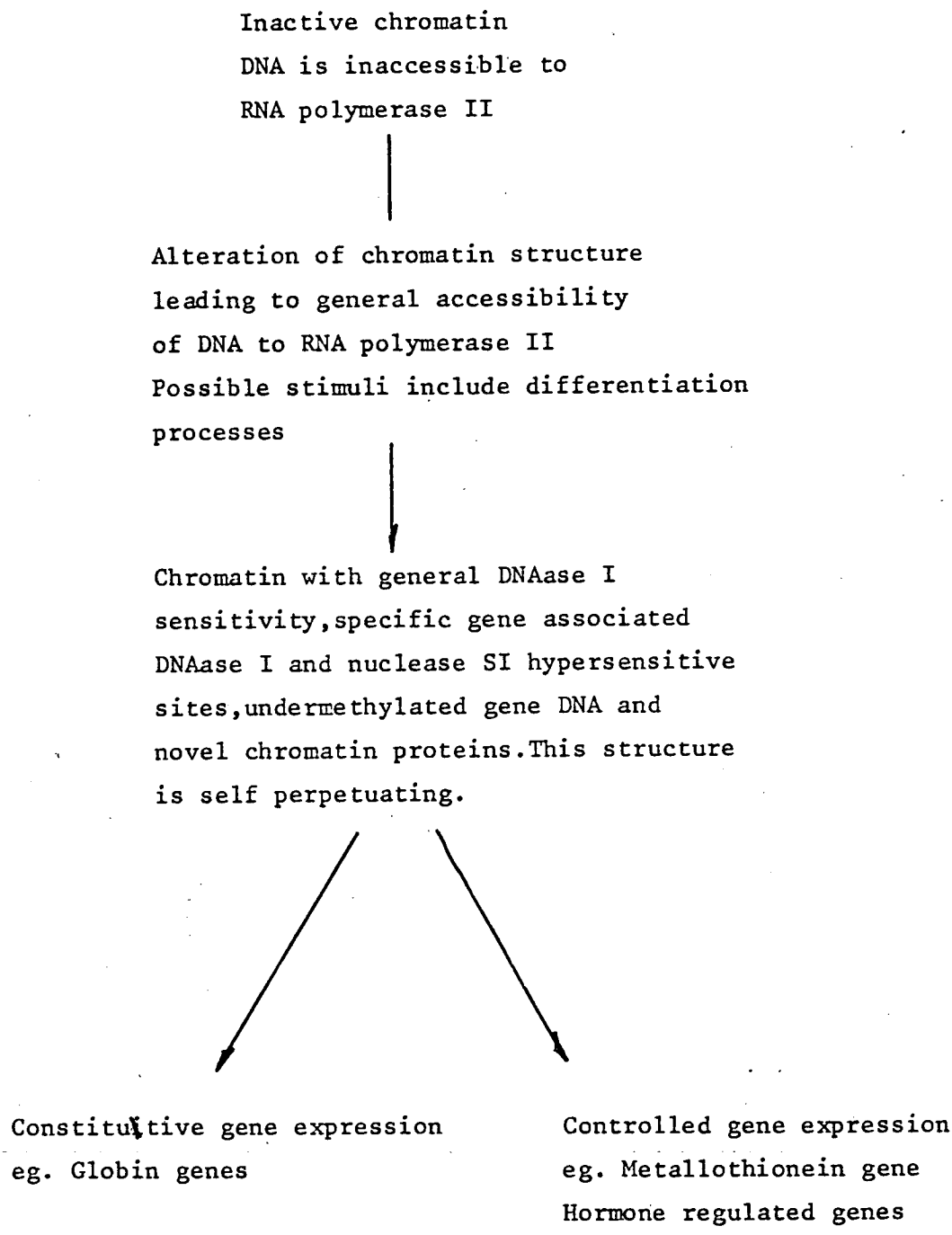
The similarity of the TATA box homology to the *E.coli* Pribnow box (Siebenlist et al., 1980) suggests that the ~~consensus sequence for the Pribnow box is the~~ functions of the two sequences may be analogous. The Pribnow box functions as a promoter for E. coli RNA polymerase. Although the consensus sequences for the TATA and Pribnow box homology are very similar their spatial relationship to the transcription initiation site is different. The Pribnow box is found only 11 to 14 nucleotides upstream from the transcriptional start site compared with the 24 to 34 nucleotides between the TATA box and the cAP site. However as

there are 10.5bp per turn of the DNA helix the two sequences are in the same orientation with respect to their initiation sites. The *E.coli* RNA polymerase appears to initiate by first binding randomly to DNA and then translocating to a promoter (Park et al., 1982). Binding to a promoter causes unwinding of nucleotides -10 to +1. Negative control of transcription initiation in *E. coli* appears to be by direct blocking of RNA polymerase binding to or transcription from the promoter sequence. The binding of control factors, usually proteins, is mediated by operator sequences. These mediate factor binding directly. Operator and promoter sequences may overlap. Positive control mechanisms, such as the cAMP/CRP system also occur. These systems appear to facilitate transcription initiation directly, the effector molecule binding immediately adjacent to the promoter sequence. The ubiquitous nature and high degree of sequence conservation of the TATA box and its apparent control role in directing transcription in *in vitro* systems support the view that it is directly involved in the mechanism of RNA polymerase II initiation, as a promoter, rather than in a transcriptional control mechanism. Control of transcription *in vivo* must then presumably depend on controlling the access to or translocation from the TATA box by RNA polymerase II. One possibility is that control may be in a manner analogous to the control system in *E.coli*, that is by repressor or inducer binding to sequences immediately adjacent to the TATA box. The localization of the sequences involved in cadmium regulation of the metallothionein gene to within 90bp of the cAP site (Brinster et al., 1982) make this regulation system a candidate for an effector molecule/operator type control mechanism. The localization of the sequences involved in hormone regulation and of potential hormone/receptor binding sites to sequences immediately upstream from hormonally regulated genes (Mulvihill et al., 1982; Kurtz, 1981; Huang et al., 1981) make these genes candidates for similar control mechanisms. The effect of hormone/receptor complexes on *in vitro* transcription has not been analysed

However the efficient transcription of the ovalbumin and conalbumin genes in the Manley HeLa cell free system suggests that such transcription systems are not suitable for such analysis. The delineation of the sequences involved in these regulation systems will help to elucidate the exact nature of the control mechanisms involved.

A second possibility for controlling access of RNA polymerase II to promoter sequences is through chromatin structure. Inactive chromatin appears to have multiple layers of organisation and such a high degree of structure probably precludes access by RNA polymerase to DNA sequences. The association of general DNAase sensitivity with active chromatin suggests that active chromatin has a more open structure than inactive chromatin. The more specific association of DNA under-methylation, HMG proteins and DNAase and nuclease S1 hypersensitive sites with active genes also suggests that specific chromatin regions that contain active genes show novel features. The exact relationships between these features and gene activity is not known. However the evidence for the importance of DNA methylation and DNA/nuclease S1 sensitives sites is compelling. The DNAase hypersensitive sites associated with the chicken α and β-globin genes are self propagating and therefore show one of the essential properties for sites involved in differential gene control (Groudine and Weintraub, 1982). However such sites are not sufficient to cause transcription. In the *Drosophila sqs*4 glue protein gene there appears to be a hierarchy of interaction and formation between the five tissue specific gene associated DNAase I hypersensitive sites. What exactly these sites represent in terms of chromatin structure and how they are involved in transcription is not known. However it is clear that control of transcription in eukarotes can be mediated in a number of ways. A possible scheme for gene control is shown in Fig.29.

Fig.29 A possible scheme for control of transcription.



(3) The AATAA consensus, polyadenylylation and transcription termination

The protamine gene sequences in pTP3A and 4A contain the sequence 5'AAATAAAA3' 20bp (to the T residue) upstream from the site of mRNA polyadenylylation. The same sequence is found in all the full length (3') cDNA clones (Fig. 2). The consensus AATAAA sequence, or a closely related sequence, is ubiquitous to all non yeast polyadenylylated mRNAs. The sequence is always found 15 to 20 nucleotides upstream from polyadenylylation site. Histone mRNAs, which lack a polyadenyl tail, do not contain the sequence. Yeast mRNAs are all polyadenylylated indicating a non specific polyadenylylation mechanism. Yeast mRNAs do not contain the sequence. All this evidence suggests that the AATAAA sequence directs polyadenylylation. Direct evidence for this role has also been produced by deletion analysis of the sequence in SV40 (Fitzgerald and Shenk, 1981). It is not known if polyadenylylation and termination are linked in any way, the available evidence for specific termination sequences in eukaryotes is very limited. The available evidence suggests that polyadenylylation occurs after endonucleolytic cleavage of the nascent transcript at the site of polyadenylylation. Termination at a downstream site could then occur randomly with respect to sequence by linkage to this cleavage or at a specific downstream site. As the cleavage/polyadenylylation reaction and degradation of the transcript 3' to the polyadenylylation site are extremely rapid most techniques are not sensitive enough to detect these presumed downstream transcripts. However, by labelling RNA in isolated nuclei and then binding specific RNA to DNA clones immobilized on nitrocellulose, Hofer et al., (1982) have shown that most transcripts of the mouse β-globin major gene appear to terminate approximately 1.4kb downstream from the polyadenylylation site. A yeast deletion mutant has been described that causes readthrough of the normal termination site (presumed to correspond to the 3' end of the mRNA in yeast) of the CYC1

gene (Zaret and Sherman, 1982). The deletion occurs immediately prior to the normal termination site. Immediately 5' to the deletion is a region of dyad symmetry. In *E. coli*, *rho* factor independent termination of transcription is associated with the presence of a region of dyad symmetry near the 3' end of the transcription unit. Termination occurs in a T rich (coding strand) sequence (Platt, 1981). A sequence containing a dyad symmetry has been directly demonstrated to be necessary for termination of the sea urchin H2A histone gene (Birchmeier et al., 1982). Termination of the H2A gene normally occurs 4 nucleotides downstream from a16 nucleotide hyphenated (by 4bp) palindrome. An analogous sequence is found at the 3' of the protamine mRNA sequence in pTP3A and 4A (Fig. 30). This is a 17 nucleotide hyphenated (by 5bp) palindrome. The 5' half of the palindrome includes the TAAAA pentanucleotide from the polyadenylylation consensus. The distance between the end of this sequence and the polyadenylylation site is 4 nucleotides. This is the same as the distance between the H2A palindrome and termination site. The terminal 4 nucleotides in the protamine and H2A mRNAs are almost identical being ACTA and ACCA respectively. The protamine sequence may then be a termination signal extendin the extremely compact nature of the gene to fusion of the polyadenylylation and termination sites. In other protamine cDNA sequences (Fig. 30) identical or closely related sequences occur. The minimum number of bases in any cDNA sequence that can pair to form a hairpin structure in the mRNA is five. The importance of these sequences could be tested by *in vitro* mutagenesis and expression analysis in eukoryotic viral/cell transfection systems (the only systems to support polyadenylylation of a clone transcripts

The sequence 3' to the polyadenylylation site in pTP3A and 4A also contains several palindrome sequences including a 16bp sequence 5'ATTAAAAGGTTTTAATA3' between residues 356 and 372. If termination occurs

Fig.30 Interrupted palindrome at the 3' end of the protamine gene

	a b
pTP3A/4A,P8[1],131[2]	CTTAAATAAAAGATGGGC-TTTTAACTAAAA
59,43[2]	CTTAAATAAAAGATGGGCGTTTTAACT(A)$_N$
178[2]	CTTAAATAAAAGATGAACGTTTTAACT(A)$_N$
6b[3]	CTTAAATAAAAGATGGGCG-TTTAACT(A)$_N$
Histone H2A[4]	AACAACGGCCCT-TAT--AGGGCCACCA

P8,131,59,43,178 and 6b are cDNA sequences

1. Jenkins,(1979).
2. Gedamu et al.,(1981a).
3. Sakai et al.,(1981).
4. Birchmeier et al.,(1981).

a Polyadenylylation consensus sequence

b Polyadenylylation site

c 3' end H2A mRNA

3' to the polyadenylylation site it would be interesting to see if any such sequence features are associated with the site. The site of termination could possibly be determined by alteration of the polyadenylylation consensus sequence allowing large amounts of full length transcripts to be synthesised. Again this analysis would best be done in an eukaryotic viral clone/transfection system as such systems are most likely to support specific termination.

