A Thesis Entitled

AN ANALYSIS OF SPECIFIC MOUSE LIVER cDNA CLONES

by

ALISTAIR CHAVE-COX

Submitted to the University of Edinburgh
in Candidature for the Degree of
Doctor of Philosophy

EX BIBL. UNIV. EDINBURGEN.

July 1986

DECLARATION

This thesis has been composed by the author and has not previously been submitted for any other degree in this or any other University.

All work described in this thesis is original work carried out by the author. Some of this work has been published in Clark, Chave-Cox, Ma and Bishop, (1985).

A. Chave-Cox (Author)

DEDICATION

I dedicate this Thesis to
my Wife

ACKNOWLEDGEMENTS

I would like to thank the University of Edinburgh for the opportunity to pursue the work embodied in this thesis and in particular to Dr. J. O. Bishop for his help and interest throughout my period in Edinburgh.

I would also like to thank my colleagues for their instruction in many of the techniques I used and to my wife and Mr. and Mrs. A. Davies for their help in the preparation of the manuscript. While I was at Edinburgh I was in receipt of a Science and Engineering Research Council studentship grant.

CONTENTS

PREFACE

METHODS

ABBREVIATIONS

ABS	Absorbance units.
AMV	Avian myeloblastosis virus.
dATP	Deoxyadenosine 5´,-triphosphate.
ddATP	Dideoxyadenosine 5´,-triphosphate.
Bisacrylamide	N,N´,-Methylene bisacrylamide.
BPB	Bromophenol blue.
BSA	Bovine serum albumin.
dCTP	Deoxycytidine Triphosphate.
ddCTP	Dideoxycytidine triphosphate.
DMSO	Dimethylsulphoxide.
DNA	Deoxyribonucleic acid.
cDNA	Complementary deoxyribonucleic acid.
DTT	D,L-Dithiothreitol.
E.coli.	Escherichia coli.
EDTA	Ethylenediaminetetraacetic acid.
E.R.	Endoplasmic reticulum.
dGTP	Deoxyguanosine triphosphate.
ddGTP	Dideoxyguanosine triphosphate.
IPTG	Isopropyl-β-D-thiogalactopyranoside.
MOPS	Morpholinopropanesulphonic acid.
dNTP	Deoxynucleotide triphosphates.
OD	Optical density.
Oligo $(dT)_{10-12}$	Oligo (deoxythymidylic acid)$_{10-12}$
PBS	Phosphate buffered saline.
PEG	Polyethylene glycol.
POPOP	1,4-Di-[2-(5-phenyloxazoyl)]-benzene.
PPO	2,5-Diphenyloxazole.
Poly (A)	Poly (adenodylic acid).
RNA	Ribonucleic acid.
mRNA	Messenger ribonucleic acid.
rRNA	Ribosomal ribonucleic acid.
tRNA	Transfer ribonucleic acid.

RNase	Ribonuclease.
SDS	Dodecyl sodium sulphate.
TCA	Trichloracetic acid.
TEMED	N,N,N´,N´-Tetramethyl-ethylenediamine.
Tris	Tris (hydroxymethyl)aminomethane.
dTTP	Deoxythymidine triphosphate.
ddTTP	Dideoxythymidine triphosphate.
UWGCG	University of Wisconsin Genetics Computer Group.
X-gal	5-Bromo-4-chloro-3-indolyl-β -D-galactoside.

BUFFER AND SOLUTION ABBREVIATIONS

BSB	Borax Saline Borate Buffer. 0.01M $Na_2B_4O_7$/0.15M NaCl/pH 8.0 at 20°C.
Eco R1 Mix	*Eco* R1 Restriction enzyme Mix. 10 mM Tris/10 mM $MgCl_2$/10 mM β-Mercapto-ethanol/100 mM NaCl/pH 7.5 at 20°C.
Enzyme diluent	Restriction enzyme diluent. 10 mM KPO_4/200 mM NaCl/1 mM Na_2 EDTA/0.5 mM DTT/0.2% triton X-100/100 µg/ml BSA/50% glycerol/pH 7.0 at 20°C.
"Kirby phenol"	Kirby phenol. 89% v/v Water saturated phenol. 1/11.5% v/v *m*-Cresol/0.08% w/v 8-Hydroxyquinaline.
"Kirby salts"	Kirby salts. 6% w/v 4-aminosalicylate/6% v/v 1-Butanol/1% w/v Na Tri-isopropylnaphalene sulphate.
L.B. Top/Bottom	L.B. top agarose (0.7%)/L.B. bottom agarose (1.5%). 1% w/v Tryptone/1% w/v NaCl/0.5% w/v Difco yeast extract/0.7% (Top) or 1.5% (Bottom) w/v Agar.

L.Broth	L.Broth. 1% w/v Tryptone/1% w/v NaCl/0.5% w/v Yeast extract.
SET	SET. 0.15 M NaCl/30 mM Tris/2 mM Na_2 EDTA/pH 8.0.
SSC	SSC. 0.15 M NaCl/0.015 M Sodium Citrate.
STKM	STKM. 0.4 M Sucrose/0.2 M Tris acetate/50 mM KCl/5 mM $Mg(OA_c)$/pH 7.9 at 22^{o}C.
TA	Tris Acetate. 0.5 M Tris/0.2 M $Na_2(OA_c)$/0.1 M NaCl 20 mM Na_2 EDTA/~2% v/v Glacial acetic acid to pH 7.9 at 20^{o}C.
TB	Tris Borate pH 8.3. 89 mM Borate/87.5 mM Tris/2.5 mM Na_2EDTA/HCl to pH 8.3 at 20^{o}C.

REAGENTS

All reagents were of analar grade unless otherwise specified. In addition the following were obtained from specific suppliers: A.M.V. Reverse transcriptase from Anglian Biotechnology; E.coli DNA Polymerase I, Radioactive substrates, restriction enzymes and T4 DNA ligase from Amersham International; Restriction enzymes; Biolabs (New England); Creatine kinase, "Klenow" fragment of E.coli DNA polymerase I, Calcium dependant Micrococcal nuclease, Restriction enzymes and T4 Polynucleotide kinase from Boehringer; Oligo $(dT)_{10-12}$, Restriction enzymes and T4 DNA Polymerase from B.R.L.; E.coli DNA Polymerase I and "Klenow" fragment of E.coli DNA Polymerase I, from Cambridge Biotechnology; Antibodies from D.A.K.O. Labs; Sheets of Nitrocellulose, pore size 0.45 nM and Elutip-d-columns, from Schleicher and Schull; Acrylamide, Bisacrylamide, DNase, D.M.S.O., Lysozyme, S1 Nuclease, TEMED and Transferrin (human) were obtained from Sigma.

LOCATION OF FIGURES

LOCATION OF TABLES

ABSTRACT

Several liver secretory protein cDNAs were isolated from a female BALB/c mouse liver cDNA library. The mouse major urinary proteins (MUPs) are encoded within a multigene family of about 35 genes. Most MUPs are members of either Group 1 or Group 2 sequences, which can be distinguished by DNA sequence divergence. Two of the sequenced clones, MUP 8 and MUP 11 were of the Group 1 type. A third MUP clone, MUP 15, has diverged from both Group 1 and Group 2 sequences (i.e. BS 6 and BS 2,3) by 15% and 17.4% respectively. The divergence is twice as great over exons 1-3 and the 3´ terminal 68 nucleotides of the comparison, as it is over the intervening sequence. This suggests that an ancestral conversion event has occurred. MUP 15, like some Group 2 genes, has a longer signal peptide than Group 1 genes and differs from both Groups in having a probable N-linked glycosylation site and a different splice configuration between exons 6 and 7.

Transferrin is the major iron binding protein in vertebrate serum. Transferrin cDNA clones corresponding to 1.16 Kb of the 3´ half of the mRNA were isolated. The clones were identical where they overlapped, which implies that there is one predominantly expressed transferrin gene in mouse liver. Comparison of the mouse exonic sequence with human and chicken transferrins showed 18.0% and 35.5% replacement and 38.4% and 99.0% silent site divergence respectively. There are also small areas of higher homology within the domains, which may define iron binding sites. Preliminary investigations into two other cDNA clones are discussed. These correspond to the 3´ end (950 Bp) of the third component of mouse complement and the N-terminal half, (810 Bp) of mouse contrapsin, which is homologous to human $alpha_1$-antichymotrypsin.

INTRODUCTION

DNA Duplication

Since the discovery of genetic duplications by the early Drosophila geneticists (Sturtevant, 1925; Bridges, 1935) questions have been asked about their origins. Probably the most intriguing question is the nature of the events leading to the first duplication in any particular system. There are two sorts of duplication events which have contributed to the diversity of protein encoding gene types in higher eukaryotes. There is the intragenic amplification of primordial domains, which has been proposed as a sequence of events in the 1 2 collagen gene of the chick. This contains multiple 54 Bp exons (Yamada et al, 1980), multiple erythrocyte spectrin 106 amino acid domains (Speicher and Marchesi 1984), the mammalian alpha-fetaprotein and serum albumin ancestral gene triplicated domains (Eiferman et al, 1981), the chicken ovomucoid triplication (Stein et al, 1980) and the transferrin gene duplication (Jeltsch and Chambon, 1982 and Williams et al, 1982). An alternative type of duplication involves the multiplication of usually complete functional genes, which are referred to as families. These can be divided into two groups, multigene families and super families. The multigene families contain genes of the same or similar structure or function but which may be expressed at different stages of development, tissues or sexes. The "super" families contain genes in which the functional relationship between the encoded proteins may be obscure,

but for which structural homologies imply joint ancestry usually greater than 300 my years B.P. Examples of the former category multigene families would include amylases (Schibler et al, 1982), the preproinsulin genes (Perler et al, 1980), the ovalbumin gene family (Heilig et al, 1980), vitillogens (Wahli et al, 1982), actin genes (Nudet et al, 1982),the glycoprotein hormone family (Talmadge et al, 1984), the rodent major urinary protein genes (Kurtz, 1981 and Bishop et al, 1982),alpha and beta globin families (Efstratiadis et al, 1980), the vertebrate eye crystallins (King and Piatigorsky, 1983), the mammalian transferrins (MacGillivray et al, 1983), the acute phase $alpha_1$-protease inhibitors (Chandra et al, 1983 and Hill et al, 1984) and alpha fetoprotein and serum albumin family (Eiferman et al, 1981).

The "super" families include: the serine protease inhibitors, contrapsin, angiotensinogen and the ovalbumin genes superfamily (see Hill et al, 1984 and references therein), the mammalian alpha-crystallin and four small Drosophila heat shock proteins superfamily (Ingolia and Craig, 1982), the transferrin, melanomal associated antigen P97 and the chicken beta-cell lymphoma transforming gene lambda Ch Blym - 1 superfamily (Goubin et al, 1983). The later superfamily status is based on significant homologies between the amino termini of the encoded proteins. Other examples of super families may include vertebrate nerve growth factor and insulin (Frazier et al, 1972) and the alpha-lactalbumin protein and lysozyme enzyme (Hill et al, 1969). It should be noted that the above mentioned are only a few examples of super families, for in 1978, 181 protein super families had already been described (Dayhoff et al, 1978) and the number has been continuously added to since then.

It is generally presumed that the rare exchange event of unequal crossover first proposed by Sturtevant, (1925) and later corroborated by Bridges, (1936) was involved in

the events leading to the initial duplication of sequences. A series of similar genes adjacent to one another is called an array.

As the time from the duplication lengthens, mutations within the duplicated genes accumulate, and may lead to a change of function. The differences between such gene sequences (and thus function within or between arrays) may be reduced by the process of gene conversion. In gene conversion the whole or part of one gene replaces that of another. Unequal crossover within an existing array may also result in homogeneity between genes. Such mechanisms are thought to account for the similar sequences found within multigene families, eg. the human fetal globin genes (Slightom et al, 1980 and Shen et al, 1981) and certain chorion genes in the silkmoth (Jones and Kafatos, 1980).

Gene Expression

One of the more interesting findings which has emerged from the explosion of information on eukaryotic gene structure, is that there seem to be relatively few examples of truly "single copy" genes in higher eukaryotes. This implies that much of the eukaryotes DNA have arisen from previous gene duplications and subsequent modification events. This in itself may have been to some extent instrumental in the development of the extremely complex controls of gene expression found in eukaryotic genes (for reviews see Brown, 1981, Breathnach and Chambon, 1981 and North, 1984). The complete regulation of expression for any single higher eukaryotic gene remains to be elucidated. Much of the work on identifying the factors involved in the control of gene expression, has centered on the transcriptional control signals 5´ to the structural gene, (see eg. McKnight, 1982) and further upstream hormone receptor binding sites (Von der Ahe et al, 1985 and references therein). Altered chromatin

structure, as assayed by DNase hypersensitive sites, has also been invoked as an important means of affecting gene expression. Factors effecting such changes include SV40 enhancer and promoter elements (Jongstra et al, 1984), the immunoglobulin tissue specific enhancers (Mills et al, 1984 and references therein) and induction/repression by steroid hormones (Fritton et al, 1984 and references therein).

The role of enhancer sequences in immunoglobulin gene expression has recently been brought into question. No mechanisms have been proposed to account for the effects of changes of chromatin structure on gene expression, although it should be noted that there is much current debate and speculation on the role that enhancers and steroid hormones play in tissue specific gene expression (See eg. North, 1984 and Velcich and Ziff, 1984). The importance of DNA methylation in gene control remains unsolved (Bird, 1984).

Evidence is also accumulating for possible gene expresssion control mechanisms, which may operate after transcription has been initiated. It has been recently proposed that transcription termination and the formation of mRNA 3´ ends are separate processes. (Proudfoot, 1984). Cleavage of some transcripts occurs at alternative polyadenylation sites, and thereby results in mRNAs with different 3´ ends (see, eg. mouse dihydrofoliate reductase gene, Setzer et al, 1982 and rat alpha$_{2\mu}$-globulin cDNAs (Unterman et al, 1981)). Recently additional sequences 3´ to the polyadenylation signal have been shown to be required for correct 3´-end formation of rabbit beta globin. (Gil and Proudfoot, 1984).

The actual protein product of some gene transcripts in some instances, is itself determined by alternative mRNA splicing (Crabtree and Kant, 1982; .King and Piatigorsky, 1983 and Nawa et al, 1984) Finally the

sequence surrounding the initiation codon may be important in determining initiation efficiency by eukaryotic ribosomes (Lomedico and McAndrew, 1982).

Each of the above observations illustrates a point in the process of gene expression at which control could be exerted. The possibility that such post transcriptional control points exist in eukaryotic cells, is enhanced by the findings of Vernice et al, (1984). These show that the induction of alpha$_1$-acid glycoprotein by glucocorticoids is due to the stabilization or reduction in degradation of its mRNA, rather than an increase of transcription rate. It is clear that much is yet to be learnt about the control events of eukaryotic genes (Reudelhuber, 1984).

Usefulness of Complementary DNA

The synthesis of complementary DNA (cDNA) of mRNAs provides extremely useful tools for analysing the structure, organization and expression of eukaryotic genes (Okayama and Berg, 1982; and references therein). The nucleotide and protein sequences derived from cDNAs can be compared with the rapidly expanding nucleotide and protein sequence data bases.

Comparison of sequences showing homology may provide information on the evolution of genes by duplication and divergence (eg. Perler et al, 1980 and Hill et al, 1984) and the structure/function relationship of related sequences (see eg. MacGillivray et al, 1983 and Ullrich et al, 1985). Complete and even truncated cDNAs are also very useful as hybridization probes for isolating additional sequences. (Maniatis et al, 1978 and Wensink et al, 1979).

Liver Specific Proteins

The mammalian liver is the major site of synthesis for many proteins, several of which are secreted into the blood plasma. These serum proteins include albumin, fibrinogen, transferrin, the serine protease inhibitors and $alpha_2$-macroglobulins, and lipoproteins, several complement proteins, some of the blood coagulation factors, other hormones and trace element transport proteins (for reviews see Putnam, 1975; Aisen and Listowsky, 1980 and Morgan, 1983). Some major murine plasma proteins do not have a human plasma homologue. The major urinary proteins (MUPs) (Rumke and Thung, 1964) are synthesized in the liver (Finlayson et al, 1965) and are of unknown function, although a role in chemical communication has been suggested. (Shaw et al, 1983). Contrapsin is a protease inhibitor of trypsin (Takahara and Sinohara, 1982) which has been shown to be a liver specific protein, descended from the same ancestral gene as $alpha_1$-antichymotrypsin but has evolved a different antiprotease activity (Hill et al, 1984).

The sequences within the mouse liver correspond to one of three RNA abundance classes. Approximately 25% of the poly (A) mRNA encodes information for abundant sequences (Hastie and Bishop, 1976). The abundant mRNAs correspond to abundant proteins (Hastie and Held, 1978). There are approximately 12 moderately abundant mRNAs in mouse liver, 5,000-30,000 copies per cell, which are under tissue specific and developmental expression. Some of the abundant liver mRNAs were also detected in brain and kidney tissue. However their concentrations in these tissues were 1/5th to 1/500th lower than in the liver. (Derman et al, 1981 and Barth et al, 1982). Approximately 7 of the 12 most abundant sequences, largely confined to the E.R. ribosome fraction, encode secretory proteins. The

most abundant mRNA in adult male mouse liver corresponds to MUP and in female liver it encodes albumin (Clissold and Bishop, 1981 and Barth et al, 1982).

THE MAJOR URINARY PROTEINS

Rodents secrete a set of closely related proteins known in the rat as the alpha$_{2\mu}$-gobulins (Roy and Neuhaus, 1966), and in the mouse as the major urinary proteins (MUPs) (Finlayson et al, 1966). The MUPs are synthesized on the endoplasmic reticulum of the liver (Clissold et al, 1981), secreted into the plasma, and rapidly excreted in the urine (Finlayson et al, 1965; Rumke and Thung, 1964). The peptides are small, molecular weight 19,500, after cleavage of the signal peptide (Szoka et al, 1980) of 18 amino acids (Clarke et al, 1984a). Cleavage of the signal peptide reduces the $_{p}I$ of the secreted MUPs by 1 unit to 4.5-4.8 (Clissold and Bishop, 1982). Electrophoresis, particularly Isolectric focusing, was used to analyse this small, acidic group of related proteins. The urinary MUPs showed about 15 distinct components, (7 major and 7 minor) and in vitro translated unprocessed MUPs could be separated into about 20 components (Clissold and Bishop, 1982). The urine of different inbred strains gave different characteristic patterns, presumably due to different sets of structural genes within the strains. (Hainey and Bishop, 1981; Clissold and Bishop, 1982). However there are also considerable differences in the expression of individual MUPs within a strain. As well as containing much less MUP (~1/5-1/20th), the female mouse displays a different and simpler pattern than the adult male (No MUPs are detectable in juvenile mice). Administration of testosterone to females induces a level and pattern of MUPs in the urine that is similar to that found in the male. Some MUPs expressed in the male are only poorly induced in females by testosterone administration, especially in the C57 BL strain (Szoka and Paigen, 1980; Clissold et al, 1984). In C57 BL and BALB/c strains one particular MUP component is dominant especially in C57 BL (Shaw et al, 1983; Clissold et al, 1984).

The Complex Control of MUP Expression

Initial studies on female mice showed that MUP could be induced upon testosterone administration (Szoka and Paigen, 1978). It was already known that thyroxine and growth hormone are involved in regulating alpha$_{2\mu}$-globulin mRNA levels, which is the rat homologue to the mouse MUP (Kurtz and Feigelson, 1977). Shortly after this period the revelation that MUP mRNA is expressed in the submaxillary gland (Hastie *et al*, 1979), plus the above observations, prompted a detailed investigation into the regulation of MUP gene expression in the liver and other tissues.

The level of MUP mRNA in tissues has been determined by hybridization to MUP specific cDNA probes (see eg. Shaw *et al*, 1983). The different MUPs, encoded by the mRNAs of different tissues, have been analysed by *in vitro* translation of tissue mRNA with subsequent MUP isolation by immuno precipitation with urinary MUP antisera (Shahan and Derman, 1984). Alternative analysis has been undertaken by cDNA hybrid selection of MUP mRNA and subsequent *in vitro* translation and processing (Shaw *et al*, 1983 and Kuhn *et al*, 1984). Several differences in the number of MUPs and their level of expression in different tissues have been reported. These reported differences are probably due to some extent to mouse strain variation in the major MUPs expressed (Clissold and Bishop, 1982). However the use of hybrid selection to isolate MUP mRNAs could result in the loss of some MUP mRNAs which show considerable divergence to the cDNA used for selection (Kuhn *et al*, 1984). Similarly the *in vitro* processing of translation products would not reveal the differences between urinary proteins with the same $_{p}$I but with different leader sequences (Clissold and Bishop, 1982; Shahan and Derman, 1984; Ghazal *et al*, 1985). A compilation of the results obtained by Clissold and Bishop, (1982); Shaw *et al*, (1983) and Shahan and Derman (1984) is presented in Table I.

TABLE I

Differences in Tissue Specific Expression of MUPs in the Mouse.

The collated data is from several sources, where different mouse strains and methods of study have been used. As a guide to the Table, data under a heading or followed by a subscript C, S or B refer to studies on C57BL/J6 mice (Shaw et al, 1983 or Kuhn et al, 1984), swiss NCS mice (Shahan and Derman, 1984) or C57/BL, JU and BALB/c mouse strains (Clissold and Bishop, 1982) respectively. Suffix p indicates data obtained from hybrid selected in vitro translated and processed, (signal peptide removed) C57BL/J6 mRNA products (Shaw et al, 1983). Suffix L indicates in vitro translated whole MUPs from swiss NCS mRNA precipitated with MUP antisera (Shahan and Derman, 1984). The exception is where P or L is followed by B, in which case the proteins analysed were either urinary MUPs or unprocessed in vitro translation products of hybrid selected MUP ER poly (A) mRNA (Clissold and Bishop, 1982).

TABLE I

Tissue	Sexual Dimorphism mRNA (: =5:1) (C=S)	Relative Level MUP mRNA (Liver=1) (C+S)	Hormonal Regulation (C)	Influencing Hormones (C) T/T4/GH	PI of Major Proteins (Relative to Liver)	Size of Protein (with Leader)	Number of Proteins (IEF)	First Detectable mRNA (C)
Liver (Male)	+	1	+	T,T4,GH	5.25-5.7$_L$ 4.4-4.8$_p$ (+1 unit with Leader)	20-23kd	15p)$_B$ 20L)	WK 3
Lachrymal (Male)	+	1/5	+	T	1-2 units more basic ($_{P+L}$)	20kd	3-6 ($_{L+P}$)	WK 2
Submaxillary Gland	-	1/25	-	None	Acidic Liver ($_L$)	22kd	1_p-4_L	WK 1
Sublingual Gland	-	1/ 30-120	-	None	Acidic Liver ($_L$)	23kd	$?_p$-4-6_L	?
Parotid Gland	-	1/<30-120	?	Unknown	?	?	?	?
Mammary Gland	+ (Female only)	1/30-120	Probable	Unknown	Middle Liver ($_p$)	?	1_p-$?_L$	First Pregnancy

From the data presented in Table I it would appear that the major proteins expressed in each tissue can be distinguished from the major MUP proteins expressed in the other tissues either by size, $_{p}I$ or their induction by hormones. These major proteins are therefore either encoded by structurally dissimilar genes (lachrymal?) or a subset of similar genes expressed under different hormonal controls in the various tissues (mammary, submaxillary, sublingual and parotid glands).

Liver and Lachrymal Glands

The IEF separation techniques used to analyse liver proteins demonstrated that the high level of MUP mRNA in liver was due, in part at least, to the expression of many genes (9-15) in this tissue, some of which have been shown to be under different hormonal controls (Kuhn *et al*, 1984). The influence of testosterone has been shown to be largely responsible for the sexual dimorphism of MUP expression in the liver and lachrymal gland (Shaw *et al*, 1983; Clissold *et al*, 1984). There is a discrepancy between the ratio of liver mRNA and MUP *in vitro* translation products 1:5 (normal female when compared to male mice) when compared to the urinary MUP levels 1:20-30 (normal female to male mice). The discrepancy may be accounted for by other, probably post transcriptional controls which occur in the intact organism (Clissold *et al*, 1984). Circulating levels of thyroxine, testosterone and growth hormone (the action of which is mainly synergistic to thyroxine) control the synthesis of the majority of MUPs made in the liver (Kuhn *et al*, 1984 and references therein). The expression of MUPs in normal mice appears to be developmentally regulated in association with sexual maturity and is therefore probably due to the level of (plasma) sex steroids (Palmiter and Lee, 1980 and Barth *et al*, 1982).

The MUPs expressed in the lachrymal gland are transcribed from genes other than those expressed in the liver. Indeed although the mRNA of lachrymal MUPs exhibit similarities with two quite different members of the MUP multigene family, the $_{p}I$ of the encoded proteins indicate that they comprise a further phenotypically distinct subset of MUP genes to those already isolated or possibly a novel splicing arrangement of the MUP genes (Bishop et al, 1982; Shaw et al, 1983; Kuhn et al, 1984 and Ghazal et al, 1985). Neither circulating growth hormone or throxine are required for lachrymal MUP expression. Hypophysectomised animals do not express lachrymal MUPs and although administration of testosterone (in male and female animals) yield MUP mRNA levels equivalent to those in normal male mice, the level of MUP mRNA had already reached the adult level in the earliest detectable gland tissue (at 2 weeks of age) of normal animals (Shaw et al, 1983). It is therefore clear that the levels of circulating sex steroids are not involved in the onset of MUP expression in the lachrymal glands of normal mice, as could be the case for liver MUPs. However, possible involvement of sex steroids in the sexual dimorphism of MUP in lachrymal and liver tissue is not excluded.

The salivary glands express MUPs of a similar size and $_{p}I$ to the range of MUPs found in the liver. It is therefore not possible to determine whether these are identical with liver polypeptides and represent products of the same genes. However the dominant MUP like proteins of the submaxillary and sublingual glands are of different sizes, which are in turn both larger that the major MUP polypeptide of the liver and therefore probably represent different gene products (Shahan and Derman, 1984). The size differences could be due to either larger signal peptides or alternative splicing arrangements (which result in a longer open reading frame for translation), or both, as no distinction is possible with the current data.

The levels of MUP synthesis in the salivary glands are essentially unaffected by endocrine manipulations, which may account for their lack of sexual dimorphism. Submaxillary gland synthesis follows a distinct developmental regime, rising from the first week to a peak between weeks 4 and 7. Thereafter it declines to approximately 1/200th of the adult male liver level (Shaw et al, 1983). This apparent lack of hormone modulation MUP or genes in the mouse salivary tissue, may not be due to the tissue being unresponsive to hormones. The same tissue in the rat has not only been shown to be hormone responsive, but that the duct cells of the salivary gland synthesize androgen-responsive proteins (i.e. epidermal growth factor) and alpha$_{2\mu}$-globulin; which is also unmodulated by external hormones (Laperche et al, 1983 and references therein). The question as to whether salivary MUPs are encoded by genes different from those in the liver, is again raised by the concommitant rat study, wherein the 4 salivary alpha$_{2\mu}$-globulins represent a distinct subgroup on the basis of their pIs relative to the multiple liver polypeptides (Laperche et al, 1983), as has been demonstrated for the lachrymal MUPs (Shaw et al, 1983 and Shahan and Derman, 1984).

A fourth type of developmental MUP expression is apparent in the mammary gland, whereby synthesis occurs for most of the pregnancy. There appears to be mainly one type of polypeptide synthesized in the mammary gland, which may be the same as the main MUP synthesized in the adult female liver (Shaw et al, 1983), although this is by no means certain. Mature MUP-like proteins from different tissues (submaxillary and sublingual), which appear to be the same in one strain, C57 BL/6J (Shaw et al, 1983) have been shown to be considerably different when compared as their unprocessed precursors from white Swiss (N.C.S.) mice (Shahan and Derman, 1984). The implication being that submaxillary and sublingual MUP-like proteins may be transcribed from different genes, or at least have

alternative gene splicing arrangements in the two tissues. However, strain variation may account for the latter observation and the proteins may be identical in C57 BL/6J mice.

MUP Function

The main MUPs expressed by the strains differ within the same tissue and between different tissues and are subject to alternative degrees of hormonal controls. However, the onset of developmental expression within tissues and the relative amount of total MUP synthesis (mRNA levels), including sexual dimorphism within the main productive tissues, appears to be the same in the different strains. (Clissold and Bishop, 1982; Shaw et al, 1983; Kuhn et al, 1984; Shahan and Derman, 1984 and Clissold et al, 1984).

The physiological role of these proteins, suggested by their widespread occurrence, structural similarities and complex hormonal and developmental control is not known. Studies on the puberty-accelerating effects on juvenile females of adult male mouse urine, and the coincidence of the onset of MUP excretion in the urine of adolescent males, have suggested that liver MUPs may play a role in sexual development. (Vandenbergh, 1975). All MUP polypeptides are secreted into external cavities or excreted in the urine. It has been recently suggested that the most probable common function of these secretions is that of a behavioural clue, for which there is considerable circumstantial evidence (for review see Shaw et al, 1983). It has also been suggested that the "active" part of the MUP molecule may be the six N-terminal amino acids, of the mature protein (Clark et al, 1985).

The MUP Genes

It is clear that the urinary protein and related genes of rodents (rats and mice) represent large multigene families, which are exceptional in so much as many of the individual family members are under very different hormonal and developmental controls. It has been estimated that there are at least 34 MUP genes in the BALB/c mouse, and that these may be split into two main subfamilies, Group 1 and Group 2, on the basis of hybridization. Furthermore there are approximately 15 Group 1, 12 Group 2 and at least 7 MUP genes outside the group classifications (Bishop et al, 1982). The main group of genes expressed in mouse liver are Group 1 sequences with very little Group 2 mRNA synthesis (Clark et al, 1982 and Clark et al, 1984a). Isolation of MUP liver cDNAs have not resulted in the recovery of any Group 2 clones (Clissold and Bishop, 1981 and Kuhn et al, 1984). Recently a second type of cDNA clone $_{p}$199 (5′ half cDNA) has been isolated from male liver cDNAs. The exonic sequence of $_{p}$199 differs from the Group 1 genes by 15.6% It is expressed in the liver and possibly the lachrymal glands, and may be representative of a third less homogeneous group of genes than Group 1 and 2 genes. Examples of other exon sequence comparisons are Group 1/1 ~0.5%, Group 2/2 ~3-5%, Group 1/2 ~10%, Group 1 or 2/ alpha$_{2\mu}$-globulin ~19-25% (Clark et al, 1984b, Kuhn et al, 1984 and Ghazal et al, 1985). Similarly a second class or subset of alpha$_{2\mu}$-globulin genes (divergence 5% compared to 2%) has recently been described in the rat (Laperche et al, 1983).

Evidence for a third more divergent Group of MUP genes expressed in the mouse comes from several sources. Primarily $_{p}$199-like sequences comprise a considerable amount of liver (1/5th) and lachrymal gland (~1/2) total MUP mRNA content, although the lachrymal gland $_{p}$199-like mRNA encodes a different protein (Novel $_{p}$I type) to the liver message (Kuhn et al, 1984; Shaw et al, 1983 and

Table I). Similarly the MUP mRNAs expressed in the submaxillary gland (Table I) are neither Group 1 or $_{p}$199-like, as determined by mRNA/MUP cDNA hybridization (Kuhn et al, 1984). The total number of MUP polypeptides expressed in the liver and other tissues (~23 + Table I), is greater than the estimated number of Group 1 genes in the mouse (~15) (Bishop et al, 1982) and must therefore be due to either Group 2 or other MUP genes. It is however unlikely that the Group 2 genes could encode the additional proteins, because all the Group 2 genes sequenced so far (4), have been shown to contain the same stop codon, at the position of the seventh amino acid of the mature protein, and are thus pseudogenes (Ghazal et al, 1985). That a considerably more divergent subset of MUP genes, relative to Groups 1 and 2, was not detected earlier is not unexpected, given the stringent experimental hybridization and other conditions required for quantitative analysis. The isolation and manipulation of the genomic clones encoding this novel set of MUP sequences remains an exciting prospect, given their apparent diversity in sequence and modulation (Table I). The isolation of tissue specific MUP cDNA's and their respective genomic clones is currently underway in several laboratories.

The Genomic Organization of the MUP Genes

A definitive study of the chromosomal locus of the MUP genes using somatic-cell hybrids, recombinant inbred strains and lower stringency hybridization conditions, has assigned all the MUP genes to the MUP-a locus on chromosome 4. (Bennett et al, 1982).

By utilizing the genomic sequences 5' to the MUP genes isolated by Clark et al, (1982) as probes to screen a charon 4A library of mouse DNA, (Clark et al, 1984b) have isolated genomic clones which contain both a Group 1

and a Group 2 gene divergently orientated, with 15kb of DNA between the 5′ ends of the genes. Flanking sequence probes showed that ~4kb of DNA 5′and ~12kb of DNA 3′ of Group 1 and Group 2 genes were homologous. There have been several insertion/deletion events within the flanking regions and the sequence divergence is more pronounced than within the coding regions, although the 5′ flanking regions and unduplicated ~7kb separating these are more uniform than 3′ flanking regions. The above information, taken together with Genomic DNA analysis, suggests that the majority of Group 1 and 2 genes are arranged in a pair wise manner of divergently arranged units spanning a total of ~45kb, Figure 1, (Clark et al, 1984b, Bishop et al, 1985).

Many of the 45kb units, of which there are 12 to 15, are adjacent to one another and as such appear to be the predominant mode of organization for MUP genes within the mouse genome (Bishop et al, 1985).

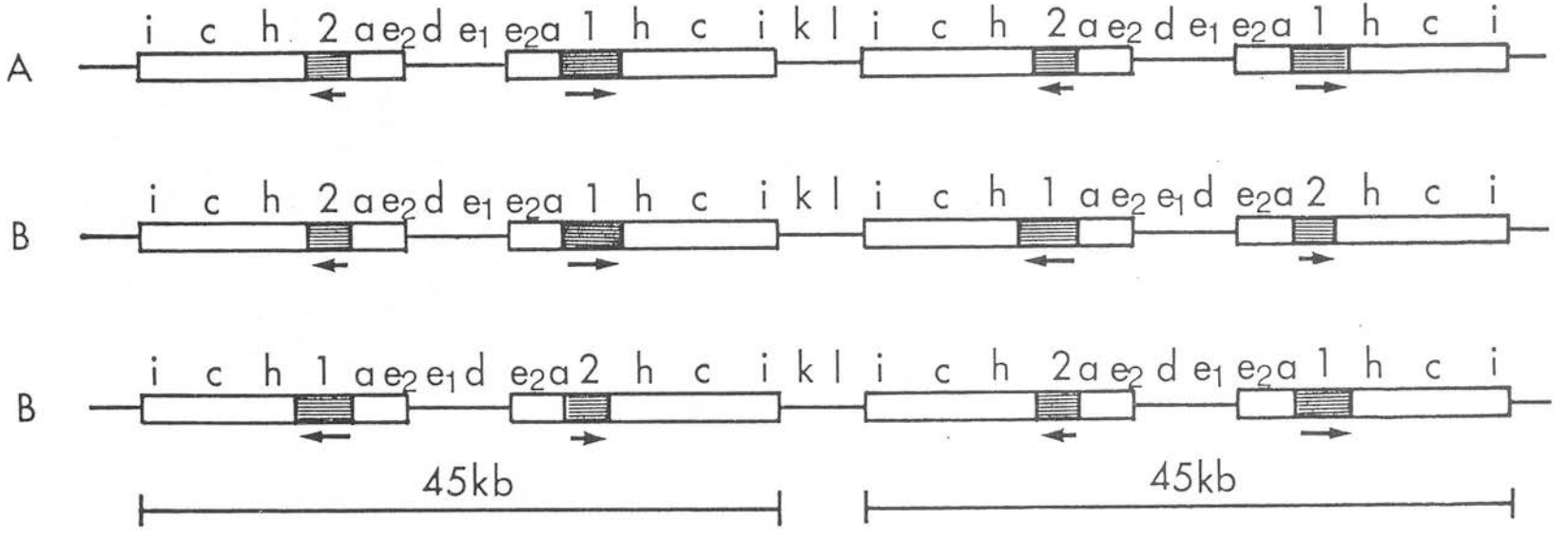
Evolution of the MUP Array

Different Group 1 genes and most $alpha_{2u}$-globulin genes differ within each species by ~0.5% of their nucleotides, whereas other MUP sequences, Group 2, differ within themselves by ~3-5% and from Group 1 by about 10%. Some $alpha_{2\mu}$-globulin sequences differ by 5% relative to most of the rat $alpha_{2\mu}$-globulin sequences. The MUP and $alpha_{2\mu}$-globulin sequences which must have evolved from a common ancestor differ by ~20% (Dolan et al, 1982; Laperche et al, 1983; Kuhn et al, 1984 and Ghazal et al, 1985). The mechanisms proposed to account for the apparent constraint of sequence divergence within species, have been repeated unequal crossing over and/or gene conversion (Dolan et al, 1982; Clark et al, 1984b; Bishop et al, 1985 and Ghazal et al, 1985). In unequal crossing over, one sequence unit

FIGURE 1

Alternative Arrangements of MUP Genes as 45-kb MUP Gene Pairs (Clark et al, 1984b).

The 20-kb units comprising MUP genes and their flanking sequences are shown as boxes, the Group 1 and Group 2 gene sequences are shaded and labelled 1 and 2 respectively. Arrows show the direction of transcription. a, c, d, e1, e2, h, i, k and l show the approximate positions of the regions that hybridize with various probes (see, Clark et al, 1984b). (A) Direct tandem repetition; (B) pairs of inverted 45-kb units.


A
i c h 2 a e2 d e1 e2 a 1 h c i k l i c h 2 a e2 d e1 e2 a 1 h c i
B
i c h 2 a e2 d e1 e2 a 1 h c i k l i c h 1 a e2 e1 d e2 a 2 h c i
B
i c h 1 a e2 e1 d e2 a 2 h c i k l i c h 2 a e2 d e1 e2 a 1 h c i
45kb
45kb

replaces a second by chromosome slipage and exchange. The unit in the case of MUP would be the 45kb between the 3' flanking region between a Group 1 - Group 2 gene pair (Clark et al, 1984b). The gene conversion scenario for the MUPs would invoke frequent conversion events between Group 1 genes. Group 2 conversions would be less frequent and Group 1/Group 2 conversions least frequent (Clark et al, 1984b). In either scenario the more divergent members of the multigene families, the arrays of which probably pre-date the rat/mouse speciation (Dolan et al, 1982), represent genes which have not undergone recent exchange events with the predominant, similar, multigene family members.

A sequence of events, to account for the generation of the mouse MUP and rat alpha$_{2\mu}$-globulin multigene families and their intra family homogeneity, has been proposed by Ghazal et al, (1985). It suggests that an ancestral array of genes existed prior to the rat/mouse speciation, sometime after which an inversion event between two MUP genes occurred generating the 45kb unit. The 45kb unit subsequently replaced much of the ancestral array of genes in the mouse.

At some point one of the genes in the pair of the 45kb unit, the progenitor of the Group 2 genes, developed a stop codon mutation. This pseudogene was then carried passively through the array by subsequent exchange events. Such a model would allow equal divergence between the rat alpha$_{2\mu}$-globulin genes and the Group 1 or Group 2 genes, (presumably similar exchange events with a single gene unit maintained homogeneity within the alpha$_{2\mu}$-globulin family). Secondly the predominance of the 45kb unit, with divergently arranged genes, would not favour exchanges between Group 1 and Group 2 genes, thus allowing their separate divergence. Finally the pseudogene nature of the Group 2 MUP genes, and therefore reduced selective pressure, would enable a higher net rate of mutation than

in the Group 1 genes, and as such may account for the greater heterogeneity within Group 2 genes compared to Group 1 genes (Ghazal et al, 1985).

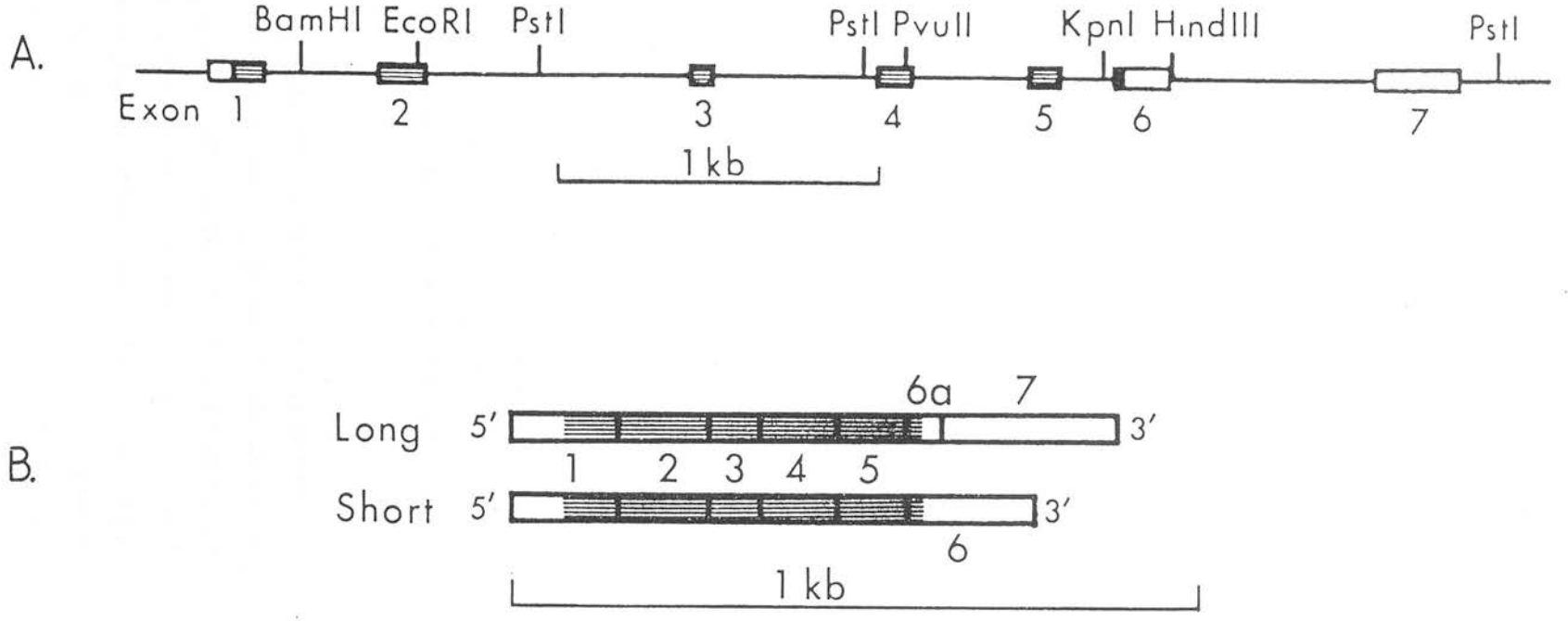
Rodent Urinary Protein Genes and Alternative Transcript Processing

The MUP and alpha$_{2\mu}$-globulin genes have a similar 7 exon structure, with exons 2-6a identical in length (Dolan et al, 1982; Clark et al, 1984a), encompassing approximately 4kb of DNA (Figure 2), which has been corroborated by the sequence of a near full length MUP cDNA $_{p}$1057 (Kuhn et al, 1984). All splice functions conform to the general consensus of Breathnach and Chambon, (1981) and the intronic sequences immediately adjacent to the splice junctions (10Bp) show extensive homology (Clark et al, 1984a). The cap site is 31 nucleotides downstream from the TATA Box (Clark et al, 1984a, Ghazal et al, 1985), which places it close to the alpha$_{2\mu}$-globulin proposed cap site (Laperche et al, 1983). The Group 1 gene encodes a 543-nucleotide open reading frame, identical in position and length to that of alpha$_{2\mu}$-globulin, except for deletion/insertions in the signal peptide (within exon 1) and 3′ untranslated regions of exon 7 (Kuhn et al, 1984). Variation in the length of the signal peptide also occurs at a similar position within the Group 2 pseudo genes (Ghazal et al, 1985). Homology between the MUP gene and alpha$_{2\mu}$-globulin coding and mRNA untranslated sequences is ~80% at the nucleotide level and ~66% at the amino acid level (Clark et al, 1984a, Kuhn et al, 1984 and Ghazal et al, 1985). Group 2 nucleotide sequences are ~90% homologous (Ghazal et al, 1985) and the N terminal MUP cDNA sequence ($_{p}$199) is ~84% homologous (Kuhn et al, 1984) to the Group 1 like Gene coding region.