Are the protamine genes in CH4A/TP3A and 4A allellic or members of a repeated gene family?

Comparison of the restriction sites in CH4A/TP3A and 4A and the corresponding subclones, pTP3A and 4A show that the two sequences are very closely related but not identical. Sequence analysis of the corresponding gene regions confirms this. The restriction mapping data from the charon 4A clones shows a single *Xba*I site polymorphism. Use of this data to calculate a figure for the nucleotide substitution rate gives a value of 0.0176 substitutions/nucleotide site. However the limited size of the data sample gives a correspondingly large variance value, 0.259. The sequence data gives a nucleotide substitution rate of 0.014 for the whole of the region sequenced in both clones (649bp). The corresponding figures for the non transcribed and transcribed sequences taken separately are 0.036 and 0.0088 respectively. The question is whether these differences represent polymorphic differences between gene alleles or members of a repeated gene family. The relevant points to be considered are

(1) The copy number of the CH4A/TP3A and 4A sequences in the trout genome

The large size of the trout genomic fragment in CH4A/TP3A and 4A and the fact that they both contain no internal *Eco*RI sites means that both

clones must be representative of the largest genomic *Eco*RI fragment. The sequence of pTP3A reveals an internal *Eco*RI* site at a position analogous to the *Eco*RI site at the boundary between the genomic sequence and the Charon 4A right arm at the foreshortened end of CH4A/TP4A. This suggests that CH4A/TP3A represents the 16.6kb genomic *Eco*RI fragment in its entirety and CH4A/TP4A this fragment cleaved at the *Eco*RI* site found in CH4A/TP3A. The copy number of CH4A/TP3A and 4A is therefore the copy number of this genomic *Eco*RI fragment. This fragment is one that hybridizes strongly in genomic Southern transfers. However it is not known whether this increased signal is due to increased copy number or other factors. If it is assumed that all protamine genes have the same structure (i.e. do not contain introns) it is highly likely that the CH4A/TP3A and 4A clones are representatives of a repeated gene rather than gene alleles as the increased signal from the 16.6kb genomic fragment becomes difficult to explain without resorting to increased copy number. However an increased copy number of this fragment would be expected to lead to an increase in the number of corresponding clones isolated from the genomic library. Two clones were isolated from a library of approximately 1.75 genome equivalent superficially suggesting a copy number of unity. However as the library appears to be distinctly non random and as the exact cause of this non randomness is not known this data cannot be used to argue for or against increased copy number..

(2) Comparison of the nucleotide substitution rate in allelic, recently diverged and repeated gene sequences

Although no systematic study of the level of nucleotide sequence polymorphism in eukaryotes has been made some data is available. A study of restriction site polymorphisms in the human β-globin gene cluster (Jeffreys, 1979) suggests that in this region as many as 1 out of every 80

to 100 nucleotides may be polymorphic (this is equivalent to a nucleotide substitution rate of 0.01 to 0.0125 substitutions/nucleotide). All the variant restriction sites detected were within intron sequences. A more detailed analysis, including nucleotide sequencing (Orkin et al., 1982) extends this analysis and shows the same average rate of polymorphisms. Many of the polymorphisms found appear to be non randomly associated. Only one sequence polymorphism is found in the translated sequence of the β globin gene, this is a C to T transition in the second codon of the translated sequence. The change is in the third base of the codon and does not alter the amino acid sequence. The human α globin gene cluster also shows polymorphic variation. However these appear to be due to DNA deletion/insertion events rather than nucleotide substitutions (Higgs et al., 1981). Deletion polymorphisms appear to be associated with short direct repeats (Efstratiadis et al., 1980; Shen et al., 1981). A sequence analysis of the duplicated γ-globin genes show that in the 5' two thirds of the gene there is more difference between the two alleles of the A_γ gene than between the A_γ and G_γ genes. The converse is true for the remaining one third of the gene. This surprising finding seems to be due to a gene conversion event between a A_γ and a G_γ gene making one of the A_γ alleles a $5'G_\gamma$, $3'A_\gamma$ hybrid. The apparently high rate of nucleotide substitution between the two A_γ alleles in the 5' two thirds of the gene is therefore actually representative of the substitution rate between duplicated genes. The original duplication event occurred approximately 34 million years ago. The gene conversion event is extremely recent, probably with the last one or two million years. Comparison of the duplicated sequence shows an average nucleotide substitution rate of 0.154. The local base substitution rate appears to be related to the AT content of the DNA (Smithies et al., 1981).

The positive correlation between AT content and the nucleotide

substitution rate may be due to the fact that AT rich sequences can tolerate distortion of the DNA helix more easily because of the relative weakness of AT over GC base pairs. The value of 0.014 for the nucleotide substitution rate between the sequenced regions of pTP3A and 4A (which, from the restriction site data, is probably of the same magnitude for the whole of the sequences in the Charon 4A clones) is therefore more consistent with the genes being alleles rather than duplicated genes. This nucleotide substitution rate is somewhat higher than that estimated for the human β globin gene cluster (0.01 to 0.0125). The difference may be partly due to the high AT content of the DNA sequences surrounding the protamine gene. The high AT content may also explain the higher number of transversions (5) than transitions (4). The data of Smithies et al., suggesting that high AT content increases the number of transversions more than the number of transitions (Gojobori et al., 1982). Comparison of globin genes from a number of organisms (Efstratiadis et al., 1980) has suggested that non coding DNA evolves at a rate of about 1% every 0.6 to 1.3 million years. Assuming an average rate of polymorphism of 1% this means that the sequences in pTP3A and 4A either duplicated, or were genetically matched,very recently ($\sim$ 0.5 million years ago) or are in fact alleles. A gene matching event is unlikely given the large size of the sequence involved (13kb). The G_γ/A_γ gene conversion involved only 1.5kb of DNA. The only non allelic non functional sequences that show extreme sequence conservation are tandemly repeated sequences such as ribosomal genes RNA or the histone genes of sea urchin or *Drosophila*. This may be due to gene matching events that operate by crossing over. Such a system would select for conservation of homology. Histone genes that are not tandemly repeated such as the human (Heintz et al., 1981) and mouse (Sittman et al., 1981) do not show conservation of gene order in different clusters or restriction sites around individual genes.

(3) Does a difference in the amino acid sequence encoded in pTP3A and pTP4A mean they are different genes?

The coding region in pTP3A encodes an amino acid sequence that has previously been described both by amino acid sequencing of purified protamine and also (from amino acid residue 7 on) by nucleotide sequencing of cDNA clones. The amino acid sequence coded for has been termed CII by Gedamu et al., (1981a). In pTP4A amino acid residue 17 is changed from a proline (in pTP3A) to an alanine residue by a C to G transversion. The corresponding codons are CCC (proline, pTP3A) and GCC (alanine, pTP4A). The resulting amino acid sequence has not been described in *Salmo gairdneri* protamine either by amino acid sequencing or by nucleotide sequencing of cDNA clones. All such sequences show a proline residue at an analogous position. This suggests that the sequence i pTP4A may represent a mutational event. Such a mutation may not be deleterious to the fish carrying it as although the proline residue is conserved in *Salmo gairdneri* protamines comparison with all known fish protamine amino acid sequences (Fig. 31) show it is not ubiquitous. In fact an amino acid sequence containing an alanine at this position is found in iridine II. This sequence is virtually identical to the sequence encoded in pTP4A. The iridines are the protamines of *Salmo irideus* which is classified as being taxonomically synonymous with *Salmo gairdneri*. Amino acid sequence comparisons of different protamines therefore show that not only is the proline residue not absolutely conserved in the molecule but that an identical amino acid substitution is found in a very closely related species. It is possible that such a change would be maintained as a gene polymorphism in *Salmo gairdneri* rather than being eliminated by selection. The previous discovery of an analogous sequence in *Salmo irideus* but not *gairdneri* may possibly be due to a higher frequency of such a polymorphism in the former species

Fig.31 Fish protamine amino acid sequences

Thynnin Y2[1]	Pro - Arg_3 - ArgGlnAlaSerArgProValArgArg_4TyrArgArgSerThrAlaAlaArgArg_4Val ValArg_4
Thynnin Z[2]	Pro - Arg_3 - ArgArgSerSerArgProValArgArg_4TyrArgArgSerThrValAlaArgArg_4Val ValArg_4
Clupeine Y1[3]	Ala - Arg_3 - ArgSerSerSerArgProIle - Arg_4ProArgArgThrThr - - Arg_4AlaGlyArg_4
Clupeine Y2[3]	Pro - Arg_3ThrArgArgAlaSerArgProVal - Arg_4ProArgArg - ValSer - - Arg_4Ala - Arg_4
Clupeine Z[4]	AlaArgArg_3SerArgArgAlaSerArgProVal - Arg_4ProArgArg - ValSer - - Arg_4Ala - Arg_4
Salmine A1[5]	ProArg_3 - - SerSerSerArgProValArgArg_4ProArg - - ValSerArgArgArg_4GlyGlyArg_4
Iridine Ia[5]	ProArg_3 - - SerSerSerArgProValArgArg_4ProArgArg - ValSerArgArgArg_4GlyGlyArg_4
Iridine Ib[5]	ProArg_3ArgArgSerSerSerArgProIle - Arg_4ProArgArg - ValSer - ArgArg_4GlyGlyArg_4
Iridine II[5]	ProArg_3 - - SerSerSerArgProVal - Arg_4AlaArgArg - ValSerArgArgArg_4GlyGlyArg_4
Protamine CI[6]	ProArg_3 - Arg - AlaSerArgArgValArgArg_4ProArg - - ValSer - ArgArg_4GlyGlyArg_4
Protamine CII[6]	ProArg_3 - - SerSerSerArgProValArgArg_4ProArg - - ValSerArgArgArg_4GlyGlyArg_4
Protamine CIII[6]	ProArg_3 - - - AlaSerArgProValArgArg_4ProArg - - ValSer - ArgArg_4GlyGlyArg_4
pTP4A	ProArg_3 - - SerSerSerArgProValArgArg_4AlaArg - - ValSerArgArgArg_4GlyGlyArg_4

1.Bretzel,(1972a)
2.Bretzel,(1973)
3.Ando and Suzuki,(1966)
4.Ando and Suzuki,(1967)
5.Ando and Watanabe,(1969)
6.Gedamu et al.,(1981a)