FIGURE 2

Structure of MUP Gene BS-6 and Main mRNA Splicing Arrangements.

The size and arrangement of the exons of MUP gene BS 6 are shown as boxes, coding regions are shaded. (A), mouse DNA insert of plasmid pBS6-1;(B), the two main MUP mRNA types showing their structural relationship to BS 6.


A.
BamHI
EcoRI
PstI
PstI
PvuII
KpnI
HindIII
PstI
Exon 1
2
3
4
5
6
7
1kb
B.
6a
7
Long
5'
3'
1
2
3
4
5
Short
5'
3'
6
1 kb

The coding region ends at nucleotide 26 of exon 6. There are however two sets of polyadenylation signals downstream of this position and alternative splicing arrangements utilize one or the other, generating two main size classes of mRNA, long and short. Short mRNA contains exons 1-6, whereas long message contains a shorter 5′ region of exon 6, (6a) and exon 7 (Figure 2). The greater part of liver MUP mRNA is of the long type and is encoded by Group 1-like genes (Clark et al, 1984a Kuhn et al, 1984 and Shahan and Derman, 1984). However Group 1-like MUP mRNA may not be the predominant message type in other tissues expressing MUPs (Kuhn et al, 1984 and Table I). The majority of the Group 2 gene transcripts are of the short type (Clark et al, 1984a). It would appear from the data presented in Clark et al (1982 and 1984a), that the amount of Group 2 (pseudogene) message is very low when compared to the total Group 1 type message in liver. It is therefore probable, that the majority of the short message is also derived from Group 1 type transcripts. This is substantiated by the fact that both of the short message cDNAs cloned (Clissold and Bishop, 1981) correspond to Group 1 type genes (Clark et al, 1984a). Furthermore, short type message is only detected in the liver and possibly the mammary gland tissues. (Shaw et al, 1983; Kuhn et al, 1984 and Shahan and Derman, 1984).

The short mRNA, in addition to being generated by alternative splicing, has two polyadenylation signals, both of which are used, one of which is the more unusual ATTAAA (Clark et al, 1984a).

In the rat alpha$_{2\mu}$-globulin genes alternative splicing and polyadenylation sites have also been observed. When several alpha$_{2\mu}$-globulin cDNAs were isolated and sequenced, it was observed that in one cDNA (alpha 6) a section of IVS 6 remained in the cDNA (Dolan et al, 1982). An alpha$_{2\mu}$-globulin cDNA expressed in the rat submaxillary gland has been cloned. The cDNA contains a mutation at the

3′ exon 6 splice junction (from AG/gt to AG/at), which renders this junction inoperative, activating a cryptic splice junction sequence 121 Bp downstream, the result of which, is the inclusion of 121 Bp of intron VI into the mRNA. The developmental expression and hormonal regulation are known to differ considerably from the liver sequences, as has been discussed above, (Laperche et al, 1983). One of the $alpha_{2\mu}$-globulin genes expressed in the rat liver (probably the major transcription product), contains two polyadenylation sequences (AATAAA and ATTAAA). Both sequences are used with AATAAA being the preferred site, generating a mRNA 74 nucleotides larger than the other (Unterman et al, 1981).

The alternative splicing and polyadenylation configuration found within the MUP 3′ non-coding regions, have been proposed as possible factors involved in the expression of those genes (Clark et al, 1984a). Recently there has been speculation that factors which affect the formation of the 3′ end of mRNA may be additional control points for regulating gene expression, (Proudfoot, 1984 and Gil and Proudfoot, 1984).

TRANSFERRINS

Function

Transferrin is the protein which binds and transports non haem iron in the plasma of higher vertebrates and it is the second most abundant protein in human serum (Aisen and Listowsky, 1980). It is a two sited protein, with an exceptionally high binding constant for Fe^{3+} at physiological levels of pH and HCO_3 (Aisen and Brown, 1975). Three functions have been proposed for transferrin. Most of the transport of Fe^{3+} (which would otherwise rapidly form insoluble hydroxide precipitates) takes place from the intestine, the site of uptake, to the liver for storage, or to reticulocytes, a major site of utilization (Putman, 1975; Aisen and Listowsky, 1980 and Morgan 1983). However Fe^{3+} and therefore transferrin are essential for the growth of all cell types (Mather and Sato, 1979). Iron delivery to cells has been studied by Karin and Mintz (1981) and is thought to comprise four basic stages. Firstly the transferrin binds to the transferrin receptor on the cell surface. Secondly the iron-containing transferrin receptor complex is internalized by receptor mediated endocytosis. Thirdly the endocytotic vesicle fuses with a lysosome, where the low pH causes Fe^{3+} and concomitant bicarbonate release. Finally the transferrin evades digestion in the lysosome and is released to the plasma, whereas the Fe^{3+} is taken up by ferritin for subsequent utilization by the cell (Karin and Mintz, 1981 and Morgan, 1983).

The very tight binding of iron by transferrin results in extremely low plasma levels of free iron. Given that iron is an essential nutrient required for the growth of most bacteria or fungi and a complex mechanism is required

for its release from the protein, then the circulating levels of transferrin mopping up any released iron must effect considerable antimicrobial activity (Aisen and Brown, 1975).

More recently transferrin has been related to proliferative processes by two lines of evidence. One line of evidence relates the NH_2 terminal sequence of transferrin with gene products active in cancer cells (Goubin et al, 1983). The major regions of homology extend over the first 19 amino acids of the mature transferrin and the chicken B-cell lymphoma lambda Ch Blym-1, and by inference, human Burkitt's lymphoma non- transforming genes (Diamond et al, 1983) and the NH_2 terminal sequence of human melanoma-associated antigen $_{p}97$ (Brown et al, 1982; Goubin et al, 1983). Contradictions within the published predicted protein and protein sequence comparison data for lambda Ch Blym-1, could reduce the published homology between lambda Ch Blym-1 and other sequences compared in Goubin et al, (1983). lambda Ch Blym-1 would also possess a very different secondary and tertiary structure to the other sequences compared in Goubin et al, (1983). However the homology of the $_{p}97$ protein to transferrin (MacGillivray, 1982 and Williams, 1982) appears to be much more extensive than just the NH_2 terminus, and although there appear to be differences in secondary and/or tertiary structure, $_{p}97$ also exhibits functional homology in that it binds iron (Brown et al, 1982). In man, the transferrin and $_{p}97$ genes are located on the same chromosome (Plowman et al, 1983).

It has been shown that the addition of transferrin to tissue culture stimulates the cell proliferation and differentiation of chicken muscle (Beach et al, 1983) and nephrogenic mesenchyme cells (Ekblom et al, 1983). However, whether differentiation is a reflection of improved cell viability and growth remains to be determined. Many tissue cultures which exhibit depression

of serum proteins or exhibit more primitive cell types synthesize transferrin (Stencher and Thornbeck, 1967) and as such may be expected to survive better in a transferrin limited culture system (Mather and Sato, 1979). The detection of transferrin mRNAs in many rat fetal tissues, especially muscle, the expression of which peaks at ~1-3 days prior to birth, has been proposed as further evidence of a role of transferrin in tissue differentiation (Levin et al, 1984). It has recently been demonstrated that the visceral yolk sac in the mouse foetus, expresses many of the proteins characteristic of the near term foetal liver including transferrin. It is believed that these proteins are involved in nurturing the growing embryo before the liver takes over these functions (Meehan et al, 1984). Transferrin synthesis increases very rapidly in the few days before birth in the rat, by which time adult levels of synthesis are achieved (Levin et al, 1984). It is therefore possible that the synthesis of transferrin in other tissues in the prenatal animal, is more a reflection of the need to synthesize the required amount of protein for rapid foetal growth prior to the liver assuming adequate transferrin synthesis; than it is in controlling tissue differentiation.

It therefore remains to be proved whether transferrin is, or is not, involved in some way with the control of tissue differentiation. However, there is no doubt that the basis of the protein homologies between transferrins, lambda Ch Blym-1 and p97, is a promising area of current research and speculation.

The Primary Structure of Transferrin

The primary structure of transferrin has been well characterised both in the chicken (Jeltsch and Chambon, 1982 and Williams et al, 1982) and human (MacGillivray et

al, 1982 and 1983, Uzan et al, 1984 and Yang et al, 1984). Attempts have also been made to define the amino acid residues important in the iron binding structure-function of human transferrin (Aisen and Brown, 1975; Shewale and Brew, 1982 and Williams et al, 1982).

The polypeptide chain of serum transferrin contains 679 amino acid residues, and has a molecular weight of 79,570 (MacGillivray et al, 1983, figure 32). Prior to the secretion it also has a 19 nucleotide leader sequence (Yang et al, 1984). This includes two N-linked bianternary glycans (Molecular weight 2207), each of which terminates in a sialic acid residue. These are attached to the COOH half of the protein at asparagine residues 417 and 610 (Aisen and Listowsky, 1980). Human transferrin consists of two homologous domains (1-336 and 337-679), each of which is associated with a single iron binding site. In the two domains, 41.7% of the amino acid sequence is identical when appropriate gaps are inserted for the comparison (MacGillivray et al, 1982 and 1983). A very similar set of observations have been made about chicken transferrin. It has a leader of 19 amino acids, leaving after cleavage 686 residues in two domains, (1-332 and 342-686) which exhibit 37% homology when aligned with gaps (Williams et al, 1982). In addition to homologies between the two iron binding domains, a much weaker fourfold homology is present in the protein sequence, which has been confirmed by comparison of the equivalent nucleotide sequence (Jeltsch and Chambon, 1982). This suggests that the ancestral gene, which was duplicated to generate the present transferrins, may itself have been the product of an earlier primordial gene duplication. Comparison of the amino acid homologies between the transferrin domains, (41.7% and 37.4%) and between the human and chicken transferrins, (49.8%) suggests that the event which gave rise to the two domain structure of transferrins, pre-dates the earliest common ancestor of chickens and humans (MacGillivray et al, 1983).

Structure Function Relationships of Transferrins

Early spectroscopic studies provided compelling evidence for two separate iron binding sites and directly implicate tyrosyl phenolic groups in the metal binding function. The number of tyrosyl groups involved in each site is 2 or 3, (probably 2), and 1 or 2 nitrogen ligands, (possibly histidine) are involved, although as many as 4 nitrogen ligands have been proposed. In addition to these protein ligands a bicarbonate and a water molecule are also thought to be involved (for review see Aisen and Brown, 1975).

Delineation of the residues involved can be achieved either by comparative structural analysis or chemical modification studies on iron binding. The similarity of the two iron binding domains is supported by the predicted secondary structure of human transferrin, which indicated that the two domains have generally the same secondary structure conformation type and distribution (MacGillivray et al, 1982). More compelling evidence was provided by an analysis of the disulphide cross linkages within human and chicken transferrins. This showed that none of the disulphide bridges linked residues from different domains, and perhaps more importantly, the pattern of linkages was very similar in each domain within each protein, which would generate protein domains with very similar tertiary structures (Williams et al, 1982). The specific residues which function in iron binding, would be expected to be conserved in both domains of all species variants and possibly in the homologous lactoferrins. Amino acid sequence comparisons limit the number of histidyl residues to 1 or 2 of 3 pairs: 139/487, 233/574, 275/624 (although other amino acids may also provide the nitrogen ligands), and tyrosines to 2 or 3 of the 4 candidate pairs: 105/443, 115/462, 211/553, 214/556 (MacGillivray et al, 1983). In addition to these direct ligands, conserved arginine residues occur at positions 144/492, 258/607, 280/629, one

of which probably interacts with the bicarbonate iron ligand (Williams et al, 1982). The residue numbers of the amino acids mentioned above and below have been altered to correspond with the appropriate residue in figure (32). Differential kinetic labelling studies suggest that, histidines 233 and 574 co-ordinate with the Fe^{3+} iron because the lysines adjacent to these residues undergo extreme changes in acetylation rates upon iron binding (Shewale and Brew, 1982). Such studies rely on conformational changes within a protein, iron binding altering the microenvironment of the amino group and the distribution of lysine residues. Therefore no conclusion may be drawn from negative results, because large conformational changes may be limited by other local strong interactions, or the lack of a lysine residue near the liganding amino acid. Similarly, it is sometimes difficult to distinguish conformational effects of modification from more specific influences at the metal binding site (Aisen and Brown, 1975).

Tyrosine nitration protection experiments have shown that all four pairs of potential tyrosine ligands are protected by iron binding (Williams et al, 1982). In particular tyrosines 211/553 and 214/556 (Figure 32) are favoured as candidate ligands, because they could form a locus for weak association prior to the formation of the complete binding site by conformational adjustments (MacGillivray et al, 1983), although presumably this would also be true for any other closely associated potential ligand donating residues.

Rodent Transferrins

Mouse transferrin has a molecular weight of 77,500 (Sawatzki et al, 1981), and rat transferrin has a plasma molecular weight of 76,500, and accounts for 6% of rat

plasma protein and is glycosylated (Schreiber et al, 1979). The leader peptide of rat transferrin is 20 residues in length and the sequence of the first 45 protein residues is known, with an additional 29 residues of later sequence (Aldred et al, 1984). Within the first 45 residues of the rat sequence, 36 are identical to the human sequence (positions 19-65, Figure 32), representing 80% homology. The homology of rat transferrin to the equivalent portion of the chicken sequence is only 36% The additional known 29 amino acid sequence of rat transferrin (Aldred et al, 1984) exhibits 79% and 50% homology to the equivalent human and chicken transferrin respectively (positions 274 to 302, Figure 32).

Over the limited amount of rat transferrin sequence known, there would appear to be approximately 80% homology between rat and human transferrins, which compares with a value of 50% between human and chicken transferrin (MacGillivray et al, 1983) and 50% between rat and chicken. The reason why the homology between the first area of amino acid comparison between rat and chicken transferrins is 36% compared to 50% is unknown. However, the first region does include the 20th to the 40th amino acids of the mature proteins, which exhibit a lower degree of conservation relative to the rest of the proteins within the transferrin superfamily of lambda Ch Blym-1, human transferrin, lactotransferrin and ovotransferrin (Goubin et al, 1983).

The rat nucleotide sequence (Aldred et al, 1984) shows homology to 9 of the 28 known nucleotides of lambda Ch Blym-1. This compared with 11 nucleotides in human transferrin and 6 in both lactotransferrin and ovotransferrin (Schreiber et al, 1979 and Goubin et al, 1983).

Chromosomal Location and Structure of Transferrins

Electrophoretic separation of mouse serum transferrins revealed two genetic variants in laboratory strains of mice, both of which exhibited microheterogeneity (Cohen, 1960 and Shreffler, 1960). There are approximately 20 such rare variants in man, found in approximately 1% of the population (Aisen and Brown, 1975). The two variant types of mouse transferrins were shown to be controlled by a pair of co-dominant alleles, located on linkage group II (Shreffler, 1963), which corresponds to chromosome 9 (Womack, 1979). The microheterogeneity exhibited additional variable changes, associated with development, stress and disease, all of which affect the relative mobility of transferrins (Cohen, 1960; Shreffler, 1963 and Klein et al, 1966). Microheterogeneity in rats has been shown to be due to the glycan chains attached to transferrin (Schreiber et al, 1979). More recently genomic DNA hybridization have shown that both in the rat (Levin et al, 1984) and human, (Uzan et al, 1984) transferrin appears to be encoded by a single gene; which is located on chromosome 3 in humans (Yang et al, 1984).

The chicken transferrin gene has been cloned and its basic organization elucidated. The oviduct and liver transferrins are the same protein. The sequence is unique in the chicken genome and therefore there is probably only one chicken transferrin gene (Cochet et al, 1979 and references therein). The gene consists of 17 exons, approximately 60-200 bases long, encoded within 10.3kb of genomic DNA. There is a 76 nucleotide 5′ untranslated region in the mRNA, which is preceeded 30 base pairs upstream, by an AT rich promoter region, which exhibits a 12 nucleotide identity with the equivalent position in the adenovirus-2 major late genes CTATAAAAGGGG and two further upsream homologies, TAGGT at position -71 and of CAAGGAAGG at position -91 (Cochet et al, 1979).

Comparison of the first 250 bp upstream from the capsite to the corresponding estrogen/progesterone inducible egg white protein ovalbumin gene, revealed no significant sequence similarities (Cochet et al, 1979 and references therein). Similarly there were no significant homologies to the glucocorticoid or progesterone receptor binding sites (Von der Ahe et al, 1985) in the 250 bp upstream sequence of chicken transferrin.

Two overlapping human transferrin genomic clones have been isolated and the 12 exons sequenced. These correspond to exons 3-14 by analogy with ovotransferrin. However, although the human exons are of a similar length to those of ovotransferrin, the lengths of the introns are significantly different. Since the 12 exons of the human transferrin gene are contained in a 24kb DNA segment, the total length will be much greater than the 10.3kb of the ovotransferrin gene (Park et al, 1985). Comparison of the exons of the two genes would suggest that human transferrin also has 17 exons.

A simple model describing the origin of present day human transferrin and ovotransferrin has been proposed. It postulates an ancestral gene of 10 exons, with exons 1 and 10 encoding the signal and 3′ regions of the gene respectively. Sequences from exons 2 to 9 were duplicated by unequal crossover between exons 1-2 and 9-10 and led to an 18 exon intermediate gene. This subsequently lost one exon from the 5′ half, equivalent to exon 4 of the ancestral gene, to form the present day 17 exon structure. An alternative model would propose a 9 exon ancestral gene of 16 exons, with exon 17 being inserted later in the 3′ moety (Park et al, 1985 and references therein).

Transferrin cDNAs

Near full-length cDNAs for the chicken (Jeltsch and Chambon, 1982) and human (Yang et al, 1984) transferrin sequences have been cloned and sequenced. The terminal halves of these are presented in Figure(31). The chicken transferrin mRNA is known to have a 76 nucleotide 5′ untranslated region. It has been noted that nucleotides 2-10 of this sequence could form a very stable 9 base pair stem-loop structure, very much like nucleotides 3-10 of the ovalbumin mRNA 5′ untranslated region. Both of these sequences immediately preceed regions of mRNA which could interact with 3′ 18S rRNA. Considering the similarities in the tissue specific hormonal induction of these two genes, it is thought that these structures may be involved in the control of their expression (Cochet et al, 1979). Comparison of the equivalent sequence in human transferrin is not possible, due to the cDNA being truncated close to the start of the leader peptide (Yang et al, 1984). The similarity of the proteins encoded by human and chicken transferrin mRNAs has already been discussed. The 3′ untranslated regions are 171 (human) and 182 (chicken) nucleotides in length and the polyadenylation signal AATAAA occurs 29 and 18 nucleotides from the ends of human and chicken mRNAs respectively. (Jeltsch and Chambon, 1982 and Yang et al, 1984).

Developmental and Hormonal Controls of Transferrin Expression

The visceral yolk sac of the mouse foetus is thought to provide several of the hepatic functions involved in nurturing and protecting the growing foetus, until the foetal liver takes over these functions, which includes the synthesis of transferrin (Meehan et al, 1984). In

foetal rat liver, the level of transferrin synthesis doubles within the last three days of gestation to the adult level (Levin et al, 1984), where it is the major site of synthesis (Morgan, 1983; Levin et al, 1984 and Meehan, 1984).

Several rat foetal tissues were also examined for transferrin synthesis (Levin et al, 1984). Transferrin mRNA was detected in all the foetal tissues tested which included: muscle, small intestine, spleen, lung, kidney, heart and brain. In general the levels rose to a peak concentration at 1 to 3 days before birth, at 1/10th - 1/20th the level present in the foetal liver, and decreased rapidly thereafter to approximately 1/100th the liver level in the adult or was not detectable (heart and small intestine). The major exception to this pattern is the brain where no prenatal increase in expression occurs. However a linear post natal increase occurs in the rat, attaining on reaching maturity, a level of synthesis equivalent to 1/10th the rate in adult liver (Levin et al, 1984). Measurement of transferrin mRNA levels in adult tissues of the mouse, have revealed levels of synthesis equivalent to 1/50th - 1/100th the rate of adult liver in the brain, spleen, testes and small intestine. The mRNA of the latter is possibly smaller in size than those of the other tissues (Meehan et al, 1984). However, no transferrin synthesis has been detected in the adult rat small intestine (Levin et al, 1984). Similarly, cultured rat sertoili cells have been shown to secrete large amounts of iron binding protein (molecular weight 71,000), antibodies to which preciptitate serum transferrin (Skinner and Griswold, 1980).

Speculation on the developmental expression of transferrin within foetal tissues has suggested that transferrin is involved in tissue differentiation (Levin et al, 1984 and references therein). However, it would seem equally, if not more plausible, that the synthesis in

extra embryonic tissue acts as a supplement to that produced by the visceral yolk sac, given the rapidity of foetal growth and the relatively late development of hepatic function in the foetal liver. The expression of transferrin (or like proteins) in the adult testes and brain is to be expected, considering the essential requirement for Fe^{3+} and thus a transporting protein to all living tissues; since plasma proteins are excluded from the brain C.S.F. and the lumen of seminiferous tubules by specialised membranes. The low levels of transferrin (or transferrin like) protein expression in the small intestine and spleen, may reflect their greater involvement with iron metabolism, being the major sites of iron absorbtion from food and iron turnover from old reticulocytes respectively.

Most of the work on the hormonal regulation of transferrin has concentrated on the chicken gene, because the gene is expressed constitutively in the liver and is unaffected by estrogen or progesterone. Whereas conalbumin (transferrin) synthesis in the oviduct tubular gland cells is inducible with a variety of steroid hormones (estrogens, progestins, glucocorticoids and androgens) (Palmiter et al, 1981; and references therein). In particular, estrogen causes a very rapid (almost instantaneous) rise in conalbumin synthesis, whereas the response to progesterone is slower (onset of increased protein synthesis, 2 hours after hormone administration) and rises gradually. The action of progesterone is dominant over that of estrogen, so much so, that progesterone administration abruptly inhibits estrogen stimulated transcription and requires ~2 hours to restart transcription. Furthermore in chick oviduct, 8 hours after estrogen stimulation, the rate of conalbumin synthesis parallels the number of nuclear estrogen receptors.

One model proposes that conalbumin has a single steroid receptor binding site and that progesterone

receptors displace estrogen receptors due to a higher affinity for the receptor site. The replacement of one receptor type by another does not allow transcriptional continuity from the conalbumin gene. Receptor binding to a specific DNA sequence is coupled to transcriptional activation by one or more proteins, with short half lives, which may effect time dependent receptor activation of transcription. Such a model would not only explain the transient inhibition of estrogen stimulated transcription, but also the slow onset of conalbumin transcription by progesterone (Palmiter et al, 1981 and references therein).

The effects of insulin on the synthesis of specific plasma proteins in cultured chick hepatocytes have been investigated. Although most of the main secretory proteins of hepatocytes showed either rapid (within 1 hour) or delayed (1-2 days) increases in protein synthesis on insulin administration, transferrin production was unaffected (Liang and Grieninger, 1981). However one factor that is known to affect transferrin gene expression in the liver is nutritional iron deficiency, which causes a 2-4 fold increase in mRNA and protein synthesis. The mechanism of action however remains obscure (McKnight et al, 1980).

METHODS

Reagent Purifications

Recrystallizations

Acrylamide and bisacrylamide were dissolved in chloroform and acetone respectively at 50°C, filtered while hot and allowed to recrystallize at -20°C. The recrystallized reagents were recovered by filtration and dried under vacuum. Formamide was purified by three recrystallizations of the liquid at 0°C for 12 hours and stored at -20°C.

Distillations

M-Cresol and Dimethylsulphoxide (DMSO) were distilled under reduced pressure at 50°C and stored at -20°C. Phenol was distilled at atmospheric pressure and saturated with water when it had cooled to 40°C. Saturated phenol was stored at 4°C.

Agarose Gels

Horizontal 0.6 - 2.0% Gels

The gels were 26 cm x 19 cm (wide) x 0.5 cm and prepared in the following manner. Agarose was refluxed in distilled water for 20 minutes, cooled to 50°C and adjusted to 1 x Tris Acetate (T.A.) and 1 µg/ml ethidium bromide and the gel cast. Buffer reservoirs (each 400 ml) were connected to the gel with Whatman 3 MM paper wicks. Samples were adjusted to 1 x TA/3% w/v Ficoll 400/0.005% BSA/10 mM Na_2EDTA. After sample loading the gel was covered with a sheet of polythene and electrophoresed at

40-150 V. DNA was visualized using a short wavelength UV transilluminator and photographed using a Polaroid camera with a red filter.

Vertical 0.8 - 1.5% Gels

The gels were made as before except that Tris Borate buffer (T.B.) was used and ethidium bromide was not added to the gel. Samples were adjusted to 1 x TB/3% w/v Ficoll 400/0.005% BPB/10mM Na_2 EDTA. Electrophoresis was at 30 - 60 V overnight. Gels were stained with ethidium bromide at 1 µg/ml in 1 x TBE for 30 - 60 minutes to visualize the DNA.

Vertical Formaldehyde Gels

Denaturing 1.4% agarose gels were prepared as described for a vertical agarose gel except that the gel was made 1 x MOPS Buffer/6% formaldehyde (Rave, Crkvenjakou and Boedtker, 1979). Samples were incubated at 60^{o}C for 5 minutes in 45% formamide/6% formaldehyde/0.9 x MOPS buffer, chilled rapidly and made 6% Ficoll 400/0.01% BPB by addition of a 5 x stock solution. The electrode buffer was 1 x MOPS and samples were run at 35 volts overnight (Clissold and Bishop, 1982).

Polyacrylamide Gels

Native Vertical

The gels 19.5 cm x 16 cm (wide) x 0.2 cm were made using a stock solution containing 14.7% recrystallized acrylamide and 0.375% recrystallized bis-acrylamide. The gels were made 1 x in TA/0.033% wt/vol ammonium persulphate/0.066% wt/vol TEMED. The application, electrophoresis and visualization were the same as those for horizontal TA agarose gels. The gels were stained by

immersion in 1 x TA/ethidium bromide 5 µg/ml for 15 minutes, to minimize the diffusion of small DNA bands.

SDS Polyacrylamide Gels

Denaturing SDS polyacrylamide gels 19.5 cm x 16 cm (wide) x 0.2 cm, for separating proteins were prepared as described by Laemmli (1970) except that the stacking gel was 1 cm/pH 6.8/5% polacrylamide and the separating gel was 18.5 cm/pH 8.8/13% polyacrylamide. Samples (20 - 30 µl) were boiled for 3 minutes and electrophoresis was for 3 hours at 100 volts.

Gels were fixed by shaking gently for 1 hour in 2 changes of 500 ml 10% acetic acid/40% methanol and then overnight in 7% acetic acid/20% methanol. The gel was impregnated with PPO in DMSO as described by Bonner and Laskey (1974) and autoradiographed (Laskey and Mills, 1975).

Polyacrylamide Sequencing Gels, pH 8.8

Thin sequencing gels 40 cm x 20 cm (wide) x 0.4 mm of the type described in Sanger and Coulson (1978) were used except that Amberlite MB3 was used to deionize the acrylamide/bisacrylamide/urea stock solution. The TBE buffer was made pH 8.8 (Winter and Coulson, 1982) and the gel mix was degassed and polymerized by the addition of 0.06% w/v ammonium persulphate and 0.08% v/v TEMED. Plasticard well formers were either 2.5 mm or 5 mm wide set 2 mm apart. Sequencing reaction samples were made 45% formamide/0.01% BPB and xylene-cyanol green/15 mM Na_2 EDTA, then heated to 100^{o}C for 6 minutes. Sample aliquots of 1 to 2 µl were analysed on the gel by electrophoresis at 25 - 27 watts (~ 1000 volts) as required. Gels were fixed in 1L 10% acetic acid/10% methanol for 20 minutes and dried onto Whatman 3MM paper at 80^{o}C for 2 hours under vacuum.

Polyacrylamide Gradient Sequencing Gels, pH 8.3

Gradient sequencing gels were prepared as described in Biggin, Gibson and Hong (1983) except that the buffers used were 0.5 or 2.5 x TBE pH 8.3. A crude gradient of 6 ml/0.5 TBE gel mix and 6 ml 2.5 TBE gel mix was formed in a 25 ml pipette. This was run down the edge of the glass plate and the remainder of the gel was formed with 0.5 TBE gel mix. Sample wells and samples were prepared as described for sequencing gels pH 8.8. The top electrode buffer was 0.5 x TBE, the lower one 2.5 x TBE and electrophoresis was carried out at 27 watts for 2.5 to 3 hours. Gels were fixed and dried down as described for sequencing gels pH 8.8.

Formamide Polyacrylamide Gels

Denaturing 5% polyacrylamide gels, 19.5 cm x 16 cm (wide) x 0.2 cm comprising 98% formamide/4.25% acrylamide/0.7% bisacrylamide/20mM Na PO_4 pH 7.5 were cast (Maniatis, Jeffrey and Van-de-Sande, 1975). Samples were made 93% formamide/0.01% BPB/20mM Na PO_4 pH 7.5 and heated to 100°C for 3 minutes. Electrophoresis was at 150 volts for 4 - 6 hours with the electrode buffer (20mM Na PO_4 pH 7.5) circulating. Gels were fixed by shaking gently for 30 minutes with 2 changes of 500 ml 40% methanol/10% acetic acid and dried onto Whatman 3MM paper at 70°C overnight under vacuum.

Restriction of DNA with Enzymes

Restriction digests using Eco RI, Bam HI, Hind III, Hinf I, Pst I, Sal I and Sst I were performed in Eco RI buffer. All other enzymes were used in the buffers recommended by the suppliers. Where the dilution of a restriction enzyme was required this was accomplished using restriction enzyme diluent.

Phenol/Chloroform Extraction of Nucleic Acids

Nucleic acids were deproteinized with a 1:1 mixture of neutralized phenol/0.1% 8 hydroxyquinoline/0.2% beta-mercaptoethanol and chloroform/4% isoamyl alcohol as described in Maniatis, Fritsch and Sambrook (1982). The solution to be deproteinized was made 50% v/v with the "phenol/Chloroform" mixture and mixed thoroughly for 5 - 10 minutes at 37°C. The aqueous and organic phases were separated by centrifugation. The aqueous layer was passed either twice over an equal volume of chloroform or 5 volumes of ether. The organic phase was routinely back extracted with aqueous buffer to minimize losses.

Ethanol Precipitation of Nucleic Acids

DNA was precipitated by the addition of NaCl to a concentration of 0.15M and 2.5 volumes of ethanol, unless otherwise stated. The solution was mixed thoroughly and the DNA precipitated for either 1 hour in a cardice/ethanol bath or 3 hours at -70°C or at -20°C overnight. The DNA was pelleted by centrifugation (Sorval HB4, 10K rpm, 30 minutes or MSE micro-centaur, 13K rpm, 20 minutes at 0°C and 4°C respectively). The DNA pellet was routinely rinsed with ice cold ethanol prior to drying under vacuum at room temperature. RNA was precipitated in a similar manner by the addition of 0.3M NaOAc pH 5.0 and 2.5 volumes of ethanol, unless otherwise stated.

Electroelution of DNA from Gels

To extract DNA from an agarose gel, a slice of agarose was removed from immediately in front of and to either side of the DNA to be eluted. Dialysis membrane was then placed in front of and underneath the gel containing the fragment. The trough was then filled with buffer and the DNA electrophoresed onto the membrane.

The dialysis membrane and DNA was rapidly removed to either 1 ml of 100 mM Tris-HCl pH 8.0 or 3 ml elutip-d-column low salt buffer. The DNA was recoved after "Phenol/Chloroform" extraction of the Tris/DNA Solution by ethanol precipitation or passed over an elutip-d-column as described by the suppliers, except that the DNA was eluted in 0.2 ml of the high salt buffer and ethanol precipitated in 0.5M NaCl/2.5 volumes of ethanol overnight at -20°C.

Extraction of DNA from acrylamide gels was accomplished in a similar manner except that the DNA containing gel fragment was placed in a small bag of dialysis tubing and filled with gel buffer. After electrophoresis the DNA was recovered from the buffer and tubing as described for DNA from agarose gels.

Transfection of E.coli HB101

An aliquot of a static overnight culture of E.coli HB101 (Boyer and Roulland-Dussoix, 1969) was diluted 1:50 in L Broth. The bacteria were grown and made competent as described in Mandel and Higa (1970). The competent cells were kept at 4°C until required.

Transfection was initiated by the incubation of 100 µl of competent cells with 50 ul TMC containing 1.25 or 12.5ng/µl plasmid DNA on ice for 15 minutes. The transformation mix was warmed at 37°C for 2 minutes after which 1 ml of L Broth, at 37°C was added and incubation at 37°C continued for a further 45 minutes. The cells were then plated on LB in 2.5 ml of LB top agar. The LB plates contained antibiotics to select for transformants. Half strength antibiotic was also included in the LB top agar.

Bulk Preparation of Plasmid DNA

Bacteria transformed with the appropriate plasmid were grown under constant selection with the antibiotics

to which they were resistant. Bacteria were cultured and plasmid DNA isolated as described in Bishop (1979) except that small DNA fragments were removed by passage over a Sepharose 2B column and the plasmid DNA ethanol precipitated.

Preparation of RNA from the Endoplasmic Reticulum (ER) of Female BALB/c Mouse Liver

The rapidly sedimenting (RS) ER membranes were prepared from the livers of 12 week old female BALB/c mice by the method of Shore and Tata (1977) except glutathione 3mM was substituted for the beta-mercaptoethanol and the concentration of magnesium acetate in the STKM solutions were 5mM.

RNA was extracted from the RS ER by the method of Parish and Kirby (1966) with the modifications that no NaCl was added to the aqeous phases and m-cresol was not present when the RNA was ethanol precipitated. The RS ER was diluted with 2 volumes of the "Kirby" aqueous solution. This was extracted twice with an equal volume of "Kirby" phenol, shaken at 150 rpm for 15 - 30 minutes at room temperature and the two phases separated by centrifugation. After the second extraction the RNA was precipitated with sodium acetate and ethanol.

Preparation of Poly(A) mRNA

Poly(A) mRNA was prepared from R.S. E.R. RNA by passage over an oligo dT_{10-12} cellulose column in a manner similar to that described by Aviv and Leder (1972). The RNA was dissolved in 20mM Tris HCl pH 7.5/0.5% SDS. This was heated to 70°C for 5 minutes and made 0.5M LiCl (Clissold, Mason and Bishop, 1981) before being passed over the column three times. The column was then washed thoroughly

with 50mM Tris-HCl pH 7.5/0.5% SDS/0.4M LiCl. The poly(A) mRNA was eluted from the column with 20mM Tris-HCl pH 7.5/0.1% SDS and ethanol precipitated.

The poly(A) mRNA was then applied to a 4 - 20% linear sucrose gradient and fractions containing 4S RNA and smaller were discarded (Clissold, Mason and Bishop, 1981). The Poly(A) mRNA was precipitated twice from ethanol, taken up in distilled water and frozen in aliquots at $-196^{o}C$.

The above procedure was modified wehere the poly(A) mRNA was to be used in the synthesis of cDNA. The SDS was omitted from the last 5 ml of washing buffer and the elution buffer. The eluate was immediately made 0.3M NaCl/2.5 volumes ethanol and stood at $-20^{o}C$ overnight. The RNA was recovered by centrifugation, washed with ethanol and stored as described for other poly(A) mRNA preparations.

Preparation of Diazaobenzyloxymethyl (DBM)-Paper

DBM paper discs, 13 mm diameter were made as described by Alwine, Kemp and Stark (1977) with the exception that the last DBM-paper wash was 1M NaOAc pH 4.0.

Attachment of Plasmid DNA to DBM-Paper

After conversion to DBM-paper and washing, the filter paper discs were blotted and incubated with sonicated denatured recombinant plasmid DNA 660 µg/ml/80% DMSO/20% 25 mM $NaPO_4$ pH 6.0 in a total volume of 30 µl per disc. The filters and DNA solutions were incubated, washed, treated with NaOH and stored as described in Stark and Williams (1979).

Annealing Poly(A) mRNA to DBM-Paper/DNA Discs

Poly(A) mRNA was annealed to the DBM-paper/DNA discs as described in Clissold and Bishop (1981), except that the poly(A) mRNA concentration was 350 - 400 µg/ml and the annealing took place at 55°C for 16 hours. Up to 20 discs were washed for 15´at 45°C twice in 200 ml 150mM NaCl/15mM Na Citrate/0.1% SDS, then twice in 200 ml 20mM Tris-HCl pH 7.5/0.1% SDS. The annealed mRNA was eluted from each filter with two treatments of 80 µl distilled water for 1 1/2 minutes at 90°C. Pooled eluates from 4 - 5 filters were precipitated from ethanol and NaOAc in the presence of 2 µg carrier Guinea-pig liver tRNA, gift of Dr. P. M. Clissold.

Preparation of Message Dependent Reticulocyte Lysate (MDL)

Rabbit reticulocyte lysate (Pelham and Jackson, 1976) was a gift of Dr. P. J. Mason. The lysate was made mRNA dependent as described by Pelham and Jackson (1976), except that no amino acids were included in the master mix and micrococcal nuclease treatment was at 4 µg/ml at 20°C for 15 minutes.

Translation of mRNA by MDL

Either half the hybrid-selected mRNA and tRNA or 1 µg total poly(A) mRNA/1 µg tRNA was lyophilized together with 10 µCi ^{35}S methionine (~400Ci/m mole). The dried mixture was resuspended in 1 µl of an amino acid cocktail (each amino acid 40mM, except methionine) and 20 µl of MDL. Reaction mixtures were incubated at 30°C for 30 minutes. Incorporation was monitored by decolourising 2 µl aliquots of MDL in 300 µl 0.3M NaOH/0.8mg/ml BSA/8mM methionine/1.2% H_2O_2 at 37°C for 15 minutes and precipitating the proteins by the addition of 100 µl 50% w/v TCA and chilling to 0°C for a further 15 minutes.

Proteins were collected over Whatman GFC filter paper, washed, dried for 1 hour at 70°C under vacuum and counted in toluene/POPOP/PPO scintillant using a Packard Tri Carb Liquid Scintillation Spectrophotometer.

Immuno Precipitation and Recovery of Translation Products

MDL incubation mixtures were diluted in PBS/0.5% SDS/1% triton x -100 and polyribosomes were removed by centrifugation (Clissold, Mason and Bishop, 1981). Antibody precipitation was performed by either of two methods. The modified method of Kessler (1975) was to add 1/10th volume of the appropriately diluted antibody to the supernatant and incubate for 1 hour at 37°C and overnight at 4°C. Immune complexes were concentrated by incubation with 1/20th volume pre-swollen protein A sepharose CL-4B beads for 3 hours at room temperature. The sepharose beads were washed with 1/3rd volume of PBS/0.1% SDS/1%Triton x-100 three times. Alternatively, the method of Bostain, Lemire, Cannon and Halvorson (1980) was used except the antibody precipitation conditions were modified to 13.3% diluted lysate/6.7% diluted antibody/80% BSB and the immune complex aggregates were washed with 1/2 volume BSB/0.01% SDS/0.1M methionine/0.3mg/ml BSA three times.

The immune precipitate/bead complexes were dissociated by boiling for 3 minutes in an equal volume of 2 x Laemmli sample buffer (Laemmli, 1970). Beads were removed by filtration and the translation products analysed using 13% polyacrylamide/SDS gels as described previously.

Southern Transfers

Southern transfers were carried out essentially as described by Southern (1975), with the following modifications. Gels were treated whole and the neutralization step extended to 45 minutes. The stack was

arranged in a slightly different manner in that there were 2 sheets of 2 x SSC soaked filter paper under the gel, 3 sheets of 2 x SSC wetted filter paper directly on top of the nitrocellulose membrane, followed by a 7 cm high stack of paper towels. Efficient contact was ensured by the application of a 2 kg weight to the top of the stack. Transfer of DNA in this manner was allowed to proceed for 24 hours. The stack was then dismantled and the membrane washed in 2 x SSC, blotted dry and baked at 80°C under vacuum for 1 1/2 to 2 hours.

Northern Transfers

RNA was transferred to nitrocellulose membranes from denaturing agarose gels as described for Southern transfers except the recommendations of Thomas (1980), not to stain the gel, pretreat it with alkali and neutralization solutions or wash the membrane before baking were observed.

Labelling RNA

Phosphate free ends were generated by treating RNA with 100mM Na_2CO_3 at 40°C for 1 hour. The cleaved RNA was end labelled as described by Donis-Keller, Maxam and Gilbert (1977), except the RNA was heated to 60°C and the buffer pH dropped to 7.5 to suit the T4 polynucleotide kinase and the period of incubation extended to 1 hour.

Labelling DNA by Nick Translation

Nick translation was performed by a modified method of Rigby, Dieckmann, Rhodes and Berg (1977). One volume of DNase, appropriately diluted in 50mM Tris-HCl pH 7.5/100 µg/ml BSA to generate one nick per 0.5 - 2.0 kb of DNA,

was added to 4 volumes of DNA in 66mM Tris-HCl pH 7.5/6mM $MgCl_2$. The reaction was incubated at 20°C for 7 minutes. DNA was recovered after "Phenol/chloroform" extraction by ethanol precipitation.

The nicked DNA was labelled using E.coli DNA polymerase I in 66mM Tris-HCl pH 7.5/6mM $MgCl_2$/5mM DTT/30mM dNTP 's (dGTP, dATP and dTTP) and 2.5 - 20 µCi alpha-^{32}P-dCTP (~400Ci/mmol) per µg of nicked DNA. Incubation was at 30°C for 30-60 minutes.

The End Labelling of DNA Fragments

Reagents were added to the various restriction digests to bring conditions near enough to the enzyme supplier's specified levels for functional T4 DNA polymerase activity. The final reaction conditions varied between 10 - 40mM Tris-HCl pH 7.5 - 8.0/50-100mM NaCl/8-10mM $MgCl_2$/either 1mM DTT or 5-10mM beta-mercaptoethanol. Nucleotide triphosphates (dATP, dGTP and dTTP) 60 µM were added when required together with alpha-^{32}P-dCTP (400Ci/mmol) 10 - 20 µCi/µg DNA. Incubation was at 30°C with T4 DNA polymerase until satisfactory incorporation had been achieved.

Estimation of Radioisotope Incorporation

The incorporation of radioactive substrates into nucleic acids was determined in the following manner. Samples were added to 1 ml of ice cold 0.2M $Na_4PO_2O_7$/250 µg/ml BSA then 0.5ml of 50% w/v TCA was added and the solution mixed. After standing on ice for 15 minutes the precipitated DNA was collected on Whatman glass filters by vacuum filtration, washed with 5% TCA and dried under vacuum at 75°C for 1 hour. The radioactivity in a sample was estimated by immersion of the filter in scintillation fluid and counting in a Packard Tri Carb liquid scintillation spectrometer.

EX BIBL. UNIV. EDINBURGEN.

Removal of Unincorporated Nucleotides

Unincorporated nucleotides were removed from nucleic acids by column chromatography. Two methods were used, either reaction mixtures were layered on to a 8 cm (long) x 1 cm column of Sephadex G50 and the column developed with 0.3M NaCl/50mM Tris-HCl pH 7.5 or the reaction mixture was applied to a spun Sephadex G50 column as described in Maniatis, Fritsch and Sambrook (1982)

Nitrocellulose Membrane Hybridizations

Nitrocellulose membranes were hybridized to labelled probes as described by Maniatis et al, (1978) except that the pre-treatment of filters was at 0.5 - 5.0 cm^2/ml and the prehybridization, hybridization steps were at 10 - 30 cm^2/ml. Dextran sulphate 10% w/v was included in the hybridization (Wahl, Stern and Stark, 1979) and poly(A) was omitted from the washing procedure. Occasionally the stringency of washing was increased by adding a final wash of 0.1 - 0.5 x SET.

cDNA Synthesis

The conditions allowing the synthesis of long DNA copies of female BALB/c liver ER poly(A) mRNA were investigated using AMV reverse transcriptase (Clissold and Bishop, 1981; Lawn et al, 1981; Retzel, Collette and Faras, 1980; Winter et al, 1981; and Woo et al, 1977). In addition the following reagents were tested in the final system to determine whether they could influence the length of cDNAs obtained, Vanadyl ribonucleoside complex (BRL) (Berger and Birkenmeier, 1979), Polyethylene glycol (Chan et al, 1980) and RNasin (de Martynoff, Pays and Vassart, 1980).

The final protocol conditions for cDNA synthesis were as follows. The mRNA and $dpT_{(10-12)}$ were incubated for 5' at 37°C, 54 µg/ml mRNA/2.7 µg/ml $dpT_{(10-12)}$/200mM NaCl. Further reagents were added to bring the reaction conditions to those set out in Table II. The reaction mixture was incubated at 37°C for 30 minutes.

Second Strand Synthesis of cDNA

The mRNA in the cDNA preparation was destroyed by treatment with 25mM NaOH at 65°C for 1 hour and the pH adjusted to 7.0 - 7.5 by the addition of HCl and Tris-HCl pH 7.0. The cDNA was purified by "Phenol/chloroform" extraction, ethanol precipitation followed by passage over a spun Sephadex G50 column and reprecipitated from ethanol a second time.

Second-strand synthesis and digestion with nuclease SI were performed as described in Maniatis, Fritsch and Sambrook (1982) except that the reverse transcriptase reaction conditions were as those set out in Table II with the modifications that Actinomycin D was omitted and all the dNTP's were 500 µM.

Cloning dScDNAs

The cDNA products were separated on a 5% acrylamide gel and cDNAs between the size ranges of 900Bp to 3000Bp were electroeluted and the cDNA ends were made flush with a brief E.coli DNA polymerase I "Klenow Fragment" treatment (Maniatis, Fritsch and Sambrook, 1982).

The cDNAs were ligated into the plasmid vector pUC 8 (Vieira and Messing, 1982) linearized with the restriction enzyme Sma I. The vector (25ng) was ligated with a 2.5 molar excess of dScDNAs under the following conditions 50mM Tris-HCl pH 7.5/10mM $MgCl_2$/1mM ATP/1mM DTT and T4 DNA ligase 1.25 units in 20 µl at 15°C overnight.

TABLE II

Table to show the reaction conditions used for the synthesis of sS cDNA

Substance	Concentration
♀ BALB/c ER poly(A) mRNA (liver)	40µg/ml
poly $dT_{(10-12)}$	2µg/ml
NaCl	140mM
Tris-Hcl pH 8.3 at 25°C	50mM
$MgCl_2$	6mM
DTT	2mM
BSA (nuclease free)	100µg/ml
dATP, dGTP, dTTP	500µM
dCTP	75µM
α-^{32}P-dCTP 400Ci/mmol	0.19µM
Actinomycin D	50µg/ml
AMV Reverse Transcriptase (nuclease free)	300units/ml

Aliquots of the ligation mix were used to transform the E. coli K-12 strain JM 83 (Messing, 1979) by the method of Hanahan (1983). Colonies transformed with recombinant plasmids, ampicillin resistant and white were transferred to master plates 22.5 cm x 22.5 cm. They were arranged so that a 96 prong replicator tool could be used to transfer colonies, either to other plates for analysis or to 96 well microtitration plates for long term storage at -25°C (Gergen, Stern and Wensink, 1979).

Preparation of cDNA Library Filter Replicas

Filter replicas of the collection were made from duplicate LB agar replicas of the master plate colonies as described in Gergen, Stern and Wensink (1979), except that the selective antibiotic used was ampicillin, 100 µg/ml and the colonies were transferred on autoclaved Whatman 541 filter paper for plasmid amplification on tetracycline, 50 µg/ml plates.

Hybridization Screening of cDNA Library Filter Replicas

Recombinant cDNA plasmids of the LVA series (Clissold and Bishop, 1981) and the MUP genomic sub clone BS 6-5 (Clark et al, 1984) were digested with restriction enzymes to remove plasmid sequences homologous to pUC 8. The resulting DNA restriction fragments were nick translated to a specific activity of $5 \times 10^6 - 5 \times 10^7$ dpm/µg.

Filter hybridizations were peformed at $30cm^2$/ml of hybridization mix at 68°C for 4 hours. Hybridization conditions were 2 x SET/5 µg/ml poly(A) 50 µg/ml denatured Salmon sperm DNA/1.25 - 2.5ng/ml denatured probe. Filters were washed twice at $3cm^2$/ml in 2 x SET/0.02% $Na_4P_2O_7$ at 68°C for 15 minutes and a further two washes at room temperature for 15 minutes (J. O. Bishop, personal communication). Filters were blotted, allowed to dry

overnight at room temperature and used to expose Kodak Xomatic S X-ray film using intensifying screens at -70°C.

Preparation and Analysis of cDNA Library Clones

Plasmid DNA mini-preparations were made by the method of Burnboime and Doly (1979) as modified by Maniatis, Fritsch and Sambrook (1982); except that the cell pellet was thoroughly resuspended in 20 µl of the culture supernatant prior to treatment with lysozyme.

Aliquots of plasmid DNA prepared by the above method were incubated with a four fold excess of restriction enzymes and the fragments analysed on a 1.5% TA agarose gel. After staining and photography some of the DNAs were transferred to nitrocellulose membranes by "Southern" transfer and the filters hybridized to nick translated probes.

Preparation of cDNA Subclones for Sequencing

The cDNAs in pUC 8 that were to be sequenced were digested with the appropriate restriction enzymes, analysed on agarose gels and the fragment elecroeluted for ligation into M13 mp8 or 9 (Messing and Vieira, 1982) or M13 tg 130 or 131 (Kieny, Lathe and Lecocq, 1983). Sub clones for sequencing MUP cDNAs were generated from the replicative form of MUP cDNA/M13 mp 9 constructs by the method of Hong (1982), except that the DNase concentration was 1.3 pg/µl, the digestion was carried out at 15°C for 6 or 12 minutes and linearised DNA was extracted from the gel by electroelution. The DNA was labelled and the end made flush by treatment with DNA polymerase I in the presence of alpha-^{32}P-dCTP and a spun Sephadex G50 column step was introduced after the PEG/NaCl precipitation to remove un-incorporated alpha-^{32}P-dCTP.

The ligation conditions used were the same as those described for the ligation of dScDNA and pUC 8, except that the T4 DNA ligase concentration and incubation time were reduced to 0.5 units/20 µl and 6 hours respectively whenever the DNAs to be ligated possessed complimentary protruding DNA sequences.

Transfection of E. coli JM101

E. coli JM101 (Messing, 1979) was made competent and transfected with recombinant phage DNA in the same manner as has been described for E. coli HB101, except for the following modifications. The heat pulse was at 42°C and immediately prior to the addition of the molten agar 250 µl of an IPTG/X-gal/Exponential JM101 cocktail was added (2 volumes X-gal 20 mg/ml/3 volumes IPTG 24mg/ml/40 volumes Exponentially growing JM101 in L broth). There were no antibiotics in the agars.

Preparation of M13 Replicative Form

A 250ml culture of JM101 cells infected with the desired recombinant virus was prepared as described by Winter and Coulson (1982). The cells were harvested, washed, lysed and the replicative form of the recombinant bacteriophage purified by the method described for the bulk preparation of plasmid DNA.

Preparation of Single Stranded M13 "Templates"

For convenience 12 - 14 "templates" were prepared at the same time as described by Winter and Coulson (1982) except that for some preparations the concentrations of PEG/NaCl were halved and the incubation step extended to overnight at 4°C. The templates were washed with 200 µl ethanol after centrifugation and the DNA resuspended in 30

- 35 µl. Aliquots of the template preparations were analysed for DNA concentration, purity and approximate size on 0.6% agarose horizontal TA gels.

Chain Terminator Sequencing

The sequences of M 13 recombinant bacteriophage templates were determined using the method of Winter and Coulson (1982) except that the following modifications were implemented. The synthetic "Universal" primer GTAAAACGACGGCCAGT was used and the annealing conditions were 8mM Tris-HCl pH 7.5/8mM $MgCl_2$/40mM NaCl/1 µg template DNA/1ng primer, in 12 µl at 70°C for 7 minutes. After cooling to room temperature 1 µl of each 100mM DTT, 24 µM alpha-^{32}P-dCTP (400Ci/mmol) and 0.5 units of E. coli DNA polymerase I "Klenow Fragment" were added. Three microlitres of the template/polymerase mix were added to each of the termination mixes Table III. The reaction was incubated at 30°C. After 15 minutes 2 µl of 0.5mM dCTP (cold chase) was added and the incubation resumed. A further modification was that some of the polymerization reactions were terminated by the addition of 1 µl 200mM EDTA and kept on ice; 2 µl aliquots of this were added to 1.5 µl formamide/DYE/EDTA mix. The sample was denatured and loaded onto the gels at 2 hour intervals in the usual manner.

Computer Sequence Analysis

The analysis of sequence data and their comparison with other sequences was performed with the assistance of UWGCG software and the NBPF and EMBL/Genbank data bases.

TABLE III

Table to Show Constitution of Termination Mixes.

Solutions	T mix	A mix	G mix	C mix
0.5mM dTTP	1µl	20µl	20µl	20µl
0.5mM dATP	20µl	1µl	20µl	20µl
0.5mM dGTP	20µl	20µl	1µl	20µl
50mM Tris HCl 1mM EDTA pH 8.0	5µl	5µl	5µl	5µl
0.5mM ddTTP	46µl			
0.5mM ddATP		46µl		
0.5mM ddGTP			46µl	
0.25mM ddCTP				65µl

RESULTS AND DISCUSSIONS

Characterization of LVA cDNA Clones

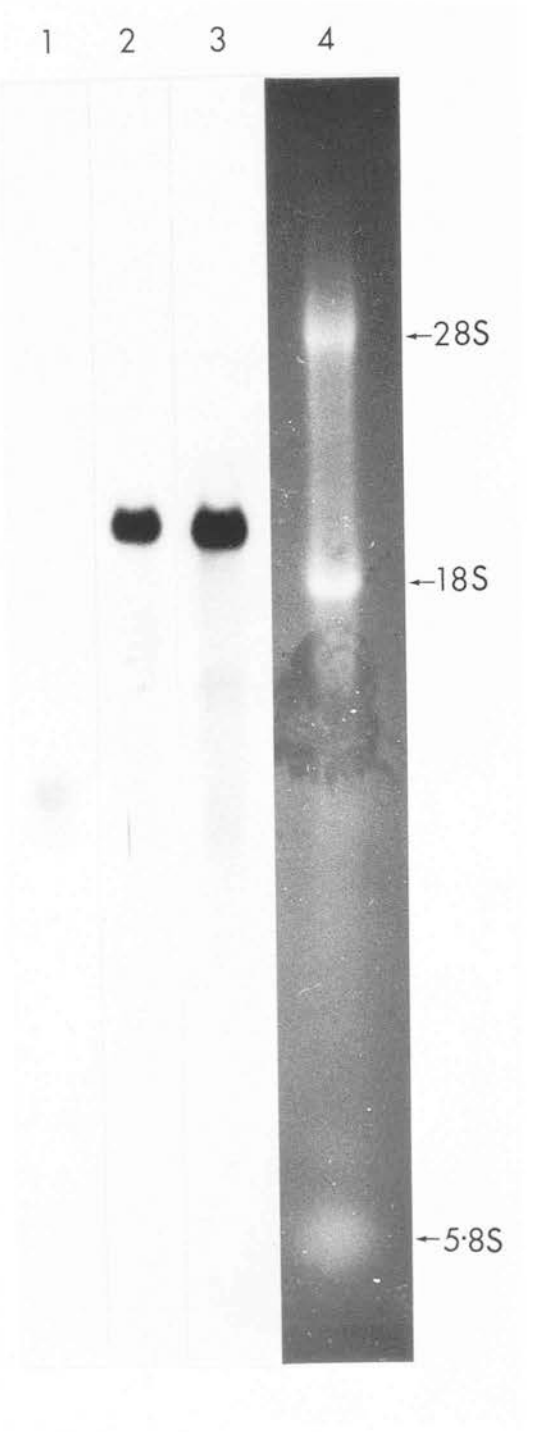
Three of the cDNA clones isolated by Clissold and Bishop, (1981) were subjected to further characterization. Clone LVA 301 had been identified as corresponding to $alpha_1$-antitrypsin and the mRNA complementary to LVA 321 has been shown to be relatively abundant in female liver E.R. poly(A) mRNA (Clissold and Bishop, 1981). The mRNA complementary to clones LVA 321 and 329 was 2.3 kb in length (Figure 3). This size, together with its abundance and its association with the liver endoplasmic reticulum, suggested that the mRNA could code for transferrin, a large secretory protein synthesized mainly in the liver (Putnam, 1975; Schreiber et al,1979 and Clissold et al,1981).

Specific identification of clones LVA 321 and 329 was attempted by hybrid selection of mRNA, translation of the mRNA in a template-dependant system, specific immuno-precipitation and polyacrylamide gel electrophosesis. The system was tested by using LVA 301 and $alpha_1$-antitrypsin antibodies as a positive control. Although the immunoprecipitated translation products of LVA 301 selected mRNA (333 $\pm$12*dpm) were visualized as a band of the expected size by fluorography of the SDS gels, no image was detectable for the equivalent LVA 321 or 329 selected mRNA products (396 $\pm$25*dpm). These findings are the same as those reported by Clissold and Bishop (1981). The amount of the product precipitated with transferrin antibodies (396 $\pm$25*dpm) were, however, significantly

FIGURE 3

Northern Blots of Female Mouse Liver E.R. mRNA probed with Recombinant Plasmids Containing Liver cDNAs.

Female Liver E.R.poly(A) mRNA (2 µg; lanes 2 and 3) was subjected to electrophoresis under denaturing conditions, transferred to nitrocellulose membrane, and probed with either of the hepatic cDNA plasmids LVA 321; track 2 or LVA 329; track 3. Track 1 contained 0.5 µg poly(A) mRNA probed with the nick-translated MUP probe LVA 325 (Clissold and Bishop, 1981). The final wash was 0.2 x SET at 68°C and exposure was for 2 days. Track 4 shows the positions of marker ribosomal RNAs stained with acridine orange.


1
2
3
4
←28S
←18S
←5·8S

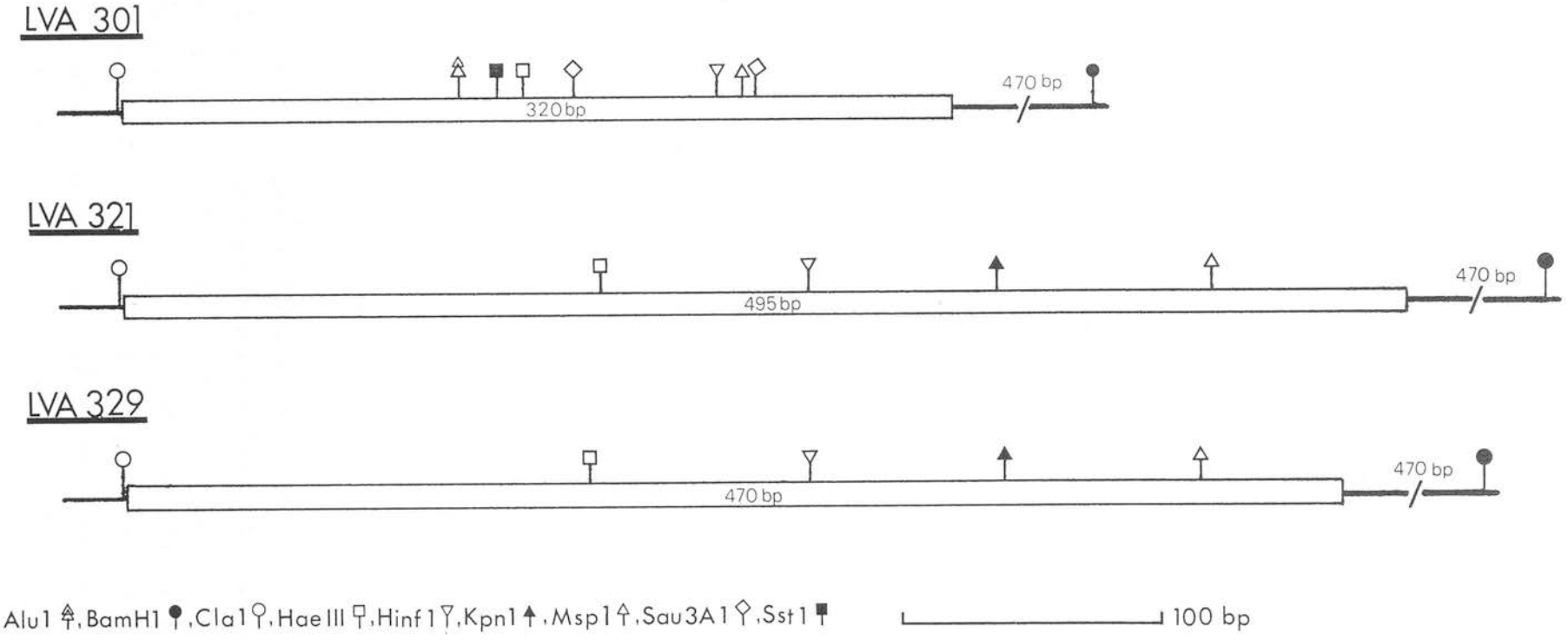
higher ($P<0.05$) than the background level of the pPH 207 negative control (270 $\pm$16*dpm). [Asterisks indicate the values for the standard error of the mean from 3 independant determinations, normalised for the total level of incorporation into translation products.] A possible explanation of these results is that the large mRNA species that corresponds to these clones is susceptible to degradation and is poorly transcribed by the cell-free translation system used. This would result in the generation of a range of truncated translation products which would not be possible to discern.

Sequential endonuclease digestions, DNA fragment isolations and an end labelling procedure were performed on the recontinant plasmids LVA 301, 321 and 329. The DNA fragments contained the cDNA insert and 470 Bp of plasmid DNA on one side (pPH 207 Hind III to Bam HI) and 6 Bp of end labelled plasmid DNA on the other, (pPH 207 Cla I to Hind III). The fragments were then partially digested with each of the following restriction enzymes Alu I, Bam HI, Cla I, Eco RI, Hae III, Hine II,Hind III, Hinf I, Kpn I, Msp I, Pst I, Pvu II, Sau 3lA, Sst I and Taq I to test for the presence of restriction sites in the cDNAs. The positions of these restriction sites are summarized in Figure 4. The detailed size analysis of LVA s 301, 321 and 329, revealed that the cDNA inserts were smaller than would have been expected from the initial estimates of Clissold and Bishop (1981) and represent approximately 1/6 of their complementary mRNAs.

FIGURE 4

Restriction Endonuclease cleavage Maps of the Cloned cDNA Fragments of these Three Plasmids; LVA 301 ($alpha_1$-antitrypsin) and LVA 321 AND LVA 329.

The cDNA fragments are schematically represented by open boxes and the cloning plasmid pPH 207 is represented by a broad line. The positions of restriction enzyme cleavage sites were determined by end labelling the Cla I site of the Cla I/Bam HI fragment of the recombinant plasmids. The labelled fragments were subjected to further partial restriction enzyme digestions. The resultant fragments were analyzed on 7.5% polyacrylamide denaturing gels.


LVA 301
320 bp
470 bp
LVA 321
495 bp
470 bp
LVA 329
470 bp
470 bp
Alu1, BamH1, Cla1, Hae III, Hinf 1, Kpn1, Msp1, Sau3A1, Sst1
100 bp

Synthesis of cDNAs Using Female Liver E.R. poly(A) mRNA

A cDNA library was made based on female BALB/c liver E.R. poly(A) mRNA. It was hoped to make a comprehensive library of cDNA clones corresponding mainly to liver secretory proteins. The library was made using female liver mRNA for two reasons: (i) The level of the abundant mRNA (MUP) in male liver is much lower in female liver and this could improve the likelihood of isolating cDNAs corresponding the the more moderately abundant messages (Hastie et al, 1979; Derman et al, 1981 and Clissold et al, 1981). (ii) There was the possibility of isolating cDNAs for MUP genes with a different pattern of hormonal regulation from those expressed in male liver (Hastie et al, 1979; Unterman et al, 1981; Shaw et al, 1983 and Clissold et al, 1984).

Several of the experimental parameters relating to the synthesis of cDNA were investigated to determine the optimum conditions for the production of long cDNA transcripts from the female mRNA preparations. The variables which had the most profound effect on the size distinction of the cDNA synthesized, were the concentrations of monovalent and divalent cations and the pH (results not shown). The omission of the sucrose gradient size selection method together with the minimal use of SDS during the mRNA preparation procedures resulted in more rapid mRNA isolation with less manipulations, which increased the yield and improved the size distribution of the cDNA synthesis products.

Several reagents have been reported to improve the yield and/or length of cDNA transcripts on addition to the reverse transcription reaction. Accordingly these were included in the reaction mix (Table II), with the following results. The addition of 12% polyethylene glycol 6000 greatly reduced the cDNA synthesis by AMV reverse transcriptase in the system used, although it has been

reported to have a stabilizing effect on many reverse transcriptases (Chan et al, 1980). The ribonuclase inhibitors ribonucleoside-vanadyl complex (Berger and Birkenmeier, 1979) and RNasin (de Martynoff, Pays and Vassart, 1980) were added to the reaction mix and increased neither the length or yield of cDNA transcribed from the poly(A) mRNA. A comparison of the cDNA synthesized at 5 and 60 minutes (Figure 5) demonstrated that most of the cDNA synthesis was completed within five minutes, with a subsequent decrease in small cDNA and concomitant rise in larger cDNA products. The implication is that there was very little ribonuclease contamination in the cDNA synthesis reaction.

A considerable improvement in the size distinction of the cDNA product arose from improvements in the quality of the mRNA preparation. The sucrose gradient step was originally included in the preparation of the mRNA to remove small RNAs, which could act as potent inhibitors of protein synthesis in reticulocyte translation systems (Leroux and London, 1982).

This procedure is redundant in preparations which are not destined to be translated and was therefore omitted. The inclusion of SDS in the solutions used for preparing poly(A) mRNA by the passage of total RNA solutions over an oligo (dT) column arose because of its properties as a surfactant and a potent inhibitor of enzymes. However SDS is only a partial inhibitor of RNase at the concentration at which it is normally used, (Berger and Birkenmeier, 1979) and its detegent properties are not necessary for the elution of poly(A) mRNA from the oligo (dT) column. SDS could therefore be omitted from the elution stages of poly(A) mRNA preparation, thus obviating the multiple ethanol precipitation procedures required to remove it from nucleic acid preparations. The use of the latter two procedures (-SDS) in the preparation of liver E.R. poly(A)

FIGURE 5

Densitometer Scan of sScDNA Products Synthesized after 5 and 60 minutes.

Female liver E.R.poly(A) mRNA was transcribed by AMV reverse transcriptase as described in Materials and Methods except that aliquots were removed at intervals and immediately frozen at -196°C. The samples were electrophoresed on a 5% polyacrylamide denaturing gel, and an autoradiograph of the fixed and dried gel made by an overnight exposure at -70°C. The autoradiograph was subsequently analyzed using a scanning microdensitometer. The size distribution profile of cDNAs after 5 minutes synthesis is indicated by the dotted line, and after 60 minutes synthesis by the solid line. Vertical lines and numbers refer to the position and size (Bp) of end-labelled pBR322 Taq I markers.

O.D.
60 min
5 min
141
313
368
475
616
1000
1307
1444
DISTANCE MIGRATED

mRNA for use in cDNA synthesis, clearly resulted in transcription product of larger and more discrete size classes (Figure 6), which presumably reflected the complete transcription of intact poly(A) mRNA (Figure 7, -SDS trace). The amount of cDNA synthesized using either of the two poly(A) mRNA preparations (+SDS or -SDS) was similar. However, the amount of DNA synthesized by the negative control which contained RNAs $\leq$4S (Figure 6, track 2) was eight fold higher than the corresponding negative control which did not contain RNAs $\leq$4S (not shown). The increase may have been due to small RNA fragments acting as random primers for cDNA synthesis.

Synthesis of the Second cDNA Strand

Second strand synthesis was determined by the incorporation of ^{3}H-dCTP and was routinely found to be equivalent to 90 - 100% of the first strand synthesized. Denaturing gel electrophoresis did not show a doubling of the DNA fragment size distribution relative to the first strand, which would have been expected if the two strands were covalently linked by a loop of DNA. (Figure 8, tracks 7 and 8). (Efstratiadis *et al*, 1976). Native polyacrylamide gel electrophoresis produced similar size distribution profiles for the sScDNA and dScDNA synthesis products (Not shown). These results are consistent with the hypothesis that single stranded breaks were being introduced into the dScDNA. Such DNA nicks could have been introduced into either strand of the DNA or the loop, by any residual endonuclease (nicking) activity of the *E. coli* DNA polymerase I "Klenow Fragment" present during the 15 hour second strand synthesis reaction.

FIGURE 6

The Size Distribution of sScDNA Reverse Transcription Products using Various poly(A) mRNA Preparations.

Reverse transciptase (AMV) was used to synthesize sScDNA using different poly(A) mRNA preparations as templates. The products were electrophoresed on a 5% polacrylamide formamide denaturing gel and the gel autoradiographed as described in the Materials and Methods Track 1 contained sScDNA products synthesized from poly(A) mRNA prepared in the presence of SDS, and from which RNAs greater than 28S and less than 4S had been removed by passage through a 4-20% sucrose gradient. Track 3 contained sScDNA products synthesized using unfractionated poly (A) mRNA from which SDS had been excluded since its elution from an oligo (dT) column. The reaction conditions of Track 2 were the same as those of Track 3, except that the Oligo $(dT)_{8-12}$ primer was omitted. Tracks 4 and 5 contained end labelled pBR322 Taq 1 and Msp I markers respectively. The sizes of the Taq I markers are shown.

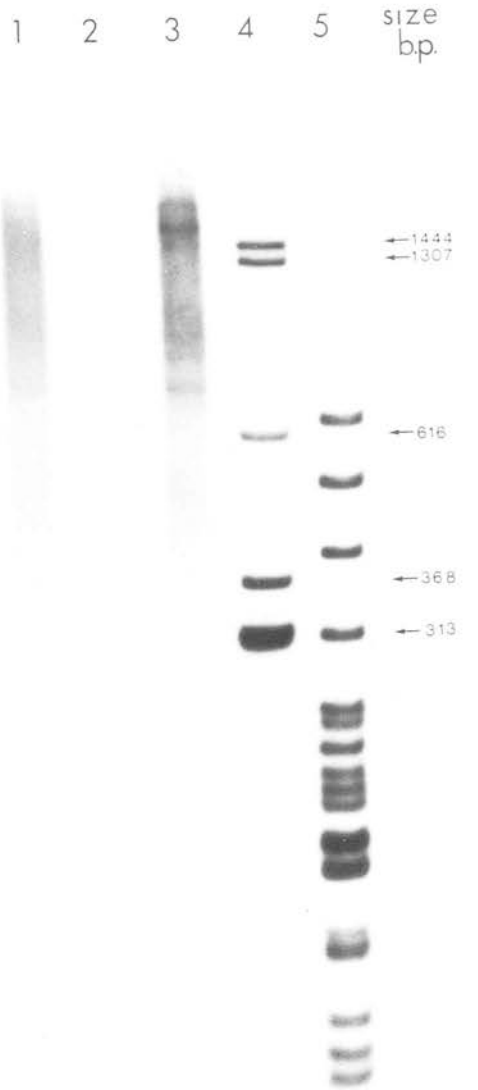

1
2
3
4
5
size
b.p.
1444
1307
616
368
313

FIGURE 7

Densitometer Scans of sScDNA Reverse Transcription Products which Utilized poly(A) mRNA prepared by Alternative Methods.

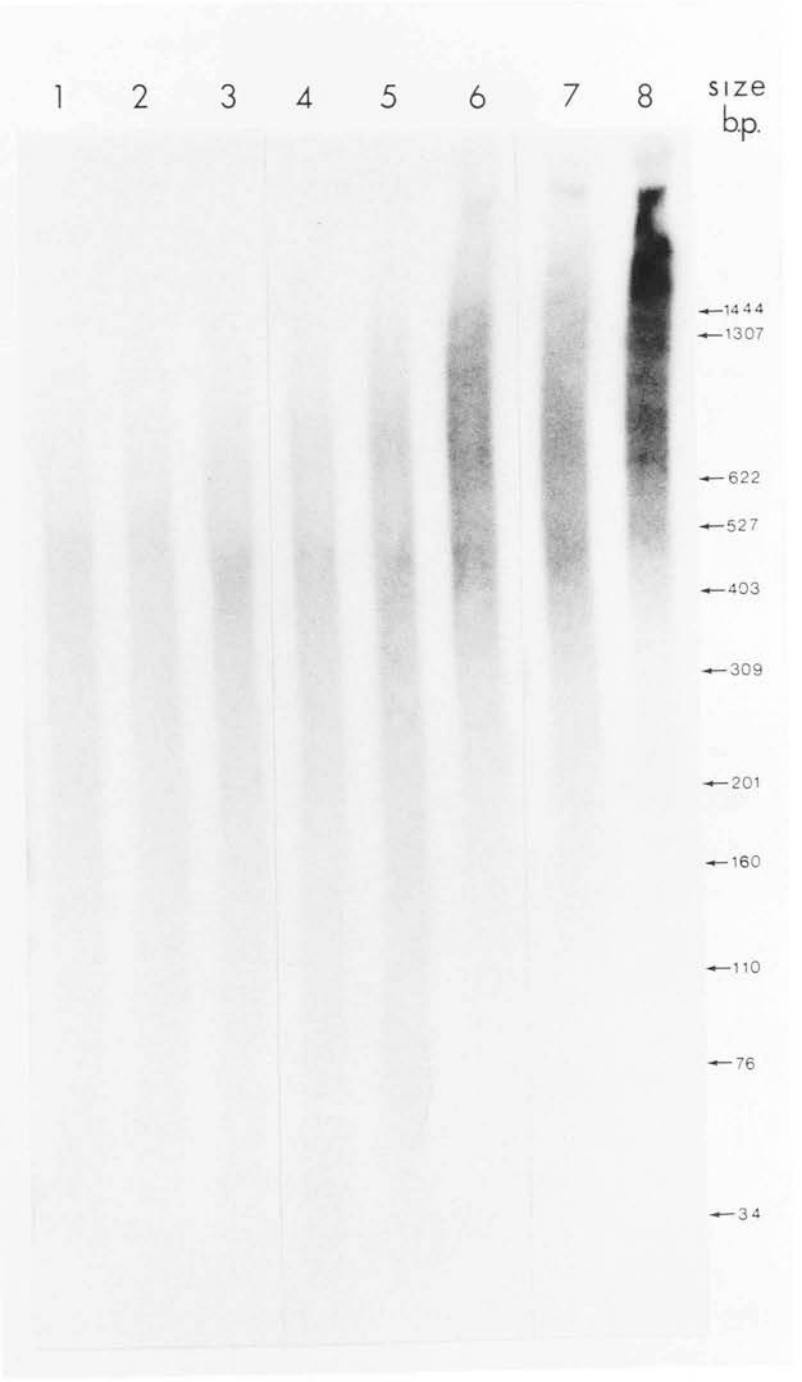
Using AMV reverse transcriptase, sScDNAs were synthesized using different poly(A) mRNApreparations, electrophoresed on a 5% polyacrylamide denaturing gel and an autoradiograph of the fixed and dried gel made by an exposure of 2 days at -70 ^{o}C, as described in Materials and Methods. The autoradiograph was subsequently analyzed using a scanning microdensitometer. The trace indicated by the (+SDS) arrow refers to the sScDNA products synthesized using poly(A) mRNA prepared in the presence of SDS and from which RNAs greater than 28S and smaller than 4S had been removed by passage through a sucrose gradient. The trace indicated by the (-SDS) arrow refers to the sScDNA products synthesized using unfractionated poly(A) mRNA, from which SDS had been excluded since the elution stages of poly(A) mRNA isolation from the oligo (dT) columns. Small vertical arrows and numbers refer to the position and size (Bp) of end labelled pPR322 Taq I markers.

O.D.
- SDS
+ SDS
313
368
475
616
1000
1307
1444
DISTANCE MIGRATED

FIGURE 8

Determination of SI Nuclease Required to Cleave the Hairpin Loops in dScDNA.

The dScDNA was treated with SI nuclease as described by Maniatis, Fritsch and Sambrook, (1982). The digestion products were electrophoresed on a 5% polycrylamide gel in 98% formamide, fixed, dried and autoradiographed. Tracks 1-6 contained 12 ng dScDNA digested with 10, 5, 1, 0.5, 0.1 or 0.0 Vogt units of SI nuclease respectively. Track 7 contained 12 ng of untreated dScDNA and Track 8 contained an equivalent amount of sScDNA. The positions of end labelled pBR322 Taq I and Msp I markers (not shown) are indicated.


1
2
3
4
5
6
7
8
size
b.p.
1444
1307
622
527
403
309
201
160
110
76
34

Nuclease SI was used to remove the hairpin loops which are a feature of dScDNA synthesis by the above method. This was necessary to expose the 3´ hydroxyl groups used in ligation and cloning of cDNA. The recovery of the dScDNA, as estimated by either TCA or ethanol precipitation was between 60 and 70% for the following range of digestion conditions; 3.5 - 12ng dScDNA/1-5 Vogt units nuclease SI per 20µl at 37°C for 20 minutes. The effect of increasing SI Nuclease concentration on the size distribution of the dScDNA is shown in Figure 8 (tracks 1-6). Increasing the SI nuclease level above 1 Vogt unit (Figure 8, tracks 1-3) had a reduced and more linear effect on the cDNA size distribution than levels between 0.1 and 1.0 units (Figure 8, tracks 3-6). The change over point between the two rates of decline was taken as the point at which most of the loops had been cleaved. It was expected that the single-stranded loops would be cleaved in preference to nicks elsewhere in the cDNA. This appears to be what had taken place when denaturing and native polyacrylamide gels of SI nuclease treated dScDNA samples were compared (Figure 8, track 3 and Figure 9). The densitometer scans of 5% TA polyacrylamide gels (Figure 9) enabled the size distinction of SI nuclease treated dScDNA to be determined. Portions of the gel were removed and the cDNA extracted for cloning. The modal size of the cDNA selected for cloning, as estimated by gel electrophoresis after extraction was 1,090 Bp.

A summary of the efficiencies of the procedures required for the synthesis of dScDNA are presented in Table IV. Typical yields of dScDNA suitable for cloning were 1.1 to 3.6% of the first strand synthesis, which is equivalent to ~105ng of dScDNA from 12.5µg of poly(A) mRNA template.

FIGURE 9

Densitometer Scan of dScDNA Products after Nuclease SI Treatment.

Double stranded cDNA was synthesized and treated with SI nuclease as described in Materials and Methods. The dScDNA was then electrophoresed on a 5% native polyacrylamide gel. After making a direct autoradiograph of the gel at 4°C overnight, the part of the gel containing DNA fragments between 900 Bp and 300 Bp was removed and the DNA electroluted. A sample of the extracted DNA was electrophoresed on a second 5% native polyacrylamide gel. The upper trace indicates the size distinction of the dScDNA after SI nuclease treatment, the lower scan, and hatched area, refers to the size distribution of the electroluted fragments. The position and size of end labelled pBR322 Taq I fragments are indicated by the vertical arrows and numbers, (Bp) respectively.

DISTANCE MIGRATED
313
368
616
1000
1307
1444
OD

TABLE IV

THE EFFICIENCY OF EACH PROCEDURE AND SUBSEQUENT RECOVERY OF DNA DURING THE BULK SYNTHESIS OF dS cDNA

PROCEDURE/RECOVERY	EFFICIENCY OF STEPS DURING TWO PREPARATIONS
FIRST STRAND SYNTHESIS sS cDNA	12.5 - 17.0%
SECOND SYNTHESIS dS cDNA	95 - 100%
S1 NUCLEASE DIGESTION OF dS cDNA	60 - 70%
SIZE SELECTION OF dS cDNA (5% T.A. NATIVE POLYACRYLAMIDE GEL)	6.5 - 13.0%
FLUSHING ENDS OF dS cDNA FOR LIGATION	75 - 80%

TABLE V

TRANSFORMATION OF *E. coli* K12 STRAINS BY PLASMIDS USING THE METHODS OF MANDEL AND HIGA (1970), MODIFIED OR HANAHAN (1983)

TRANSFORMATION PROTOCOL	STRAIN OF *E. coli* K12	TRANSFORMING DNA	NUMBER OF TRANSFORMANTS (STANDARD ERROR)
MANDEL AND HIGA (1970), MODIFIED	HB101	C.C.C. $_{P}$AT153	1.66 ($\pm$ 0.28)$^{\boxplus}$ x 10^6/µg DNA
HANAHAN (1983)	HB101	C.C.C. $_{P}$AT153	2.26 ($\pm$ 0.66)$^{\boxplus}$ x 10^7/µg DNA
	RRI	C.C.C. $_{P}$AT153	9.5 ($\pm$ 2.5)$^{\boxplus}$ x 10^7/µg DNA
	JM83	RE-LIGATED SmaI DIGESTED $_{P}$UC8*	BLUE 3.34 ($\pm$ 0.87)$^{\boxplus}$ x 10^7/µg DNA & WHITE 1.34 ($\pm$ 0.31)$^{\boxplus}$ x 10^6/µg DNA

* RE-LIGATED IN THE PRESENCE OF AN EQUAL AMOUNT OF dScDNA, MODAL SIZE 1Kb

$^{\boxplus}$ STANDARD ERROR OF THE MEAN

Cloning dScDNA

Transformation procedures were compared in relation to efficiency of transformation and the ease of analysis of recombinant plasmids. The results of these investigations are summarized in Table V. Although the E. coli. k12 strain RRI gave the highest transformation efficiencies by the Hanahan, (1983) protocol, the pUC 8/JM 83 system offered better combination of efficiency and analysis. An advantage of the pUC 8/JM 83 combination was the identification system for transformed colonies harbouring recombinant plasmids (white from blue) on the initial plating. Another was the presence of polylinker restriction sites adjacent to the cloned cDNA insertion site, which facilitated subsequent DNA manipulations (Vieira and Messing, 1982).

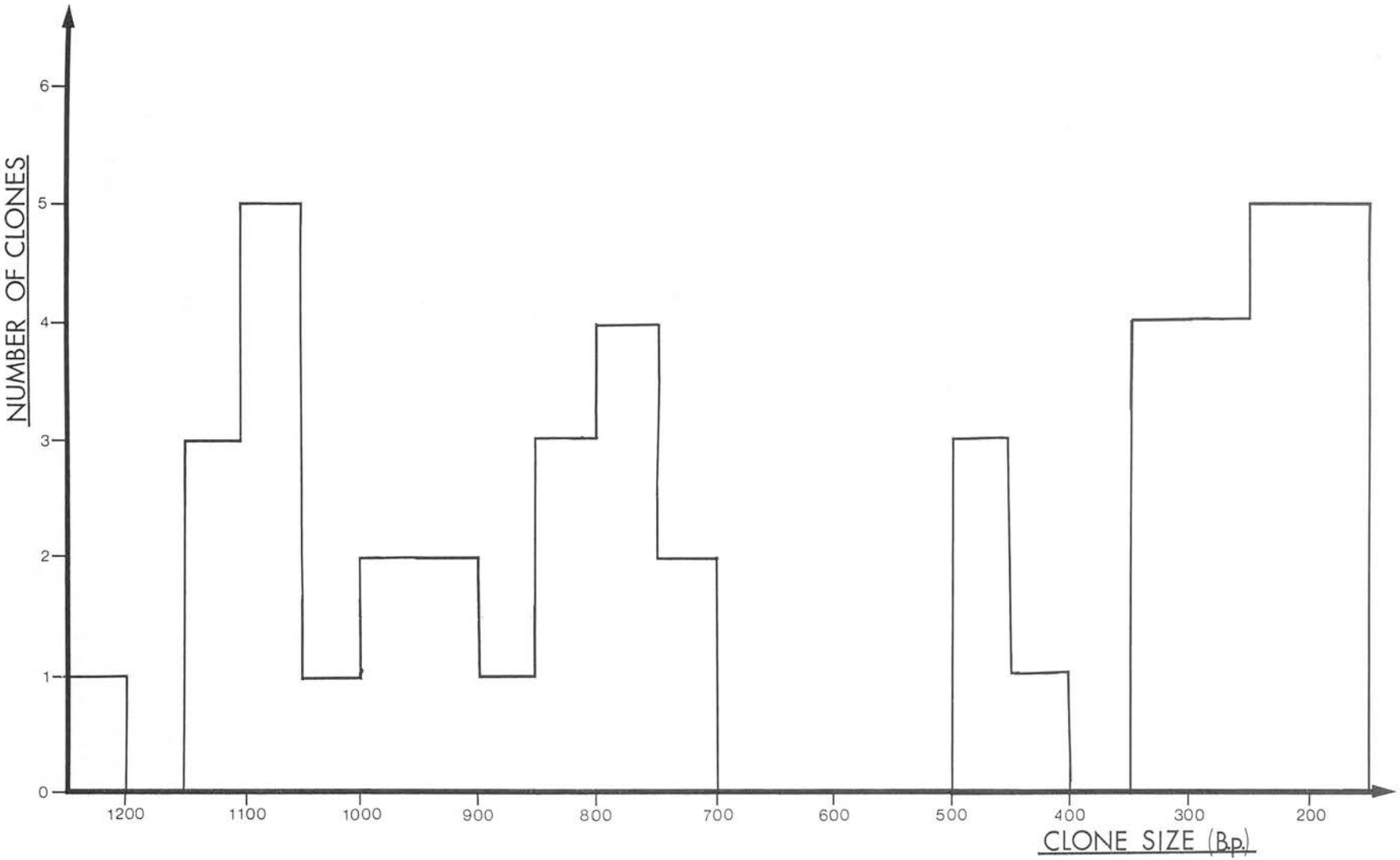
A cDNA library of 2000 recombinant plasmids was screened by the hybridization of library filter replicas, (Gergen, Stern and Wensink, 1979) to nick translated plasmid fragments of BS 6-5 (MUP), LVA 301 ($alpha_1$-antitrypsin) and LVA 321 ("Transferrin"), from which the DNA that would have cross-hybridized to pUC 8 had been removed. Fifteen MUP, fourteen $alpha_1$-antitrypsin and 59 "Transferrin" hybridization positive clones were found (0.75, 0.7 and 2.9%). Most colony hybridizations fell into two distinct classes of signal strengths, designated strong or weak. Mini-plasmid preparations were made from the positive colonies and the cDNAs removed from the vector (pUC8) by cleavage with restriction enzymes. Some clone preparations resisted enzymatic digestion and were not analysed further.