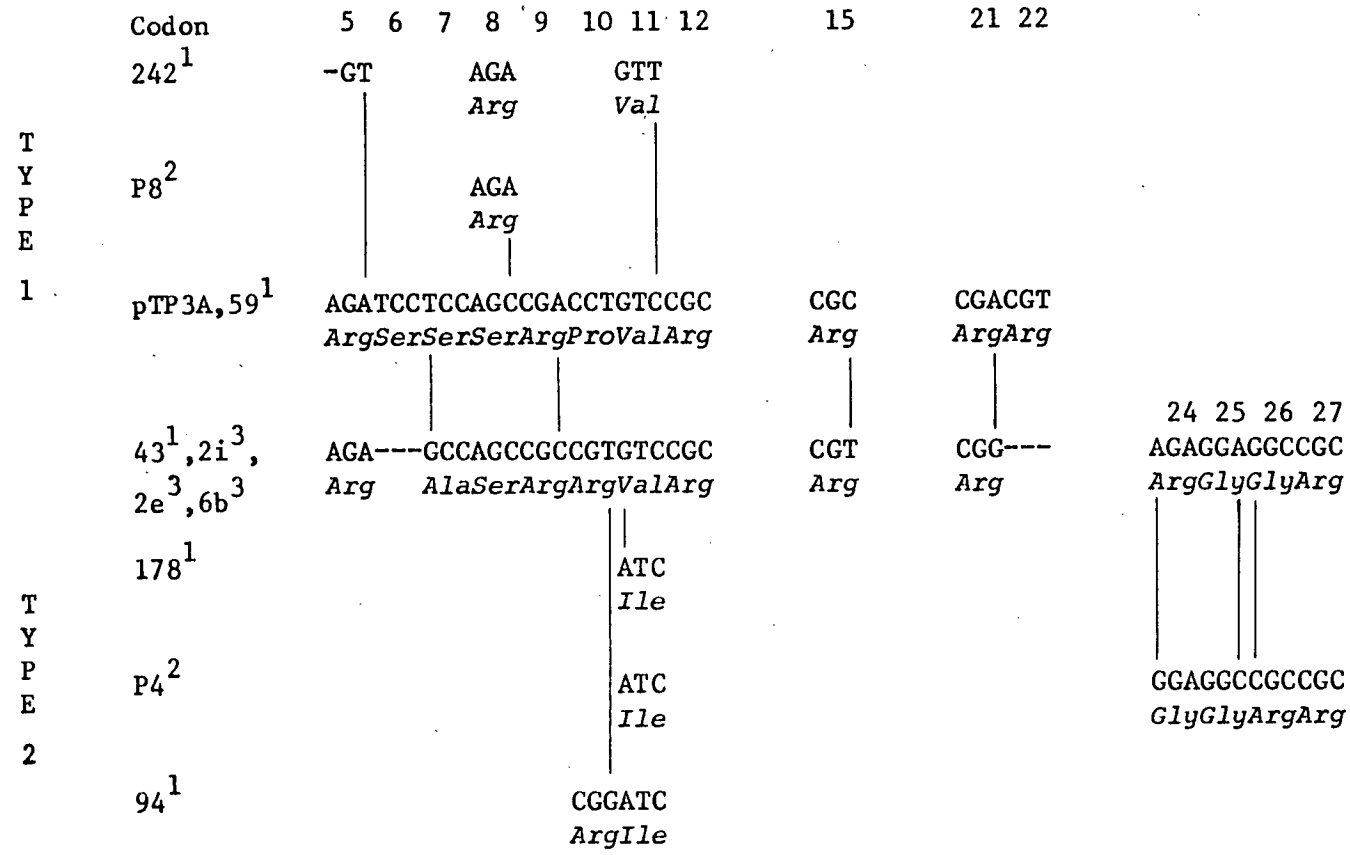
rather than the presence of a novel gene. Alternatively the two sequences in pTP3A and pTP4A may represent different genes. However such a change does not argue absolutely for the sequences being different genes.

In conclusion the evidence would appear to suggest that two two genes may in fact be allelic. However without cloning and analysis of the remaining protamine genes in *Salmo gairdnerii* and accurate determination of gene repetition frequency the possibility that they represent two members of a repeated gene family remains open.

Sequence conservation in protamine gene sequences

Protamine amino acid sequences show that their structure is very similar. Analysis of genomic and cDNA sequences show that this similarity is also found at the nucleotide level and extends to the 3' non translated part of the gene sequence. The main area of sequence variation occurs between amino acid residues 5 and 11 (Fig. 32). This area corresponds to the major site of serine phosphorylation of the protamines. The sequence variation in this region allow the cDNA sequences two be divided into two types (Fig. 2). The partition of the sequences into two groups is reinforced by analysis of the sequences outside the region of most variation. As well as being one amino acid shorter than type 1 sequences in the above region type 2 sequences are also an amino acid shorter in the carboxy terminal half of the coding sequences (residue 22 in the type 1 sequence is deleted in type 2 sequences).

Apart from this change the amino acid sequences,with one exception, are identical in this region. However type 1 and 2 sequences can be distinguished by use of synomonous arginine codons at amino acid residues 15 and 21 (numbered in type 1 sequence). At residue 15 the type 1



1. Gedamu et al.,(1981a)
2. Jenkins,(1979)
3. Sakai et al.,(1981)

Fig.32 Coding region sequence variation in protamine genes

sequence has a CGC codon, the type 2 sequence a CGT codon. At residue 21 the type 1 sequence has a CGA codon, the corresponding type 2 codon is CGC. In addition to this variation between types there is a novel type 2 variant in the cDNA pTP4A (Jenkins, 1979) sequence. While all other type 2 sequences encode Arg_5 Gly_2 Arg_5 in the carboxy terminal part of the protamine molecule pTP4A encodes Arg_4 Gly_2 Arg_5. In addition all type 1 cDNA sequences terminate at the end (amino terminal/5') of codon six or in the first two base pairs of codon five. Only a single type 2 cDNA sequence terminates at an analogous point. The remainder include a complete coding sequence. Some also extend into the 5' non translated part of the mRNA sequence. Comparison of the pTP3A/4A gene sequences with full length type 2 cDNA clones suggest that the amino acid and nucleotide sequence of the amino terminal part of the molecule is the same in both type 1 and type 2 sequences. The 3' non translated part of the gene sequence can also be used to divide the sequences into the same two types. However there is one exception to the type 1 and 2 gene sequences. This is the cDNA pTP11 (Jenkins, 1979) sequence which has a type 1 coding sequence and a type 2 non translated 3' sequence. This could represent a gene sequence that arose by recombination between a type 1 and type 2 gene with the recombination occurring in the identical sequences in the last ten codons of the coding regions. The pTP11 sequence could also be a cloning artifact resulting from template switch during the first stage of cDNA synthesis. This could be caused by a RNAase activity degrading the first mRNA template allowing the cDNA to reinitiate on a second mRNA.

The sequence conservation of the protamine gene 3' non translated sequences at the nucleotide level is surprising. Only the β globin mRNA sequences of higher primates show a similar sequence conservation (Martin et al., 1981). Presumably such conservation of nucleotide sequence reflects

a functional role for the 3' non translated part of the protamine mRNA. The sequence homology is found not only within the two sequence types but also between the two types. After alignment sequences representative of the two types (clones pRTP43 and 59, Gedamu et al., 1981a) show a nucleotide substitution rate of 0.034. The function of the 3' non translated sequence may be represented by a secondary structure requirement. The sequences of protamine mRNA ribonuclease T1 resistant oligoribunucleotides (Davies et al., 1979) show that they are all derived from the 3' non translated sequence (Fig. 33). T1 resistant oligoribonucleotides representative of both gene sequence types are found. One possible functional role for the sequence conservation/ secondary structure of this part of the mRNA may to be allow mRNA storage in the cytoplasmic ribonucleoprotein particles found in trout testis cells (Gedamu et al., 1977a). RNA from these particles only directs the synthesis of protamine in a wheat germ translation system. RNA storage in these particles therefore appears to be specific. This specificity presumably resides in both the protein and mRNA parts of the particles. This storage system may also select for nucleotide sequence conservation of the coding region of the mRNA. Alternatively the whole sequence may be under some other form of nucleotide sequence variation restraint. One probable restraint on the coding usage is selection for preferential codon usage. The high GC content of the coding region may also make it more evolutionary stable as GC rich sequences appear to mutate less rapidly than AT rich sequences (Smithies et al., 1981).

Evolution of the protamine gene family

Comparison of known cDNA sequences shows that between them they describe six different amino acid sequences. The number of different coding region sequences is increased to seven by a nucleotide substitution

Fig.33 Ribonuclease T1 resistant oligoribonucleotides from protamine mRNA.

Type 1 gene sequence $TAGN_{10}GTAGAACCTACCTGA\overset{A}{C}CTA\overset{T}{A}CCGCCCCCTCCGGGTTC\overset{A}{T}CCCTCCCGACCCCTGGT\overset{G}{A}GTGN_{45}(A)_N$

T1 28,29 AACCTACCTGACCTATCCGCCCCCTCCGG

T1 54 AACCTACCTGACCTATCCGCCCCCTCCGGGCTCTCCCTCCCGACCCCTGXNNG

Type 2 gene sequence $TAGN_9GTA\overset{G}{A}C\overset{T}{C}TACCT\overset{A}{G}AACTAACC\overset{A}{G}CCCCCTACCGGTTCTCCCTCCAGACTN_{61}(A)_N$

T1 31,32 AACTAACCGCCCCCTACCGGTTCTCCCTCCAG

T1 43 TAACCTACCTGAACTAACCGCCCCCTACCGGTTCTCCCTCCAG

Internal G residues are underlined

Type 1 and 2 gene sequences are from fig.2

T1 oligonucleotide sequences are from Davies et al., (1979)

X Purine

TAG Coding region termination codon

in the wobble base at codon 10 (Fig.32) between the cDNA sequences 178 and 94. However there are also amino acid sequences for protamines which have not been described as cDNA nucleotide sequences. These are the CI type protamine and a third variant of CIII type protamine containing a proline residue at position 10. This increase the number of amino acid sequences described to eight. In addition the pTP4A sequence described in this study also encodes a ninth protamine sequence. Whether all these sequence variants represent different genes or not is unknown and this can only be determined by cloning all the protamine genes from a single fish. The different sequences are shown in Fig. 34. Also given is the number of coding and 3' non translated base substitutions and deletion insertions. On the basis of these changes an evolutionary scheme for the protamine genes can be proposed (Fig. 35). Again, without cloning and analysis of all protamine genomic sequences such a scheme is only an estimate. With respect to the evolution of the protamine genes it would be interesting to determine the number and structure of the clupeine genes of herring. The *Salmonidae*, of which the rainbow trout is a member, apparently arose from the *Clupeidae* by a tetraploidisation event. The clupeine gene structure and number should therefore allow the determination of a more detailed evolutionary scheme for the protamine genes and division of the evolutionary history of the genes into pre and post tetraploidisation events.

Conclusion

In this study two protamine genes from the rainbow trout (Salmo gairdnerii) were cloned from a lambda genomic library. The structural analysis of these genes show that they contain no intervening sequences. One clone, CH4A/TP3A, contains a protamine gene which encodes a previously described amino acid sequence. In addition a cDNA sequence

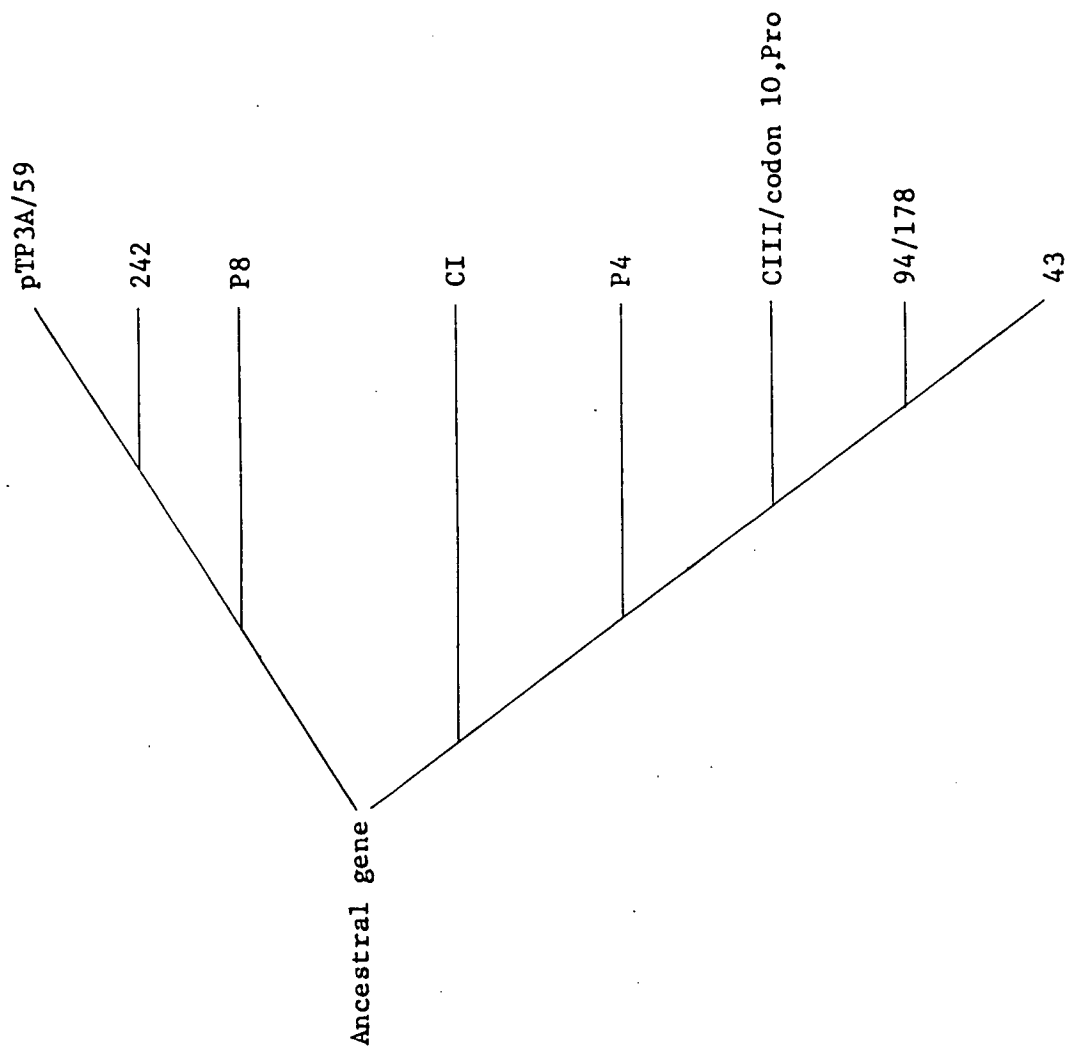
Fig.34 Different protamine genes deduced from nucleotide and amino acid sequences.