The cloned cDNA size distribution is presented as a histogram in Figure 10. There are two main size classes, those greater than 700 Bp ($x = 943 (\pm 32$, standard error of the mean)) and those below 500 Bp ($x = 276 (\pm 23$, standard error of the mean)).

FIGURE 10

Histogram to Show the Cloned cDNA Size Distribution in the pUC 8 Library.

Filter replicas of 2,000 cDNA clones were probed with nick-translated fragments of the recombinant plasmids LVA 301, LVA 325 and LVA 321; which hybridize to alpha$_1$-antitrypsin, MUP (Clissold and Bishop, 1981) and transferrin derived sequences. The hybridization stringency wash was 2 x SET at 68^{o}C. Mini plasmid preparations were made of all clones that demonstrated any hybridization to the probes. The clone inserts were removed by enzymic cleavage from the plasmids and sized by electrophoresis on 1.5% agarose gels. Full practical details are described in Materials and Methods.


NUMBER OF CLONES
0
1
2
3
4
5
6
1200
1100
1000
900
800
700
600
500
400
300
200
CLONE SIZE (B.p.)

The small cDNAs should have been excluded from cloning by the size selection process (Figure 9). However, the non denaturing conditions required for this process may have enabled some of the small cDNAs to become retained by, or associated with the larger DNAs, although shortening of larger cDNAs during transformation or *in vivo* cannot be excluded as causes of short cDNA inserts. The size distribution of the cDNAs from the colonies which generated strong hybridization signals is shown in Figure 11.

By comparing Figure 10 and Figure 11 it can be seen that most of the smaller cDNAs were from clones showing weak hybridization to the probes. These were not analysed further. A few of the larger cDNAs hybridized weakly to the "Transferrin" probe. Due to the possibility that these may have corresponded to a domain of the protein different from the probe domain, they were also included in the more detailed study. Transferrin is known to contain two domains that exhibit considerable homologies, and which could have arisen by an ancestral gene duplication (MacGillivray *et al*, 1982).

The limited size of the LVA 321 probe, relative to the length of the mRNA, means that it can contain only one domain or the non-homologous parts of two. Consequently, some of the cloned cDNAs might be expected to give a weak hybridization signal.

Analysis of the Cloned cDNAs

The position of restriction enzyme sites were determined in cDNA clones over 700 Bp in length that hybridized to the restriction fragment probes BS 6-5

FIGURE 11

Histogram to Show the Size Distribution of cDNA lanes that Hybridized Strongly with LVA 325, and LVA 321 Probes.

Filter replicas of 2,000 cDNA clones were probed for sequences which hybridized strongly to nick translated fragments of plasmids LVA 301, LVA 325 or LVA 321, which contain alpha$_1$-antitrypsin, MUP (Clissold and Bishop, 1981) and transferrin partial cDNAs respectively. The hybridization stringency wash was 2 x SET at 68^{o}C. The clone inserts were removed by restriction enzymes from the pUC8 vector and sized on 1.5% agarose gels. Full practical details are described in Materials and Methods.

NUMBER OF CLONES
0
1
2
3
4
5
CLONE SIZE (B.P.)
200
300
400
500
600
700
800
900
1000
1100
1200

Hind III/Bam HI and LVA 321 Cla I/Cla I ("Transferrin"). To confirm the identity of these clones, the cDNAs were isolated from the plasmids, transferred to nitrocellulose and hybridized with plasmid BS 6-5 or LVA 321. Confirmed MUP clones were designated by the prefix pUC MUP, strong "Transferrin" clones by pUC TRF and clones still demonstrating weak hybridization with LVA 321 derived probes were given the prefix pUC T (Figure 12).

Analysis of the Cloned Female MUP cDNAs

pUC MUP clones with inserts greater than 700 Bp were mapped with restriction enzymes and the orientations of their cDNAs in the pUC 8 vectors determined by restriction site position and Southern Blot hybridization (Clark et al, 1984 and Kuhn et al, 1984), (Figure 13). Two of the cloned cDNAs, pUC MUP 6 and 15, were estimated to be ~950 Bp long, close to the length of the long MUP mRNA transcripts. Similarly the stretch of insert past the Pvu II site in all 6 clones was sufficient to contain the complete 3′ end of long messages, providing only short poly(A) tails (<20Bp) were present in some of them (Clark et al, 1984a). Clones pUC MUP 6, 9 and 15 extend approximately 50 Bp further 3′ than the other clones.

Four clones contain an Eco RI site 200 Bp 5′ to the common Pvu II site. Only one clone (pUC MUP 8) extends beyond this position but does not contain the site. Most cloned MUP sequences contain the site, but two are already known which do not, namely the BALB/c genomic clone BL 1 and the C57 BL/6 cDNA, p499. (Bishop et al, 1982 and Kuhn et al, 1984).

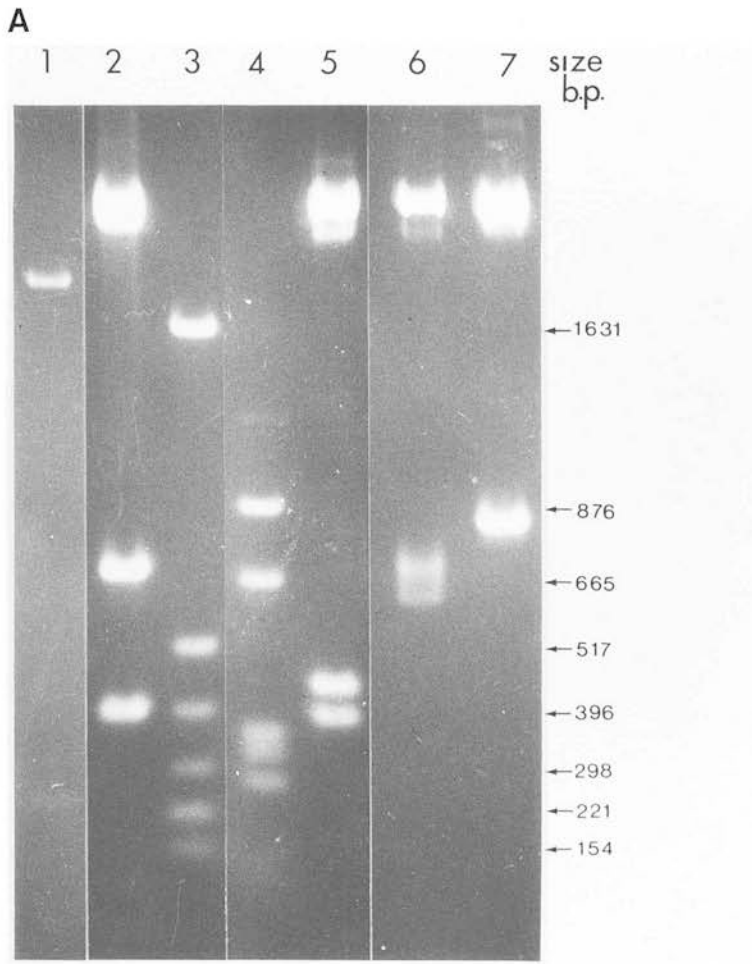
FIGURE 12

Isolation and Size Determination of Several cDNAs(A)s and their Hybridization after Southern Transfer to LVA 321(B).

Recombinant plasmids were digested with restriction enzymes, electrophoresed on a 1.5% agarose gel, stained with Et Br, photographed (A) and transferred to nitrocellulose filters. The filters were annealed with nick-translated LVA 321 DNA, final wash 1 x SET, 68°C and exposed (B).

The Track Contents Were:			
Track	DNA	Restriction Enzyme Digestion	Exposure [(B) only]
1	LVA 321 (Control)	Cla I/Kpn I	1x
2	pUC TRF 14	Eco RI/Bam HI	1x
3	pAT153 (Marker)	Hinf I	1x
4	pAT153 (Marker)	Sau 3AI	1x
5	pUC TRF 21	Eco RI/Bam HI	1x
6	pUC T39	Eco RI/Pst I	2x
7	pUC T40	Eco RI/Pst I	2x

The size of the plasmid bands are shown adjacent to (A).


A
1 2 3 4 5 6 7
size b.p.
1631
876
665
517
396
298
221
154

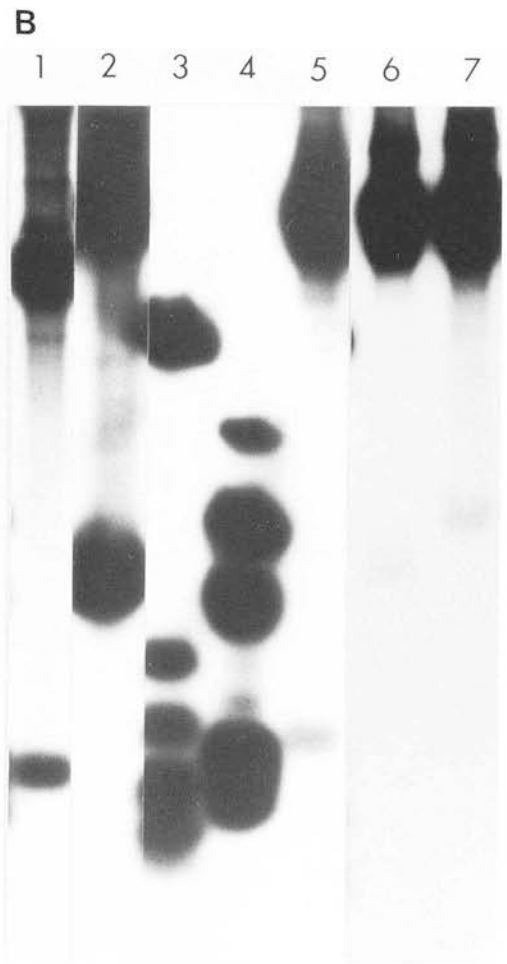
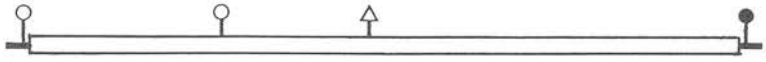

B
1 2 3 4 5 6 7

FIGURE 13

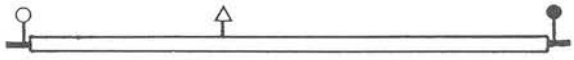
Restriction Endonuclease Cleavage Maps of the MUP cDNA clones.

Clones from the BALB/c female liver E.R. library which hybridized to the partial MUP genomic clone BS 6-5 were isolated as mini DNA preparations, and digested with several pairwise combinations of restriction enzymes. The resultant digest products were electrophoresed on 1.5% agarose gels and stained with Et Br. The cDNAs of clones with inserts greater than 700 Bp are schematically represented by open boxes and the flanking plasmid pUC8 by a broad line. Vertical lines with symbols refer to restriction enzyme sites; Eco RI (♀), Pvu II (♦) and Bam HI (♥).

pUC MUP 6



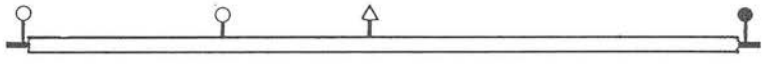
pUC MUP 8



pUC MUP 9

pUC MUP 11

pUC MUP 15



pUC MUP 16

The MUP genes of the BALB/c mouse have been analysed by cross-hybridization. Most of the genomic clones fall into either of two groups (Group 1 or Group 2) while a few clones are equally removed from either group using this criterion. (Bishop et al, 1982). The female MUP cDNAs >700 Bp were analysed for Group 1 and Group 2 cross-hybridization in precisely the same manner. (Table 6). Clones pUC MUP 8, 9, 11 and 16 gave the hybridization profile expected of Group I cDNA clones, whereas pUC MUP 6 and 15 gave weak signals with both the Group 1 or the Group 2 probe and could on this basis correspond to one or more of the estimated 7 MUP genes that do not lie within the Group 1/Group 2 classifications (Bishop et al, 1982). None of the cDNAs were of Group 2 type. These results are in good agreement with Clissold and Bishop, (1981) and Bishop et al, (1982) who cloned 6 BALB/c MUP cDNA fragments of the Group 1 type and a recent paper by Kuhn et al, (1984) who isolated 3 MUP cDNA clones from either C57 BL/6 or BALB/c, two of which were Group 1 type and one, pl99 was like pUC MUP 6/15 (see below).

Analysis of the cDNA Clones that Hybridized to LVA 321 Probes

pUC TRF clones with inserts greater than 700 Bp long were also mapped with restriction enzymes. The orientation of the inserts, relative to LVA 321, was determined by restriction site position and Southern Blot hybridization (Figure 14). The cDNAs of the pUC TRF clones overlapped extensively six of the seven cDNAs were ~1,100 Bp long and in total comprised ~1,200 Bp, or approximately half of the Transferrin mRNA sequence (Figure 3 and Jeltsch and Chambon, 1982). The cloned "Transferrin" cDNAS exhibit two unusual phenomena , 6 of the 7 cDNAs share the same orientation in pUC 8 and these 6 cDNAs correspond to

TABLE VI

Hybridization of MUP cDNA Inserts to Group 1 and Group 2 MUP Gene Probes.

The insert fragment of each pUC MUP clone, greater than 700 Bp long, was separated from the cloning vector pUC8 by digestion with the restriction enzymes Eco RI and Bam HI, followed by electrophoresis on two 1.5% agarose gels. The DNA was transferred to nitrocellulose membranes and hybridized with either nick translated pBS 6-5-5 (Group 1) or pBS 2-2-2 (Group 2) probes (Bishop et al, 1982). The final was 0.2 x SET at 68^{o}C. After an exposure of 7 hours a qualitative comparison of the autoradiographs was made. The notation used was: very strong (V.S.), strong (S), moderate (M), weak (W) and not visible (N/D).

TABLE VI

CLONE INSERT FROM	RELATIVE HYBRIDIZATION OF PROBE	
	GROUP 1 ($_{P}$BS655)	GROUP 2 ($_{P}$BS 2-2-2)
$_{P}$UC MUP 6	W	W
$_{P}$UC MUP 8	S	N/D
$_{P}$UC MUP 9	S	N/D
$_{P}$UC MUP 11	S	N/D
$_{P}$UC MUP 15	W	W
$_{P}$UC MUP 16	S	N/D
$_{P}$BS 6-5-5	VS	S
$_{P}$BS 2-2-2	S	VS

the same 1.1 ±0.1 Kb of the mRNA. The former phenomenon also occurs in the pUC MUP clones. The only clone deviating from the above is pUC TRF21, which lacks a 300 Bp region of the sequence shared by the other cDNAs and has been cloned in the opposite orientation.

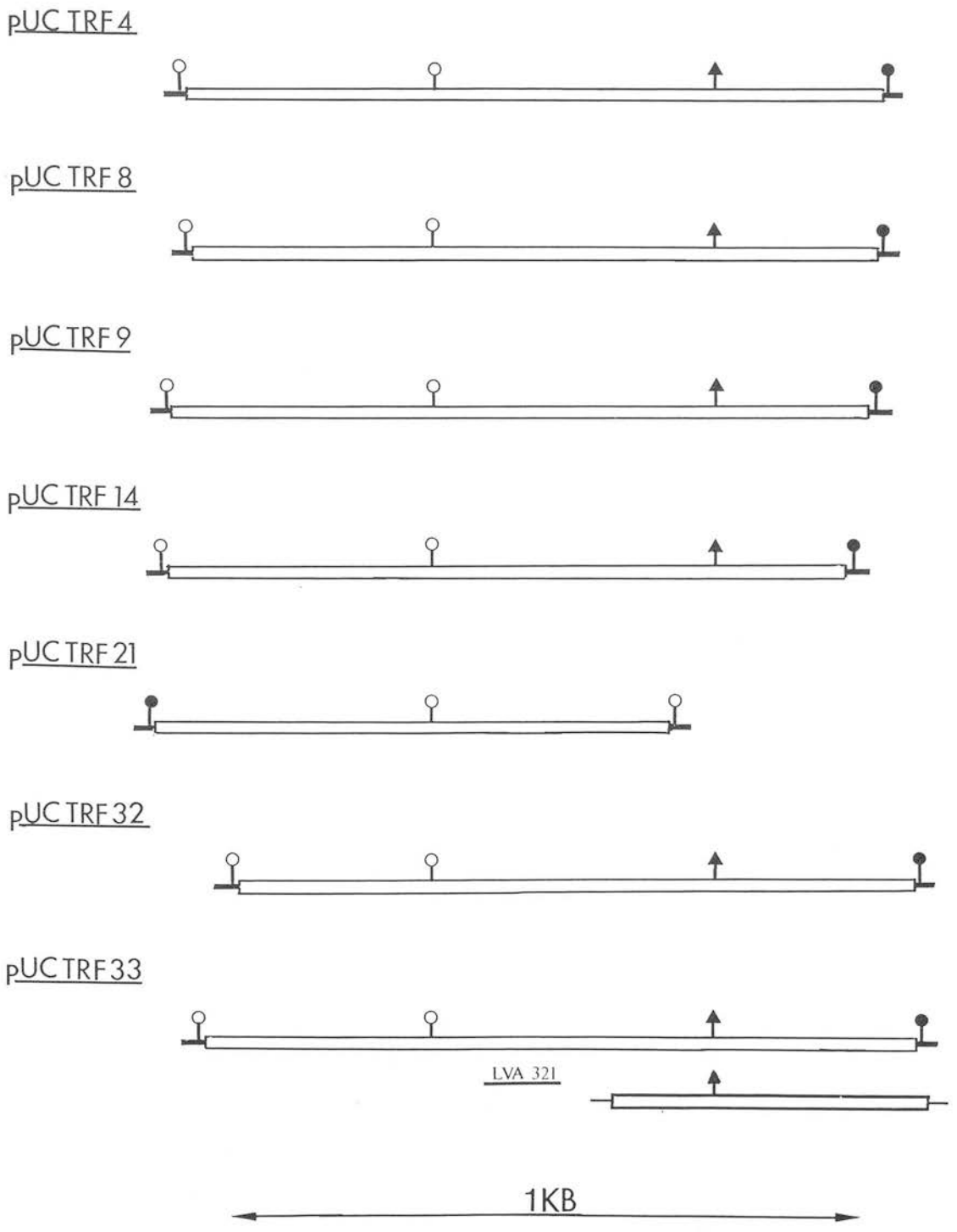
A possible explanation of why the sequences have been inserted in the same orientation, is that they possess a specific structure or sequence at one end that would kinetically or sterically, promote or inhibit the plasmid/cDNA ligation process. The structure or sequence could then interact with other sequences adjacent to the Sma I insertion site in pUC 8. Such a structure or sequence would have to be present in 12 of the 13 pUC MUP and TRF clones. The most likely candidate sequence/ structure to be present at the end of cDNAs synthesized by oligo (dT) priming able to cause this effect, is a poly(A) tail. It is hoped that the sequencing of several of these clones will provide an insight into this phenomenon.

There are several possible explanations as to why the pUC TRF cDNAs all terminate within ~100 Bp of each other. One possibility was the presence of a barrier to reverse transcription adjacent to the 5′ end of the cloned cDNAs. Monomethylation of guanine to N^2-methylguanine (m^2G) has been shown to cause attenuated transcription intermediates of RNA by reverse transcriptase (Youvan and Hearst, 1979) and m^6_2 adenine in RNA is an efficient terminator of reverse transcription (Hagenbuchle et al, 1978). Alternatively there may be a secondary structure/sequence that either promotes second strand synthesis, or is sensitive to digestion by SI nuclease in the region of the cDNA 5′ to that eventually cloned. Speculation as to which, if any, of the above scenarios was the cause of the homologous cDNAs would probably require the cloning and sequencing of a more complete mouse transferrin cDNA.

FIGURE 14

Restriction Endonuclease Cleavage Maps of the Putative Transferrin cDNA Clones.

Clones from the BALB/c female liver E.R. library which hybridized strongly to the cDNA clone LVA 321 (Clissold and Bishop, 1981) were isolated as mini DNA preparations, digested with several restriction enzymes, electrophoresed on 1.5% agarose gels, transferred to nitrocellulose membrane and hybridized to whole nick-translated plasmid LVA 321. The final wash was 1 x SET at 68°C. The cloned cDNAs greater than 700 Bp in length that hybridized strongly to LVA 321 or LVA 321 fragments are represented by open boxes and the flanking plasmid pUC 8 by a broad line. (LVA 321 cDNA open box, pPH207 flanking regions narrow line). Vertical lines and symbols refer to restriction enzyme site, Eco RI (⚲), Kpn 1 (⍋) and Bam HI (📍).


pUC TRF 4
pUC TRF 8
pUC TRF 9
pUC TRF 14
pUC TRF 21
pUC TRF 32
pUC TRF 33
LVA 321
1KB

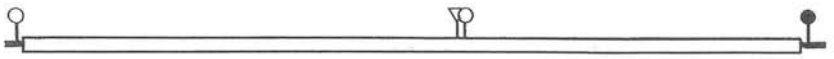
The pUC T clones with inserts greater than 700 Bp were mapped with restriction enzymes and hybridized to the LVA 321 probe. Clones which continued to exhibit weak hybridizations to the probe e.g. (Figure 12) are shown in Figure 15. No correlation between the restrictions site positions and the regions of cDNA to which the LVA 321 probe hybridized was apparent (Figure 15; pUC T37 Eco RI/Pst I; pUC T39, pUC T42, pUC T57 Pst I/Bam HI and pUC T40, pUC T52 Eco RI/Bam HI). Furthermore the clone inserts, on the basis of their restriction site and hybridization positions, appear to have originated from different sequences. A relatively small area of homology (~30-80Bp), between the "Transferrin" cDNA of pUC TRF21 and the probe LVA 321 (Figure 14), was sufficient to generate a hybridization signal several fold in excess of that shown by the strongest pUC T clone (Figure 12 tracks 5 and 7). Thus the possibility that the LVA 321/pUC T clone hybridizations merely represent fortuitous weak homologies between the sequences cannot be excluded.

FIGURE 15

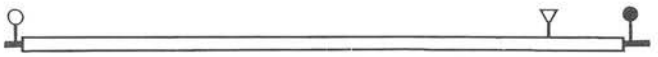
Restriction Endonuclease Cleavage Maps of cDNA Clones which Demonstrateed Weak Anomalous Hybridization to the Putative Transferrin cDNA LVA 321.

Clones from the BALB/c female liver E.R. library which hybridized weakly to the cDNA LVA 321 (Clissold and Bishop, 1981), were isolated as mini DNA preparations, digested with a variety of restriction enzymes, electrophoresed on 1.5% agarose gels, transferred to nitrocellulose membrane and hybridized to whole nick-translated plasmid LVA 321. The final wash was 1 x SET at 68°C. The cloned cDNAs greater than 700 Bp which yielded weak or very weak hybridizations to LVA 321 are represented by open boxes and the flanking plasmid PUC 8 by a broad line. Vertical lines and symbols represent restriction enzyme sites; Eco RI (⚲), Kpn I (♦), Pst I (▽) and Bam HI (♦).

pUC T37



pUC T39



pUC T40

pUC T42

pUC T52

pUC T57

The Sequencing, Analysis and Comparison of Female MUP cDNA Clones.

MUP cDNA sequencing

Four of the six pUC MUP cDNAs were excised from the plasmid pUC8 by partial Eco RI and Bam Hl restriction enzyme digestion, and ligated into the sequencing vector Ml3mp9, as described in the Methods section. The female liver E.R. cDNA sequences removed fom the pUC MUP clones will hereafter be referred to as MUPs, followed by the number of the clone in which it was isolated: e.g. the cDNA insert from pUC MUP 15 will be called MUP 15.

The four MUP cDNAs were chosen for sequencing for the reasons discussed earlier: MUPs 6 and 15 because they did not fall within the Group 1/Group 2 MUP gene classifications on hybridization criteria; MUP 8 because it contained an Eco RI restriction enzyme site polymorphism and MUP 11 because it contained the most complete group I type female MUPcDNA. (Figure 13 and Table VI).

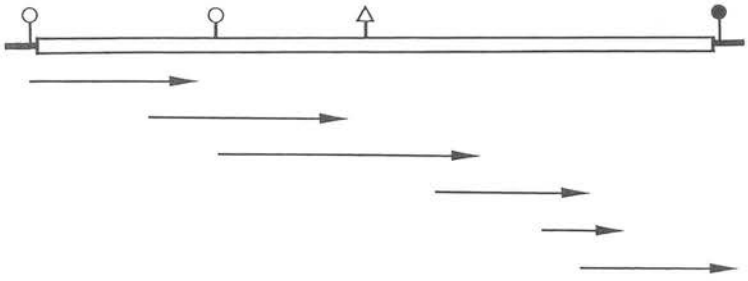
Overlapping sub-clones representing the mRNA-like strand of MUPs 6, 8, 11 and 15 for sequence determination were prepared (Figure 16 and 17). The derivation of the subclones relied on the random digestion of each DNA insert by an endonuclease, followed by circularisation of the r.f., (Hong, 1982). Approximately 350 Bp of the sequence 3´ to each cleavage site may be determined by the procedures described in the Methods section (Figure 18).

FIGURE 16

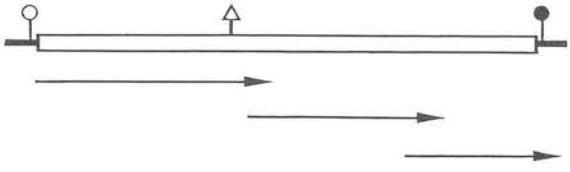
Sequence-Analysis Strategy for Cloned Female Major Urinary Protein cDNAs.

The cDNA sequences are schematically represented by open boxes. The M13mp9 flanking sequence is represented by a broad line. Horizontal arrows indicate the extent and polarity of sequence data obtained from constructs generated by the method of Hong (1982), or from recloned restriction fragments (arrows indicated by asterisks). Vertical lines and symbols refer to the location of restriction enzyme sites; Eco RI ( ), Pvu II( ) and Bam HI ( ).

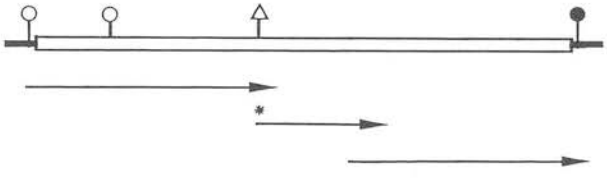
M13 MUP 6



M13 MUP 8



M13 MUP 11



M13 MUP 15

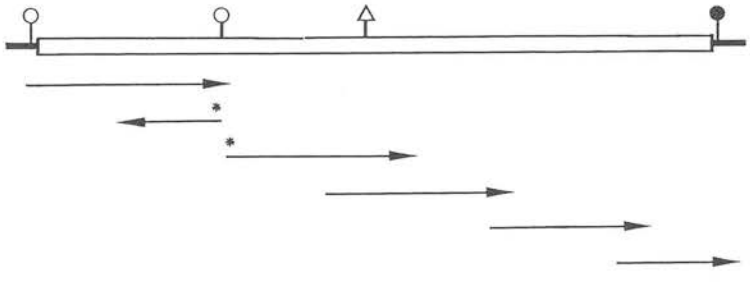


FIGURE 17

Example of a Typical 2 hour Sequencing Gel.

The autoradiograph is of five recombinant M13 bacteriophage sequences determined by the chain terminator sequencing method and after electrophoresis through a pH8.8 polyacrylamide sequencing gel, in which 2.5 mm wide slot formers were employed, as described in Materials and Methods.

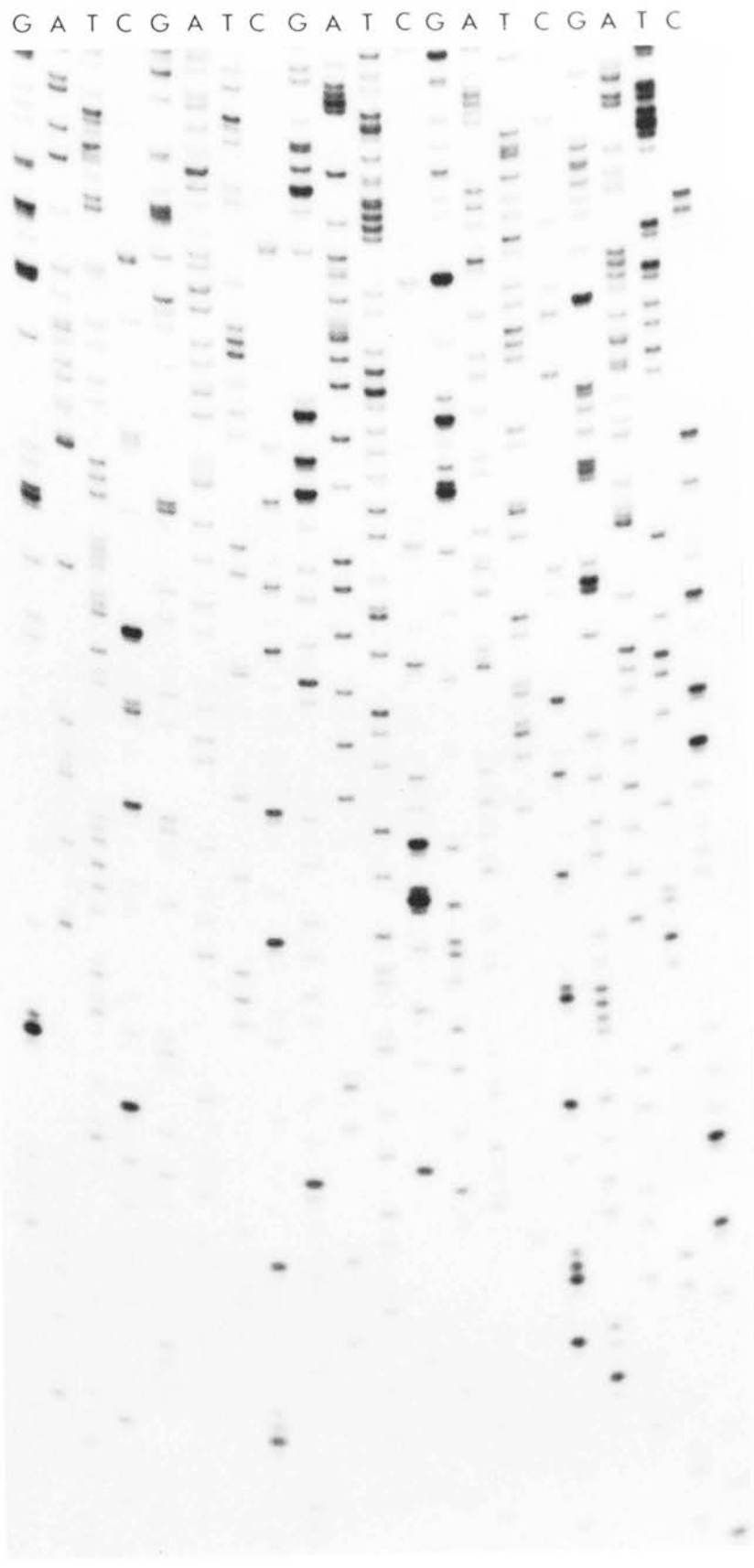
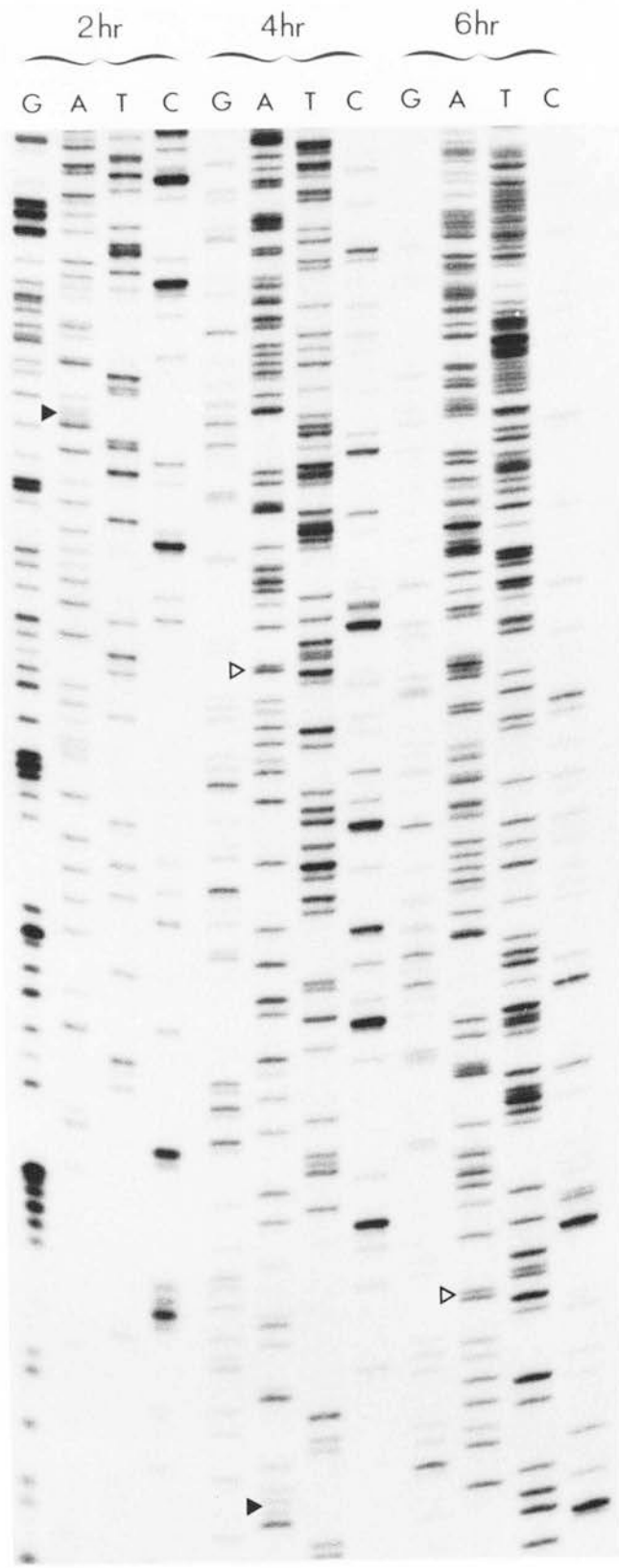

G A T C G A T C G A T C G A T C G A T C

FIGURE 18

An Example of the Gel System Used to Determine the Sequence for an Average of 350 Nucleotides from the Priming Site of One Set of Chain Terminator Sequencing Reactions.

The chain terminator sequencing reactions were arrested by the addition of EDTA. Aliquots were removed from the reaction mixes, denatured and loaded into well flushed out 5 mm wide slots at 2 hourly intervals as described in Materials and Methods. Open and closed triangles indicate the position of identical sequences after the various periods of electrophoresis. The run time is shown above each set of tracks.


2hr
4hr
6hr
G A T C G A T C G A T C

some necessary segments of the MUP 11 and 15 sequences escaped subcloning by random digestion. These segments were sequenced by cloning specific restriction fragments within the sequencing vectors M13mp8 or M13mp9 (Figure 16).

The sequences of MUPs 6, 8, 11 and 15 were determined. Each contained a single long reading frame. The sequences of MUP 6 and MUP 15 were identical. MUP 8 is 682 Bp long (Figure 19). MUP 11 is 726 Bp long (Figure 20) and MUP 15 (and MUP 6) are 883 Bp long (Figure 21).

Nucleotide and Protein Sequence Comparisons of the Female MUPs

The DNA and deduced amino acid sequences of the 3 different MUP clones MUP 8, 11 and 15, were aligned for maximum homology (Figures 22 and 23). All 3 MUP sequences show extensive homology. MUPs 8 and 11 are greater than 99% homologous, with two nucleotide changes in the 3′ untranslated region and two replacement site differences; one of which is a T to A replacement responsible for the Eco RI restriction site polmorphism of MUP 8.

In contrast to the strong homology between MUP 8 and 11, MUP 15 shows considerably more divergence in the nucleotide and deduced amino acid sequences. The divergence is particularly pronounced in the first half of the mature protein, which includes the creation of a potential N-glycosylation site (Figure 23) and towards the end of the 3′ untranslated region (Figure 22). MUP 15 also has a 31 Bp insertion relative to MUPs 8 and 11 near the start of the 3′ untranslated region at the end of exon 6 (Clark et al, 1984a), but is identical over the regions surrounding the translation termination codon TGA and the polyadenylation sequence AATAAA which is followed 16 Bp later by a poly(A) tail in all three sequences.

FIGURE 19

Sequence of the Mup cDNA Cloned in pUC MUP 8.

DNA sequence analysis was performed by the dideoxy chain termination method. Nucleotides are numbered at the left of each line. This indicates the size and position of the cDNA clone relative to the coding sequence of a Group 1 MUP gene (Clark et al, 1984a), where one refers to the A of the translation initiation codon ATG. The predicted protein sequence is shown below the DNA. Numbers below the amino acids refer to their distance from the start of the mature protein (Clark et al, 1984a). The predicted sequence stops at the translation stop codon TGA. The 3′ polyadenylation signal AATAAA (Proudfoot and Brownlee, 1976) is underlined. The extent of the poly(A) tail is indicated at the end of the sequence.

```
132 TGACAAAAGAGAAAAGATAGAAGATAATGGCAACTTTAGACTTTTTCTGGAGCAAATCCA
     AspLysArgGluLysIleGluAspAsnGlyAsnPheArgLeuPheLeuGluGlnIleHi
                30                                  40
192 TGTCTTGGAGAAATCCTTAGTTCTTAAATTCCATACTGTAAGAGATGAAGAGTGCTCCGA
    sValLeuGluLysSerLeuValLeuLysPheHisThrValArgAspGluGluCysSerGl
                50                                  60
252 ATTATCTATGGTTGCTGACAAAACAGAAAAGGCTGGTGAATATTCTGTGACGTATGATGG
    uLeuSerMetValAlaAspLysThrGluLysAlaGlyGluTyrSerValThrTyrAspGl
                70                                  80
312 ATTCAATACATTTACTATACCTAAGACAGACTATGATAACTTTCTTATGGCTCATCTCAT
    yPheAsnThrPheThrIleProLysThrAspTyrAspAsnPheLeuMetAlaHisLeuIl
                90                                 100
372 TAACGAAAAGGATGGGGAAACCTTCCAGCTGATGGGGCTCTATGGCCGAGAACCAGATTT
    eAsnGluLysAspGlyGluThrPheGlnLeuMetGlyLeuTyrGlyArgGluProAspLe
               110                                 120
432 GAGTTCAGACATCAAGGAAAGGTTTGCAAAACTATGTGAGGAGCATGGAATCCTTAGAGA
    uSerSerAspIleLysGluArgPheAlaLysLeuCysGluGluHisGlyIleLeuArgGl
               130                                 140
492 AAATATCATTGACCTATCCAATGCCAATCGCTGCCTCCAGGCCCGAGAATGAAGAATGGC
    uAsnIleIleAspLeuSerAsnAlaAsnArgCysLeuGlnAlaArgGluEnd
               150                                 160
552 CTGAGCCTCCAGTGTTGAGTGGAGACTTCTCACCAGGACTCCACCATCATCCCTTCCTAT

612 CCATACAGCATCCCCAGTATAAATTCTGTGATCTGCATTCCATCCTGTCTCACTGAGAAG

632 TCCAATTCCAGTCTATCCACATGTTACCTAGGATACCTCATCAAGAATCAAAGACTTCTT

732 TAAATTTCTCTTTGATATACCCATGACAATTTTTCATGAATTTCTTCCTCTTCCTGTTCA

792 ATAAATGATTACCCTTGCACTT Poly(A)25
```

FIGURE 20

Sequence of the MUP cDNA cloned in pUC MUP 11.

DNA sequence analysis was performed by the dideoxy chain termination method. The conventions used are given in the legend to Figure 19.

```
                .         .         .         .         .         .
 88 AGAGAAAAGATTAATGGGGAATGGCATACTATTATCCTGGCCTCTGACAAAAGAGAAAAG
    ArgGluLysIleAsnGlyGluTrpHisThrIleIleLeuAlaSerAspLysArgGluLys
                              20                                30

                .         .         .         .         .         .
148 ATAGAAGATAATGGCAACTTTAGACTTTTTCTGGAGCAAATCCATGTCTTGGAGAATTCC
    IleGluAspAsnGlyAsnPheArgLeuPheLeuGluGlnIleHisValLeuGluAsnSer
                              40                                50

                .         .         .         .         .         .
208 TTAGTTCTTAAATTCCATACTGTAAGAGATGAAGAGTGCTCCGAATTATCTATGGTTGCT
    LeuValLeuLysPheHisThrValArgAspGluGluCysSerGluLeuSerMetValAla
                              60                                70

                .         .         .         .         .         .
268 GACAAAACAGAAAAGGCTGGTGAATATTCTGTGACGTATGATGGATTCAATACATTTACT
    AspLysThrGluLysAlaGlyGluTyrSerValThrTyrAspGlyPheAsnThrPheThr
                              80                                90

                .         .         .         .         .         .
328 ATACCTAAGACAGACTATGATAACTTTCTTATGGCTCATCTCATTAACGAAAAGGATGGG
    IleProLysThrAspTyrAspAsnPheLeuMetAlaHisLeuIleAsnGluLysAspGly
                             100                               110

                .         .         .         .         .         .
388 GAAACCTTCCAGCTGATGGGGCTCTATGGCCGAGAACCAGATTTGAGTTCAGACATCAAG
    GluThrPheGlnLeuMetGlyLeuTyrGlyArgGluProAspLeuSerSerAspIleLys
                             120                               130

                .         .         .         .         .         .
448 GAAAGGTTTGCACAACTATGTGAGGAGCATGGAATCCTTAGAGAAAATATCATTGACCTA
    GluArgPheAlaGlnLeuCysGluGluHisGlyIleLeuArgGluAsnIleIleAspLeu
                             140                               150

                .         .         .         .         .         .
508 TCCAATGCCAATCGCTGCCTCCAGGCCCGAGAATGAAGAATGGCCTGAGCCTCCAGTGTT
    SerAsnAlaAsnArgCysLeuGlnAlaArgGluEnd
                             160

                .         .         .         .         .         .
568 GAGTGGAGACTTCTCACCAGGACTCCACCATCATCCCTTCCTATCCATACAGCATCCCCA

                .         .         .         .         .         .
628 GTATAAATTCTGTGATCTGCATTCCATCCTGTCTCACTGAGAAGTCCAATTCCAGTCTAT

                .         .         .         .         .         .
688 CCACATGTTACCTAGGATACCTCATCAAGGATCAAAGACTTCTTTAAATTTCTCTTTGAT

                .         .         .         .         .         .
748 ATACCCATGACAATTTCTCATGAATTTCTTCCTCTTCCTGTTCAATAAATGATTACCCTT

808 GCACTT Poly(A)5
```

FIGURE 21

Sequence of the MUP cDNA Cloned in pUC MUP 15.

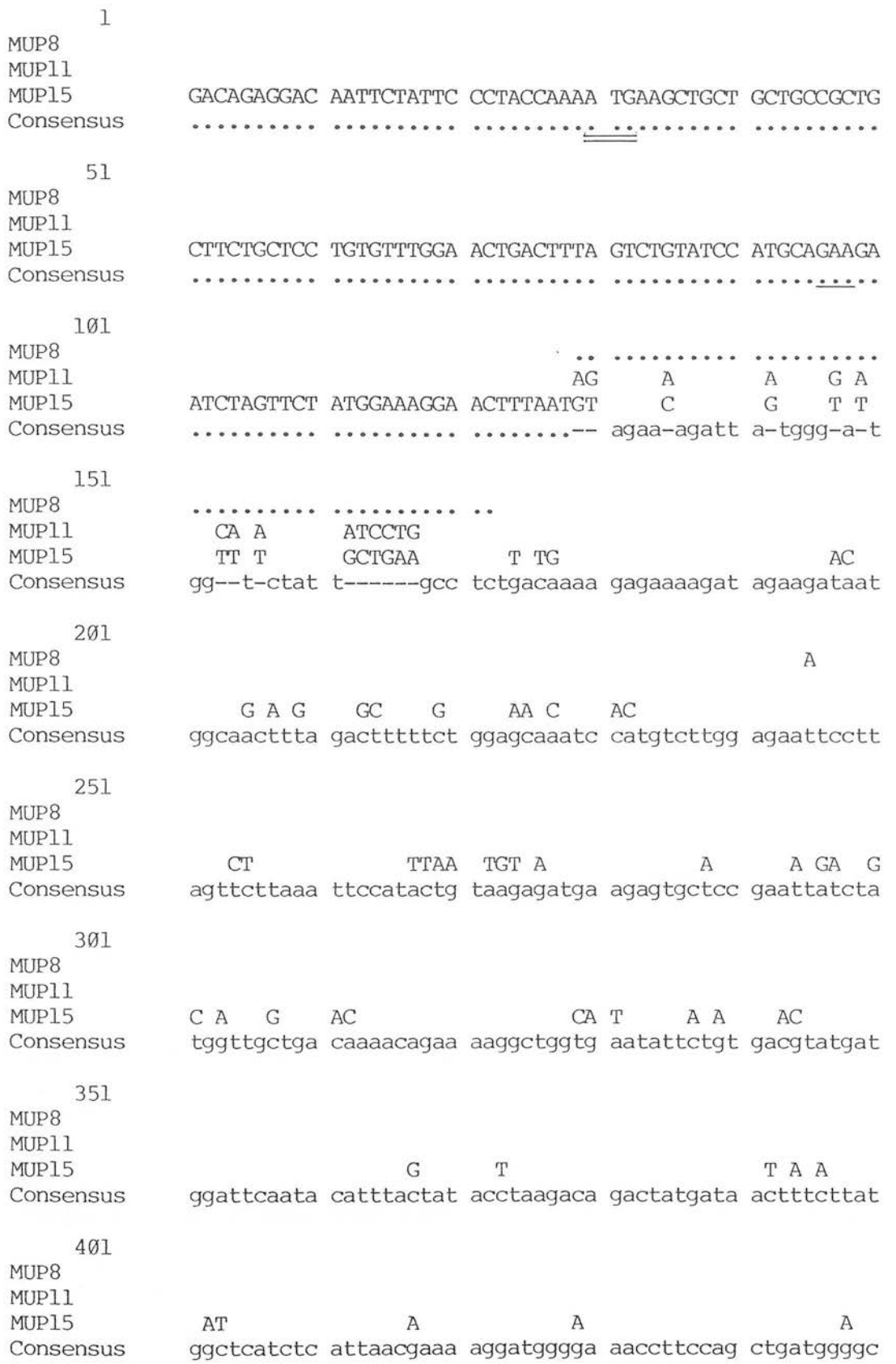
Nucleotides are numbered to the left of each line. One refers to the A of the translation initiation codon ATG. Negative numbers refer to the 5′ untranslated sequence, (Clark et al, 1984a) which extends ~36 nucleotides further 5′ than position -29 shown here (P. Ghazal, personal communication). The predicted protein sequence is shown below the DNA. Numbers below the amino acids are -22 for the start of the signal peptide, and 1 for the beginning of the mature protein (Clark et al, 1984a). The predicted sequence stops at the translation termination codon, TGA. The 3′ polyadenylation signal AATAAA (Proudfoot and Brownlee, 1976) is underlined and the extent of the poly(A) tail is indicated at the end of the sequence.

```
                                                                    .
-29  GACAGAGGACAATTCTATTCCCTACCAAAATGAAGCTGCTGCTGCCGCTGCTTCTGCTCC
                                                 MetLysLeuLeuLeuProLeuLeuLeuLeuL
                                                    -20

                                                                    .
 32  TGTGTTTGGAACTGACTTTAGTCTGTATCCATGCAGAAGAATCTAGTTCTATGGAAAGGA
     euCysLeuGluLeuThrLeuValCysIleHisAlaGluGluSerSerSerMetGluArgA
        -10                                1

                                                                    .
 92  ACTTTAATGTAGAACAGATTAGTGGGTATTGGTTTTCTATTGCTGAAGCCTCTTATGAAA
     snPheAsnValGluGlnIleSerGlyTyrTrpPheSerIleAlaGluAlaSerTyrGluA
       10                                20

                                                                    .
152  GAGAAAAGATAGAAGAACATGGCAGCATGAGAGCTTTTGTGGAAAACATCACTGTCTTGG
     rgGluLysIleGluGluHisGlySerMetArgAlaPheValGluAsnIleThrValLeuG
       30                                40

                                                                    .
212  AGAATTCCTTAGTCTTTAAATTCCATTTAATTGTAAATGAAGAGTGCACCGAAATGACTG
     luAsnSerLeuValPheLysPheHisLeuIleValAsnGluGluCysThrGluMetThrA
       50                                60

                                                                    .
272  CGATTGGTGAACAAACAGAAAAGGCTGGCATATATTATATGAACTATGATGGATTCAATA
     laIleGlyGluGlnThrGluLysAlaGlyIleTyrTyrMetAsnTyrAspGlyPheAsnT
       70                                80

                                                                    .
332  CATTTAGTATACTTAAGACAGACTATGATAATTATATTATGATTCATCTCATTAACAAAA
     hrPheSerIleLeuLysThrAspTyrAspAsnTyrIleMetIleHisLeuIleAsnLysL
       90                               100

                                                                    .
392  AGGATGGGAAAACCTTCCAGCTGATGGAGCTCTATGGCCGAGAACCAGATTTGAGTTTAG
     ysAspGlyLysThrPheGlnLeuMetGluLeuTyrGlyArgGluProAspLeuSerLeuA
      110                               120

                                                                    .
452  ACATCAAGGAAAAGTTTGCAAAACTATGCGAGGAGCATGGAATCATTAGAGAAAATATCA
     spIleLysGluLysPheAlaLysLeuCysGluGluHisGlyIleIleArgGluAsnIleI
      130                               140

                                                                    .
512  TTGACCTAACCAATGTCAATCGCTGCCTCGAGGCCCGAGAATGAAGAATGGCCTGAGCCT
     leAspLeuThrAsnValAsnArgCysLeuGluAlaArgGluEnd
      150                               160

                                                                    .
572  CCAGGTGGGCAATATACAATGAGAGCAAGGAGGAGTGTTGAGTGGAGACTTTTCACCAGG

                                                                    .
632  ACTCCAGCATCATCCCTTCCTATCCATACAGACTCCCATGCCAAGGTCTGTGATCTGCTC

                                                                    .
692  TCCACCTGTCTCACAGAGAAGTGCAATCCCGTTCTCTCCAGCATGTTACCTAGGATAACT

                                                                    .
752  CATCAAGAATCAAAGATTTCTTTAAATTTCTCTTTGCCAACACATGGAAATTCTCCATTG

                                                                    .
812  ATTTCTTTCCTGTCCTGTTCAATAAATGATTACACTTGCACTT Poly(A)67
```

FIGURE 22

Nucleotide Sequence Comparison of the MUP cDNA Clones from Female BALB/c mice.

The cDNA sequences from the pUC plasmid MUP 8, MUP 11 and MUP 15 were compared and aligned for maximum sequence homology to generate a consensus sequence. The consensus is numbered at the ends of each line. Lower case letters refer to the consensus sequence; a dot indicates where a gap has been placed to maximize the homology; a dash indicates a lack of homology at that position and upper case letters illustrate deviations in the cDNA sequence from that of the consensus. The codons for the start of the mature protein and translation termination are underlined. The signal peptide initiation codon ATG and the polyadenylation signal sequence are double underlined.


1
50
MUP8
MUP11
MUP15 GACAGAGGAC AATTCTATTC CCTACCAAAA TGAAGCTGCT GCTGCCGCTG
Consensus ....
51
100
MUP8
MUP11
MUP15 CTTCTGCTCC TGTGTTTGGA ACTGACTTTA GTCTGTATCC ATGCAGAAGA
Consensus ....
101
150
MUP8 .. ..
MUP11 AG A A G A
MUP15 ATCTAGTTCT ATGGAAAGGA ACTTTAATGT C G T T
Consensus ....-- agaa-agatt a-tggg-a-t
151
200
MUP8 ....
MUP11 CA A ATCCTG
MUP15 TT T GCTGAA T TG AC
Consensus gg--t-ctat t------gcc tctgacaaaa gagaaaagat agaagataat
201
250
MUP8 A
MUP11
MUP15 G A G GC G AA C AC
Consensus ggcaacttta gacttttct ggagcaaatc catgtcttgg agaattcctt
251
300
MUP8
MUP11
MUP15 CT TTAA TGT A A A GA G
Consensus agttcttaaa ttccatactg taagagatga agagtgctcc gaattatcta
301
350
MUP8
MUP11
MUP15 C A G AC CA T A A AC
Consensus tggttgctga caaaacagaa aaggctggtg aatattctgt gacgtatgat
351
400
MUP8
MUP11
MUP15 G T T A A
Consensus ggattcaata catttactat acctaagaca gactatgata actttcttat
401
450
MUP8
MUP11
MUP15 AT A A A
Consensus ggctcatctc attaacgaaa aggatgggga aaccttccag ctgatggggc

```
     451                                                                500
MUP8
MUP11
MUP15                                        T                   A
Consensus      tctatggccg agaaccagat ttgagttcag acatcaagga aaggtttgca

     501                                                                550
MUP8
MUP11          C
MUP15                   C                  A                           A
Consensus      aaactatgtg aggagcatgg aatccttaga gaaaatatca ttgacctatc

     551                                                                600
MUP8
MUP11
MUP15               T              G
Consensus      caatgccaat cgctgcctcc aggcccgaga atgaagaatg gcctgagcct

     601                                                                650
MUP8
MUP11
MUP15              GTGGGC AATATACAAT GAGAGCAAGG AGGAG
Consensus      ccag..... .......... .......... .....tgttg agtggagact

     651                                                                700
MUP8
MUP11
MUP15           T                G                              . CT    T
Consensus      tctcaccagg actccaccat catcccttcc tatccataca gcatccccag

     701                                                                750
MUP8
MUP11
MUP15          GCC  GG              T C    .           A        G    C
Consensus      tataaattct gtgatctgca ttccatcctg tctcactgag aagtccaatt

     751                                                                800
MUP8
MUP11                                                     G
MUP15            GT   C     G                  A                    T
Consensus      ccagtctatc ca.catgtta cctaggatac ctcatcaaga atcaaagact

     801                                                                850
MUP8
MUP11                                                    C
MUP15                         CC   .  A    G A    C C    TG
Consensus      tctttaaatt tctctttgat atacccatga caatttttca tgaatttctt

     851                                                                886
MUP8
MUP11
MUP15          T CTG                        A
Consensus      cctcttcctg ttcaataaat gattaccctt gcactt Poly(A)X
```

Most of the amino acid substitutions are conservative, with little change in the overall distribution of charged or polar residues. Most MUP sequences are very polar and on average 34% charged amino acid, and have a net positive charge due to a preponderance of acidic (20%) over basic (14%) amino acids. MUP 8, MUP 11 and MUP 15 have one more basic, one less basic and one more acidic residue than the consensus respectively (Figure 23). There is only one location of considerable change in sequence polarity between the Group 1 type genes and MUP 15. The difference occurs between amino acids 108 and 112 (Figure 23), where a net charge of -2 becomes a net charge of +2 in MUP 15 due to two changes of glutamate to lysine residues. Sequence comparisons, as described in the Methods section were made between some or all of the following sequences. The cDNAs from female BALB/c liver MUP 6, MUP 8, MUP 11 and MUP 15. The male liver cDNAs of BALB/c mice LVA 132, LVA 325 (Clissold and Bishop, 1981 and Clark et al, 1984b) and p1057 and those of C57BL/6 mice, p499 and p199 (Kuhn et al, 1984). (Several BALB/c genomic MUP clones BL 1, BS 1, BS 5 and BS 6 (Group 1) and BS 2,3 (Group 2) have been sequenced over the exonic regions, (Clark et al, 1984a; Clark et al 1985a; Clark et al, 1985b and Ghazal et al, 1985) and compared with the above. Finally, representative clones of each type of MUP sequence were aligned with the rat alpha$_{2\mu}$-Globulin cDNA sequences from male liver (Unterman et al, 1981) and the submaxillary gland (Laperche et al, 1983).

The Group 1 type cDNAs MUP 8 and MUP 11 were found to have the same high degree of sequence conservation, greater than 99% that has been reported for the Group 1 type male cDNAs (Shaw et al, 1984 and Clark et al, 1984a) and genomic sequencs (Bishop et al, 1982 and Clark et al, 1984a). The genomic clone BL 1 and male cDNAs LVA 325 and LVA 318 were found to be identical (over the comparable exonic sequences). The cDNAs were of the short type which

FIGURE 23

Homology between the Predicted Protein Sequences of Female MUP cDNAs.

The deduced amino acid sequences from the pUC plasmids MUP 8, MUP 11 are compared. The consensus between two or more sequences is shown in lower case. The comparison is numbered to the left of each line and corresponds to the position of the amino acid in the mature protein. Dots indicate gaps in a sequence; dashes show a lack of homology and non homologous amino acids are shown in bold type, potential N-glycosylation sequence (Asn-x-Thr) in MUP 15 is boxed.

```
      13
MUP8          .......... ....                          K
MUP11          K N E HT  IL
MUP15          Q S Y FS  AE  YE      EH SM A V  N T        F   LIVN
Consensus     e-i-g-w--i --asdkreki edngnfrlfl eqihvlensl vlkfhtvrde

      63
MUP8
MUP11
MUP15           T MTAIGE Q     I YM N       S  L      YI  I    K   K
Consensus     ecselsmvad ktekageysv tydgfntfti pktdydnflm ahlinekdge

     113
MUP8
MUP11                                 Q
MUP15             E          L     K          I       T  V    E
Consensus     tfqlmglygr epdlssdike rfaklceehg ilreniidls nanrclqare*
```

terminate within exon 6 (Clark et al, 1984a). However MUP 8 is identical to the long message form of BL 1, with most of exon 6 spliced out and exon 7 added (Clark et al, 1984a). This is the first evidence that both long and short MUP mRNA transcripts can be produced from one gene. However the possiblity the MUP 8 and LVA 325 transcripts are derived from MUP genes with identical exons, but different splice sites, cannot be excluded due to the large number of Group 1 genes (~ in BALB/c mice, Bishop et al, 1982), their very high sequence homology to the consensus, (Figure 24) and where the differences do occur, they frequently do so in more than one sequence (Figure 24 and Clark et al, 1985b). Therefore, the question as to whether more than one type of spliced transcript can be generated from a single gene, could perhaps be resolved by the introduction and expression of the BL 1 gene in a eukaryotic expression system.

Some of the sequences compared in Figure 24 carry identical mutations relative to the consensus sequence. These are, BS 1 and p1057 which have the same mutations at positions 537 and 804, the latter being present in p499 as well. The clone p499 also has the same mutations as p1057 at position 191 (Figure 24) and MUP 8/BL 1 at position 269 (not shown). Sequences BS 6 and MUP 11 also share a common mutation at position 782, although BS 5 shares no such mutations (Figure 24). The pair of nucleotide differences at the start of the MUP 11 sequence are probably cloning artifacts.

Of all the sequence differences within the Group 1 MUP genes Figures 22 and 24, only two result in non conservative amino acid changes Both these differences are present in the MUP 8/BL 1 sequence and the first is also present in the p499 sequence. The first is the T to A change at position 269, which results in the Eco RI restriction enzyme difference and the replacement of an asparagine (amino acid 50) with lysine. The second is a C

FIGURE 24

Comparison of Group 1 Type MUP Sequences from either Genomic or cDNA Clones.

The sequences of the Group 1 type genes BS 6, (Clark et al, 1984a); BS 1 and BS 5, (Clark et al, 1985a) and cDNAs p499 and p1057, (Kuhn et al, 1984) were aligned to illustrate their differences and compared to the cDNA MUP 11. Dots indicate unknown 5'-terminal sequences, and upper case letters indicate differences from the consensus, which is shown in lower case. The box indicates nucleotide differences from the consensus shared by BS 6 and MUP 11. Nucleotide differences between BS 6 and MUP 11 are circled. The translation initiation, mature protein start, translation termination and polyadenylation sequences are underlined. The consensus sequence is numbered at either side of each line, position 1 refers to the transcription initiation point.

```
      1                                                                  50
BS-1
BS-5
BS-6
P499          .......... .......... .......... .......... ..........
P1057         .......... .......... .......... .......... ..........
MUP11         .......... .......... .......... .......... ..........
Consensus     gagtgtagcc acgatcacaa gaaagacgtg gtcctgacag acagacaatc

     51                                                                 100
BS-1
BS-5
BS-6
P499          .......... .......... .......... .........G
P1057         .......... .......... .......
MUP11         .......... .......... .......... .......... ..........
Consensus     ctattcccta ccaaaatgaa gatgctgctg ctgctgtgtt tgggactgac

    101                                                                 150
BS-1
BS-5
BS-6
P499
P1057               G
MUP11         .......... .......... .......... .......... ..........
Consensus     cctagtctgt gtccatgcag aagaagctag ttctacggga aggaacttta

    151                                                                 200
BS-1
BS-5
BS-6
P499                                                        A
P1057                                                       A
MUP11         ..AG
Consensus     atgtagaaaa gattaatggg gaatggcata ctattatcct ggcctctgac

    201                                                                 250
BS-1
BS-5                                                  A
BS-6
P499
P1057
MUP11
Consensus     aaaagagaaa agatagaaga taatggcaac tttagacttt ttctggagca

    251                                                                 300
BS-1
BS-5
BS-6
P499                          A
P1057                                                 G
MUP11
Consensus     aatccatgtc ttggagaatt ccttagttct taaattccat actgtaagag
```

```
     301                                                                350
BS-1                          G
BS-5
BS-6
P499
P1057
MUP11
Consensus      atgaagagtg ctccgaatta tctatggttg ctgacaaaac agaaaaggct

     351                                                                400
BS-1
BS-5
BS-6
P499
P1057
MUP11
Consensus      ggtgaatatt ctgtgacgta tgatggattc aatacattta ctatacctaa

     401                                                                450
BS-1
BS-5
BS-6
P499
P1057
MUP11
Consensus      gacagactat gataactttc ttatggctca tctcattaac gaaaaggatg

     451                                                                500
BS-1
BS-5
BS-6
P499
P1057
MUP11
Consensus      gggaaacctt ccagctgatg gggctctatg gccgagaacc agatttgagt

     501                                                                550
BS-1                                                  A
BS-5
BS-6
P499
P1057                                                 A
MUP11
Consensus      tcagacatca aggaaaggtt tgcacaacta tgtgaggagc atggaatcct

     551                                                                600
BS-1
BS-5
BS-6
P499
P1057
MUP11
Consensus      tagagaaaat atcattgacc tatccaatgc caatcgctgc ctccaggccc
```

```
     601                                                                650
BS-1
BS-5
BS-6
P499
P1057                                                                   T
MUP11
Consensus      gagaatgaag aatggcctga gcctccagtg ttgagtggag acttctcacc

     651                                                                700
BS-1
BS-5
BS-6
P499
P1057
MUP11
Consensus      aggactccac catcatccct tcctatccat acagcatccc cagtataaat

     701                                                                750
BS-1
BS-5
BS-6
P499
P1057
MUP11
Consensus      tctgtgatct gcattccatc ctgtctcact gagaagtcca attccagtct

     751                                                                800
BS-1
BS-5
BS-6                                                    G
P499                                                    |
P1057                                                   |
MUP11                                                   G
Consensus      atccacatgt tacctaggat acctcatcaa gaatcaaaga cttctttaaa

     801                                                                850
BS-1               T
BS-5
BS-6
P499               T
P1057              T
MUP11                                          ©
Consensus      tttctctttg atatacccat gacaattttt catgaatttc ttcctcttcc

     851                                                                878
BS-1
BS-5
BS-6
P499
P1057
MUP11
Consensus      tgttcaataa atgattaccc ttgcactt
```

to A change at position 525, which results in the replacement of glycine (amino acid 136) with another lysine (Figures 23 and 24). The replacement of neutral amino acids with basic ones may form a basis in this very polar acidic protein, upon which the forces of natural selection could act. The apparent lack of such putative selection mechanisms, to account for the four other shared nucleotide differences, suggest that the more homologous sequences represent closer evolutionary relationships between certain members of the Group 1 MUPs . No simple series of pairwise recombinants between postulated ancestral genes can account for the distribution of identical mutations within family members relative to the consensus. One explanation that could account for the origin of the identical mutations, would be to assume an ancestry of gene duplication and subsequent gene conversion during Group 1 gene evolution (Clark et al, 1985b).

The MUP 15 clone, which is considerably different from the Group 1 type sequences (Figures 22 and 23), is identical to the shorter cDNA sequence p199 prepared from male C57 BL/6 liver, except that the third nucleotide of the p199 sequence is a G, whereas the equivalent position in MUP 15 is an A. It is possible that the p199 sequence from C57BL/6 is an polymorphism of the MUP 15 gene. However it would seem more probable that the nucleotide in question underlined, $^{5'}G_n$ TGGAAGA$^{3'}$ is either a misincorporation by reverse transcriptase close to the 5' end of the cDNA (see e.g. MUP 11, Richards et al, 1979 and Talmadge et al, 1984) or else it is part of the poly G tale, where the preceeding T represents the introduction of an unexpected nucleotide into the linker segment, due to the incomplete removal of the dNTPs used in second strand synthesis (see, e.g. Okayama and Berg, 1982). The differences in nucleotide sequence between MUP 15 and the other known MUPs and alpha$_{2\mu}$-globulin sequences are discussed elsewhere.

Two different splicing configurations for the Group 1 and 2 MUP genes have been reported, ("long" and "short") together with alternative polyadenylation sites and mRNA termini for the shorter type of transcript (Clark et al, 1984a). The MUP 15 sequence possesses yet a third type of MUP splicing configuration, resulting in mRNA 31 nucleotides longer than the "long" mRNA type. This will be referred to as "extra long" mRNA. MUP 15 is identical to both Group 1 and 2 sequences over the 10 nucleotide zone to either side of the intron 6 5′ splice junction (Figure 25, position 60/61). The inclusion of the 31 nucleotides distal to exon 6a in "extra long message", (Exon 6_{15}) could be due to a G to A mutation in MUP 15 at position 90. This would generate an apparently preferential 5′ intron 6 splice site, if the unknown MUP 15 intron sequence immediately after position 92, (Figure 25) is similar to the Group 1 or 2 sequences.

Should the sequence of the MUP 15 gene immediately after Exon 6_{15} be the same as the equivalent region of the Group 1 genes, then both the Group 1 Exon 6a/intron 6 splice junction, CAG/GTGGGC and MUP 15 Exon 6_{15}/intron 6 splice junction, GAG/GTTGGT, would conform to a similar degree with the 5′ intron splice consensus AG/GT AGT proposed by Mount, (1982).

Several studies have revealed that although the only invariant nucleotide of the 5′ splice consensus is a G for the first intron nucleotide. For nearly all sequences the second is a T and there appears to be a considerable preference for nucleotide five to be a G, (Esumi et al, 1983; Wieringa et al, 1983; Fisher et al, 1984 and Wieringa et al, 1984). However on this basis both 5′ intron 6 splice junctions would be expected to work equally well.

The most simple explanation why the Exon 6a/intron 6 splice junction is used to generate long mRNAs in the majority of Group 1 transcripts, and the Exon 6_{15}/intron 6 junction is used in MUP 15, would be to postulate that there is no unique relationship between the 5′ splice sites normally used and the pre-mRNA. Factors other than the splice consensus, although essential for splicing, would determine the final location of the splicing event. Several such factors have been proposed. These include (1) the secondary or tertiary structure of the RNA and possibly proteins (Wieringa et al, 1983), (2) the intron length (Wieringa et al, 1984) through thermodynamic considerations, (3) the terminus of the RNA moiety of UlsnRNP (Krämer et al, 1984), (4) unknown proteins presumed to interact with an additional consensus sequence near the 3′ intron splice junction, which result in the formation and final excision of the intron as a lariat structure (Ruskin et al, 1984 and Weissmann, 1984). As may be expected with a system of intron excision of this complexity, it appears that there are additional cellular regulatory mechanisms that control the factors involved, as demonstrated by the production of alternatively spliced mRNAs from a single gene transcript.

Most alternatively spliced transcription products result in the production of novel protein structures. This may be brought about by the utilization of different exons specifying novel signal and amino half protein sequences. The different light chain myosin proteins are produced in this way. Alternative splicing has also been demonstrated in the alpha A crystallin gene, where in a minority of cases, an additional short exon is expressed from part of what is usually an intron (King and Piatigorsky, 1983). The most common use of alternative splice sites appears to be the generation of proteins with different c-terminal ends, see e.g. membrane and secreted forms of Ig M (Alt et al, 1982; Early et al, 1980) and the differentially charged terminal regions of $gamma^B$ and $gamma^A$ fibrinogen

(Crabtree and Kant, 1982). Alternative splicing mechanisms have also been used to unite the different transcription initiation, and thereby 5′ untranslated regions used in different tissues to the same structural gene, as described in Young et al, (1981) for alpha amylase.

Since the demonstration that introns interrupt the coding exons, the suggestion has been made that the introns often "mark" subdomains or segments of protein structure with some function (Gilbert, 1974; 1981 and Craik et al, 1982). The presence of introns within the 5′ untranslated region also supports this view of functional domain demarcation. It has been shown that the spliced 5′ untranslated region of preproinsulin is probably important in the correct initiation of protein translation (Lomedico et al, 1982), and possibly in other genes with spliced 5′ untranslated regions (Lomedico et al, 1979). The rate of translation by ribosomes may also be determined by different 5′ untranslated regions (Young et al, 1981).

From the above, it seems that alternative splicing mechanisms for a particular gene transcript, would imply that there are either, alternative modes of expression, or different protein functions are possible for the encoding gene and resultant transcript. It is therefore possible that the use of the Exon 6_{15}/intron 6 splice junction in MUP 15, in preference to the Exon 6a/intron 6 junction used in the majority of Group 1 MUP gene transcripts, (Figure 25), is indicative of an as yet unknown function for the 3′ untranslated region of MUP mRNAs.

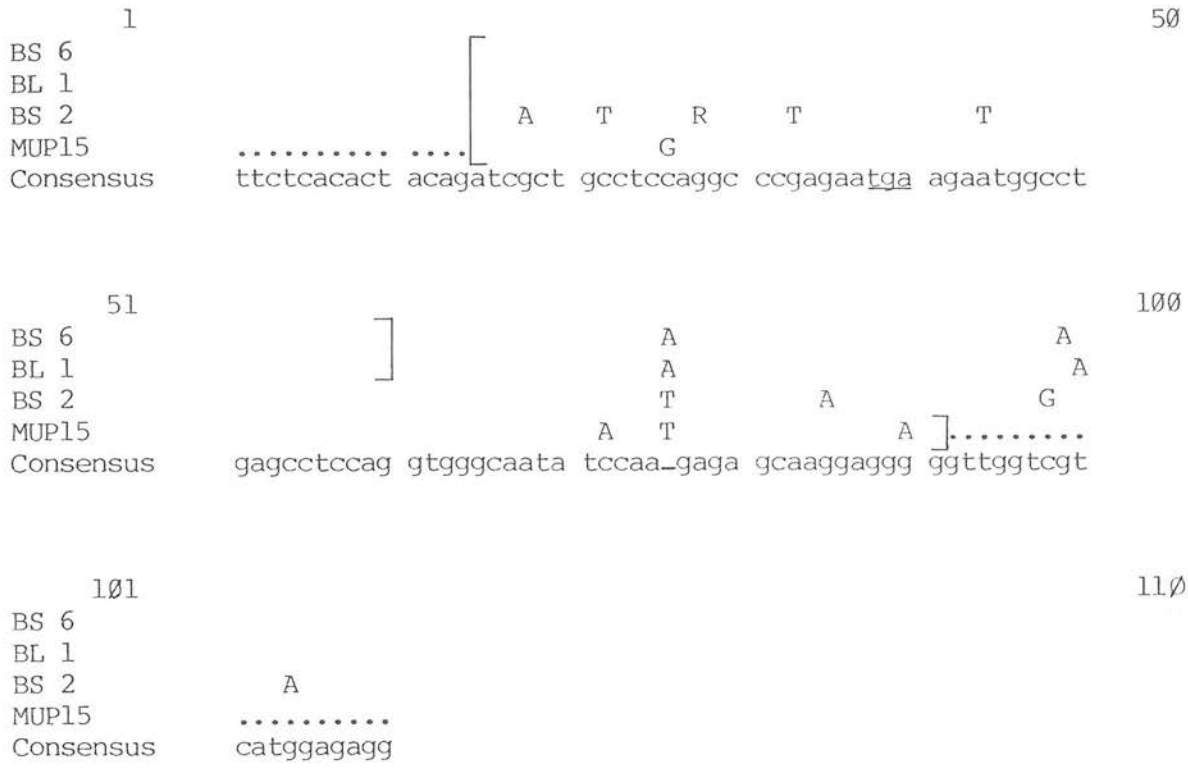
Tissue Levels of p199/MUP 15 Like mRNA

The presence of p199/MUP 15 type mRNAs have already been established in C57BL/6 mouse liver mRNA in substantial amounts. Differences in tissue specific expression and hormonal regulation of p199/MUP 15 and

FIGURE 25

Different Splicing Configurations in the 3´ Non Coding Region of Several MUP Sequences.

Exon 6 and the immediately adjoining intron sequences from the MUP genes BS 6, BL 1 and BS 2,3 (Clark et al, 1984a and Clark et al, 1985a), are aligned with exon 6 from the cDNA MUP 15. Dots indicate gaps introduced to maximize homology, and upper case initials show differences from the consensus lower case initials. The start of exon 6 is indicated by the bracket, ([) and exon 6/intron 6 boundary positions present in cloned cDNA sequences are shown by the brackets (]). The translation termination codon TGA is underlined. The consensus sequence is numbered at the ends of each line.

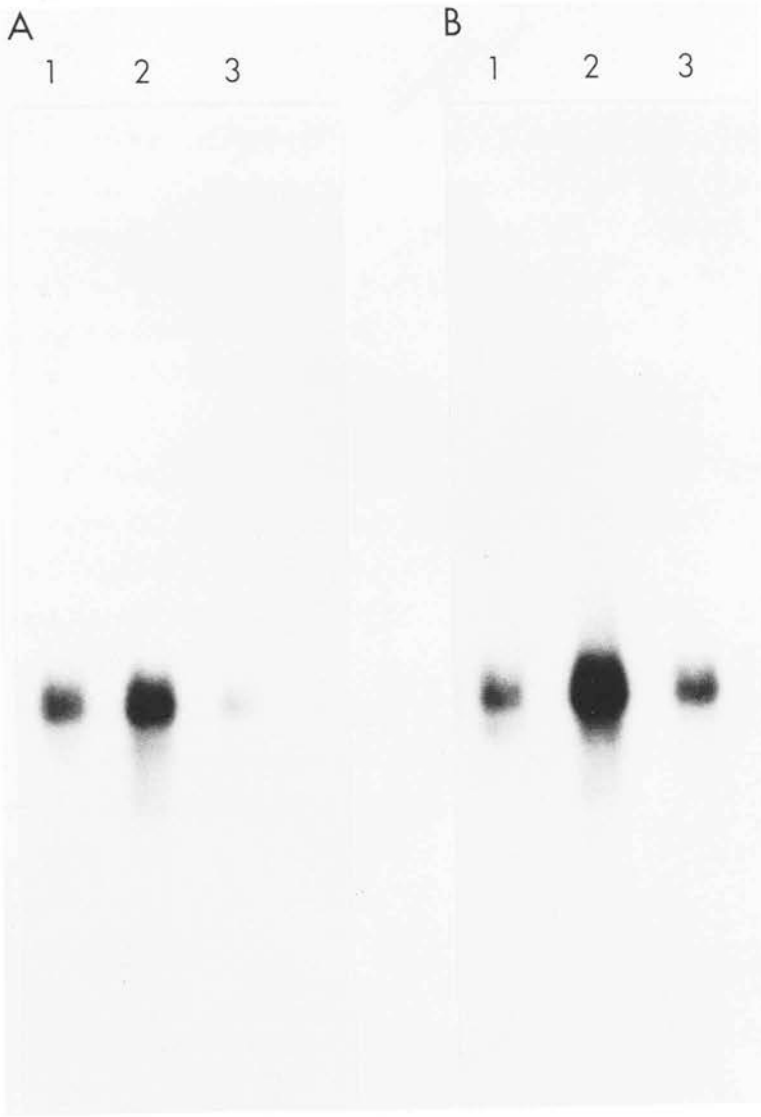

1
50
BS 6
BL 1
BS 2
MUP15
Consensus
A T R T T
G
ttctcacact acagatcgct gcctccaggc ccgagaatga agaatggcct
51
100
BS 6
BL 1
BS 2
MUP15
Consensus
A
A
T A G
A T A
gagcctccag gtgggcaata tccaa_gaga gcaaggaggg ggttggtcgt
101
110
BS 6
BL 1
BS 2
MUP15
Consensus
A
catggagagg

p499/Group 1 like mRNA levels were also demonstrated (Kuhn et al, 1984). The northern blots (of BALB/c mRNA) in Figure 26 were washed down to a stringency of 0.2 x SET at 68°C, which has been shown to minimize cross hybridization between Group 1 and Group 2 sequences (Clark et al, 1984a) and cross hybridization between MUP 15 type and Group 1 and Group 2 sequences (Table VI). It appears that there is half to one fifth as much MUP 15 type, as there is Group 1 type mRNA in male BALB/c liver. There was no detectable short type message corresponding to the MUP 15 type probe. The level of expression of MUP 15 type mRNA in female mice is similar to the amount of Group 1 type message, of which a greater proportion corresponds to the short type, relative to male liver. These sex and MUP type dependant splicing differences, may also reflect a possible role of the 3′ untranslated region in MUP expression.

FIGURE 26

Northern Blots of Male and Female Mouse Liver E.R.poly(A) mRNA Probed for MUP Sequences.

Six Male and six Female BALB/c mice, 12 weeks of age, were sacrificed and the Liver E.R.poly(A) mRNA fractions isolated. Two fractions of the female (2µg; lane 1) and male preparations (2µg; lane 2 and 0.4 µg, lane 3) were subject to electrophoresis under denaturing conditions, transferred to nitrocellulose and probed with nick translated cDNA inserted into the sequencing vector M13mp9. Set A were probed with M13 MUP 15 and Set B were probed with M13 MUP 11. The final wash was 0.2 x SET at 68 °C and the filters were exposed for 4 days. Equal loading of sets A and B was confirmed by hybridization to a second probe, transferrin, (results not shown).


A
1
2
3
B
1
2
3

Comparative Sequence Divergence

The mRNA sequence of one representative member of each type of mouse MUP type Group 1: BS 6, Group 2: BS 2,3 MUP 15 and the analogous rat alpha$_{2\mu}$-globulin rat sequence were aligned in Figure 27. Simple pairwise comparisons of these sequences, Table VII revealed that unlike the BS 6 sequence comparison with BS 2,3 (~13%) the differences between BS 6 and MUP 15 were distributed unevenly across the length of the comparison (see also Clark et al, 1985a and b).

The sequence divergence between BS 6 and MUP 15 was nearly twice as great in the first three exons of the mature protein and terminal quarter of the 3′ untranslated mRNA region as it was over the intervening portion of the sequence comparison (Table VII). A similar sequence divergence pattern to the above was also observed in the BS 2,3 and MUP 15 comparison and to a much lesser extent between the rat and mouse MUP sequences, which will be discussed later.

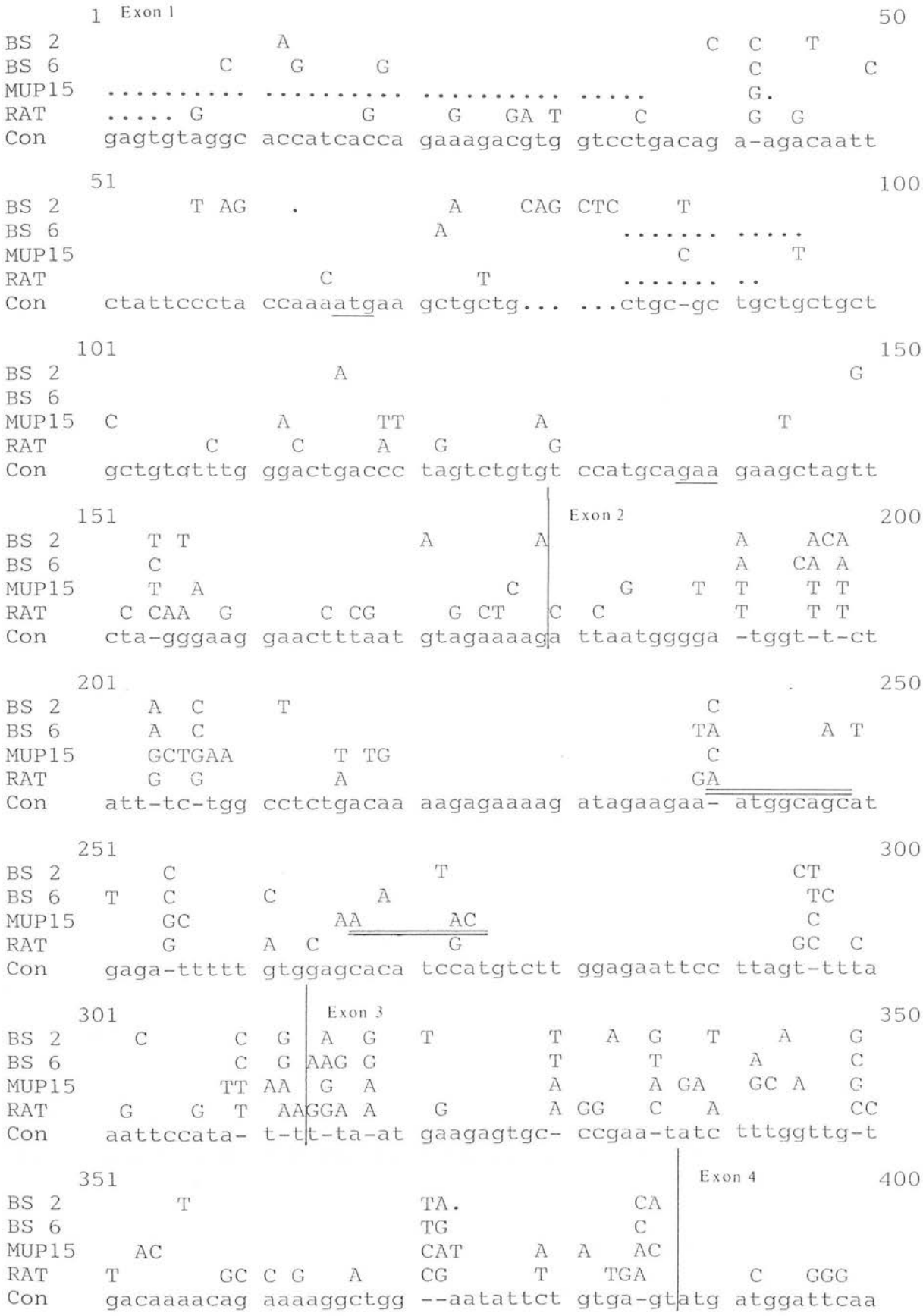
Selective Constraint and Non-Uniform Sequence Divergence

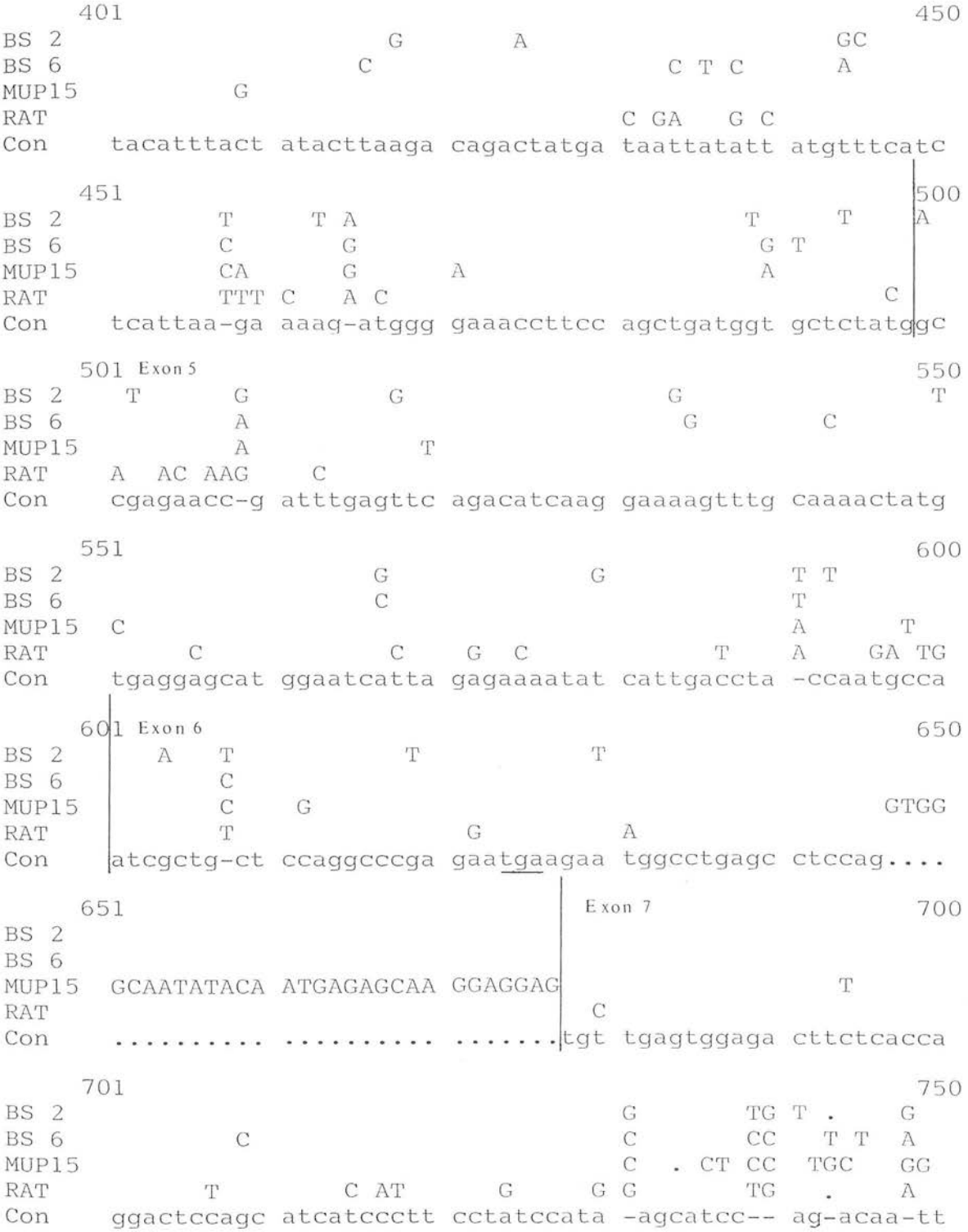
One possible explanation for the differences in sequence divergence along the length of the comparison is that the first three exons and terminal quarter of untranslated region (end regions) are subject to fewer (or weaker) selective constraints than the remainder of the sequence (central region).