pTP3A,59[1]	$Pro Arg_4 Ser_3 Arg Pro Val Arg_5 Pro Arg Ser Arg_6 Gly_2 Arg_4$				
242[1],P8[2,5]		$Ser_2 Arg_2$			
CIII[1],43[1],2i[3]	$Pro Arg_4 Ala Ser Arg_2 Val Arg_5 Pro Arg Val Ser Arg_5 Gly_2 Arg_4$				
CIII[1],94[1],178[1,6]			Ile		
P4[2]			Ile		$Arg_4 Gly_2 Arg_5$
CIII[1,4,7]			Pro		
CI[4]		Arg_5	Val		

1. Gedamu et al.,(1981a).
2. Jenkins,(1979).
3. Sakai et al.,(1981)
4. Descibed only as amino acid sequences
5. 242 and P8 differ by two wobble position base substitutions (codons 5 and 11)and by three base substitutions and three deletion/insertion events in the 3' non translated region.
6. 94 and 178 differ by one wobble position base substitution (codon 10) and by one base substitution in the 3' non-translated region.
7. An Ile or Val to Pro amino acid substitution involves at least two base substitutions.

Fig.35 Evolution of the protamine gene family.

Arranged by minimum change between adjacent genes(see figs 31 and 33)
Distances between branch points is not to scale.



which is identical to the CH4A/TP3A sequence in the coding (translated region) has previously been described. The second clone CH4A/TP3A although almost identical to the other in structure, contains a novel protamine gene sequence. The similarity in structure of the two clones suggest that they are either alleles showing polymorphisms or very closely related repeated genes.

REFERENCES

ALT, F.W., ROSENBERG, N. , CASANOVA, R.J., THOMAS, E. and BALTIMORE,D. (1982). Immunoglobulin heavy chain expression and class switching in a murine leukemia cell line. Nature 296, 325-331.

ANDO, T. and HASHIMOTO, C. (1958). Studies on protamines. V. Changes of the proteins in the cell nuclei of the testic during the formation of spermatozoa of the rainbow trout. Journal Biochemistry (Tokyo) 45, 529-540.

ANDO T. and SUZUKI, K. (1966). The amino acid sequence of the second component of clupeine. B.B.A. 121, 427-429.

ANDO, T. and SUZUKI, K. (1967). The amino acid sequence of the third component of clupeine. B.B.A. 140, 375-377.

ANDO, T. and Watanabe, S. (1969). A new method for fractionation of protamines and the amino acid sequences of one component of salmine and three components of iridene. International Journal Protein Research 1, 221-224.

ANTONARAKIS, S.E., BOEHM, C.D., GIARDINA, P.J.V. and KAZAZIAN, H.H. (1982). Non random association of polymorphic restriction sites in the B-globin gene cluster. P.N.A.S. (USA) 79, 137-141.

AVIV, H. and LEDER, P. (1972). Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid cellulose. P.N.A.S. (USA) 69, 1408-1412.

BAILEY, G.S., COCKS, G.T. and WILSON, A.C. (1969). Gene duplication in fishes: Malate dehydrogenases of salmon and trout. B.B.R.C. 34, 605-612.

BAILEY, G.S. and WILSON,A.C.(1968). Homologies between isoenzymes of fishes and verterbraetes. J.B.C. 243, 5843-5853.

BAIRD, M., DRISCOLL, C., SCHREINER, H., SCIARRATTA, G.V., SANSONE, G., NIAZI, G., RAMIREZ, F. and BANK,A. (1981). A nucleotide change at a splice junction in the human B globin gene is associated with B^{O} thalassemia. P.N.A.S. (USA) 78, 4218-4221.

BALHORN, R. (1982). A model for the structure of chromatin in mammalian sperm. Journal cell Biology 93, 298-305.

BANERJI, J., RUSCONI, S. and SCHAFFNER, W. (1981). Expression of a B globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299-308.

BARNES, W.M. (1977). Plasmid detection and sizing in single colony lysates. Science 195, 393-324.

BELLVE, A.R., ANDERSON, E. and HANLEY-BOWDOIN, L. (1975). Synthesis and amino acid composition of basic proteins in mammalian sperm nuclei. Developmental Biology 47, 349-365.

BELLVE, A.R. and CARRAWAY, R. (1978). Characterization of two basic chromosomal proteins isolated from mouse spermatozoa. Journal Cell Biology 79, Abstracts, 177, G1006.

BENTON, W.D. and DAVIS, R.W. (1977). Screening λgt recombinant clones by hybridization to single plaques in situ. Science 196, 180-182.

BERK, A.J., and SHARP, P.A. (1978). Spliced early mRNAs of simian virus 40. P.N.A.S. (USA) 75, 1274-1278.

BIRCHMEIER, C., GROSSCHEDL,R. and BIRNSTEIL, M.L. (1982). Generation of authentic 3' termini of an H2A mRNA in vivo is dependent on a short inverted repeat and on spacer sequences. Cell 28, 739-745.

BISHOP, J.M. (1981). Enemies within: the genesis of retrovirus oncogenes. Cell 23, 5-6.

BISHOP, J.O. (1979). A DNA sequence cleaved by restriction endonuclease R. EcoRI in only one strand. J.M.B. 128, 545-559.

BISHOP, J.O., PEMBERTON, R. and BAGLIONI, C. (1972). Reiteration frequency of haemoglobin genes in the Duck. Nature New Biology 235, 231-234.

BLAKE, C.C.F. (1979). Exons encode protein functional units. Nature 277, 598.

BLATTNER, F.R., WILLIAMS, B.G., BLECHL, A.E., DENNISTON-THOMPSON, K., FABER, H.E., FURLONG, L.A., GRUNWALD, D.J., KIEFER, D.O., MOORE, D.D., SCHUMM, J.W., SHELDON, E.L. and SMITHIES, O. (1977). Charon phages: Safer derivatives of bacteriophage lambda for DNA cloning. Science 196, 161-169.

BOLS, N.C., BOLISKA, S.A., RAINVILLE, and KASINSKY, H.E. (1980). Nuclear basic protein changes during spermiogenesis in the long-nose skate and spiny dogfish. Journal Experimental Zoology 212, 423-433.

BOUCHÉ, J.P. (1981). The effect of spermidine on endonuclease inhibition by agarose contaminants. Analytical Biochemistry 115, 42-45.

BOYER, H.W. and ROULLAND-DUSSOIX, D. (1969). A complementation analysis of the restriction and modification of DNA in Escherica coli. JMB 41, 459-472.

BRACK, C., HIRAMA, M., LENHARD-SCHULLER, R. and TONEGAWA, S. (1978). A complete immunoglobulin gene is created by somatic recombination. Cell 15, 1-14.

BREATHNACH, R. and CHAMBON, P. (1981). Organization and expression of eucaryotic split genes coding for proteins. Annual Review of Biochemistry 50, 349-383.

BRETZEL, G. (1972a). Über thynnin, das protamin des Thunfisches. Die vollständige aminosauresequenz von thynnin Y2. Hoppe-Seyler's Zeitschrift für physiologische chemie 353, 933-943.

BRETZEL, G. (1972b). Über thynnin, das protamin des Thunfisches. Die sequenz der komponente Y1. Hoppe-Seyler's Zeitschrift für physiologische chemie.353, 1362-1364.

BRETZEL, G. (1973). Über thynnin, das protamin des Thunfisches. Die aminosäuresequenz von thynnin Z1. Hoppe-Seyler's Zeitschrift für physiologische chemie. 354, 312-320.

BRINSTER, R.L., CHEN, H.Y., WARREN, R., SARTHY, A. and PALMITER, R.D. (1982). Regulation of metallothionein-thymidine kinase fusion plasmids injected into mouse eggs. Nature 296, 39-42.

BRITTEN, R.J. and KOHNE, D.E. (1968). Repeated sequences in DNA. Science 161, 529-540.

BROWN, D.D., WENSINK, P.C. and JORDAN, E. (1972). A comparison of the ribosomal DNAs of Xenopus laevis and Xenopus mulleri: the evolution of tandem genes. J.M.B. 63, 57-73.

BROWN, W.M., GEORGE, M. and WILSON, A.C. (1979). Rapid evolution of animal mitochondrial DNA. P.N.A.S. (USA) 76, 1967-1971.

BUETTI, E. and DIGGELMANN, H. (1981). Cloned mouse mammary tumor virus DNA is biologically active in transfected mouse cells and its expression is stimulated by glucocorticoid hormones. Cell 23, 335-345.

BUNICK, D., ZANDOMENI, R., ACKERMAN, S. and WEINMANN, R. (1982). Mechanism of RNA polymerase II - specific initiation of transcription in vitro: ATP requirement and uncapped runoff transcripts. Cell 29, 877-886.

BUSSLINGER, M., MOSCHONAS, M. and FLAVELL, R.A. (1981). β^+ thalassemia: Aberrant splicing results from a single point mutation in an intron. Cell 27, 289-298.

CAMERON, J.R., PHILIPPSEN, P. and DAVIS, R.W. (1977). Analysis of chromosomal intergration and deletions of yeast plasmids. N.A.R. 4, 1429-1448.

CHENG, H.L., BLATTNER, F.R. , FITSMAURICE, L., MUSHINSKI, J.F. and TUCKER, P.W. (1982). Structure of genes for membrane secreted murine IgD heavy chains. Nature 296, 410-415.

CHIRGWIN, J.M., PRZYBYLA, A.E., MACDONALD, R.J., and RUTTER, W.J. (1979) Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry 18, 5294-5299.

CLARKE, L. and CARBON, J. (1976). A colony bank containing synthetic Col El hybrid plasmids representative of the entire E. coli genome. Cell 9, 91-99.

COCHET, M., GANNON, F., HEN, R., MAROTEAUX, L., PERRIN, F., and CHAMBON, P. (1979). Organisation and sequence studies of the 17-piece chicken conalbumin gene. Nature 282, 567-574.

COELINGH, J.P., MONFOORT, C.H., ROZIJN, T.H., GEVERS-LEUVEN, J.A., SCHIPHOF, R., STEYN-PARVÉ, E.P., Braunitzer, G., SCHRANK, B. and RUHFUS, A. (1972). The complete amino acid sequence of the basic nuclear protein of bull spermatozoa. B.B.A. 285, 1-14.