FIGURE 27

Homology Between the Major Urinary Protein Genes of Mice and the Corresponding Rat Protein alpha$_{2\mu}$-Globulin.

The exon sequences of a Group 1 MUP gene BS 6 (Bishop et al, 1982 and Clark et al, 1985b), a Group 2 MUP gene BS 2,3 (Bishop et al, 1982 and Clark et al, 1985b), an alpha$_{2\mu}$-Globulin cDNA (Unterman et al, 1981 and Laperche et al, 1983) and the cDNA MUP 15 were aligned. Dots indicate gaps introduced to maximize homologies. Capital letters indicate differences with respect to the consensus sequence which is depicted by either lower case letters or a dash where there is no homology. The translation initiation, mature protein start and translation termination codons are underlined. The polyadenylation sequence AATAAA and two potential N-glycosylation sequences are double underlined. The consensus sequence is numbered at either side of each line, position 1 refers to the transcription initiation point. The vertical lines delineate the exons.


1 Exon 1 50
BS 2 A C C T
BS 6 C G G C C
MUP15 .... G.
RAT ..... G G G GA T C G G
Con gagtgtaggc accatcacca gaaagacgtg gtcctgacag a-agacaatt
51 100
BS 2 T AG . A CAG CTC T
BS 6 A ....
MUP15 C T
RAT C T ....
Con ctattcccta ccaaaatgaa gctgctg... ...ctgc-gc tgctgctgct
101 150
BS 2 A G
BS 6
MUP15 C A TT A T
RAT C C A G G
Con gctgtgtttg ggactgaccc tagtctgtgt ccatgcagaa gaagctagtt
151 Exon 2 200
BS 2 T T A A A ACA
BS 6 C A CA A
MUP15 T A C G T T T T
RAT C CAA G C CG G CT C C T T T
Con cta-gggaag gaactttaat gtagaaaaga ttaatgggga -tggt-t-ct
201 250
BS 2 A C T C
BS 6 A C TA A T
MUP15 GCTGAA T TG C
RAT G G A GA
Con att-tc-tgg cctctgacaa aagagaaaag atagaagaa- atggcagcat
251 300
BS 2 C T CT
BS 6 T C C A TC
MUP15 GC AA AC C
RAT G A C G GC C
Con gaga-ttttt gtggagcaca tccatgtctt ggagaattcc ttagt-ttta
301 Exon 3 350
BS 2 C C G A G T T A G T A G
BS 6 C G AAG G T T A C
MUP15 TT AA G A A A GA GC A G
RAT G G T AAGGA A G A GG C A CC
Con aattccata- t-tt-ta-at gaagagtgc- ccgaa-tatc tttggttg-t
351 Exon 4 400
BS 2 T TA. CA
BS 6 TG C
MUP15 AC CAT A A AC
RAT T GC C G A CG T TGA C GGG
Con gacaaaacag aaaaggctgg --aatattct gtga-gtatg atggattcaa


401 450
BS 2 G A GC
BS 6 C C T C A
MUP15 G
RAT C GA G C
Con tacatttact atacttaaga cagactatga taattatatt atgtttcatc
451 500
BS 2 T T A T T A
BS 6 C G G T
MUP15 CA G A A
RAT TTT C A C C
Con tcattaa-ga aaag-atggg gaaaccttcc agctgatggt gctctatggc
501 Exon 5 550
BS 2 T G G G T
BS 6 A G C
MUP15 A T
RAT A AC AAG C
Con cgagaacc-g atttgagttc agacatcaag gaaaagtttg caaaactatg
551 600
BS 2 G G T T
BS 6 C T
MUP15 C A T
RAT C C G C T A GA TG
Con tgaggagcat ggaatcatta gagaaaatat cattgaccta -ccaatgcca
601 Exon 6 650
BS 2 A T T T
BS 6 C
MUP15 C G GTGG
RAT T G A
Con atcgctg-ct ccaggcccga gaatgaagaa tggcctgagc ctccag....
651 Exon 7 700
BS 2
BS 6
MUP15 GCAATATACA ATGAGAGCAA GGAGGAG T
RAT C
Con ....tgt tgagtggaga cttctcacca
701 750
BS 2 G TG T . G
BS 6 C C CC T T A
MUP15 C . CT CC TGC GG
RAT T C AT G G G TG . A
Con ggactccagc atcatccctt cctatccata -agcatcc-- ag-acaa-tt

```
     751                                                          800
BS 2                T         A                  T           T  G
BS 6                 A                    T         C      T  A    A
MUP15                  C     .                                  T
RAT         C        A            C G          A             T
Con      ctgtgatctg ctttccatcc tgtctcacag agaagtgcaa tcccggtctc

     801                                                          850
BS 2         CG                  .
BS 6         .                                C           G          .C
MUP15                                                               .
RAT                CT TCCCTA        C     C     A     G       T    AG
Con      tccagcat.. ......gtta cctaggataa ctcatcaaga atcaaagact

     851                                                          900
BS 2         A           .A     A T                    A     T
BS 6                           A T T                    T T
MUP15                          C C .  A      GA             TG
RAT          .                G C CC           TC     CG    CA    A
Con      ttctttaaat ttctctttg- -acacccatg acaattctcc atgaatttct

     901                                                          937
BS 2         T          A                   A     A  T
BS 6                                        C
MUP15        T CTG                          A
RAT          G     G A                      C
Con        tcctcttcct gttcaataaa tgattac-ct tgcactt
                      ========
```

TABLE VII

Total Percentage Divergence Between the Various Regions of the Mouse MUPs and Rat alpha$_{2\mu}$-Globulin Sequences.

The total number of differences between the sequence pairs, as aligned in Figure 27, were determined over the mature protein and 3′ untranslated regions. All insertions/deletions were scored as one base pair change. The differences are expressed as percentages and are not corrected for multiple point mutations.

TABLE VII

Sequences Compared	Mature Protein Region		3´Untranslated Region	
	Exons 1-3 (250 Bp)	Exons 4-6 (239 Bp)	Exons 6-7 (204 Bp)	Exon 7-End (last 68 Bp)
BS 6-BS 2,3	12.8	12.1	14.2	13.2
BS 6-MUP15	23.6	7.5	11.8	22.0
BS 2,3-MUP15	23.6	12.6	13.7	26.5
RAT-BS 6	24.0	18.0	19.6	22.0
RAT-MUP15	28.0	17.2	19.1	27.9
RAT-BS 2,3	28.0	20.1	17.2	29.4

As a consequence of greater selective pressure the central regions are not as free to diverge as the end regions. If this were the case then the ratio of replacement to silent site differences would be greater in exons 1-3 than in exons 4-6. The corrected frequencies of silent and replacement site differences were calculated for the pairs of sequences compared (Perler et al, 1980). (Table VIII). BS 2,3 is a pseudogene (Clark et al, 1985a and Ghazal et al, 1985) and as such comparisons including this sequence are not suitable for the above type of analysis. The disproportionate conservation of exons 4-6 over exons 1-3, would initially seem to be due to selection, as the ratio of replacement to silent site differences is 2.8 for exons 1-3 and 1.5 for exons 4-6 between MUP 15 and BS 6 and 1.1 to 0.6 and 1.4 to 0.6 for BS 6-rat and MUP 15-rat respectively.

A similar analysis comparing the silent & replacement site divergence of sequences from the consensus sequence in Figure 27 was done in an attempt to see what has occurred in individual genes, (Table IX). The ratio of replacement to silent sites for MUP 15 over exons 1-3 and 4-6 was very similar. Therefore it is unlikely that selection alone could account for the four fold increase in both silent and replacement sites within exons 1-3 over those in exons 4-6 of MUP 15 from the consensus sequence.

Any postulated selection mechanism to account for the uniform sequence divergence between the MUP 15 and BS 6 or BS 2,3 sequences compared in Table VII, would also have to include the distribution of "silent site" differences of the coding and 3´ non coding regions. Within the 3´ untranslated region there are local fluctuations in sequence common to all the mouse/mouse and some of the mouse/rat comparisons, (Table X). The first seventy base pairs of the 3´ untranslated region long type mRNA show

TABLE VIII

Silent and Replacement Site Sequence Divergence Between Mouse MUPs and Rat alpha$_{2\mu}$-Globulin.

The mature protein coding sequences of mouse Group 1 gene (BS 6), MUP 15 and rat alpha$_{2\mu}$-globulin liver cDNA were aligned as described in Figure 27 (positions 138-626). The percentage sequence divergence between each pair of sequences was calculated, for both silent and replacement sites and corrected for multiple point mutations, as described in Perler et al, (1980).

TABLE VIII

Pairwise Comparison of Mature Protein Sequences	Corrected Percentage Change			
	Exons 1-3		Exons 4-6	
	Silent	Replacement	Silent	Replacement
BS 6-MUP15	10.4	29.8	4.7	7.1
BS 6-RAT	24.9	27.1	23.5	13.7
MUP15-RAT	23.6	32.0	22.7	13.7

TABLE IX

Silent and Replacement Site Sequence Divergence Between Mouse MUPs, Rat alpha$_{2\mu}$-Globulin and the Figure 27 Consensus Sequence.

Individual mouse MUP sequences and the rat alpha$_{2\mu}$-globulin sequence were aligned and compared with the consensus sequence as described in Figure 27. The percentage sequence divergences between the consensus and the mouse or rat sequences were calculated, for both silent and replacement sites and corrected for multiple point mutations (Perler et al, 1980).

TABLE IX

Consensus Sequence Compared With	Corrected Percentage Change			
	Exons 1-3		Exons 4-6	
	Silent	Replacement	Silent	Replacement
BS 6	2.0	5.3	3.9	3.3
MUP15	6.2	13.5	1.3	3.3
RAT	18.7	14.2	14.5	9.5
BS 2,3 "Pseudogene"	6.9	5.6	2.6	7.2

TABLE X

Divergence Within the 3′ Untranslated Region Between MOUSE MUPs and RAT alpha$_{2\mu}$-Globulin.

The 3′ untranslated region sequences of a Group 1 MUP gene BS 6, a Group 2 MUP gene BS 2,3, the MUP 15 cDNA and a rat alpha$_{2\mu}$-globulin cDNA were aligned as described in Figure 27 (positions 627-937). The percentage sequence divergence between each pair of sequences was calculated and corrected for multiple point mutations as described in Perler et al, (1980); except that point insertions or deletions were treated as nucleotide changes and longer insertions or deletions were excluded from the comparison.

TABLE X

Pairwise Comparison Of 3′ Untranslated Regions	ZONE (Percentage Change)					
	1 627-725 (68Bp)	2 726-793 (68Bp)	3 794-869 (68Bp)	Sub-Total 1-3 (204Bp)	4 870-937 (68Bp)	Total Untranslated Region (272Bp)
BS 6-BS 2,3	3.0	24.1	14.2	15.7	14.5	15.5
BS 6-MUP15	3.0	26.1	11.8	12.8	26.1	15.9
BS 2,3-MUP15	3.0	28.2	13.7	15.1	30.4	18.7
RAT-BS 6	12.8	26.1	19.6	22.7	26.1	23.6
RAT-MUP15	12.8	34.9	19.1	22.0	37.3	25.1
RAT-BS 2,3	12.8	16.4	17.2	19.5	34.9	23.6

some of the greatest sequence homologies of any part of the comparisons between the different MUPs (3.0%) and MUPs and the rat sequence (12.8%).

Where highly homologous non-coding sequences occur, it is usually due to recent divergence or a conversion event (Talmadge et al, 1984; Jones and Kafatos, 1984 and Shen et al, 1981) and in these cases the coding regions show similar degrees of divergence to the non-coding regions. Specific homologous zones within untranslated regions have only been found within sequences where the RNA has a specific function. Examples of such structures occur within the μ intron of the immunoglobulin heavy chain genes, where a portion of the intron has been shown to have enhancer activity (Mills et al, 1983 and Picard and Schaffner, 1984) and within the 5′ untranslated regions of two preproinsulin genes where a putative secondary structure (Lomedico et al, 1979) may ensure correct translation initiation (Lomedico et al, 1982).

It is therefore possible that this region immediately after the coding region of MUPs and alpha$_{2\mu}$-globulin has some unknown functional role within urinary protein gene expression. However it should be noted that the distribution of mutations is usually random within non coding sequences (Perler et al, 1980) and that the small area of sequence comparison (70Bp) does mean that chance could play a large part in the generation of such areas of low sequence diversity. Indeed similarly sized areas of high sequence homology may be found in the sequence comparison of some BS 6 and BS 2,3 introns (Clark et al, 1985a), which are presumably due to the random nature of silent site mutation events. Alternatively if there were a sequence conservation/functional relationship to the mRNA just after the translation stop sequence, then the various possible splicing arrangements both within specific gene transcripts, e.g. MUP 15 and possibly BL 1 or between MUP (Clark et al, 1984a) and alpha$_{2\mu}$-globulin

genes (Unterman et al, 1981; Dolan et al, 1982 and Laperche et al, 1983), would interrupt or delete the conserved sequence and as such could represent a post translational control mechanism of any function thereof. A possible but as yet unknown function for the 3´ untranslated regions of MUPs and alpha$_{2\mu}$-globulin has also been postulated by Clark et al, (1984a), Ghazal et al, (1985) and Dolan et al, (1982) respectively.

There is insufficient data to determine whether the first 70 Bp of the 3´ untranslated region of high sequence homology between MUP sequences is a chance event. Only the testing of the premise, possibly by in vitro-mutagenesis will resolve which of the alternative explanations is correct. Until then the sentiment expressed by Proudfoot, (1984), that "the view that all the controls that modulate eukaryotic gene expression centre on the promotor, may well prove too narrow", may apply to the MUP multi gene family.

It is therefore apparent that in consideration of the possibilities by which differential selective constraints could have generated the non uniform sequence divergences, there are some mechanisms which could account for the differences over parts of the compared sequences. These include some of the replacement sites within the coding regions and silent sites in the first quarter of the 3´ untranslated region. These may account for the slightly non uniform sequence divergence between the mouse and rat sequences. Tables VII, VIII and X. However they do not seem to account for all the differences in both silent and replacement sites across the length of the MUP 15 and BS 6 or BS 2,3 comparisons.

The use of Percentage Silent and Replacement Site Mutations as Evolutionary Clocks for MUPs and alpha$_{2\mu}$-Globulin

The percentage replacement change of nucleotides between gene sequences has been shown to be a good evolutionary clock over long periods of time. Similarly silent sites can be used as evolutionary clocks, although they are less accurate due to a wider spread of rates, and only over shorter time intervals less than 85 M.Y., after which the truly silent sites saturate. The observations (Perler et al, 1980) were made using genes, the products of which were essential to the survival of the organism and which have known structural/functional relationships. These silent sites accumulated mutations far more rapidly (5 to 7 times quicker) than the replacement sites over the first 85 MY. The replacement and silent site data presented in Tables VIII-X are unsuitable for similar comparisons from one another for a number of reasons.

(i) It is not known whether the MUP and alpha$_{2\mu}$-globulin proteins have a function.

(ii) The relationship of the translated protein structure to its possible function is therefore similarly unknown. There may therefore be differing degrees of selective pressures on the members of the multi gene family.

(iii) BS 2,3 is a Group 2 gene, which are known to be pseudogenes (Ghazal et al, 1985), and are accumulating mutations at the same rate in both introns and exons (Clark et al, 1985a).

(iv) The Group 1, Group 2 and alpha$_{2\mu}$-globulin genes are members of gene arrays, between the members of which multiple information exchange events have been proposed during their evolution (Clark et al, 1985a and Dolan et al, 1982). These could affect to an unknown degree the accumulation and distribution of mutations in both coding and non coding positions.

(v) The sequences compared by Perler et al, (1980), where the silent sites accumulate (initially) at 5-7 times the rate of replacement mutations in all exons. The MUP and alpha$_{2\mu}$-globulin sequences have frequently accumulated replacement mutations, at the same or to a greater extent than the "silent" site, some of which may have a function in gene expression, (Clark et al, 1984b, and Dolan et al, 1982).

(vi) The substitution rates between different areas of the same comparison could differ by up to a factor of two. Furthermore the sequence comparisons of lower overall silent site divergence showed the greatest fluctuations in rates of silent site mutations between different areas of the comparison. It is therefore possible that the relatively small percentage silent site differences between the MUP/MUP and MUP/alpha$_{2\mu}$-globulin sequence comparison in Tables VIII and IX are influenced by up to a factor of 2, or possibly more, by the small number and random distribution of silent mutations.

Further Observations on MUP and alpha$_{2\mu}$-Globulin Rates of Silent and Replacement Mutations

Some of the factors which may alter the apparently linear evolutionary relationships, may themselves be drawn upon in the analysis of the anomalous sequence divergence result of the MUP and alpha$_{2\mu}$-globulin genes. While still acknowledging these considerations and their potential effects, several observations and inferences about urinary proteins may be made, or previous ones confirmed or extended. The relatively high ratio of replacement site to silent site differences in comparisons of the rat alpha$_{2\mu}$-globulin gene with the Group 1 gene BS 6 (Clark et al, 1984a), other Group 1 genes and partial MUP 15 type

genes (Kuhn et al, 1984) suggests that a precise primary amino acid structure is not as important to the function of urinary protein genes as it is for several other multi gene families. (See e.g. Perler et al, 1980; Kedes, 1979; Talmadge et al, 1984) as was suggested by Kuhn et al, (1984).

It has been suggested that it is the N-terminal hexapeptide of the mature protein that is the active part of the MUP molecule. If this were so, the MUP could be the androgen regulated proteinacious urine product which stimulates the onset of puberty in young female mice, (Clark et al, 1985a). A comparison of the sequences of the two hexapeptides of Group 1 and Group 2 genes revealed that they were quite similar. This implies that the truncated product of the Group 2 gene could possibly still have a function, (Clark et al, 1985a). The N-six terminal amino acid sequences of Group 1, Group 2, mup15 and rat alpha$_{2\mu}$-globulin sequences are also compared (Figure 28). The major urinary protein components of both mice Group 1 and rat alpha$_{2\mu}$-globulin (liver) are identical over this region. MUP 15 is very similar having the same number and types of changed and polar R groups on its amino acid. The BS 2,3 N-terminal sequence is also similar as it is very polar and has 4 out of 6 residues identical to Group 1 and rat alpha$_{2\mu}$-globulin sequences, but unlike MUP 15 it has a different net change to the others. The structure of the N-terminal six amino acids has been well conserved among the urinary proteins; as would be expected if the suggestion of Clark et al, (1985a) was correct. Similarly the hypothesis that the N-terminal part is the active part of the MUP molecule would also explain how a more variable protein structure could be better tolerated within urinary protein genes, than it was for other multi gene families. However there may well be additional functions for the carboxyl half of the protein as reflected by its lower replacement site to silent site mutation ratio when compared to the amino half of the protein. (Tables VIII

FIGURE 28

The N-Terminal Hexapeptide Sequences of Mouse MUPs and Rat alpha$_{2\mu}$-Globulin Proteins.

The N-terminal six amino acids of the mature protein translation products that correspond to mouse MUPs and rat alpha$_{2\mu}$-globulin cDNAs were compared. MUP Group 1 sequences (Kuhn et al, 1984 and Clark et al, 1985b), MUP Group 2,3 sequences (Ghazal et al, 1985), MUP 15 (Figure 21) and rat liver alpha$_{2\mu}$-globulin (Unterman et al, 1981). The N-terminal sequences all begin with the first glutamate residue. The symbols under the amino acids indicate the polar nature of their R groups; $\oplus$, positively and $\ominus$, negatively charged between pH 5-7; P_{OH}, polar due to free hydroxyl group; NP, non polar and NP_S non polar group containing sulphur.

FIGURE 28

cDNA	N-HEXAPEPTIDE TRANSLATION	NET CHARGE
MUP Group 1	GLU(⊖) - GLU(⊖) - ALA(NP) - SER(P_{OH}) - SER(P_{OH}) - THR(P_{OH})	-2
RAT $alpha_{2\mu}$-Globulin	GLU(⊖) - GLU(⊖) - ALA(NP) - SER(P_{OH}) - SER(P_{OH}) - THR(P_{OH})	-2
MUP15	GLU(⊖) - GLU(⊖) - SER(P_{OH}) - SER(P_{OH}) - SER(P_{OH}) - MET(NP_S)	-2
MUP Group 2	GLU(⊖) - GLU(⊖) - ALA(NP) - ARG(⊕) - SER(P_{OH}) - MET(NP_S)	-1

and IX). Some of the differences in amino acid structures of the MUPs may be related to their expression in different tissues, see e.g. Table I, (Shahan and Derman, 1984). Structural differences may also reflect nuances in the function of MUPs, as has been proposed for similarly organized multi gene family of the chorion genes of the silk moth (Jones and Kafatos, 1980).

The possible ways by which differential selective constraints could account for the non-uniform sequence divergence between the MUP genes, especially BS 6 and MUP 15 have been discussed. Selective constraint on sequence divergence could account for some of the replacement site differences in the carboxyl half of the mature protein and the first quarter of the 3′ untranslated region. However the extent of the differences in silent site replacements between regions of the 3′ untranslated region and the parallel decrease in both silent and replacement site divergence over the carboxyl half of the protein sequence, would require a natural selection hypothesis of prodigious complexity to account for them.

Gene Conversion and Non-Uniform Sequence Divergence

Two mechanisms have been proposed to explain the occurrence of sequences which show unexpected similarity when compared. One is the unequal crossover between arrays of repetitions units (Smith, 1976). It has been proposed as a possible mechanism by which the different percentage divergences within and between Group 1 and Group 2 MUP genes could be explained. In the case of MUP genes, the unit is 45 Kb in length and comprised of a Group 1 and Group 2 gene, divergently orientated and their flanking sequences (Clark et al, 1984b and Bishop et al, 1985). Similarly unequal crossover between arrays of repetitious units has been proposed as a mechanism by which sequence

homogenization occurred in the rat alpha$_{2\mu}$-globulin genes (Dolan et al, 1982).

The other major mechanism by which sequences of unusual similarity may have been generated is gene conversion (Slighton et al, 1980 and Shen et al, 1981). Gene conversion, superimposed on gene duplication, has been proposed as the most likely mechanism by which the non random distribution of differences at 5 sites could have occurred when seven Group 1 genes were compared (Clark et al, 1985b).

When Clark et al, (1985b) compared Group 1 and Group 2 genes with MUP 15, they proposed that the unequal distribution of divergence between exons 1-3 and exons 4-7 could be explained if there had been a gene conversion event between MUP 15 and other MUP genes. There is a notional junction between exons 3 and 4 (with 2-3 times as much divergence to the 5′ side as the 3′ side, Table VII) which would imply that one boundary to the converted region lies within intron 3. Genomic clones for MUP 15 are not yet available therefore the precise location of the boundary cannot be made. However within intron 3 of both BS 6 and BS 2,3 genes there is a stretch of the tandemly repeated dinucleotide (TG)n (Clark et al, 1985a) which has been associated with the proposed end points of gene conversion (Proudfoot and Mariatis, 1980 and Shen et al, 1981).

Further analysis of the 3′ non-coding region of MUP and rat alpha$_{2\mu}$-globulin sequences, comparisons, Figure 27, revealed a second notional junction. Like the exon 3, exon 4 notional junction, the 3′ notional junction is defined by two regions of differing divergence, being twice as great on the 3′ side as it is on the 5′ side (Table VII) and occurs approximately 70 base pairs from the 3′ end of long type mRNA (Figure 27). Immediately 3′ to the proposed boundary is the sequence (TTCTTTAAATTTCTCTTTG), identical in BS 6 and MUP 15, within which lies the directly repeated pentanucleotide sequence

TCTTT, with comparable positions being 12 nucleotides apart.

This feature is very similar to that described by Proudfoot and Maniatis, (1980) and the 5′ boundary of a region of almost complete sequence identity which occurs on the 3′ side of the pseudo alpha$_1$ and alpha$_2$ human globulin genes, where the pentanucleotide GCCTG is repeated with comparable nucleotides 10 bases apart. Similarly at the 5′ boundary of the ^{G}gamma and ^{A}gamma foetal globin gene conversion there is the sequence TCAAAAAT, directly repeated with equivalent positions 51 nucleotides apart (Shen et al, 1981).

Although it will be necessary to study more gene conversion events in order to determine the conditions and/or structures required for higher eukaryotic gene conversion events to take place, it seems probable that either one or both of the following are required. A short directly repeated pentanucleotide with equivalent positions 10-12 nucleotides apart or a longer repeat at a greater repeat distance. Alternatively, or together with, the consecutively repeated dinucleotide $(TG)_n$ or $(CA)_n$ where n>4, which has already been implicated as a possible requirement for the related events of gene conversion and unequal crossover. In the examples cited above both types of sequence occur, although there does not seem to be any consensus as to whether they occur together (Proudfoot and Maniatis, 1980), separately, or at either end of the sequence (Shen et al, 1981).

The Origins of the Non-Uniform Divergence Between the Compared MUP Sequence

There are several possible ways by which the non uniform divergence between MUP 15 and the Group 1 or BS 2,3

genes could have arisen, as described by Clark et al, (1985b). Although all of these mechanisms are still possible, some would seem to be less likely than others in the light of further analysis of the sequence comparisons.

Primarily, the location of the 3′ notional junction, in addition to the 5′ exon 3/exon 4 notional junction, renders the hypothesis that the 5′ regions of MUP 15 or the Group 1/Group 2 ancestor were converted by a more distantly related MUP gene less likely because a similar event would have to occurred at the 3′ end too. Thus if the end regions of MUP 15 or the Group 1/Group 2 ancestor were the converted regions, two conversion events close together would have to be postulated, whereas if the portions between the notional junctions were the converted region, then one conversion event would suffice to explain the unequal divergence results (Table VII).

Two of the hypothesies that were proposed by Clark et al (1985b) would require only one gene conversion event to account for the non-uniform divergence.

(i) A gene ancestral to both, or either of the Group 1 and Group 2 genes may have converted the homologous region of MUP 15.

(ii) The homologous region of MUP 15 converted the ancestor of the Group 1 and Group 2 genes. If the latter had occurred, or the ancestor to Group 1 and Group 2 genes had converted part of MUP 15, then specific patterns of divergence between all the modern day sequences of MUP 15, BS 6 and BS 2,3 would be expected. If the conversion occurred about the time of the onset of the divergence of Group 1 and Group 2 genes then they should all be equally diverged from one another. Should the conversion have occurred prior to the onset of divergence of the Group 1 and Group 2 genes, then BS 6 and BS 2 would show greater similarities over the proposed converted region than either BS 6 and MUP 15 or BS 2,3 and MUP 15, both of which would be similar in their degrees of divergence. Neither

of these patterns of divergence were found when the MUP sequences were compared (Table VII). While it should be noted that there are many factors which may adversely influence some of the silent and replacement site mutations rates and thus sequence divergence between MUP sequences, as discussed above, there is one possible gene conversion scenario which would best fit the data presented in Table VII. It is possible that from somewhere in intron three to approximately 70 base pairs from the end of the long mRNA (i.e. positions 388-869, Figure 27), after onset of the divergence of the Group 1 and Group 2 genes, a Group 1 type gene converted the homologous region of MUP 15. From the above type of conversion event the following pattern of sequence divergence would be expected for the MUP gene comparison over the proposed region of conversion:

(i) The degree of divergence between BS 2,3 and MUP 15 would be about the same as that between BS 2,3 and BS 6. (ii) The silent site sequence divergence in both the carboxyl half of the translated region and the 3′ untranslated region of the sequence comparison between BS 6/MUP 15 be lower than either that of BS 6/BS 2 or BS 2/MUP 15. The percentage silent site divergences corrected for multiple mutations (Perler et al, 1980), for the mature protein region exons 4-6 and the proposed converted 3′ untranslated region are: BS 6/MUP 15 4.7% and 12.8%, BS 6/BS 2,3 10.5% and 15.7%, BS 2,3/MUP 15 8.3% and 15.1% respectively. It would therefore seem that the proposed conversion event between the progenators of the MUP 15 gene and a Group 1 gene is the most likely explanation for the un-uniform sequence divergence between the mouse MUP genes compared (between positions 388-869, Figure 27). The above necessary comparisons, particularly between other MUP 15/pl99 like sequences (Kuhn et al, 1984) and the Group 1 (Clark et al, 1985b) and Group 2 genes (Ghazal et al, 1985) will be essential to test the above hypothesis. The possible conversion event, discussed above, does not preclude the possibility of other

information exchanges either within or between the different types of MUP genes, except for subsequent exchanges betwen MUP 15 and the known sequences of either Group 1 or Group 2 genes. Indeed one question which remains unresolved is the relatively low percentage of silent site changes in exons 1-3 of the MUP 15/BS 6 comparison, (Table VIII) and with respect to the consensus sequence (Table IX). The overall replacement site divergence within exons 1-3 and the silent site divergence outwith the proposed conversion region, between MUP 15 and Group 1 or Group 2 genes all show levels similar to those between the rat alpha$_{2\mu}$-globulin and the Group 1 or Group 2 genes (Tables 7-10). Whether the low percentage of silent sites changes within exons 1-3 of MUP 15, reflects the wide spread in silent site mutation rates, that have been observed in other related genes (Perler et al, 1980), or some other factors, remains to be determined.

THE SEQUENCING, ANALYSIS AND COMPARISON OF TRANSFERRIN cDNA CLONES

Three transferrin cDNA clones were chosen for sequencing to obtain the maximum length of cDNA sequence for mouse transferrin from the 7 pUC TRF clones isolated (Figure 14). The three clones pUC TRF 21, 32 and 14 were used because they would provide the longest sequence overlaps between the restriction enzyme sites used in the sequencing strategy (Figure 30). Between them the clones also contain all four boundary limits (most and least 5' and 3' cloned regions) of the common 1.2 $\pm$ 0.1 Kb of cDNA present in 6 of the 7 transferrin clones (Figure 14). It was hoped that the sequence of the boundaries delimiting the common portion of cDNA would provide clues to the mechanism by which it arose.

The selected cDNAs were digested with several combinations of restriction enzymes. The cDNA segments were then ligated into the sequencing vectors and their structures determined as described in the methods section and Figure 30.

The overlapping sequences of clones pUC TRF 21, 32 and 14 are identical and on this basis seem to represent partial overlapping clones of the transcription product of one gene. The length of transferrin mRNA encoded by the clones is 1164 Bp (Figure 29), approximately half the estimated length of mouse transferrins mRNA, 2.3 Kb (Figure 3). There is one extensive open reading frame within the sequence which enables 327 amino acids of the COOH terminus to be predicted. The translation termination codon TAA

FIGURE 29

Sequence Data Assembled from Transferrin cDNA Inserts Cloned in pUC TRFs 14, 21, and 32.

Nucleotides are numbered to the left of each line, one, refers to the most 5' nucleotide of the insert in pUC TRF 21. There was only one continuous open reading frame and the predicted protein sequence is shown under the DNA. Numbers below the amino acids refer to their distance from the start of the predicted sequence. The predicted sequence stops at the translation termination codon TAA. The 3' polyadenylation signal is underlined and the extent of the poly (A) tail (pUC TRF 32) is indicated at the end of the sequence. The sequence data was obtained as shown in Figure 30; pUC TRF 21, nucleotides 1-434; pUC TRF 32, nucleotides 143-1164 and pUC TRF 14, nucleotides 771-1135 and all overlapping sequence determinations were identical.

```
                .         .         .         .         .         .
  1  CGGGGGGACCAAGTGTGACGAGTGGAGCATCATCAGTGAGGGAAAGATAGAGTGTGAGTC
      GlyGlyThrLysCysAspGluTrpSerIleIleSerGluGlyLysIleGluCysGluSe
        1                                 10

                .         .         .         .         .         .
 61  AGCAGAGACCACTGAGGACTGCATTGAAAAGATTGTGAACGGAGAAGCGGACGCCATGAC
     rAlaGluThrThrGluAspCysIleGluLysIleValAsnGlyGluAlaAspAlaMetTh
      20                                30

                .         .         .         .         .         .
 121  TTTGGATGGAGGACATGCCTACATTGCAGGCCAGTGTGGTCTAGTGCCTGTCATGGCAGA
     rLeuAspGlyGlyHisAlaTyrIleAlaGlyGlnCysGlyLeuValProValMetAlaGl
      40                                50

                .         .         .         .         .         .
 181  GTACTACGAGAGCTCTAATTGTGCCATCCCATCACAACAAGGTATCTTTCCTAAAGGGTA
     uTyrTyrGluSerSerAsnCysAlaIleProSerGlnGlnGlyIlePheProLysGlyTy
      60                                70

                .         .         .         .         .         .
 241  TTATGCCGTGGCTGTGGTGAAGGCATCGGACACTAGCATCACCTGGAACAACCTGAAAGG
     rTyrAlaValAlaValValLysAlaSerAspThrSerIleThrTrpAsnAsnLeuLysGl
      80                                90

                .         .         .         .         .         .
 301  CAAGAAGTCCTGCCACACTGGGGTAGACAGAACCGCTGGTTGGAACATCCCTATGGGCAT
     yLysLysSerCysHisThrGlyValAspArgThrAlaGlyTrpAsnIleProMetGlyMe
     100                               110

                .         .         .         .         .         .
 361  GCTGTACAACAGGATCAACCACTGCAAATTCGATGAATTTTTCAGTCAAGGCTGCGTCCC
     tLeuTyrAsnArgIleAsnHisCysLysPheAspGluPhePheSerGlnGlyCysValPr
     120                               130

                .         .         .         .         .         .
 421  GGGTATGAGAAGAATTCACCTCTGTGACCTGTGTATTGGCCCACTCAAATGTGCTCCGAA
     oGlyMetArgArgIleHisLeuCysAspLeuCysIleGlyProLeuLysCysAlaProAs
     140                               150

                .         .         .         .         .         .
 481  CAACAAAGAGGAATATAATGGTTACACAGGGGCTTTCAGGTGTCTCGTTGAGAAAGGAGA
     nAsnLysGluGluTyrAsnGlyTyrThrGlyAlaPheArgCysLeuValGluLysGlyAs
     160                               170

                .         .         .         .         .         .
 541  TGTAGCCTTTGTGAAACACCAGACTGTCCTGGATAACACCGAAGGAAAGAACCCTGCCGA
     pValAlaPheValLysHisGlnThrValLeuAspAsnThrGluGlyLysAsnProAlaGl
     180                               190

                .         .         .         .         .         .
 601  ATGGGCTAAGAATCTGAAGCAGGAAGACTTCGAGTTGCTCTGCCCTGATGGCACCAGGAA
     uTrpAlaLysAsnLeuLysGlnGluAspPheGluLeuLeuCysProAspGlyThrArgLy
     200                               210

                .         .         .         .         .         .
 661  GCCTGTGAAAGATTTTGCCAGCTGCCACCTGGCCCAAGCTCCAAACCATGTTGTGGTCTC
     sProValLysAspPheAlaSerCysHisLeuAlaGlnAlaProAsnHisValValValSe
     220                               230
```

```
721  ACGAAAAGAGAAGGCAGCCCGGGTTAAGGCTGTACTGACTAGCCAGGAGACTTTATTTGG
     rArgLysGluLysAlaAlaArgValLysAlaValLeuThrSerGlnGluThrLeuPheGl
     240                             250

781  GGGAAGTGACTGCACCGGCAATTTCTGTTTGTTCAAGTCTACCACCAAGGACCTTCTGTT
     yGlySerAspCysThrGlyAsnPheCysLeuPheLysSerThrThrLysAspLeuLeuPh
     260                             270

841  CAGGGATGACACCAAATGTTTCGTTAAACTTCCAGAGGGTACCACACCTGAAAAATACTT
     eArgAspAspThrLysCysPheValLysLeuProGluGlyThrThrProGluLysTyrLe
     280                             290

901  AGGAGCGGAGTACATGCAATCTGTCGGTAACATGAGGAAGTGCTCAACCTCACGACTCCT
     uGlyAlaGluTyrMetGlnSerValGlyAsnMetArgLysCysSerThrSerArgLeuLe
     300                             310

961  GGAAGCCTGCACTTTCCACAAACATTAAAATCCAAGAGGTGGGTTGCCACTGTGGTGGAG
     uGluAlaCysThrPheHisLysHisEnd
     320

1021 ACAGATGCTCCCTCCCGTGGCCCATGGGCTTCTCTTGGTCTTCATGCCCTGAGGGGTTGG

1081 GGCTAACTGGTGTAGTCTTCGCTGCTGTGCCTTACCACATACACAGAGCACAAAATAAAA

1141 ACGACTGCTGACTTTATATTTCCC Poly(A)33
```

precedes a 3' non-coding region of 176 nucleotides. The canonical polyadenlation signal AATAAA (Proudfoot and Brownlee, 1979) is present 31 nucleotides upstream of the poly (A) tract (Figure 29).

The clone pUC TRF 14 is truncated by 29 nucleotides from the 3' terminus of mouse transferrin mRNA. The other 5 cDNA clones which share the common ~1Kb of transferrin mRNA sequence extend 30Bp further 3' than pUC TRF 14 and probably represent clones with intact 3' untranslated regions and different length poly (A) tracts (Figure 14). pUC TRF 14 stops at the second adenine within the polyadenylation sequence, thus terminating in poly $(A)_4$ (Figure 29, position 1135). The variant polyadenylation sequence ATTAAA, that has been reported in several sequences (see, eg. Clark et al, 1984a; and references therein) is present 152 BP upstream of the poly $(A)_4$. It is considered unlikely that this represents an alternative mRNA termination/polyadenylation product, because the distance between the signal and site of adenylation would be of an unprecedented length and the putative poly $(A)_4$ occurs exactly at the same position as four consecutive adenine residues within the longer cDNA clones. A more plausible suggestion for the generation of the truncated 3' end of pUC TRF 14 at the underlined nucleotides within the sequence AAAATAAAAA, is that it is the result of SI nuclease digestion within the AT rich region during the cloning procedure. A similar proposal was made to account for the apparently truncated MUP clone LVA 132 within the sequence TTTAAATTT (Clark et al, 1984a), and may also have caused the shorter human transferrin cDNA clone of Uzan et al, (1984) relative to that of Yang et al, (1984) with the sequence ATTTATATTTC poly (A); although in this case its proximity to the polyadenylation signal and poly (A) tract may equally reflect a genuine alternative mRNA 3' end (Proudfoot, 1984).

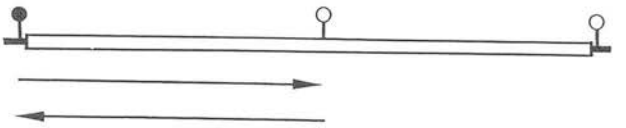
The relative homogeneity of the 3' ends of the mouse transferrin cDNA clones (Figure 14) can be explained by the

FIGURE 30

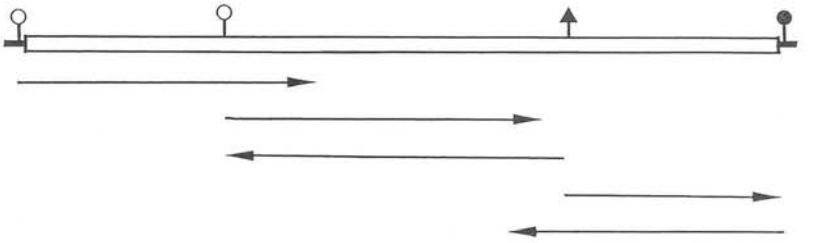
Sequence-Analysis, Strategy for Cloned Transferrin cDNAs.

The cDNA sequences are schematically represented by open boxes. The flanking pUC 8 sequence is represented by a broad line. Horizontal arrows indicate the extent and polarity of sequence data obtained from the various restriction fragments subcloned into M13 mp 8/9 or tg 130/131 sequencing vectors. Vertical lines and symbols refer to the location of restriction sites: Eco RI (⚲), Kpn I (↟) and Bam HI (↟).

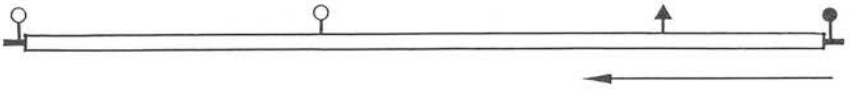
pUC TRF21



pUC TRF32



pUC TRF14



complete cloning of the 3' end of the mRNA and in one case its subsequent digestion. However such mechanisms are not applicable to the 5' ends of the clones, as the mRNA extends to a further~ 1.2 Kb 5' and there are no A/T rich sequences at or near the 5' ends of the clones pUC TRF 21 and 32. The presence of a kinetic barrier to reverse transcriptase within the mRNA at, or further 5' to the terminus of the most extensive clone is one possible explanation of how the similar clones could have been generated. A barrier to reverse transcriptase would have resulted in sScDNAs with identical 5' ends for one gene's transcripts. The different final 5' ends of the clones, within a relatively short region probably reflects the formation of hairpin loops at different positions during the synthesis of the second strand of dScDNAs. Whatever mechanism prevented the complete systhesis of mouse transferrin cDNAs or their cloning was not evident from an analysis of the sequence data. It is possible that the causative agent is present in other transferrin sequences although the effect may not be as strong. Uzan et al, (1984) have recently cloned a similarly truncated human transferrin cDNA; although a complete clone has also been isolated (Yang et al, 1984). It is therefore probable that the elucidation of this phenomenon will only be resolved by cloning the remaining carboxyl and amino domains of mouse transferrin and/or by reverse transcriptase kinetic inhibition studies (see, eg. Hagenbuchle et al, 1978 and Hearst, 1979).

Comparison of Transferrins

The mouse transferrin sequence (Figure 29) was compared with the equivalent human (Yang et al, 1984) and chicken (Jeltsch and Chambon, 1982) cDNA sequences (Figure 31). The percentage corrected sequence divergences were calculated as described in the Methods section. The silent site divergences between the mouse/human, mouse/chicken and human/chicken sequences were 38.4%, 99% and 72%

respectively. The mouse/human silent site divergence is 10% lower than the equivalent beta globin genes and some 30-40% lower than the equivalent growth hormone, alpha globin and preproinsulin gene comparisons (Perler et al, 1980). This raises the possibility that there has been some selective pressure to conserve certain amino acid codon silent sites within the mouse and human COOH transferrin domain, possibly reflecting a preferred mRNA secondary structure. However should any such constraints occur, then they are confined to the mammalian sequences, as comparisons with ovotransferrin were in very good agreement with the silent site divergence between other chicken/human and chicken/rodent gene comparisons (Perler et al, 1980).

The replacement site divergence over the COOH transferrin domain between mouse/human, mouse/chicken and human/chicken sequences was 18%, 35.5% and 37.5% respectively. The replacement changes between transferrins are all 8-15% higher than the equivalent comparisons between alpha and beta globin and preproinsulin genes. This could either reflect a less stringent structure/functional relationship or the localization of a conserved functional domain(s) within the body of the protein which has less stringent structural criteria.

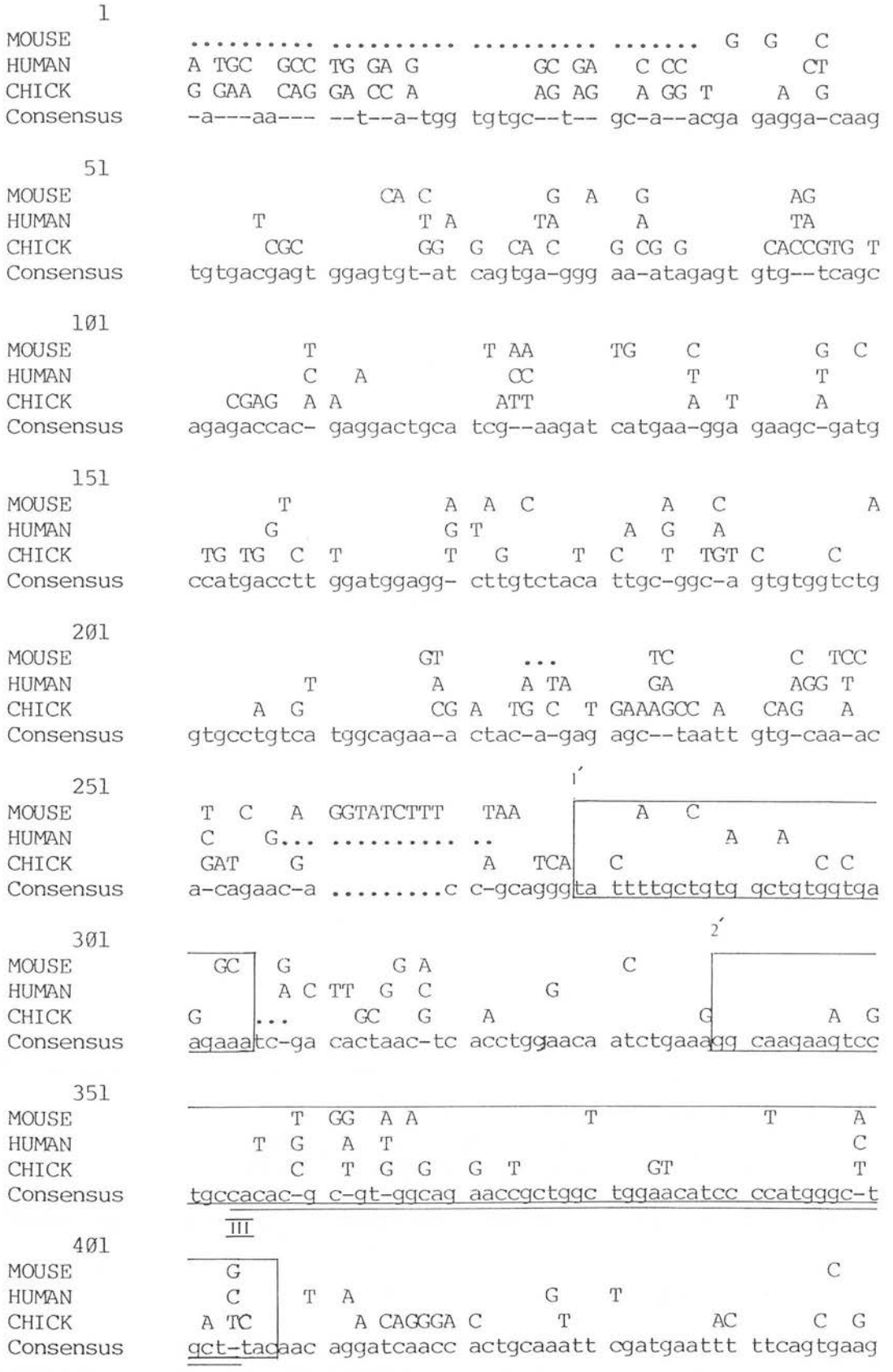
Iron Binding Domains

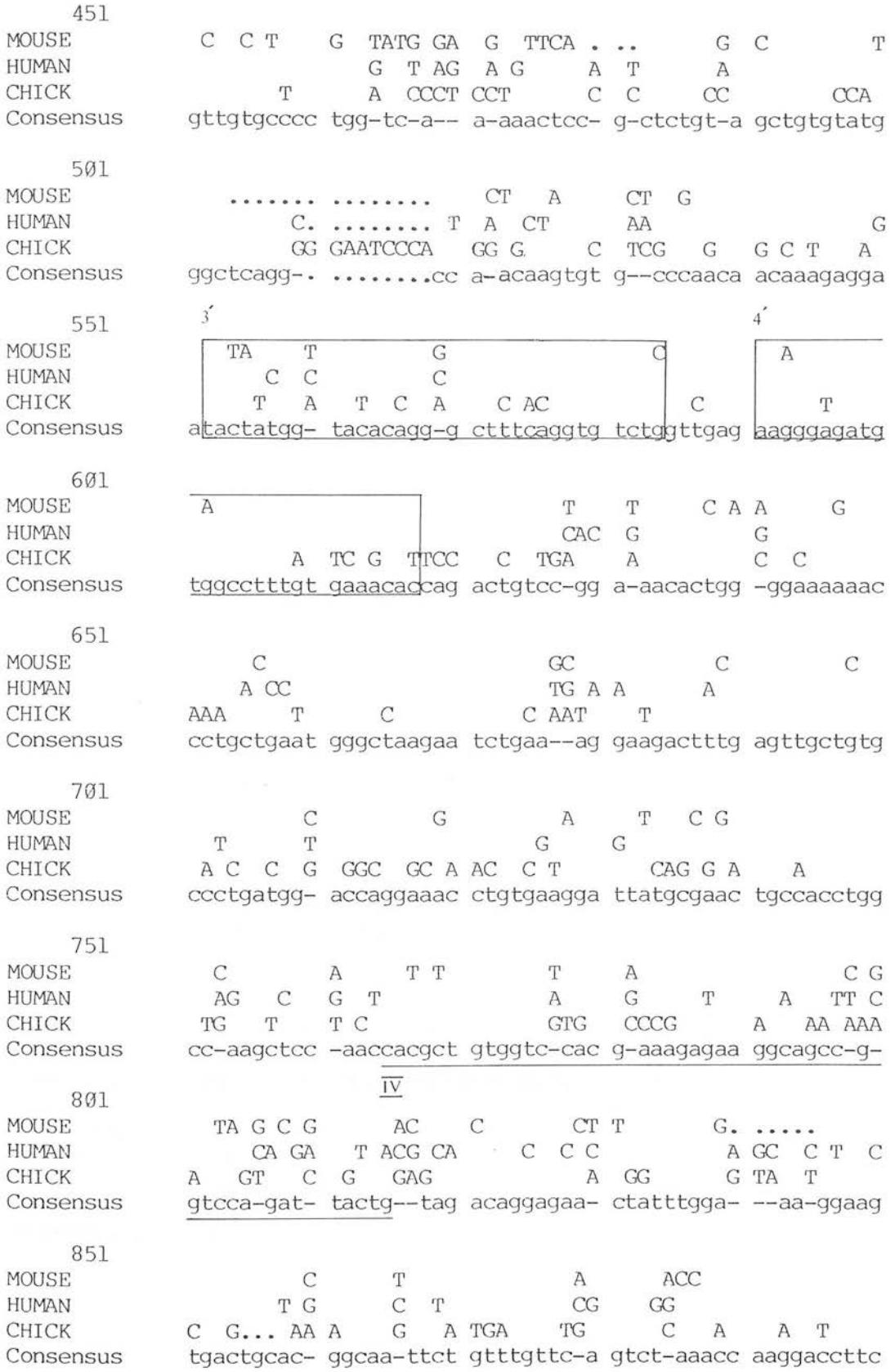
The predicted protein sequences for the full length human and chicken clones, together with the mouse transferrin sequence revealed that the mouse sequence encodes the whole COOH duplicated domain except for the first 14 amino acids. Four areas of high protein sequence homology are evident, which have been conserved both between the NH_2 (1-4) and COOH (1'-4') protein domains and between species (Figure 32, boxed areas). The boxed areas lie within larger blocks of sequence similarity between the NH_2 and COOH domains for which significant homology has been shown (Jeltsh and Chambon, 1982 and Yang et al, 1983).

FIGURE 31

Homology Between the 3'Half of Chicken, Human and Mouse Transferrin cDNA Sequences

The following sequences are aligned for homologies, the mouse cDNA sequence (Figure 29), the human cDNA sequence, starting at amino acid 338 (Yang et al, 1984) and the chicken cDNA sequence, starting at amino acid 341 (Jeltsch and Chambon, 1982). Dots indicate gaps introduced to maximize homologies. Capital letters indicate differences to the consensus sequence which is depicted by either small case letters or a dash where there is no homology. The translation termination codon positions are underlined and the polyadenylation sequence is double underlined. Sequences enclosed by the numbered boxes correspond to the highly conserved duplicated protein domains with transferrin (Figure 32). The underlined sequences show the location of the putative ancient quadruplication domains, as described by Jeltsch and Chambon, (1982). The consensus sequence is numbered at the ends of each line, position 1 corresponds to nucleotide 1155 of chicken ovotransferrin (Jeltsch and Chambon, 1982).


1
50
MOUSE .... G G C
HUMAN A TGC GCC TG GA G GC GA C CC CT
CHICK G GAA CAG GA CC A AG AG A GG T A G
Consensus -a---aa--- --t--a-tgg tgtgc--t-- gc-a--acga gagga-caag
51
100
MOUSE CA C G A G AG
HUMAN T T A TA A TA
CHICK CGC GG G CA C G CG G CACCGTG T
Consensus tgtgacgagt ggagtgt-at cagtga-ggg aa-atagagt gtg--tcagc
101
150
MOUSE T T AA TG C G C
HUMAN C A CC T T
CHICK CGAG A A ATT A T A
Consensus agagaccac- gaggactgca tcg--aagat catgaa-gga gaagc-gatg
151
200
MOUSE T A A C A C A
HUMAN G G T A G A
CHICK TG TG C T T G T C T TGT C C
Consensus ccatgacctt ggatggagg- cttgtctaca ttgc-ggc-a gtgtggtctg
201
250
MOUSE GT ... TC C TCC
HUMAN T A A TA GA AGG T
CHICK A G CG A TG C T GAAAGCC A CAG A
Consensus gtgcctgtca tggcagaa-a ctac-a-gag agc--taatt gtg-caa-ac
1′
251
300
MOUSE T C A GGTATCTTT TAA A C
HUMAN C G... .... A A
CHICK GAT G A TCA C C C
Consensus a-cagaac-a .....c c-gcagggta ttttgctgtg gctgtggtga
2′
301
350
MOUSE GC G G A C
HUMAN A C TT G C G
CHICK G ... GC G A G A G
Consensus agaaatc-ga cactaac-tc acctggaaca atctgaaagg caagaagtcc
351
400
MOUSE T GG A A T T A
HUMAN T G A T C
CHICK C T G G G T GT T
Consensus tgccacac-g c-gt-ggcag aaccgctggc tggaacatcc ccatgggc-t
III
401
450
MOUSE G C
HUMAN C T A G T
CHICK A TC A CAGGGA C T AC C G
Consensus gct-tacaac aggatcaacc actgcaaatt cgatgaattt ttcagtgaag

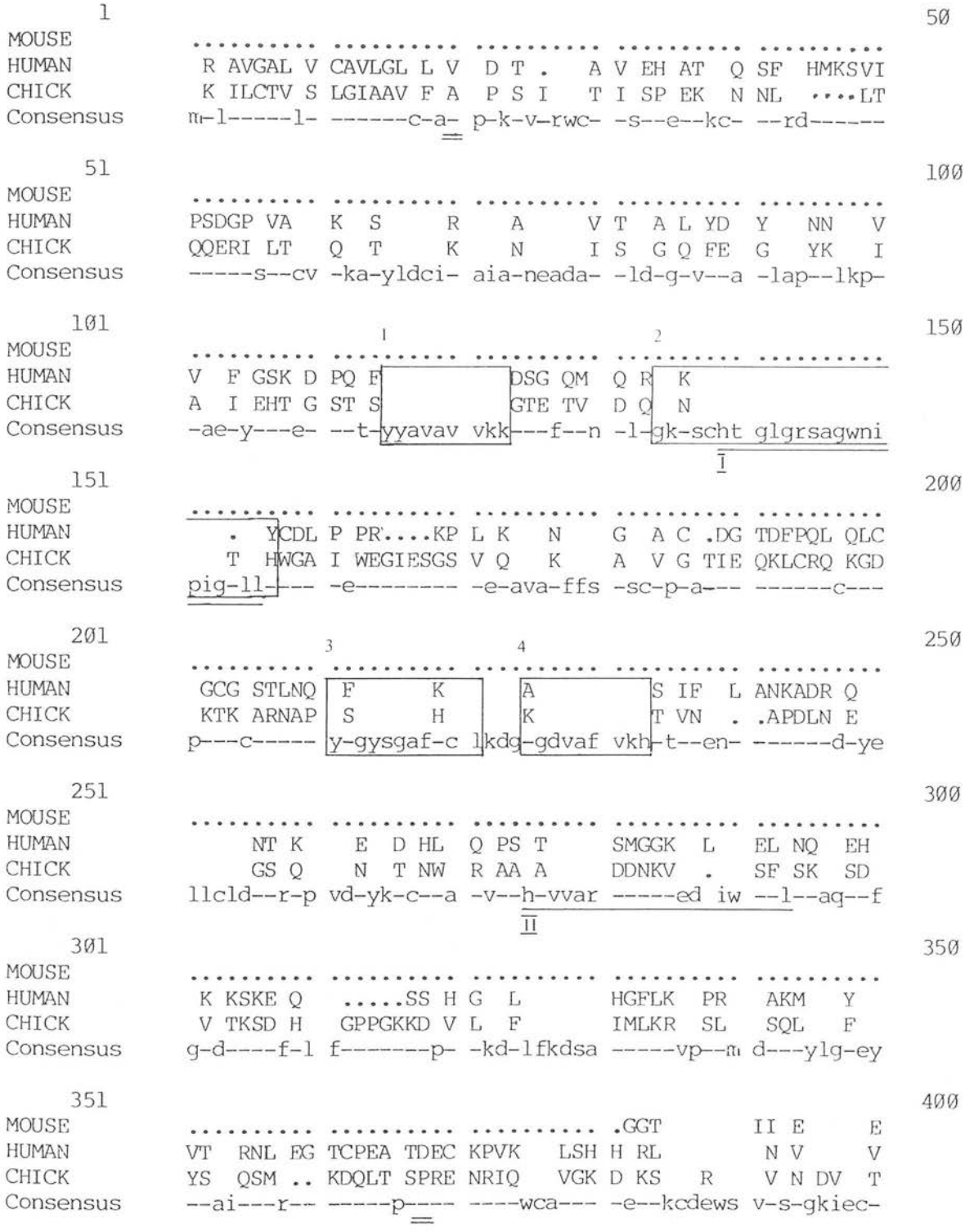

451
500
MOUSE C C T G TATG GA G TTCA . .. G C T
HUMAN G T AG A G A T A
CHICK T A CCCT CCT C C CC CCA
Consensus gttgtgcccc tgg-tc-a-- a-aaactcc- g-ctctgt-a gctgtgtatg
501
550
MOUSE .... CT A CT G
HUMAN C. ..... T A CT AA G
CHICK GG GAATCCCA GG G C TCG G G C T A
Consensus ggctcagg-. .....cc a-acaagtgt g--cccaaca acaaagagga
551
3'
4'
600
MOUSE TA T G C A
HUMAN C C C
CHICK T A T C A C AC C T
Consensus atactatgg- tacacagg-g ctttcaggtg tctggttgag aagggagatg
601
650
MOUSE A T T C A A G
HUMAN CAC G G
CHICK A TC G TTCC C TGA A C C
Consensus tggcctttgt gaaacaccag actgtcc-gg a-aacactgg -ggaaaaaac
651
700
MOUSE C GC C C
HUMAN A CC TG A A A
CHICK AAA T C C AAT T
Consensus cctgctgaat gggctaagaa tctgaa--ag gaagactttg agttgctgtg
701
750
MOUSE C G A T C G
HUMAN T T G G
CHICK A C C G GGC GC A AC C T CAG G A A
Consensus ccctgatgg- accaggaaac ctgtgaagga ttatgcgaac tgccacctgg
751
800
MOUSE C A T T T A C G
HUMAN AG C G T A G T A TT C
CHICK TG T T C GTG CCCG A AA AAA
Consensus cc-aagctcc -aaccacgct gtggtc-cac g-aaagagaa ggcagcc-g-
IV
801
850
MOUSE TA G C G AC C CT T G. .....
HUMAN CA GA T ACG CA C C C A GC C T C
CHICK A GT C G GAG A GG G TA T
Consensus gtcca-gat- tactg--tag acaggagaa- ctatttgga- --aa-ggaag
851
900
MOUSE C T A ACC
HUMAN T G C T CG GG
CHICK C G... AA A G A TGA TG C A A T
Consensus tgactgcac- ggcaa-ttct gtttgttc-a gtct-aaacc aaggaccttc

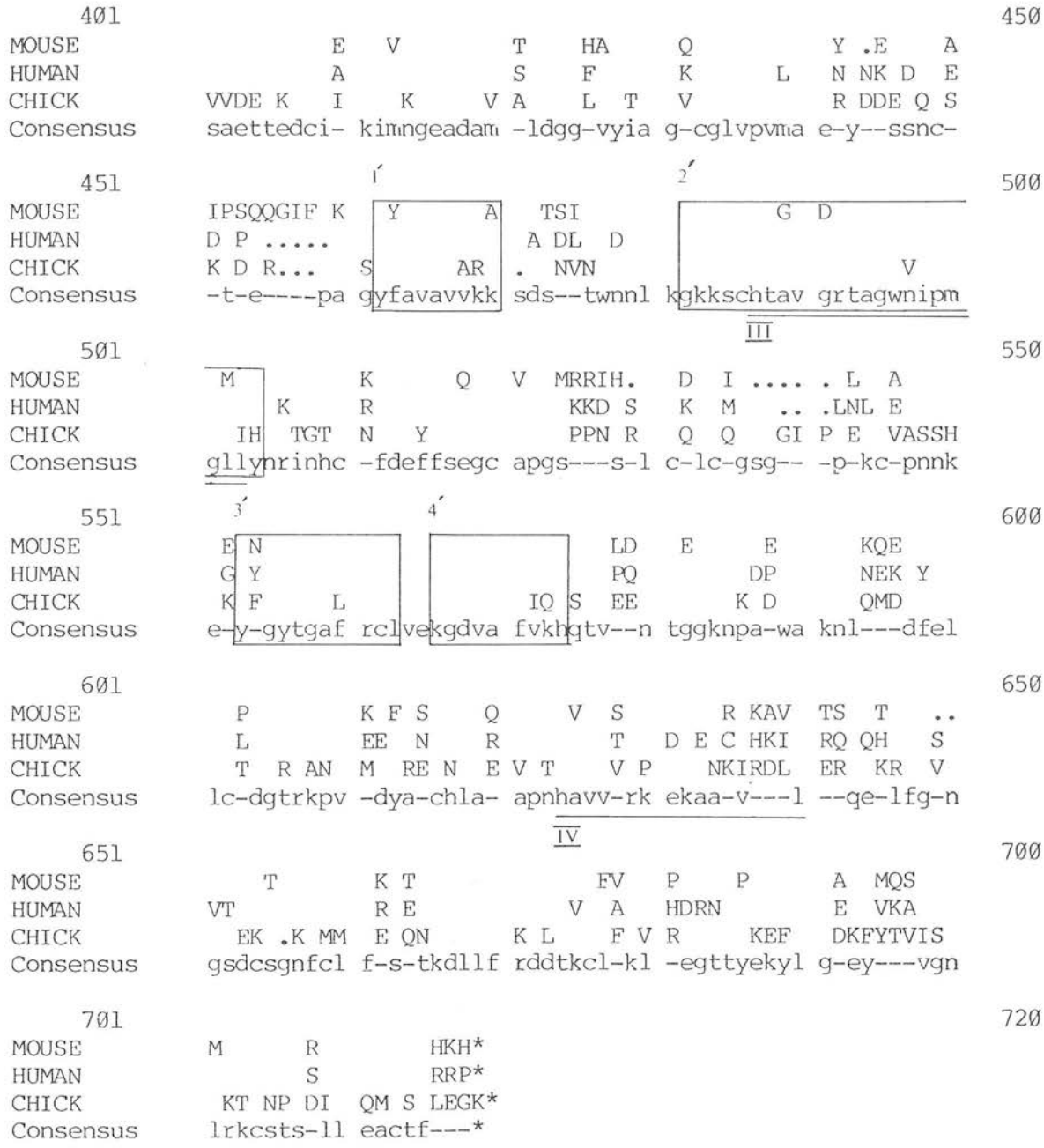
```
    901                                                          950
MOUSE                G                    C           C    G  T
HUMAN                              AGT          CC         AT   CA   A
CHICK              T A    CTTA     G   CC  T      G C G    A     A
Consensus     tgttcagaga tgacaccaaa tgtttggtta aacttc-aga -ggaaccaca

    951                                                          1000
MOUSE         CC                  C G  G  CA G C AT    C       A
HUMAN                               A      G C   G
CHICK           CA GG G  T C T      TA   T TAT  CT TGA  T CC     C A
Consensus     tatgaaaaat acttaggaga -gaatat-t- aa-gctgttg gtaacctgag

    1001                                                         1050
MOUSE         G  G     A       CG                              AC    A
HUMAN           A              TC                              G  G  C
CHICK           CC   AA  C A   GATA     CC GAT G    GC     T G GGGCA
Consensus     aaa-tgctcc acctca--ac tcctggaagc ctgcactttc c-taaac-tt

    1051                                                         1100
MOUSE               CA       .   T       TGT        AC .. .......
HUMAN               TC                   C       ........ .......
CHICK          GT AAGGGA G  A     C CTTCTG     G    G  AC TTCGCGCC
Consensus     aaaatc--ag aggtagggct gccac-aagg tggag-aa.. .......gat

    1101                                                         1150
MOUSE          CTC    C     G  ....       C  C T T    C     .  C  TG
HUMAN          GGAA G AG A  AT ....     A T  G         .     C     A
CHICK          ACT    T     C  CCCG   C  A...A  A   A G G   GA  G  C
Consensus     g---cctc-c gtg-cc.... catggg-tt- ccctggt-tt ca-tggccc-

    1151                                                         1200
MOUSE            G ..      G      TG GTGTA          T   G   ....CC  A
HUMAN              ..   T          A C T T       A A        ....    G
CHICK         TC  CCA    C    TC  C T C CT  C    TCA       AATG  GCTT
Consensus     agtg..gttg gtgctaacc- -g-c-gtctt cgc-gctctg ....tgtt-c

    1201                                                         1250
MOUSE            CA AC  A  ....    C            CG C G     C
HUMAN              G G G    .... A              TT T AT    T
CHICK           T CGC      CCAC      G        T ......  C  .... C  C
Consensus     cat-t-tact g....agcaa aaaataaaaa --a-t-ctga -tttatattt

    1251                                                         1253
MOUSE          CC
HUMAN           ..
CHICK          ...
Consensus      c..
```

FIGURE 32

Homology Between the Predicted Protein Sequences of Chicken, Human and Mouse Transferrin

The predicted protein sequences of human (Yang et al, 1984), chicken (Jeltsch and Chambon, 1982) and the carboxyl half of mouse transferrin (Figure 29) were compared. Dots indicate gaps introduced to maximize homology. Capital letters indicate differences to the consensus sequence, which is depicted either by small case letters or dashes, where there is no homology. The start of the mature protein and COOH domain are double underlined. Numbered boxes enclose highly conserved duplicated protein domains within the transferrins. The underlined sequences show the location of putative ancient quadruplication domains, as described in Jeltsch and Chambon, (1982). The consensus sequence is numbered at either end of each line, position 1 indicates the start of the signal peptides.


1
50
MOUSE
HUMAN R AVGAL V CAVLGL L V D T . A V EH AT Q SF HMKSVI
CHICK K ILCTV S LGIAAV F A P S I T I SP EK N NL ....LT
Consensus
51
100
MOUSE
HUMAN PSDGP VA K S R A V T A L YD Y NN V
CHICK QQERI LT Q T K N I S G Q FE G YK I
Consensus
101
150
MOUSE
HUMAN V F GSK D PQ F DSG QM Q R K
CHICK A I EHT G ST S GTE TV D Q N
Consensus
151
200
MOUSE
HUMAN . YCDL P PR....KP L K N G A C .DG TDFPQL QLC
CHICK T HWGA I WEGIESGS V Q K A V G TIE QKLCRQ KGD
Consensus
201
250
MOUSE
HUMAN GCG STLNQ F K A S IF L ANKADR Q
CHICK KTK ARNAP S H K T VN . .APDLN E
Consensus
251
300
MOUSE
HUMAN NT K E D HL Q PS T SMGGK L EL NQ EH
CHICK GS Q N T NW R AA A DDNKV . SF SK SD
Consensus
301
350
MOUSE
HUMAN K KSKE Q .....SS H G L HGFLK PR AKM Y
CHICK V TKSD H GPPGKKD V L F IMLKR SL SQL F
Consensus
351
400
MOUSE
HUMAN VT RNL EG TCPEA TDEC KPVK LSH H RL N V V
CHICK YS QSM .. KDQLT SPRE NRIQ VGK D KS R V N DV T
Consensus


401
450
MOUSE E V T HA Q Y .E A
HUMAN A S F K L N NK D E
CHICK VVDE K I K V A L T V R DDE Q S
Consensus saettedci- kimngeadam -ldgg-vyia g-cglvpvma e-y--ssnc-
451
500
1′
2′
MOUSE IPSQQGIF K Y A TSI G D
HUMAN D P ..... A DL D
CHICK K D R... S AR . NVN V
Consensus -t-e----pa gyfavavvkk sds--twnnl kgkkschtav grtagwnipm
III
501
550
MOUSE M K Q V MRRIH. D I .... . L A
HUMAN K R KKD S K M .. .LNL E
CHICK IH TGT N Y PPN R Q Q GI P E VASSH
Consensus gllynrinhc -fdeffsegc apgs---s-l c-lc-gsg-- -p-kc-pnnk
3′
4′
551
600
MOUSE E N LD E E KQE
HUMAN G Y PQ DP NEK Y
CHICK K F L IQ S EE K D QMD
Consensus e-y-gytgaf rclvekgdva fvkhqtv--n tggknpa-wa knl---dfel
601
650
MOUSE P K F S Q V S R KAV TS T ..
HUMAN L EE N R T D E C HKI RQ QH S
CHICK T R AN M RE N E V T V P NKIRDL ER KR V
Consensus lc-dgtrkpv -dya-chla- apnhavv-rk ekaa-v---l --qe-lfg-n
IV
651
700
MOUSE T K T FV P P A MQS
HUMAN VT R E V A HDRN E VKA
CHICK EK .K MM E QN K L F V R KEF DKFYTVIS
Consensus gsdcsgnfcl f-s-tkdllf rddtkcl-kl -egttyekyl g-ey---vgn
701
720
MOUSE M R HKH*
HUMAN S RRP*
CHICK KT NP DI QM S LEGK*
Consensus lrkcsts-ll eactf---*

The relatively shorter boxed regions of very high amino acid conservation must represent sequences against which selection may act to preserve structure, both between domains and species. One such set of selectable sequences would be the parts of the protein responsible for iron binding. Spectroscopic studies have provided compelling evidence that the transferrin iron binding ligands involve at least two tyrosyl groups (the only amino acid with a phenolic hydroxyl group) and one nitrogen ligand. The nitrogen ligand appears to be donated by a histidine residue, although other potential ligands cannot be excluded eg. tryptophan and possibly phenylalanine (Macgillivray et al, (1983), and for review see, Aisen and Brown, (1975)). Each of the boxed regions contains at least one of the above amino acid types in all species and protein domains (Figure 33), making a total of three conserved tyrosines (y), two conserved histidines (h) and one each of the potential ligand donating amino acids, tryptophan (w) and phenylalanine (f). Furthermore, the boxed regions within both human and chicken sequences are brought into the same close conformational relationship within the folded proteins by disulphide bridges between the conserved cystine residues in regions 2 and 3, Figure 33 (Bridge 3, Figures 4-6, Williams et al, (1982)). Thus the available evidence strongly implicates the boxed regions within the five domains in the binding of iron in serum transferrins. However this does not exclude the possibility of additional iron interactions with other amino acids outside the boxed regions.

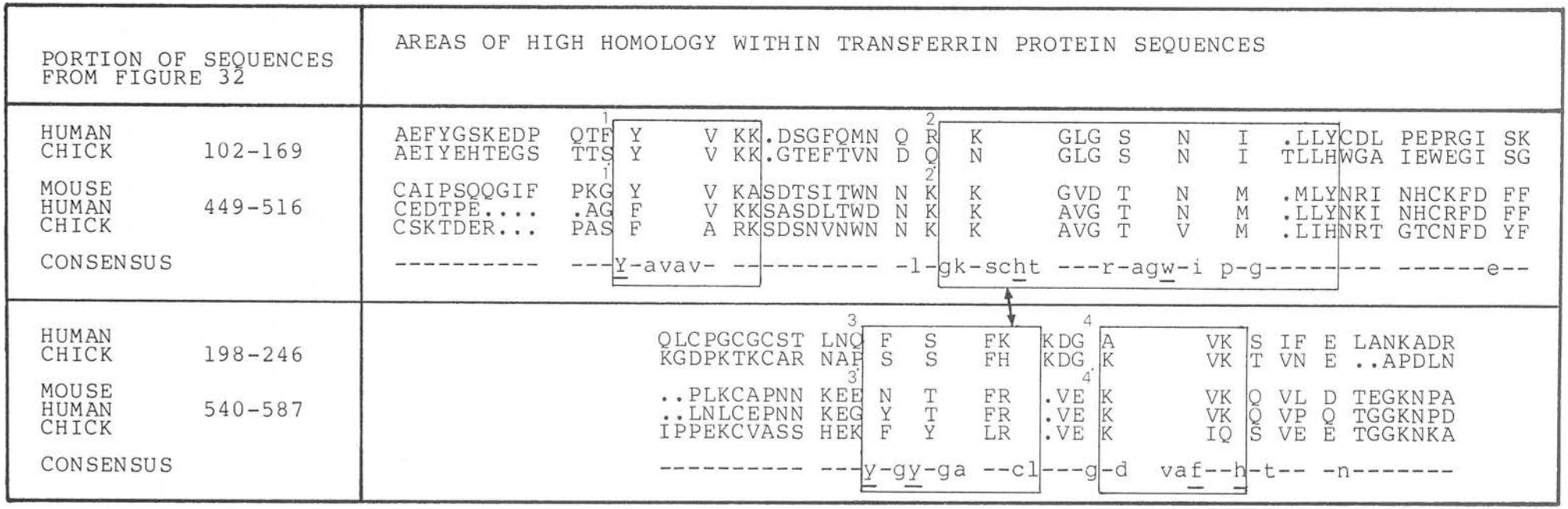
Transferrin Quadruplication of a Primordial Gene.

It is possible that the serum transferrin may have evolved by quadruplication on the basis of a weak fourfold homology present in the human protein sequence (MacGillivray, Mendez and Brew, 1977). Comparison of the ovotransferrin sequence gave further support to the quadruplication hypothesis (Jeltsch and Chambon, 1982).

FIGURE 33

Comparison of Areas of Greatest Protein Sequence Conservation Between the Amino and Carboxyl Halves of Chicken, Human and Mouse Transferrins

The four areas of high protein sequence homology from each domain of each sequence 1-4 and 1'-4' and their immediate flanking regions (Figure 32), were compared. Amino acids present in all five domains, thought to be involved in iron binding/ligand formation (MacGillivray et al, 1983) are underlined. Other putative ligand donating residues are underscored with a cross. The cystine residues which form the disulphide bridges, joining the boxed regions 2 to 3 and 2' to 3' are indicated by the arrows (Williams et al, 1982).


PORTION OF SEQUENCES FROM FIGURE 32
AREAS OF HIGH HOMOLOGY WITHIN TRANSFERRIN PROTEIN SEQUENCES
HUMAN
CHICK
102-169
MOUSE
HUMAN
CHICK
449-516
CONSENSUS
HUMAN
CHICK
198-246
MOUSE
HUMAN
CHICK
540-587
CONSENSUS

Within ovotransferrins are an additional 4 sub-domains (I-IV), of which I and III correspond to the previously described boxed regions 2 and 2' (Figure 32). Comparisons between the four regions I/II, I/IV and II/III revealed 44% homology (Jeltsch and Chambon, 1982), which is significantly higher ($P<0.001$) than the value expected for 51 nucleotide blocks taken at random. Similar comparison of regions II/III and III/IV between the regions of chicken and mouse transferrins (Figure 31 and Jeltsch and Chambon 1982) revealed 39% sequence homologies which is still significantly higher ($P<0.02$) than the expected homology of equivalent random sequences. It therefore seems probable that the ancestral domain that was duplicated and gave rise to the present transferrins, was itself a product of a primordial duplicated gene. It is to be hoped that the elucidation of the exonic organization of transferrin genes will provide conclusive evidence for the quadruplication hypothesis, as it has for the triplicated primordial gene event of the alpha-fetoprotein and serum albumin ancestral gene (Eiferman et al, 1981). If the exon organization does not validate the quadruplication hypothesis then it should at least contribute to the delineation of the promordial gene structure. (see, eg. Stone, Rothblum and Schwartz, 1985).

Within the putative primordial domains only two amino acids are conserved in all sequence comparisons; a histidine residue is present at the start of each region (I - IV) and an arginine occurs five amino acids down stream (Figure 32). Either of these amino acids, including those of regions II and IV could possibly act as nitrogen ligands, to which, the observation that regions I and III are within boxes 2 and 2' respectively which do not contain tyrosine residues may be pertinent. The tertiary structure of transferrins is not known and thus the special relationship of regions II and IV to the boxed regions in the folded protein is unknown. However, a close structural relationship within the native protein may not be necessary for some ligand-forming residues as transferrins have been

reported to undergo conformational changes upon iron binding (for review see, Aisen and Brown, (1975)). Indeed in postulating mobile amino acids that interact with iron in the process of binding; the apparent conflict between the observed conformational changes and the relative rigidity between the highly conserved boxed regions imposed by disulphide linkages (Williams et al, 1982) would be resolved. It must be emphasised at this juncture that any suggestion that the conserved histidine or arginine of the primordial domain regions II and IV are involved with the binding of iron to the boxed regions within the amino and carboxyl domains of transferrins, is based on circumstantial evidence only.

Sequence Divergence of Potential Iron Binding Regions

Previous analysis of the replacement site divergence in the COOH protein domains between the sequences of mouse, human and chicken transferrins revealed nearly twice the divergence found between the equivalent comparisons of globin and preproinsulin genes. It was thus suggested that either the transferrins had a weaker structure-function relationship or that the functional element was localised within a protein structure which was under less rigorous environmental selection. The comparison and analysis of the predicted protein sequence favours the latter interpretation. The most strongly conserved regions include the boxed sequences (1 - 4), certain cystine residues involved in disulphide linkages and at least two amino acid positions (Histidine and Arginine) within regions I - IV (Figures 31, 32; Williams et al, (1982) and Jeltsch and Chambon, (1982)). The remainder of the mature protein sequence is subject to a less stringent conservation of secondary structure (MacGillivray et al, (1982) and Williams et al, (1982)), with the possible exception of the first 18 NH_2 terminal amino acids sequence, and to a lesser extent their equivalent positions in the COOH domain. The beginnings of these domains

exhibit primary, but not secondary structure homologies to the proteins p 97 (Brown et al, 1982) and Lambda Ch Blym -1 (Goubin et al, 1983). It was therefore considered appropriate to determine the sequence divergence between the three species for what are possibly the functional regions of these sequences (Figure 31, Boxed regions). The comparisons between the mouse/human, mouse/chicken and human/chicken sequences revealed corrected percentage divergences of 55.1%, 124% and 66.4% for silent sites, which are very similar to the values of the 3' untranslated regions of 50.5%, 92.2% and 75.5% respectively, which are in good agreement with other similar different gene comparisons (Perler et al, 1980). The same three above described sequence comparisons showed 5.8%, 16.5% and 11.6% replacement site divergence respectively within the boxed regions. This compares with replacement site divergence levels of 18%, 35.5% and 37.5% over the coding portion of the COOH domain of the mouse/human, mouse/chicken and human/chicken sequences. Therefore there has been approximately 3 fold greater conservation of replacement sites within the boxed regions than there has been over the remaining COOH domain sequences (Figure 31). The percentage change and pattern of sequence divergence within the COOH protein domain of transferrin sequences is almost identical to the percentage change and pattern of silent and replacement changes of the rat, human and chicken preproinsulin genes. Particularly notable is the greater divergence (4-7 fold) within the linking peptide C, than there is between the A and B peptides which comprise the functional insulin, which is analogous to the higher divergence (3 fold) of the remaining COOH terminal sequence from the boxed regions in transferrins. The major difference between this otherwise analogous situation being that the C peptide is excised from the active protein (Perler et al, 1980) and is therefore probably under less constraint than the unboxed regions of the mature transferrin proteins.

Transferrin Transcription in the Liver

An unexpected observation regarding transferrin transcription was made when Northern blots of E.R. poly (A) mRNA were probed with the transferrin clone pUC TRF 32, where it would appear that the concentration of mRNA in female liver E.R. is approximately 20 times higher than the male level (Figure 34). The quality of the male and female E.R. poly (A) mRNA preparation was checked by hybridization with another major liver secretory protein probe MUP 11. The MUP probe hybridized to the male and female mRNAs with the expected ratio of approximately 5 to 1, as has been previously reported for adult mice (Hastie et al, (1979); Clissold et al (1981); Shaw et al, (1983) and Clissold et al, (1984)). The veracity of the results obtained with pooled E.R. poly (A) mRNA at 12 weeks of age was checked by dot-blot hybridization of individual total mRNA preparations of male and female sibling offspring from 6.5 - 7.5 weeks of age (Figure 35). It would appear that the age range 6.5 - 7.5 weeks represents a period of considerable change in the relative transferrin mRNA concentrations within male and female liver. The level of transferrin mRNA in male liver doubles from 6.5 - 7.0 weeks and remains similar from 7.0 - 7.5 weeks of age. The concentration of transferrin message in female mice is approximately half the male value at 6.5 weeks. It then increases four fold between 6.5 and 7.0 weeks, achieving parity with the male concentration and increases a further 2 fold from 7.0 - 7.5 weeks becoming twice the equivalent male level (Figure 35). If this rate of increase were sustained then it would account for the apparently large difference observed between the sexes at 12 weeks of age (Figure 34).

The variant serum protein transferrin types Trf^{a}, Trf^{b} and Trf^{b} modified, have been used to investigate the genetic control and linkage of the transferrin locus (Cohen, (1960); Shreffler, (1960); Shreffler, (1963) and Klein, Roop and Roop, (1966)). Mouse serum transferrins

FIGURE 34

Northern Blot of Male and Female Mouse Liver E.R. Poly (A) mRNA Probed for Transferrin Sequences

Male and female mice 12 weeks of age were sacrificed and the liver E.R. poly (A) mRNA fractions isolated. Fractions of the female (2 µg; lane 1) and male preparation (2 µg; lane 2) were subject to electrophoresis under denaturing conditions, transferred to nitrocellulose and probed with the nick translated transferrin cDNA plasmid pUC TRF 32. The final wash was 0.2 x SET at 68°C and the filter was exposed for 4 days.