COELINGH, J.P., ROZIJN, T.H. and MONFOORT, C.H. (1969). Isolation and characterization of a basic protein from bovine sperm heads. B.B.A. 188, 358-356.

COLLINS, J. and HOHN, B. (1978). Cosmids: A type of plasmid gene cloning vector that is packageable in vitro in bacteriophage lambda heads. P.N.A.S. (USA) 75, 4242-4246.

COMB, M., HERBERT, E. and CREA, R. (1982). Partial characterization of the mRNA that codes for enkephalins in bovine adrenal medulla and human pheochromocytoma. P.N.A.S. (USA) 79, 360-364.

COMPERE, S.J. and PALMITER, R.D. (1981). DNA methylation controls the inducibility of the mouse metallothionein-1 gene in lymphoid cells. Cell 25, 233-240.

CORDEN, J., WASYLYK, B., BUCHWALDER, A., SASSONE-CORSI, P., KEDINGER, C. and CHAMBON, P. (1980). Promoter sequences of eukaryotic protein coding genes. Science 209, 1407-1414.

CORNEO, G., GINELLI, E. and POLLI, E. (1971). Renaturation properties and localization in the heterochromatin of human satellite DNAs. B.B.A. 247, 528-534.

CRAIK, C.S., BUCHMAN, S.R. and BEYCHOK, S. (1980). Characterization of globin domains: Heme binding to the central exon product. P.N.A.S. (USA) 77, 1384-1388.

DAVIDSON, E.H. and BRITTEN, R.J. (1973). Organization, transcription and regulation in the animal genome. Quarterly Review Biology 48, 565-613.

DAVIDSON, E.H., HOUGH, B.R., AMENSON, C.S. and BRITTEN, R.J. (1973). General interspersion of repetitive with non-repetitive sequence elements in the DNA of Xenopus. J.M.B. 77, 1-23.

DAVIDSON, E.H., HOUGH, B.R., KLEIN, W.H. and BRITTEN, R.J. (1975). Structural genes adjacent to interspersed repetitive DNA sequences. Cell 4, 217-238.

DAVIES, P.L., DIXON, G.H., FERRIER, L.N., GEDAMU, L. and IATROU, K. (1976). The structure and function of protamine mRNA from developing trout testis. Progress in Nucleic Acid Research and Molecular Biology 19, 135-155.

DAVIES, P.L., DIXON, G.H., SIMONCSITS, A. and BROWNLEE, G.G. (1979). Sequences of large T1 ribonuclease-resistant oligoribonucleotides from protamine mRNA: the overall architecture of protamine mRNA. N.A.R. 7, 2323-2345.

DAVIES, P.L., FERRIER, L.N. and DIXON, G.H. (1977). Sequence analysis of protamine mRNA from the rainbow trout. J.B.C. 252, 1386-1393.

DIXON, G.H. (1972). The basic proteins of trout testis chromatin: Aspects of their synthesis, post synthetic modifications and binding to DNA. Karolinska Symposia on Research Methods in Reproductive Endocrinology. 5th Symposium. Gene Transcription in Reproductive Tissue, pp.

DUGAICZYK, A., BOYER, H.W. and GOODMAN, H.M. (1975). Ligation of EcoRI endonuclease generated DNA fragments into linear and circular structures. J.M.B. 96, 171-184.

EARLY, P., ROGERS, J., DAVIS, M., CALAME, K., BOND, M., WALL, R. and Hood, L. (1980). Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. Cell 20, 313-319.

EFSTRATIADIS, A., POSAKONY, J.W., MANIATIS, T., LAWN, R.M., O'CONNELL, C., SPRITZ, R.A., DeRIEL, J.K., FORGET, B.G., WEISSMAN, S.M., SLIGHTOM, J.L., BLECHL, A.E., SMITHIES, O., BARALLE, F.E., SHOULDERS, C.C. and PROUDFOOT, N.J. (1980). The structure and evolution of the human β-globin gene family. Cell 21, 653-668.

ENGEL, J.D., SUGARMAN, B.J. and DODGSON, J.B. (1982). A chicken histone H3 gene contains intervening sequences. Nature 297, 434-436.

FAHRNER, K., YARGER, J. and HEREFORD, L. (1980). Yeast histone mRNA is polyadenylated. N.A.R. 8, 5725-5737.

FAVALORO, J., TREISMAN, R. and KAMEN, R. (1980). Transcription maps of Polyoma Virus-Specific RNA: Analysis by two-dimensional nuclease S1 gel mapping. Methods in Enzymology 65, 718-749. (Ed. Grossman, L. and Moldave, K.).

FAYE, G., LEUNG, D.W., TATCHELL, K., HALL, B.D. and SMITH, M. (1981). Deletion mapping of sequences essential for in vivo transcription of the iso-1-cytochrome c gene. P.N.A.S.(USA) 78, 2258-2262.

FERRIER, L.N., DAVIES, P.L. and DIXON, G.H. (1977). Protamine messenger RNA from rainbow trout testis contains the nucleotide sequence AAUAAA in an untranslated region. B.B.A. 479, 460-470.

FITZGERALD, M. and SHENK, T. (1981). The sequence 5'-AAUAAA-3' forms part of the recognition site for polyadenylation of late SV40 mRNAs. Cell 24, 251-260.

GALAU, G.A. (1974). A measurement of the sequence complexity of polysomal messenger RNA in sea urchin embryos. Cell 2, 9-20.

GEDAMU, L., CHACONAS, G., van de SANDE, J.H. and DIXON, G.H. (1981b). Studies on the heterogeneity of the 5' ends of the protamine mRNAs from rainbow trout testis. Bioscience Reports 1, 61-70.

GEDAMU, L. and DIXON, G.H. (1976). Purification and properties of biologically active rainbow trout testis protamine mRNA. J.B.C. 251, 1455-1463.

GEDAMU, L. and DIXON, G.H. (1979). Heterogeneity of biologically active deadenylated protamine mRNA components isolated from rainbow trout testis. N.A.R. 6, 3661-3672.

GEDAMU, L., DIXON, G.H. and DAVIES, P.L. (1977a). Identification and isolation of protamine messenger ribonucleoprotein particles from rainbow trout testis. Biochemistry 16, 1383-1391.

GEDAMU, L., IATROU, K. and DIXON, G.H. (1977b). Isolation and characterization of trout testis protamine mRNAs lacking poly(A). Cell 10, 443-451.

GEDAMU, L., IATROU, K. and DIXON, G.H. (1979). Translation of partially purified poly(A)$^+$ protamine messenger RNA components in wheat germ and rabbit reticulocyte cell free systems. Evidence for translational control mechanisms. B.B.A. 562, 481-494.

GEDAMU, L., WOSNICK, M.A., CONNOR, W., WATSON, D.C., DIXON, G.H. and IATROU, K. (1981a). Molecular analysis of the protamine multigene family in rainbow trout testis. N.A.R. 9, 1463-1482.

GHANGAS, G.S. and WU, R. (1975). Specific hydrolysis of the cohesive ends of bacteriophage λDNA by three single strand-specific nucleases. J.B.C. 250, 4601-4606.

GILBERT, W. (1978). Why genes in pieces. Nature 271, 501.

GOJOBORI, T., LI, W.H. and GRAUR, D. (1982). Patterns of nucleotide substitutions in pseudogenes and functional genes. Journal of Molecular Evolution 18, 360-369.

GOLDBERG, R.B., CRAIN, W.R., RUDERMAN, J.V., MOORE, G.P., BARNETT, T.R., HIGGINS, R.C., GELFAND, R.A., GALAU, G.A., BRITTEN, R.J. and DAVIDSON, E.H. (1975). DNA sequence organization in the genome of five marine invertebrates. Chromosoma 51, 225-251.

GOLDBERG, R.B., GALAU, G.A., BRITTEN, R.J. and DAVIDSON, E.H. (1973). DNA sequence representation in sea urchin embryo messenger RNA. P.N.A.S.(USA) 70, 3516-3520.

GORSKI, J., MORRISON, M.R., MERKEL, C.G. and LINGREL, J.B. (1974). Size heterogeneity of polyadenylate sequences in mouse globin messenger RNA. J.M.B. 86, 363-371.

GORSKI, J., MORRISON, M.R., MERKEL, C.G. and LINGREL, J.B. (1975). Poly(A) size class distribution in globin mRNA as a function of time. Nature 253,749-751.

GRAHAM, D.E., NEUFELD, B.R., DAVIDSON, E.H. and BRITTEN, R.J. (1974). Interspersion of repetitive and non-repetitive DNA sequences in the Sea Urchin genome. Cell 1, 127-138.

GRANTHAM, R., GAUTIER, C., GOUY, M., MERCIER, R. and PAVÉ, A. (1980). Codon catalogue usage and the genome hypothesis. N.A.R. 8, r49-r62.

GREEN, P.J., BETLACH, M.C., BOYER, H.W. and GOODMAN, H.M. (1974). The EcoRI restriction endonuclease. In Methods in Molecular Biology Series: DNA Replication and Biosynthesis 7, 87-111. (Ed. Wickner, R.B.).

GREENE, P.J., POONIAN, M.S., NUSSBAUM, A.L., TOBIAS, L., GARFIN, D.E., BOYER, H.W. and GOODMAN, H.M. (1975). Restriction and modification of a self-complementary octanucleotide containing the EcoRI substrate. J.M.B. 99, 237-261.

GROSSCHEDL, R. and BIRNSTEIL, M.L. (1980). Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. P.N.A.S. (USA) 77, 1432-1436.

GROSSCHEDL, R., and BIRNSTEIL, M.L. (1981). Delimitation of far upstream sequences required for maximal in vitro transcription of an H2A histone gene. P.N.A.S.(USA) 78, 297-301.

GROSVELD, F.G., DAHL, H-H.M, de BOER, E. and FLAVELL, R.A. (1981). Isolation of β-globin-related genes from a human cosmid library. Gene 13, 227-237.

GROSVELD, G.C., KOSTER, A. and FLAVELL, R.A. (1981). A transcription map for the rabbit β-globin gene. Cell 23, 573-584.

GROSVELD, G.C., SHEWMAKER, C.K., JAT, P. and FLAVELL, R.A. (1981). Localization of DNA sequences necessary for transcription of the rabbit β globin gene in vitro. Cell 25, 215-226.

GROUDINE, M. and WEINTRAUB, H. (1982). Propagation of globin DNAase I hypersensitive sites in absence of factors required for induction: A possible mechanism for determination. Cell 30, 131-139.

HAMER, D.H., KAEHLER, M. and LEDER, P. (1980). A mouse globin gene promoter is functional in SV40. Cell 21, 697-708.

HANAHAN, D. and MESELSON, M. (1980). Plasmid screening at high colony density. Gene 10, 63-67.

HARRISON, P.R., HELL, A., BIRNIE, G.D. and PAUL, J. (1972). Evidence for single copies of globin genes in the mouse genome. Nature 239, 219-221.

HEINTZ, N., ZERNIK, M. and ROEDER, R.G. (1981). The structure of the human histone genes: Clustered but not tandemly repeated. Cell 24, 661-668.

HENIKOFF, S., TATCHELL, K., HALL, B.D. and NASMYTH, K.A. (1981). Isolation of a gene from Drosophila by complementation in yeast. Nature 289, 33-37.

HENTSCHEL, C.C. and BIRNSTEIL, M.L. (1981). The organization and expression of histone gene families. Cell 25, 301-314.

HEREFORD, L.M. and ROSBASH, M. (1977). Number and distribution of polyadenylated RNA sequences in yeast. Cell 10, 453-462.

HIGGS, D.R., GOODBOURN, S.E.Y., WAINSCOAT, J.S. and CLEGG, J.B. (1981). Highly variable regions of DNA flank the human α globin genes. N.A.R. 9, 4213-4224.

HOFER, E. and DARNELL, J.E. (1981). The primary transcription unit of the mouse β-major globin gene. Cell 23, 585-593.

HOFER, E., HOFER-WARBINEK, R. and DARNELL, J.E. (1982). Globin RNA transcription: A possible termination site and demonstration of transcriptional control correlated with altered chromatin structure. Cell 29, 887-893.

HOHN, B. and MURRAY, K. (1977). Packaging recombinant DNA molecules into bacteriophage particles in vitro. P.N.A.S.(USA) 74, 3259-3263.

HOLLIS, G.F., HIETER, P.A., McBRIDE, O.W., SWAN, D. and LEDER, P. (1982). Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. Nature 296, 321-325.

HUANG, A.L., OSTROWSKI, M.C., BERARD, D. and HAGER, G.L. (1981). Glucocorticoid regulation of the Ha-MuSV p21 gene conferred by sequences from mouse mammary tumor virus. Cell 27, 245-255.

HYNES, N.E., KENNEDY, N., RAHMSDORF, U. and GRONER, B. (1981). Hormone-responsive expression of an endogenous proviral gene of mouse mammary tumor virus after molecular cloning and gene transfer into cultured cells. P.N.A.S.(USA) 78, 2038-2042.

IATROU, K. and DIXON, G.H. (1977). The distribution of poly(A)$^+$ and poly(A)$^-$ protamine messenger RNA sequences in the developing trout testis. Cell 10, 433-441.

IATROU, K., SPIRA, A.W. and DIXON, G.H. (1978). Protamine messenger RNA: Evidence for early synthesis and accumulation during spermatogenesis in rainbow trout. Developmental Biology 64, 82-98.

INGLES, C.J. and DIXON, G.H. (1967). Phosphorylation of protamine during spermatogenesis in trout testis. P.N.A.S.(USA) 58, 1011-1018.

JEFFREYS, A.J. (1979). DNA sequence variants in the G_γ-, A_γ-, δ- and β-globin genes of man. Cell 18, 1-10.

JELINEK, W.R., TOOMEY, T.P., LEINWAND, L., DUNCAN, C.H., BIRO, P.A., CHOUDARY, P.V., WEISSMAN, S.M., RUBIN, C.M., HOUCK, C.M., DEININGER, P.L. and SCHMID, C.W. (1980). Ubiquitous, interspersed repeated sequences in mammalian genomes. P.N.A.S.(USA) 77, 1398-1402.

JENKINS, J.R. (1979). Sequence divergence of rainbow trout protamine mRNAs; comparison of coding and non-coding nucleotide sequences in three protamine cDNA plasmids. Nature 279, 809-811.

JENKINS, J.R., BISHOP, J.O. and BUTTERWORTH, P.H. (1979). Molecular cloning of three major sequence species from rainbow trout protamine mRNA. N.A.R. 6, 3805-3819.

JERGIL, B. and DIXON, G.H. (1970). Protamine kinase from rainbow trout testis. J.B.C. 245, 425-434.

JONES, K.W. (1970). Chromosomal and nuclear location of mouse satellite DNA in individual cells. Nature 225, 912-915.

KARN, J., BRENNER, S., BARNETT, L. and CESARENI, G. (1980). Novel bacteriophage λ cloning vector. P.N.A.S. (USA) 77, 5172-5176.

KEDES, L.H. and BIRNSTEIL, M.L. (1971). Reiteration and clustering of DNA sequences complementary to histone messenger RNA. Nature New Biology 230, 165-169.

KEDES, L.H., COHN, R.H., LOWRY, J.C., CHANG, A.C.Y. and COHEN, S.N. (1975). The organization of sea urchin histone genes. Cell 6, 359-369.

KEENE, M.A., CORCES, V., LOWENHAUPT, K. and ELGIN, S.C.R. (1981). DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. P.N.A.S.(USA) 78, 143-146.

KEMP, D.J., CORY, S. and ADAMS, J.M. (1979). Cloned pairs of variable region genes for immunoglobulin heavy chains isolated from a clone library of the entire mouse genome. P.N.A.S.(USA) 76, 4627-4631.

KISTLER, W.S., GEROCH, M.E. and WILLIAMS-ASHMAN, H.G. (1973). Specific basic proteins from mammalian testes. J.B.C. 248, 4532-4543.

KISTLER, W.S., NOYES, C. and HEINRIKSON, R.L. (1974). Partial structural analysis of a highly basic low molecular weight protein from rat testis. B.B.R.C. 57, 341-347.

KRAUS, J.P. and ROSENBERG, L.E. (1982). Purification of low-abundance messenger RNAs from rat liver by polysome immunoadsorption. P.N.A.S.(USA) 79, 4015-4019.

KURTZ, D.T. (1981). Hormone inducibility of rat $\alpha_{2\mu}$ globulin genes in transfected mouse cells. Nature 291, 629-631.

LARSEN, A. and WEINTRAUB, H. (1982). An altered DNA conformation detected by SI nuclease occurs at specific regions in active chick globin chromatin. Cell 29, 609-622.

LASKEY, R.A. and MILLS, A.D. (1975). Enhanced autoradiographic detection of ^{32}P and ^{125}I using intensifying screens and hypersensitized film. FEBS Letters 82, 314-316.

LEDER, A., SWAN, D., RUDDLE, F., D'EUSTACHIO, P. and LEDER, P. (1981). Dispersion of α-like globin genes of the mouse to three different chromosomes. Nature 293, 196-200.

LEDER, P., HANSEN, J.N., KONKEL, D., LEDER, A., NISHIOKA, Y. and TALKINGTON, C. (1980). Mouse globin system: a functional and evolutionary analysis. Science 209, 1336-1342.

LEDER, P., TIEMEIER, D. and ENQUIST, L. (1977). EK2 derivatives of bacteriophage lambda useful in the cloning of DNA from higher organisms: the λgt WES system. Science 196, 175-177.

LEVY, B.W. and DIXON, G.H. (1977a). Changes in the sequence diversity of polyadenylated cytoplasmic RNA during testis differentiation in rainbow trout (Salmo gairdnerii). European Journal Biochemistry 74, 61-67.

LEVY, B. and DIXON, G.H. (1977b). Reiteration frequency of the protamine genes in rainbow trout (Salmo gairdnerii). J.B.C. 252, 8062-8065.

LEWIN, B. (1975). Units of transcription and translation: sequence components of heterogeneous nuclear RNA and messenger RNA. Cell 4, 77-93.

LINDAHL, G., SIRONI, G., BIALY, H. and CALENDAR, R. (1970). Bacteriophage lambda; Abortive infection of bacteria lysogenic for phage P2. P.N.A.S. (USA) 66, 587-594.

LING, V. and DIXON, G.H. (1970). The biosynthesis of protamine in trout testis. II. Polysome patterns and protein synthetic activities during testis maturation. J.B.C. 245, 3035-3042.

LING, V., JERGIL, B. and DIXON, G.H. (1971). The biosynthesis of protamine in trout testis. III. Characterization of protamine components and their synthesis during testis development. J.B.C. 246, 1168-1176.

LOENEN, W.A.M. and BRAMMER, W.J. (1980). A bacteriophage lambda vector for cloning large DNA fragments made with several restriction enzymes. Gene 10, 249-259.

LOUIE, A.J. and DIXON, G.H. (1972). Trout testis cells. I. Characterization by deoxyribonucleic acid and protein analysis of cells separated by velocity sedimentation. J.B.C. 247, 5490-5497.

LOWY, I., PELLICER, A., JACKSON, J.F., SIM, G.K., SILVERSTEIN, S. and AXEL, R. (1980). Isolation of transforming DNA: Cloning the Hamster aprt gene. Cell 22, 817-823.

MANDEL, J.P. and CHAMBON, P. (1979). DNA methylation: organ specific variations in the methylation pattern within and around ovalbumin and other chicken genes. N.A.R. 7, 2081-2103.

MANDEL, M. and HIGA, A. (1970). Calcium-dependent bacteriophage DNA infection. J.M.B. 53, 159-162.

MAKI, R., ROEDER, W., TRAUNECKER, A., SIDMAN, C., WABL, M., RASCHKE, W. and TONEGAWA, S. (1981). The role of DNA rearrangement and alternative RNA processing in the expression of immunoglobulin delta genes. Cell 24, 353-365.

MANIATIS, T., FRITSCH, E.F., LAVER, J. and LAWN, R.M. (1980). The molecular genetics of human haemoglobins. Annual Review of Genetics 14, 145-178.

MANIATIS, T., HARDISON, R.C., LACY, E., LAUER, J., O'CONNELL, C., QUON, D., SIM, G.K. and EFSTRATIADIS, A. (1978). The isolation of structural genes from libraries of eucaryotic DNA. Cell 15, 687-701.

MANIATIS, T., JEFFREY, A. and van de SANDE, H. (1975). Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. Biochemistry 14, 3787-3794.

MANIATIS, T., KEE, S.G., EFSTRATIADIS, A. and KAFATOS, F.C. (1976). Amplification and characterization of a β globin gene synthesised in vitro. Cell 8, 163-182.

MANLEY, J.L., FIRE, A., CANO, A., SHARP, P.A. and GEFTER, M.L. (1980). DNA dependent transcription of adenovirus genes in a soluble whole-cell extract. P.N.A.S.(USA) 77, 3855-3859.

MARTIN, S.L., ZIMMER, E.A., DAVIDSON, W.S., WILSON, A.C. and KAN, Y.W. (1981). The untranslated regions of β-globin mRNA evolve at a functional rate in higher primates. Cell 25, 737-742.

MARUSHIGE, K. and DIXON, G.H. (1969). Developmental changes in chromosomal composition and template activity during spermatogenesis in trout testis. Developmental Biology 19, 397-414.

MARUSHIGE, K. and DIXON, G.H. (1971). Transformation of trout testis chromatin. J.B.C. 246, 5799-5805.

MASSARO, E.J. and MARKERT, C.L. (1968). Isozyme patterns of Salmonoid fishes: Evidence for multiple cistrons for lactate dehydrogenase polypeptides. Journal Experimental Zoology 168, 223-238.

MAURON, A., LEVY, S., CHILDS, G. and KEDES, L. (1981). Monocistronic transcription is the physiological mechanism of sea urchin embryonic histone gene expression. Journal Molecular and Cellular Biology 1, 661-671.

MAXAM, A.M. and GILBERT, W. (1980). Sequencing end labelled DNA with base-specific chemical cleavages. Methods in Enzymology 65, 499-559. (Ed. Grossman, L. and Moldane, K.).

McGHEE, J.D., WOOD, W.I., DOLAN, M., ENGEL, J.D. and FELSENFELD, G. (1981). A 200 base pair region at the 5' end of the chicken adult β-globin gene is accessible to nuclease digestion. Cell 27, 45-55.

McKNIGHT, S.L., GAVIS, E.R., KINGSBURY, R. and AXEL, R. (1981). Analysis of transcriptional regulatory signals of the HSV thymidine kinase gene: Identification of an upstream control region. Cell 25, 385-398.

McKNIGHT, S.L. and KINGSBURY, R. (1982). Transcriptional control signals of a eukaryotic protein-coding gene. Science 217, 316-324.

McMASTER-KAYE, R. and KAYE, J.S. (1976). Basic protein changes during the final stages of sperm maturation in the house cricket. Experimental Cell Research 97, 378-386.

MELLON, P., PARKER, V., GLUZMAN, Y. and MANIATIS, T. (1981). Identification of DNA sequences required for transcription of the human α1-globin gene in a new SV40 host-vector system. Cell 27, 279-288.

MEUNIER-ROTIVAL, M., SORIANO, P., CUNY, G., STRAUSS, F. and BERNARDI, G. (1982). Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. P.N.A.S. (USA) 79, 355-359.

MILCAREK, C., PRICE, R. and PENMAN, S. (1974). The metabolism of a poly(A) minus mRNA fraction in HeLa cells. Cell 3, 1-10.

MODIANO, G., BALTISTUZZI, G. and MOTULSKY, A.G. (1981). Nonrandom patterns of codon usage and of nucleotide substitutions in human α and β-globin genes: An evolutionary strategy reducing the rate of mutations with drastic effects. P.N.A.S.(USA) 78, 1110-1114.

MOLVIHILL, E.R., LePENNEC, J.P. and CHAMBON, P. (1982). Chicken oviduct progesterone receptor: location of specific regions of high-affinity binding in cloned DNA fragments of hormone-responsive genes. Cell 24, 621-632.

MONFOORT, C.H., SCHIPHOF, R., ROZIJN, T.H. and STEYN-PARVÉ, E.P. (1973). Amino acid composition and carboxyl-terminal structure of some basic chromosomal proteins of mammalian spermatozoa. B.B.A. 322, 173-177.

MORROW, J.F., COHEN, S.N., CHANG, A.C.Y., BOYER, H.W., GOODMAN, H.M. and HELLING, R.B. (1974). Replication and transcription of eukaryotic DNA in Escherichia coli. P.N.A.S. (USA) 71, 1743-1747.

MUNOZ-GUERRA, S., AZORIN, F., CASSAS, M.T., MARCET, X., MARISTANY, M.A., ROCA, J. and SUBIRANA, J.A. (1982). Structural organization of sperm chromatin from the fish Carassius auratus. Expl. Cell Res. 137, 45-53.

MURRAY, N.E., BRAMMAR, W.J. and MURRAY, K. (1977). Lambdoid phages that simplify the recovery of in vitro recombinants. Molecular and General Genetics 150, 53-61.

MUSKAVITCH, M.A.T. and HOGNESS, D.S. (1982). An expandable gene that encodes a Drosophila glue protein is not expressed in variants lacking remote upstream sequences. Cell 29, 1041-1051.

NAKANO, M., TOBITA, T. and ANDO, T. (1970). Fractionation of galline, a protamine from fowl sperm, and some characterization of the components. B.B.A. 207, 553-555.

NAGATA, S., MANTEI, N. and WEISSMAN, C. (1980). The structure of one of the eight or more distinct chromosomal genes for human interferon-α. Nature 287, 401-408.

NEI, M. and LI, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. P.N.A.S. (USA) 76, 5269-5273.

NEMER, M., GRAHAM, M. and DUBROFF, L.M. (1974). Co-existence of non-histone messenger RNA species lacking and containing poly-adenylic acid in sea urchin embryos. J.M.B. 89, 435-454.

NEVINS, J.R. and DARNELL, J.E. (1978). Steps in the processing of Ad2 mRNA:Poly(A)$^+$ nuclear sequences are conserved and poly(A) addition precedes splicing. Cell 15, 1477-1493.

NEWMARK, P. (1982). Cancer genes-processed genes-jumping genes. Nature 296, 393-394.

NISHIOKA, Y., LEDER, A. and LEDER, P. (1980). Unusual α-globin like gene that has cleanly lost both globin intervening sequences. P.N.A.S.(USA) 77, 2806-2809.

NUSSINOV, R. (1981). Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. J.M.B. 149, 125-131.

O'HARE, K., BREATHNACH, R. and CHAMBON, P. (1979). No more than seven interruptions in the ovalbumin gene: comparison of genomic and double stranded cDNA sequences. N.A.R. 7, 231-334.

OHNO, S. and ATKIN, N.B. (1966). Comparative DNA values and chromosome complements of eight species of fishes. Chromosoma 18, 455-466.

OHNO, S., WOLF, U. and ATKIN, N.B. (1968). Evolution from fish to mammals by gene duplication. Hereditas 59, 169-187.

ORKIN, S.H., KAZAZIAN, H.H., ANTONARAKIS, S.E., GOFF, S.C., BOEHM, C.D., SEXTON, J.P., WABER, P.G. and GIARDINA, P.J.V. (1982). Linkage of β-thalassaemia mutations and β-globin gene polymorphisms with DNA polymorphisms in human β-globin gene cluster. Nature 296, 627-631.

OZAKI, H. (1971). Developmental studies of sea urchin chromatin. Chromatin isolated from spermatozoa of the sea urchin Strongylocentrotus purpuratus. Developmental Biology 26, 209-219.

PARK, C.S., WU, F.Y.-H. and WU, C.-W. (1982). Molecular mechanism of promoter selection in gene transcription. II. Kinetic evidence for promoter search by a one dimensional diffusion of RNA polymerase molecule along the DNA template. J.B.C. 257, 6950-6956.

PATERSON, B.M., ROBERTS, B.E. and KUFF, E.L. (1977). Structural gene identification and mapping by DNA-mRNA hybrid arrested cell-free translation. P.N.A.S.(USA) 74, 4370-4374.

PAYVAR, F., WRANGE, O., CARLSTEDT-DUKE, J., OKRET, S., GUSTAFSSON, J.A. and YAMAMOTO, K.R. (1981). Purified glucocorticoid receptors bind selectively in vitro to a cloned DNA fragment whose transcription is regulated by glucocorticoids in vivo. P.N.A.S. (USA) 78, 6628-6632.

PETES, T.D. (1980). Molecular genetics of yeast. Annual Review of Biochemistry 49, 845-876.

PLATT, T. (1981). Termination of transcription and its regulation in the tryptophan operon of E. coli. Cell 24, 10-23.

PLOEG, L.H.T. and FLAVELL, R.A. (1980). DNA methylation in the human $\gamma\delta\beta$-globin locus in erythroid and nonerythroid tissues. Cell 19, 947-958.

POLISKY, B., GREENE, P., GARFIN, D.E., McCARTHY, B.J., GOODMAN, H.M. and BOYER, H.W. (1975). Specificity of substrate recognition by the EcoRI restriction endonuclease. P.N.A.S.(USA) 72, 3310-3314.

PROUDFOOT, N.J. and BROWNLEE, G.G. (1976). 3' non-coding region sequences in eukaryotic messenger RNA. Nature 263, 211-214.

PROUDFOOT, N.J., SHANDER, M.H.M., MANLEY, J.L., GEFTER, M.L. and MANIATIS, T. (1980). Structure and in vitro transcription of human globin genes. Science 209, 1329-1336.

RAE, P.M.M. (1972). The distribution of repetitive DNA sequences in chromosomes. Advances in Cell Molecular Biology 2, 109-150.

RAVE, N., CRKVENJAKOV, R. and BOEDTKER, H. (1979). Identification of procollagen mRNAs transferred to diazobenzyloxymethyl paper from formaldehyde agarose gels. N.A.R. 6, 3559-3567.

REANNEY, D. (1979). RNA splicing and polynucleotide evolution. Nature 277, 598-600.

RIGBY, P.W.J., DIECKMANN, M., RHODES, C. and BERG, P. (1977). Labelling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase 1. J.M.B. 113, 237-251.

RIMM, D.L., HORNESS, D., KUCERA, J. and BLATTNER, F.R. (1980). Construction of coliphage lambda Charon vectors with BamHI cloning sites. Gene 12, 301-309.

ROBERTSON, O.H. and RINFRET, A.P. (1957). Maturation of the infantile testes in rainbow trout (Salmo gairdnerii) produced by salmon pituitary gonadotrophins administered in cholesterol pellets. Endocrinology 60, 559-562.

ROBINS, D.M., PACK, I., SEEBURG, P.H. and AXEL, R. (1982). Regulated expression of human growth hormone genes in mouse cells. Cell 29, 623-231.

ROOP, D.R., NORDSTROM, J.L., TSAI, S.Y., TSAI, M.-J. and O'MALLEY, B.W. (1978). Transcription of structural and intervening sequences in the ovalbumin gene and identification of potential ovalbumin precursors. Cell 15, 671-685.

SAKAI, M., FUJII-KURIYAMA, Y. and MURAMATSU, M. (1978). Number and frequency of protamine genes in rainbow trout testis. Biochemistry 17, 5510-5515.

SAKAI, M., FUJII-KURIYAMA, Y., SAITO, T. and MURAMATSU, M. (1981). Closely related mRNA sequences of protamines in rainbow trout testis. Journal Biochemistry 89, 1863-1868.

SAKANO, H., ROGERS, J.H., HÜPPI, K., BRACK, C., TRAUNECKER, A., MAKI, R., WALL, R. and TONEGAWA, S. (1979). Domains and the hinge region of an immunoglobulin heavy chain are encoded in separate DNA segments. Nature 277, 627-633.

SANDER, M.M. and DIXON, G.H. (1972). The biosynthesis of protamine in trout testis. IV. Sites of phosphorylation. J.B.C. 247, 851-855.

SANGER, F. and COULSON, A.R. (1978). The use of thin acrylamide gels for DNA sequencing. FEBS Letters 87, 107-110.

SAMAL, B., WORCEL, A., LOUIS, C. and SCHEDL, P. (1981). Chromatin structure of the histone genes of D. melanogaster. Cell 23, 401-409.

SEILER-TUYNS, A. and BIRNSTEIL, M.L. (1981). Structure and expression in L-cells of cloned H4 histone gene of the mouse. J.M.B. 151, 607-626.

SETZER, D.R., McGROGAN, M., NUNBERG, J.H. and SCHIMKE, R.T. (1980). Size heterogeneity in the 3' end of dihydrofolate reductase messenger RNAs in mouse cells. Cell 22, 361-370.

SHEINESS, D. and DARNELL, J.E. (1973). Polyadenylic acid segment in mRNA becomes shorter with age. Nature New Biology 241, 265-268.

SHEN, C.K.J. and MANIATIS, T. (1980). Tissue specific DNA methylation in a cluster of rabbit β-like globin genes. P.N.A.S.(USA) 77, 6634-6638.

SHEN, S., SLIGHTOM, J.L. and SMITHIES, O. (1981). A history of the human fetal globin gene duplication. Cell 26, 191-203.

SHERMOEN, A.W. and BECKENDORF, S.K. (1982). A complex of interacting DNAase-I-hypersensitive sights near the Drosophila glue protein gene, Sgs 4. Cell 29, 601-607.

SIEBENLIST, U., SIMPSON, R.B. and GILBERT, W. (1980). E.coli RNA polymerase interacts homologously with two different promoters. Cell 20, 269-281.

SINGER, M.F. (1982). SINES and LINES: Highly repeated short and long interspersed sequences in mammalian genomes. Cell 28, 433-434.

SITTMAN, D.B., CHIU, I.-M., PAN, C.-J., COHN, R.H., KEDES, L.H. and MARZLUFF, W.F. (1981). Isolation of two clusters of mouse histone genes. P.N.A.S.(USA) 78, 4078-4082.

SLIGHTOM, J.L., BLECHL, A.E. and SMITHIES, O. (1980). Human fetal $^{G}\gamma$- and $^{A}\gamma$-globin genes: Complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21, 627-638.

SMITH, H.O. and BIRNSTEIL, M.L. (1976). A simple method for DNA restriction site mapping. N.A.R. 3, 2387-2398.

SMITH, M., LEUNG, D.W., GILLAM, S., ASTELL, C.R., MONTGOMERY, D.L. and HALL, B.D. (1979). Sequence of the gene for Iso-1-Cytochrome c in Saccharomyces cerevisiae. Cell 16, 753-761.

SMITHIES, O., ENGELS, W.R., DEVEREUX, J.R., SLIGHTOM, J.L., and SHEN, S. (1981). Base substitutions, length differences and DNA strand asymmetries in the human G_{γ} and A_{γ} fetal globin gene region. Cell 26, 345-353.

SOUTHERN, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J.M.B. 98, 503-517.

SPRADLING, A., PENMAN, S., CAMPO, M.S. and BISHOP, J.O. (1974). Repetitious and unique sequences in the heterogeneous nuclear and cytoplasmic messenger RNA of mammalian and insect cells. Cell 3, 23-30.

STALDER, J., LARSEN, A., ENGEL, J.D., DOLAN, M., GROUDINE, M. and WEINTRAUB, H. (1980). Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNAase 1. Cell 20, 451-460.

STARK, G.R. and WILLIAMS, J.G. (1979). Quantitative analysis of specific labelled RNAs using DNA covalently linked to diazo-benzyloxymethyl paper. N.A.R. 6, 195-204.

STEIN, J.P., CATTERALL, J.F., KRISTO, P., MEANS, A.R. and O'MALLEY, B.W. (1980). Ovomucoid intervening sequences specify functional domains and generate protein polymorphism. Cell 21, 681-687.

STEIN, R., RAZIN, A., CEDAR, H. (1982). In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. P.N.A.S. (USA) 79, 3418-3422.

STEINMETZ, M., MOORE, K.W., FRELINGER, J.G., SHER, B.T., SHEN, F.W., BOYSE, E.A. and HOOD, L. (1981). A pseudogene homologous to mouse transplantation antigens: Transplantation antigens are encoded by eight distinct exons that correlate with protein domains. Cell 25, 683-692.

STERNBERG, N., TIEMEIER, D. and ENQUIST, L. (1977). In vitro packaging of a λ Dam vector containing EcoRI DNA fragments of E. coli and phage P1. Gene 1, 255-280.

SUBIRANA, J.A., COZCOLLUELA, C., PALAU, J. and UNZETA, M. (1973). Protamines and other basic proteins from spermatozoa of molluscs. B.B.A. 317, 364-379.

SUGGS, S.V., WALLACE, R.B., HIROSE, T., KAWASHIMA, E.H. and ITAKURA, K. (1981). Use of oligonucleotides as hybridization probes: Isolation of cloned cDNA sequences for human β_2-microglobulin. P.N.A.S.(USA) 78, 6613-6617.

SUTCLIFFE, J.G. (1978). Complete nucleotide sequence of the E. coli plasmid pBR322. Cold Spring Harbor Symposia on Quantitative Biology 43, 77-90.

SUZUKI, Y., GAGE, L.P. and BROWN, D.D. (1972). The genes for silk fibroin in Bombyx mori. J.M.B. 70, 637-649.

Taya, Y., Devos, R., Tavernier, J., Cheroutre, H., Engler, G. and Fiers, W. (1982). Cloning and structure of the human immune interferon - γ chromosomal gene. The EMBO Journal 1, 953 - 958.

THOMAS, P.S. (1980). Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. P.N.A.S.(USA) 77, 5201-5205.

TOPAL, M.D. and FRESCO, J.R. (1976). Complementary base pairing and the origin of substitution mutations. Nature 263, 285-289.

TREISMAN, R., PROUDFOOT, N.J., SHANDER, M. and MANIATIS, T. (1982). A single-base change at a splice site in a β^0-thalassemic gene causes abnormal RNA splicing. Cell 29, 903-911.

TSAI, M.-J., TING, A.C., NORDSTROM, J.L., ZIMMER, W. and O'MALLEY, B.W. (1980). Processing of high molecular weight ovalbumin and ovomucoid precursor RNAs to messenger RNA. Cell 22, 219-230.

TSAI, S.Y., TSAI, M.-J. and O'MALLEY, B.W. (1981). Specific 5' flanking sequences are required for faithful initiation of in vitro transcription of the ovalbumin gene. P.N.A.S.(USA) 78, 879-883.

TSUDA, M. and SUZUKI, Y. (1981). Faithful transcription initiation of fibroin gene in a homologous cell-free system reveals an enhancing effect of 5' flanking sequence for upstream. Cell 27, 175-182.

TSUJIMOTO, Y., HIROSE, S., TSUDA, M. and SUZUKI, Y. (1981). Promoter-sequence of fibroin gene assigned by in vitro transcription system. P.N.A.S.(USA) 78, 4838-4842.

TWIGG, A.J. and SHERRATT, D. (1980). Trans-complementable copy-number mutants of plasmid Col El. Nature 283, 216-218.

VARDIMON, L., KRESSMAN, A., CEDAR, H., MAECHLER, M. and DOEFLER, W. (1982). Expression of a cloned adenovirus gene is inhibited by in vitro methylation. P.N.A.S.(USA) 79, 1073-1077.

VAUGHN, J.C. and HINSCH, G.W. (1972). Isolation and characterization of chromatin and DNA from the sperm of the spider crab, Libinia emarginata. Journal Cell Science 11, 131-152.

WAECHTER, D.E. and BASERGA, R. (1982). Effect of methylation on expression of microinjected genes. P.N.A.S.(USA) 79, 1106-1110.

WAHL, G.M., STERN, M. and STARK, G.R. (1979). Efficient transfer of large DNA fragments from agarose gels and rapid hybridization by using dextran sulphate. P.N.A.S.(USA) 76, 3683-3687.

WAHLI, W., DAWID, I.B., WYLER, T., WEBER, R. and RYFFEL, G.U. (1980). Comparative analysis of the structural organization of the two closely related vitellogenin genes in X. laevis. Cell 20, 107-117.

WALKER, P.M.B. (1971a). Origin of satellite DNA. Nature 229, 306-308.

WALKER, P.M.B. (1971b). Repetitive DNA in higher organisms. Progress in biophysics and molecular biology 23, 145-190.

WARRANT, R.W. and KIM, S.-H. (1978) α-Helix-double helix interaction shown in the structure of a protamine-transfer RNA complex and a nucleoprotamine model. Nature 271, 130-135.

WASYLYK, B., DERBYSHIRE, R., GUY, A., MOLKO, D., ROGET, A., TÉOULE, R. and CHAMBON, P. (1980). Specific in vitro transcription of conalbumin gene is drastically decreased by single-point mutation in T-A-T-A box homology sequence. P.N.A.S.(USA) 77, 7024-7028.

WEIL, P.A., LUSE, D.S., SEGALL, J. and ROEDER, R.G. (1979). Selective and accurate initiation of transcription at the Ad2 major late promoter in a soluble system dependent on purified RNA polymerase II and DNA. Cell 18, 469-484.

WEINTRAUB, H., BEUG, H., GROUDINE, M. and GRAF, T. (1982). Temperature-sensitive changes in the structure of globin chromatin in lines of red cell precursors transformed by ts-AEV. Cell 28, 931-940.

WEINTRAUB, H. and GROUDINE, M. (1976). Chromosomal subunits in active genes have an altered conformation. Science 193, 848-856.

WEINTRAUB, H., LARSEN, A. and GROUDINE, M. (1981). α-globin-gene switching during the development of chicken embryos: Expression and chromosome structure. Cell 24, 333-344.

WEISBROD, S., GROUDINE, M. and WEINTRAUB, H. (1980). Interaction of HMG 14 and 17 with actively transcribed genes. Cell 19, 289-301.

de WET, J.R., DANIELS, D.L., SCHROEDER, J.L., WILLIAMS, B.G., DENNISTON-THOMPSON, K., MOORE, D.D. and BLATTNER, F.R. (1980). Restriction maps for twenty-one Charon vector phages. Journal of Virology 33, 401-410.

WILDE, C.D., CROWTHER, C.E., CRIPE, T.P., LEE, M.G.S. and COWAN, N.J. (1982). Evidence that a human β-tubulin pseudogene is derived from its corresponding mRNA. Nature 297, 83-84.

WOODBURY, C.P., DOWNEY, R.L. and von HIPPEL, P.H. (1980b). DNA site recognition and overmethylation by the EcoRI methylase. J.B.C. 255, 11526-11533.

WOODBURY, C.P., HAGENBUCHLE, O., and von HIPPEL, P.H. (1980a). DNA site recognition and reduced specificity of the EcoRI endonuclease. J.B.C. 255, 11534-11546.

WOZNEY, J., HANAHAN, D., MORIMOTO, R., BOEDTKER, H. and DOTY, P. (1981). Fine structural analysis of the chicken pro α2 collagen gene. P.N.A.S. (USA) 78, 712-716.

WU, C. (1980). The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature 286, 854-860.

WU, C., BINGHAM, P.M., LIVAK, K.J., HOLMGREN, R. and ELGIN, S.C.R. (1979a). The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell 16, 797-806.

WU, C., WONG, Y.C. and ELGIN, S.C.R. (1979b). The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. Cell 16, 807-814.

YAMADA, Y., AVVEDIMENTO, V.E., MUDRYJ, M., OHKUBO, H., VOGELI, G., IRANI, M., PASTAN, I. and de CROMBROGGHE, B. (1980). The collagen gene: evidence for its evolutionary assembly by amplification of a DNA segment containing an exon of 54 bp. Cell 22, 887-892.

YAMATO, K.R. and ALBERTS, B.M. (1976). Steroid receptors: elements for modulation of eukaryotic transcription. Annual Review of Biochemistry 45, 722-746.

ZAKUT, R., SHANI, M., GIVOL, D., NEUMAN, S., YAFFE, D. and NUDEL, U. (1982). Nucleotide sequence of the rat skeletal muscle actin gene. Nature 298, 857-859.

ZARET, K.S. and SHERMAN, F. (1982). DNA sequence required for efficient transcription termination in yeast. Cell 28, 563-573.

ZIFF, E.B. and EVANS, R.M. (1978). Coincidence of the promoter and capped 5' terminus of RNA from the adenovirus 2 major late transcription unit. Cell 15, 1463-1475.

Abbreviations and Symbols

AEV	avien erythroblastosis virus
ATP	adenosine 5' triphosphate
BBL	BBL trypticase
bis acrylamide	N,N'-methylenebisacrylamide
β-MSH	2-mercaptoethanol
bp	base pair
BPB	bromophenol blue
BSA	bovine serum albumin
cAMP	cyclic (3' - 5') adenosine monophosphate
cDNA	DNA copy of RNA
CM	carboxymethyl
cpm	counts per minute
CRP	cAMP receptor protein
dATP	deoxyadenosine triphosphate
dCTP	deoxycytosine 5' triphosphate
dGTP	deoxyguanosine 5' triphosphate
dNTP	deoxynucleotide 5' triphosphate
DEAE	diethylaminoethyl
DNA	deoxyribonucleic acid
DNAase	deoxyribonuclease
ds	double stranded
DTT	dithiothreotol
EDTA	diaminoethanetetra-acetic acid
exo VII	exonuclease VII
GFC	glass fibre filter
HAP	hydroxyapatite
kb	kilobase (pair)
Krpm	10^3 revolutions per minute
LA	L agar
LB	L broth
M	molar (moles/litre)
mg	milligramme (10^{-3} gramme)
ml	millilitre (10^{-3} litre)
mM	millimolar (10^{-3} molar)
MMTV	mouse mammary tumour virus
MOPS	morpholinopropane sulphonic acid

mRNA	messenger RNA
M_w	molecular weight
NP40	nonidet-P40
OD_x	optical density at wavelength of x nanometres
OAc	acetate (CH_3COO^-)
P	phosphate (PO_4^-)
PEG	polyethylene glycol
pfu	plaque forming units
pH	minus $\log_{10}$ hydrogen ion concentration
PPO	2,5 - diphenyloxazole
PPOPOP	1,4-bis-2-(4-methyl-5-phenyloxazolyl)-benzene
rDNA	ribosomal RNA coding DNA
RNA	ribonucleic acid
RNAase	ribonuclease
rRNA	ribosomal RNA
SAM	S adenosyl methionine
SDS	Sodium dodecyl sulphate
SET	0.15M NaCl, 30mM TrisHClpH8.0, 2mM EDTA
ss	single stranded
SSC	0.15M NaCl, 0.015M Na Citrate
S1	nuclease S1
TBE	tris borate/EDTA gel buffer (90mM Tris, 90mM boric acid, 2mM EDTA)
TCA	trichloroacetic acid
TEMED	N,N,N',N' - tetramethylenediamine
Tris(HCl)	Trishydroxymethylamino methane (pH'ed with hydrochloric acid)
ts	temperature sensitive
TTP	thymidine 5' triphosphate
µg	microgramme (10^{-6} gramme)
µl	microlitre (10^{-6} litre)
µM	µM (10^{-6} molar)
v/v	volume/volume
w/v	weight/volume
w/w	weight/weight
xgal	5 chloro 4 bromo 3 indolyl-β-D-galactoside