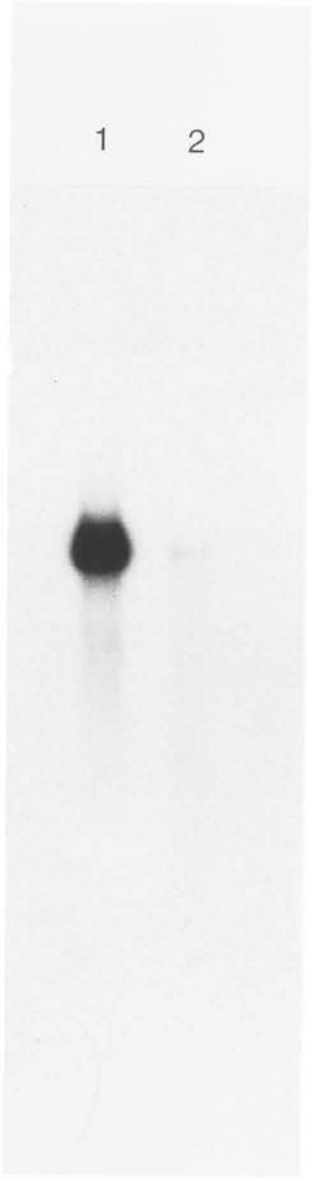
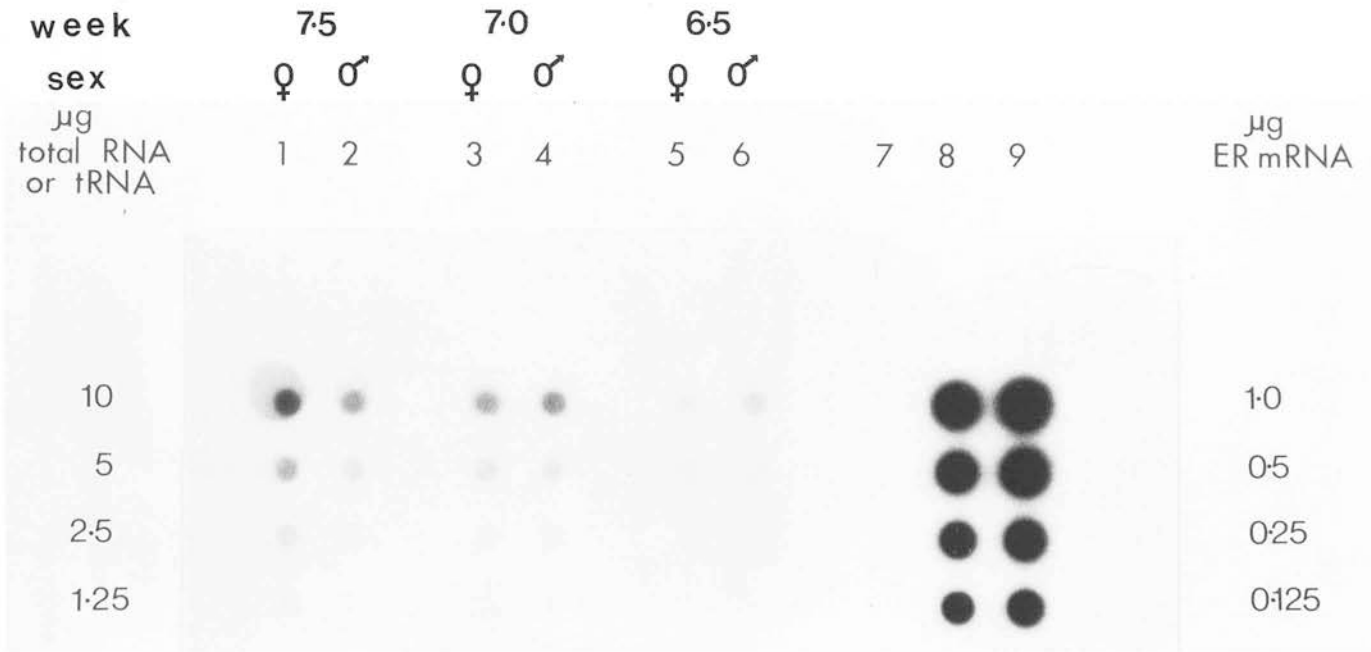

1
2

FIGURE 35

Dot Blot Analysis of Transferrin Sequences in Male and Female Liver RNA

Sequential two fold dilutions of total RNA preparations were applied to nitrocellulose membrane as described in the Experimental Procedures. The RNA was hybridized to a nick translated transferrin cDNA probe pUC TRF 32. The final wash was 1 x SET at 68 ^{o}C and exposure was for 7 days. The total RNA preparations were from individual 7.5 week (row 1), 7.0 week (row 3) and 6.5 week (row 5) females are from 7.5 week (row 2), 7.0 week (row 4) and 6.5 week (row 6) males. The individuals of each age group were siblings of different matings. Row 7 contained Guinea pig tRNA (gift of Dr. P.M. Clissold), whereas rows 8 and 9 contained pooled female liver E.R. poly (A) mRNA preparations. The amount of RNA applied to each spot is indicated on either side of the figure.


week
7·5
7·0
6·5
sex
♀ ♂ ♀ ♂ ♀ ♂
µg
total RNA
or tRNA
1 2 3 4 5 6 7 8 9
µg
ER mRNA
10
5
2·5
1·25
1·0
0·5
0·25
0·125

were reported to exhibit a number of qualitative and quantitative changes from the first to the fourth week after birth reverting to the adult pattern (also present at birth) by the sixth to eighth week (Shreffler, 1960 and 1963). It is therefore possible that the increasing mRNA levels (Figure 35), reflect on earlier disruption and subsequent recovery of transferrin synthesis, although no significant quantitative differences between animals of similar age and sex were encountered by Cohen, (1960) with animals at six weeks of age. Furthermore the analysis of serum transferrin types from individual male and female F2 and Backcross progeny (>8 weeks) revealed no significant protein differences between the sexes (Shreffler, 1960).

The developmental expression of transferrin mRNA was studied in male rats, wherein it was found that transferrin mRNA represents a nearly constant proportion of the total liver RNA from birth to 16 months of age (Levin et al, 1984). In contrast to the above a stimulation of transferrin mRNA synthesis in chicken liver by estrogen or dietary iron deficiency has been found (McKnight et al, 1980a, 1980b) and transferrin (conalbumin) mRNA synthesis induction in the chick oviduct by estrogen or progesterone administration has been well documented (see, eg. Palmiter et al, 1981).

A tentative explanation of the preliminary results presented here (Figure 34 and 35) is that there is some steroid hormonal stimulation of transferrin mRNA expression in the adult female BALB/c mouse liver, although other factors, possible pretranslational or degradative, act to maintain equivalent serum transferrin levels in either gender. Clearly a detailed study of transferrin gene expression is required to verify the apparently large sex and development dependant differences of transferrin mRNA expression in BALB/c mice and to determine (if verified) its basis and relationship to the serum protein levels.

OTHER LIVER SEQUENCES

In addition to the cDNA clones which demonstrate strong hybridization to the transferrin probe LVA 321 (Figure 14), several much weaker hybridizations were noted (Figure 15). It was considered possible that some of these cDNAs could represent the NH_2 domain of mouse transferrins.

Three of the largest weakly hybridizing cDNA clones, pUC T37, pUC T42 and pUC T57, were selected for preliminary sequencing the strategy for which is shown in Figure 39. It was assumed that any homology to mouse transferrin would be apparent from the preliminary sequence. It was therefore quite unexpected when no significant sequence homologies were found when the sequences (Figure 36, 37 and 38) were compared with the transferrin sequences (Figure 31 and 32). The inference is that the initial hybridizations reflect serendipitous homologies between other liver cDNAs and the LVA 321 probe which are not easily detectable in computer assisted sequence comparisons.

The preliminary sequences of pUC T37, 42 and 57 were compared with the data bases as described in the Methods section. Although these sequences represent mostly single strand analyses with poor overlaps (Figure 39) homologous sequences are easily detected. The sequence of pUC T57 (Figure 38) does not contain any long open reading frames although it contains a long poly (A) tail at the 3' end and a normal polyadenylation signal 24 nucleotides further upstream. It is therfore probable that pUC T57 represents a long 3' untranslated region of an as yet unknown mouse protein. A search of the current literature failed to identify any homolgous sequences.

FIGURE 36

Preliminary Sequence of the cDNA Cloned in the Plasmid pUC T37

Nucleotides are numbered to the left of each line, one, refers to the most 5' nucleotide of the insert DNA. There are no long open reading frames in the sequence shown. The double underlined sequences AAAAA$_n$ indicated a stretch of poly (A) residues approximately 50 nucleotides in length. The underlined sequence was determined from gels at the limits of their resolution and therefore require verification by additional sequencing. The position and approximate length of an area of the insert that was not sequenced is indicated by a row of x's. A computer aided search revealed identical match between the complementary strand of the cDNA cloned in pUC T37 (nucleotides 403-180)to the most 5' region of a partial mouse contrapsin cDNA clone (Hill et al, 1984).

```
1    TTCTACTCACAATTCTAGAATTTNCAGTTAGCATTAATTCAAGCYTACGTATTCACCCTC
61   CTAGTAAGCYTATATCTACATGATAATACATAAAAA(n)GGATTACCCTATTGCTGAACAAT
121  TCTAACAATTCTCATTCTAGTCTCTGTTTAACTGCTTTCCACTCTATCTTCCCTTGTTCA
181  GGGAGGATGAGCAGGGCGCTGGCATTTCCTGTGTACTTCAGCTCCAACACAGAGCACGAT
241  AGCTCCTCATCACGGAAGTGGCGTGTGGTCAGTAACTTCATTTTCATCATGGGAACCTTC
301  ACAGATCTCTTCTCATCCAAGTAGAACTCAGACTCAAATGTGTCCTGGGGGTCAAAGGAT
361  ATCTTCCATTTGCCTTTAAAGTAGATGTAATTCACCAGCACCATCAATGTCCTCTCATCC
421  AGTTCTGAGATGAGTTCCTTGATCATCCCCTGGGTCTGATTGCTCACAGAGTCATTGATG
481  AGGTTTTTGGCCTCAGTAGGCTGCTGGAAGTCTGCTGTGAAGGCCTCAGTCTGGTACAGA
541  GCCCTTGTCTTCTCATGGAATTCXXXXXXXXXXCTGCAGGACCTTTTCAATAAACATGGC
601  ATTGCTATGTTTATCTGATCCTGGTCTTCTGGCTGGCTGAGACTCTGTAGGAGGTTGCCA
661  AAGCCCTGGTGGATGTCTGCTTCAGGGGTCTCTGTGAGATTGAACTTGAGGCCTTCTAGA
721  ATCTCTTCCATGGTCTTGCCCTTTGCTCCAGGGACACAAGGGCCAAGGCAGCTGAGATGC
781  TAAGTGGGGAGAAGACAATATTTGTATCTGGATTCTTCAAAGCCAGCTTCTTGTACAGGC
841  TGAAGGCAAAGTCAGTGTTGACGGAGGCCAATGTGAGACTGTCATCTTGTGTCCCATTGT
901  CTTGGTGTTCATGGAATACAATGTCCATTTCCTTTGTGCCATCTGGTAAGCATAGGACAG
961  CAGGACAGATTCCAGCCATTAAGATCATCC  990
```

FIGURE 37

Preliminary Sequence of the cDNA Cloned in the Plasmid pUC T42

Nucleotides are numbered to the left of each line, one, refers to the most 5' nucleotide of the insert DNA. The underlined sequence was determined from gels at the limits of their resolution and therefore require verification by additional sequencing. There is an open reading frame between nucleotides 1 and the stop codon TGA (Boxed) at 742, which is interrupted only once by a stop codon (Boxed) within the underlined region. The 3' polyadenylation signal AATAAA is double underlined and the extent of the poly(A) tail is indicated at the end of the sequence. Ambiguity codon nomenclature is the same as that supported by the U.W.G.C.G. A computer aided search revealed an identical match between nucleotides 740-947 and the 3' end of the third component of mouse complement, the C3 gene (Wiebauer et al, 1982).

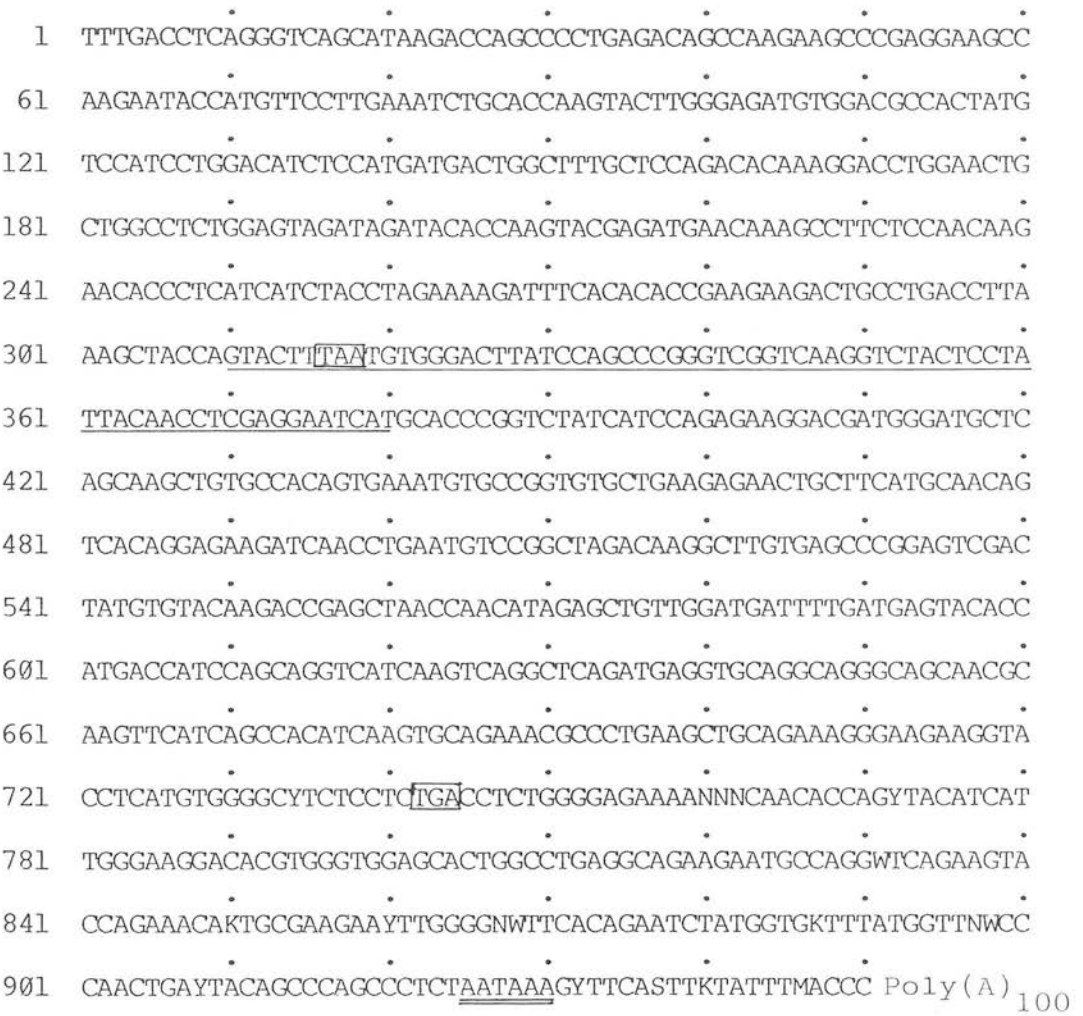

1 TTTGACCTCAGGGTCAGCATAAGACCAGCCCCTGAGACAGCCAAGAAGCCCGAGGAAGCC
61 AAGAATACCATGTTCCTTGAAATCTGCACCAAGTACTTGGGAGATGTGGACGCCACTATG
121 TCCATCCTGGACATCTCCATGATGACTGGCTTTGCTCCAGACACAAAGGACCTGGAACTG
181 CTGGCCTCTGGAGTAGATAGATACACCAAGTACGAGATGAACAAAGCCTTCTCCAACAAG
241 AACACCCTCATCATCTACCTAGAAAAGATTTCACACACCGAAGAAGACTGCCTGACCTTA
301 AAGCTACCAGTACTTTAAATGTGGGACTTATCCAGCCCGGGTCGGTCAAGGTCTACTCCTA
361 TTACAACCTCGAGGAATCATGCACCCGGTCTATCATCCAGAGAAGGACGATGGGATGCTC
421 AGCAAGCTGTGCCACAGTGAAATGTGCCGGTGTGCTGAAGAGAACTGCTTCATGCAACAG
481 TCACAGGAGAAGATCAACCTGAATGTCCGGCTAGACAAGGCTTGTGAGCCCGGAGTCGAC
541 TATGTGTACAAGACCGAGCTAACCAACATAGAGCTGTTGGATGATTTTGATGAGTACACC
601 ATGACCATCCAGCAGGTCATCAAGTCAGGCTCAGATGAGGTGCAGGCAGGGCAGCAACGC
661 AAGTTCATCAGCCACATCAAGTGCAGAAACGCCCTGAAGCTGCAGAAAGGGAAGAAGGTA
721 CCTCATGTGGGGCYTCTCCTCTGACCTCTGGGGAGAAAANNNCAACACCAGYTACATCAT
781 TGGGAAGGACACGTGGGTGGAGCACTGGCCTGAGGCAGAAGAATGCCAGGWTCAGAAGTA
841 CCAGAAACAKTGCGAAGAAYTTGGGGNWTTCACAGAATCTATGGTGKTTTATGGTTNWCC
901 CAACTGAYTACAGCCCAGCCCTCTAATAAAGYTTCASTTKTATTTMACCC Poly(A)100

FIGURE 38

Preliminary Sequence of the cDNA Cloned in the Plasmid pUC T57

Nucleotides are numbered to the left of each line, one, refers to the most 5' nucleotide of the insert DNA. The underlined sequence was determined from gels at the limits of their resolution and therefore require verification by additional sequencing. The 3' polyadenylation signal AATAAA is doubled underlined and the extent of the poly (A) tail is indicated at the end of the sequence. There are no long open reading frames in this sequence. Ambiguity codon nomenclature is the same as that supported by the U.W.G.C.G.

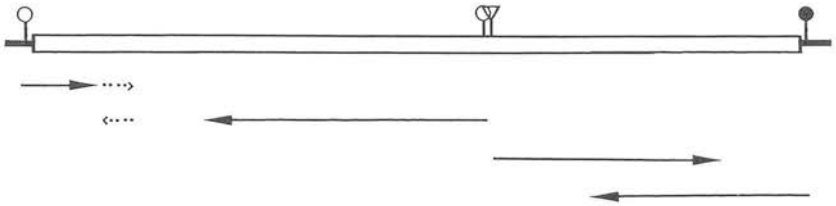
```
1    TGTCGACTGTTGAGAAATTCAGTTTGGGTCTGGAGGGAGCGCAGTAAGCAGCAGAGGCCC
61   ACAAGGCGGCAACTGTGAGCTTTCCGGTACCGCATAGATAACTGGAGATTGATGTCCGGA
121  GATTTGTCAAGATGCTTGGGAAAAGGAAGCTGTCCCCCAGGGCCCGGTCCCTGAGGGTGT
181  GATCCGAATCTACAGCATGAGGTTCTGCCCCTANTCGCANAGGGCAGCTGGTTCTAAGGC
241  TAAAGGATNAGGATGAAGTGATTAATATTAACCCTGAATAAGGAAGCTGATGGTAATATA
301  NTGGATCCTTTGGCCAATTCCTNTCTTRGAGACCATGTCAGCTGGTCTATGAATCTGTCA
361  TTGCTTGTGAGTACCTGGAGTTAGCTCTACCCGGAAGAAGACTGTTTCCGTATGACCCGT
421  ATGAACGAGCTCGCCAGAAGATGTTATTGGAGCTATTCTGTAAGGTCCGCCTTTAAGCAA
481  GGAATGTCTGATAGCGCTGAGATGCGGAAGAGACTGTACGGATCTGAAGGTCGCCCTGCG
541  TCAGGAGTTGTGCAACAGGAAGAGATTCTTGAATATCAGAACACTACCTTCTTCGGCGGA
601  GACTGTATATCCATGATTGATTACCACGTCTGGCCCTGGTTTGAGCGCCTGGACGTATAT
661  GGACTGGCTGACTGCGTGAATCACACCCCGATGCTGCGGCTCTGGATAGCCTCCATGAAG
721  CAGGACCCTGCAGTGTGTGCTCTCATGATAAGAGCGTCTTCCTGGGCTTCTTGAATCTCT
781  ATTTCCAAAACAACCCTTGTGCCTTTGATTTTGGGCTGTGTAACCCAATCATACGATAAT
841  CCAGGCATTGCCGGGACTCTCGGTCATTCTGATGTCTGCATCACGGGTCATCTTGGGGTT
901  CCGTGTATTTGTTCTTTTTTTTTTGAGGTCCTAATAAATATGGATGTGTAAAATAT POLY(A)49
```

FIGURE 39

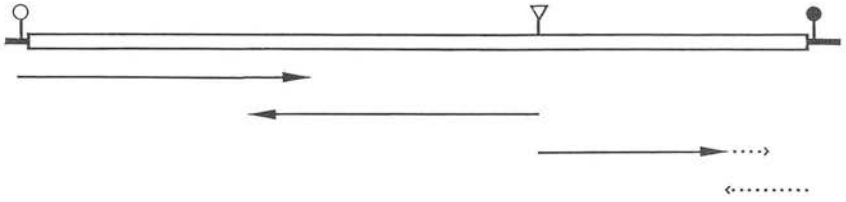
Preliminary Sequence-Analysis Strategy for Some of the Clones which Hybridized Weakly to the LVA 321 Fragment Probe

The sequences are schematically represented by open boxes. The flanking pUC 8 sequence is represented by a broad line. Horizontal arrows indicate the extent and polarity of sequence data obtained from various restriction fragments subcloned into M13 mp 8/9 or tg 130/131. Dotted horizontal arrows indicate stretches of poly (A) which prevented subsequent sequence determination. Vertical lines and symbols refer to the location of restriction enzyme sites, Eco RI (⚲), Kpn I (⏉), Pst I (▽) and Bam HI (🌳).

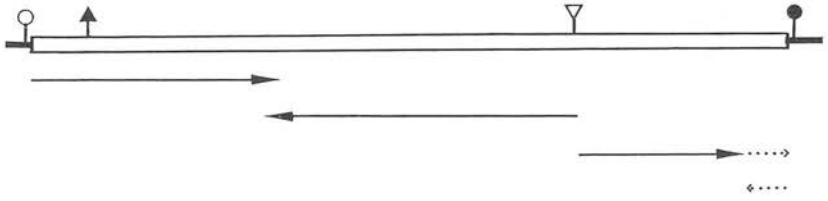
pUCT37



pUCT42



pUCT57



The sequence of clone pUC T42 (Figure 37) could not be accurately determined because of the weak overlap of the central cDNA portions and the tendency of runs of homolgous nucleotides near the 3' end to cause multiple terminations in all tracks. The sequence encodes one long open reading frame which is interrupted by a stop codon in the region of poor overlap. At the 3' end there is a long poly (A) tail preceded by a polyadenylation signal 26 nucleotides upstream.

A search of the current literature revealed an identical match between positions 740-947 (accepting ambiguity codons) to the 200 Bp cDNA clone of the 3' end of the third component of the mouse complement genes, C3 (Wiebauer et al, 1982). The cDNA pUC T42 therefore probably encodes an additional 700 base pairs of the murine C3 gene, the correct sequence for which will be determined by more exhaustive sequencing. The total length of the mouse C3 cDNA is approximately 5 Kb in length (Domdey et al, 1982).

The sequencing of pUC T37 presented few difficulties in sequencing (Figure 36) except for the region adjacent to a large stretch of adenine residues near the 5' end of the insert and the omission of a short length of sequencing due to misinterpretation of the arrangement of the two adjacent restriction sites (Figure 39). It would appear that the pUC T37 insert is in fact composed of two cDNAs which presumably became associated during the cDNA/vector (pUC 8) ligation process. The only open reading frame within the sequence is in the complementary strand of pUC T37 and extends through almost the full length of the insert until the breakdown in the resolution of the sequence near the "second" sequence (Figure 36 and 39).

The open reading frame of pUC T37 is shown in pUC T37N (Figure 40). A computer assisted homology search for sequences similar to this sequence revealed 60% homology to the monkey alpha$_I$-antitrypsin sequence. Because alpha$_I$-antitrypsin is a member of the super family of proteins

FIGURE 40

Sequence of the Contrapsin cDNA (T37N)
Cloned in pUC T37

The contrapsin sequence is the complementary DNA strand of the insert in pUC T37. Nucleotides are numbered on the left of each line, where one refers to nucleotide 990, and 812 refers to position 180 of the insert sequence of pUC T37, Figure 36. There was only one continuous open reading frame and the predicted protein sequence is shown under the DNA. Numbers below the amino acids refer to their distance from the start of the predicted sequence, however it is not known where protein translation is initiated on the mRNA. The position and approximate length of an area of the insert that was not sequenced is indicated by a row of x's.

```
                .         .         .         .         .         .
  1  GGATGATCTTAATGGCTGGAATCTGTCCTGCTGTCCTATGCTTACCAGATGGCACAAAGG
       MetIleLeuMetAlaGlyIleCysProAlaValLeuCysLeuProAspGlyThrLysG
         1                         10

                .         .         .         .         .         .
 61  AAATGGACATTGTATTCCATGAACACCAAGACAATGGGACACAAGATGACAGTCTCACAT
     luMetAspIleValPheHisGluHisGlnAspAsnGlyThrGlnAspAspSerLeuThrL
   20                          30

                .         .         .         .         .         .
121  TGGCCTCCGTCAACACTGACTTTGCCTTCAGCCTGTACAAGAAGCTGGCTTTGAAGAATC
     euAlaSerValAsnThrAspPheAlaPheSerLeuTyrLysLysLeuAlaLeuLysAsnP
   40                          50

                .         .         .         .         .         .
181  CAGATACAAATATTGTCTTCTCCCCACTTAGCATCTCAGCTGCCTTGGCCCTTGTGTCCC
     roAspThrAsnIleValPheSerProLeuSerIleSerAlaAlaLeuAlaLeuValSerL
   60                          70

                .         .         .         .         .         .
241  TGGGAGCAAAGGGCAAGACCATGGAAGAGATTCTAGAAGGCCTCAAGTTCAATCTCACAG
     euGlyAlaLysGlyLysThrMetGluGluIleLeuGluGlyLeuLysPheAsnLeuThrG
   80                          90

                .         .         .         .         .         .
301  AGACCCCTGAAGCAGACATCCACCAGGGCTTTGGCAACCTCCTACAGAGTCTCAGCCAGC
     luThrProGluAlaAspIleHisGlnGlyPheGlyAsnLeuLeuGlnSerLeuSerGlnP
  100                         110

                .         .         .         .         .         .
361  CAGAAGACCAGGATCAGATAAACATAGCAATGCCATGTTTATTGAAAAGGTCCTGCAGXX
     roGluAspGlnAspGlnIleAsnIleAlaMetProCysLeuLeuLysArgSerCysXxxX
  120                         130

                .         .         .         .         .         .
421  XXXXXXXXGAATTCCATGAGAAGACAAGGGCTCTGTACCAGACTGAGGCCTTCACAGCAG
     xxXxxXxxGluPheHisGluLysThrArgAlaLeuTyrGlnThrGluAlaPheThrAlaA
  140                         150

                .         .         .         .         .         .
481  ACTTCCAGCAGCCTACTGAGGCCAAAAACCTCATCAATGACTCTGTGAGCAATCAGACCC
     spPheGlnGlnProThrGluAlaLysAsnLeuIleAsnAspSerValSerAsnGlnThrG
  160                         170

                .         .         .         .         .         .
541  AGGGGATGATCAAGGAACTCATCTCAGAACTGGATGAGAGGACATTGATGGTGCTGGTGA
     lnGlyMetIleLysGluLeuIleSerGluLeuAspGluArgThrLeuMetValLeuValA
  180                         190

                .         .         .         .         .         .
601  ATTACATCTACTTTAAAGGCAAATGGAAGATATCCTTTGACCCCCAGGACACATTTGAGT
     snTyrIleTyrPheLysGlyLysTrpLysIleSerPheAspProGlnAspThrPheGluS
  200                         210

                .         .         .         .         .         .
661  CTGAGTTCTACTTGGATGAGAAGAGATCTGTGAAGGTTCCCATGATGAAAATGAAGTTAC
     erGluPheTyrLeuAspGluLysArgSerValLysValProMetMetLysMetLysLeuL
  220                         230

                .         .         .         .         .         .
721  TGACCACACGCCACTTCCGTGATGAGGAGCTATCGTGCTCTGTGTTGGAGCTGAAGTACA
     euThrThrArgHisPheArgAspGluGluLeuSerCysSerValLeuGluLeuLysTyrT
  240                         250

                .         .         .
781  CAGGAAATGCCAGCGCCCTGCTCATCCTCCCT  812
     hrGlyAsnAlaSerAlaLeuLeuIleLeuPro
  260                         270
```

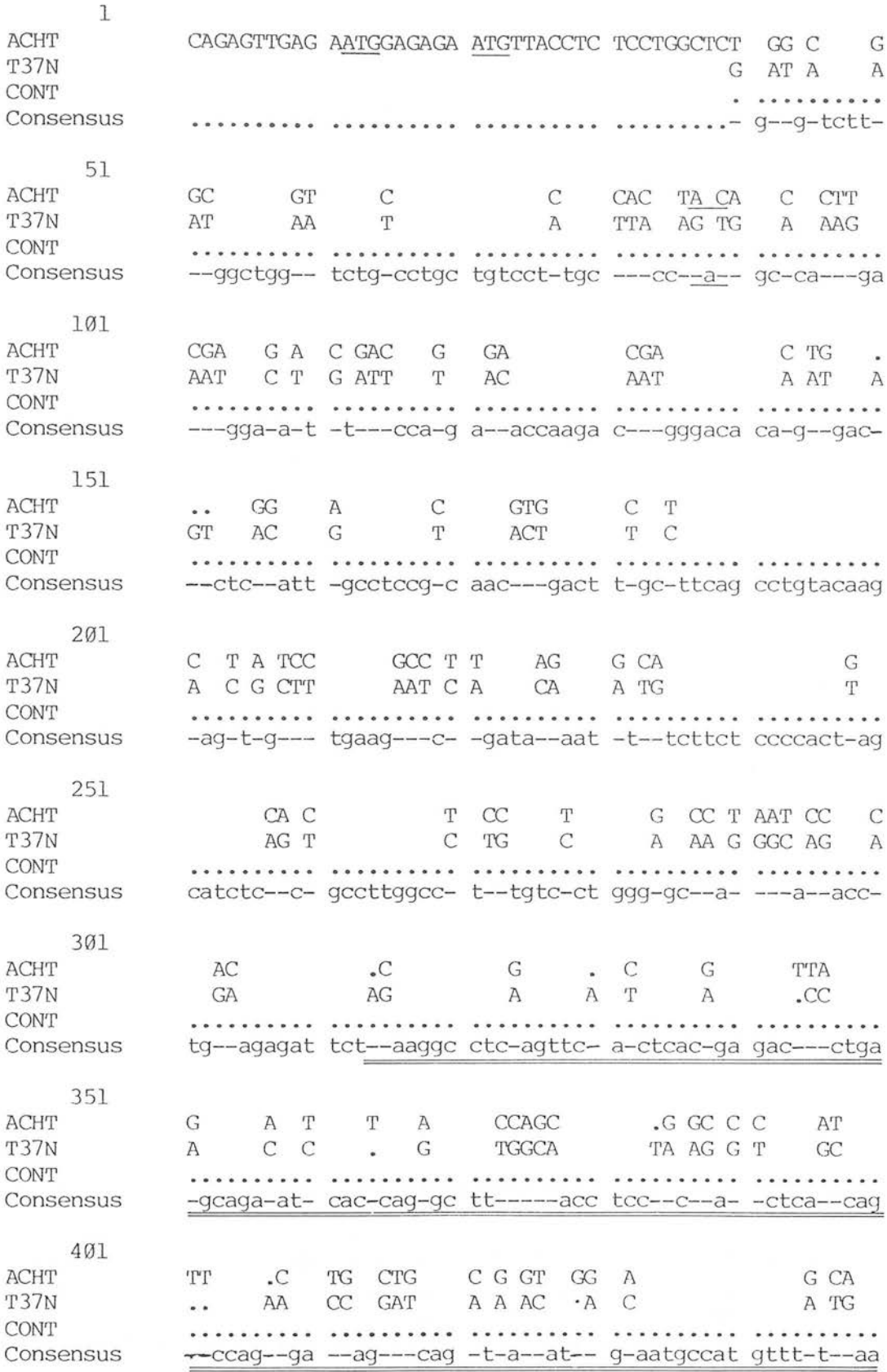
which include the ovalbumin genes serum protease inhibitors (Leicht et al, 1982 and Chandra et al, 1983) members of these families were also compared. The initial homology to alpha$_I$ - antitrypsin suggested that pUC T37N encoded part of the protease inhibitor (Chandra et al, 1983 and Hill et al, 1984).

The sequence encoded by pUC T37N corresponds to the mouse antitrypsin protease inhibitor contrapsin (Hill et al, 1984 and Takahara and Sinohara 1982). The cDNA encodes the 5' half of the contrapsin mRNA and overlaps with the 3' half cDNA of Hill et al, (1984) by an identical 220 Bp (Figure 41). Contrapsin is the equivalent sequence in the mouse to human alpha$_I$-antichymotrypsin (Hill et al, 1984). The reactive centres of these two proteins have diverged considerably, which may account for their differences in protease inhibition specificity. Human alpha$_I$-antichymotrypsin sequence (Chandra et al, 1983) is also compared with pUC T37N (Figure 41 and 42). Two things are immediately apparent from these comparisons, that the reactive site regions are very different, as observed by Hill et al,(1984) and secondly that the alignment for maximal homology between the 5' ends of contrapsin (pUC T37N) and alpha$_I$-antichymotrypsin necessitates the introduction of several single and one double nucleotide gaps (Figure 41). Therefore if the sequences are correct then this represents a novel evolutionary mechanism for generating quantum changes in the protein sequence of diverging genes (Figure 42,position 102-148). To check the sequencing data of pUC T37N for errors both the contrapsin and alpha$_I$-antichymotrypsin sequences were aligned and compared with primate alpha$_I$ antitrypsin (Kurachi et al,(1981) Figure 43) The comparison shows that all the gap differences between alpha$_I$-antichymotrypsin and contrapsin (pUC T37N) except one are due to the alpha$_I$-antichymotrypsin seqence of Chandra et al, (1983). Considerable differences also occur between the human alpha$_I$-antichymotrypsin seqences of Chandra et al, (1983) and Hill et al, (1984) near the 3' untranslated region

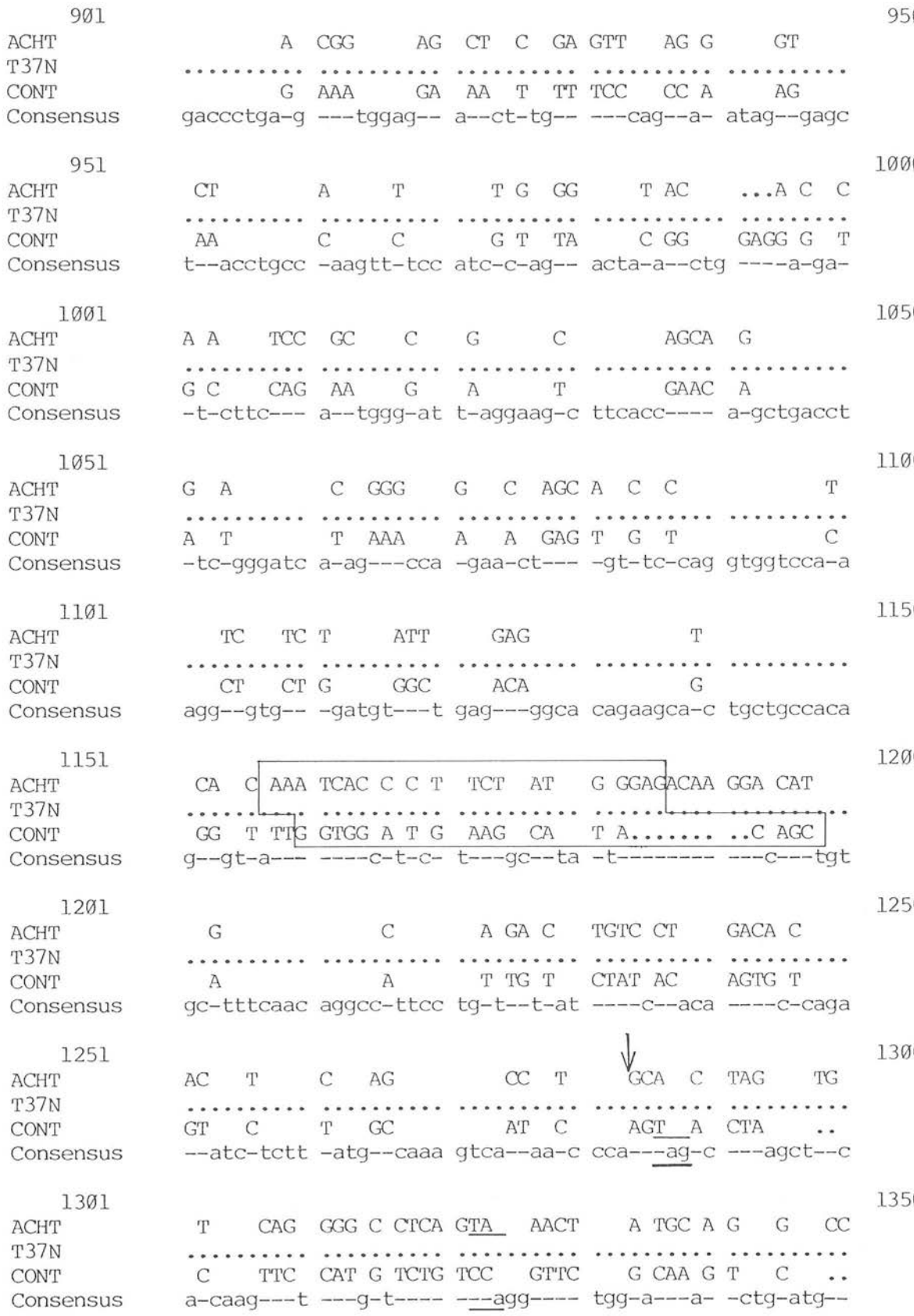
FIGURE 41

Sequence Comparison between Human Alpha $_{I}$-Antichymotrypsin and Murine Contrapsin cDNAs

The sequence of Human alpha $_{I}$-antichymotrypsin (Chandra et al, 1983) was compared with the 5' cDNA contrapsin sequence, T37N (Figure 40) and 3' cDNA contrapsin clone (Hill et al, 1984). Dots indicate gaps introduced to maximize the homology. Capital letters indicate differences to the consensus sequence, which is depicted either by all case letters or dashes, where there is no homology. The consensus sequence is numbered at either end of each line. Position 1 indicates the start of alpha $_{I}$-antichymotrypsin cDNA sequence. The row of x's represents the region between two subclones which was not sequenced. The start of mature alpha $_{I}$-antichymotrypsin (ACHT) and the putative translation initiation and termination codons are underlined. The reactive centre regions of protease inhibitors are boxed (Hill et al, 1984). Sequences within which there could be several translation reading frame shift deletions, are double underlined. An arrow indicates the position at which the sequence of Chandra et al, (1983) comes into variance with the alpha $_{I}$-antichymotrypsin data of Hill et al, (1984).


1
50
ACHT CAGAGTTGAG AATGGAGAGA ATGTTACCTC TCCTGGCTCT GG C G
T37N G AT A A
CONT . .....
Consensus ....- g--g-tctt-
51
100
ACHT GC GT C C CAC TA CA C CTT
T37N AT AA T A TTA AG TG A AAG
CONT ....
Consensus --ggctgg-- tctg-cctgc tgtcct-tgc ---cc--a-- gc-ca---ga
101
150
ACHT CGA G A C GAC G GA CGA C TG .
T37N AAT C T G ATT T AC AAT A AT A
CONT ....
Consensus ---gga-a-t -t---cca-g a--accaaga c---gggaca ca-g--gac-
151
200
ACHT .. GG A C GTG C T
T37N GT AC G T ACT T C
CONT ....
Consensus --ctc--att -gcctccg-c aac---gact t-gc-ttcag cctgtacaag
201
250
ACHT C T A TCC GCC T T AG G CA G
T37N A C G CTT AAT C A CA A TG T
CONT ....
Consensus -ag-t-g--- tgaag---c- -gata--aat -t--tcttct ccccact-ag
251
300
ACHT CA C T CC T G CC T AAT CC C
T37N AG T C TG C A AA G GGC AG A
CONT ....
Consensus catctc--c- gccttggcc- t--tgtc-ct ggg-gc--a- ---a--acc-
301
350
ACHT AC .C G . C G TTA
T37N GA AG A A T A .CC
CONT ....
Consensus tg--agagat tct--aaggc ctc-agttc- a-ctcac-ga gac---ctga
351
400
ACHT G A T T A CCAGC .G GC C C AT
T37N A C C . G TGGCA TA AG G T GC
CONT ....
Consensus -gcaga-at- cac-cag-gc tt-----acc tcc--c--a- -ctca--cag
401
450
ACHT TT .C TG CTG C G GT GG A G CA
T37N .. AA CC GAT A A AC ·A C A TG
CONT ....
Consensus --ccag--ga --ag---cag -t-a--at-- g-aatgccat gttt-t--aa

```
     451                                                                500
ACHT          G  CAA  .    TCTGCTGG ACAGG   AC G   G TG C  A AGG
T37N          A  GTC  G    XXXXXXXX XXGAA   CA T   A GA A  G GCT
CONT          .......... .......... .......... .......... ..........
Consensus     -ag---ct-c ag-------- -----ttc-- -gag-a--c- a-g---ctgt

     501                                                                550
ACHT           TGGCT C         TG C A T     T    G CT AG    CA  T  G
T37N           CCAGA T         CA A G A     C    C GC TA    AG  C  A
CONT          .......... .......... .......... .......... ..........
Consensus     a-----c-ga ggcctt--c- -c-gactt-c ag-a--c--c tg--gc-aa-

     551                                                                600
ACHT             G         C    AC     AG   GGA   TAG      AA    CA
T37N             C         T    CT     GC   CAG   CCA      TG    AG
CONT          .......... .......... .......... .......... ..........
Consensus     aa-ctcatca a-gact--gt ga--aat--- ac---gggga --atca--ga

     601                                                                650
ACHT          T  G   AAG   C CC  CT C CA    A        C
T37N          A  C   TCA   A TG  TG A AG    T
CONT          .......... .......... .......... .G
Consensus     -ct-atc--- ga-c--ga-- -g--gaca-t gatggtgctg gtgaattaca

     651                                                                700
ACHT             T         C        G    GC              A  T  TCA
T37N
CONT
Consensus     tctactttaa aggcaaatgg aagatatcct ttgaccccca ggacacattt

     701                                                                750
ACHT          C    AAG           AG CA    A AG  GG  A T    G
T37N
CONT
Consensus     gagtctgagt tctacttgga tgagaagaga tctgtgaagg ttcccatgat

     751                                                                800
ACHT            GTT  C T CAC     T  T  CTT       G  C        G  C
T37N
CONT
Consensus     gaaaatgaag ttactgacca cacgccactt ccgtgatgag gagctatcgt

     801                                                                850
ACHT            A C   G                    C            A  CT
T37N
CONT
Consensus     gctctgtgtt ggagctgaag tacacaggaa atgccagcgc cctgctcatc

     851                                                                900
ACHT                T   A A  A     G  G A          TGC  G TC
T37N               .... .......... .......... .......... ..........
CONT                C   G G  G     C  C G          GCT  A AA
Consensus     ctccctga-c a-g-ca-gat g-ag-a-gtg gaagcca--- t-c--ccaga
```


901 950
ACHT A CGG AG CT C GA GTT AG G GT
T37N
CONT G AAA GA AA T TT TCC CC A AG
Consensus gaccctga-g ---tggag-- a--ct-tg-- ---cag--a- atag--gagc
951 1000
ACHT CT A T T G GG T AC ...A C C
T37N
CONT AA C C G T TA C GG GAGG G T
Consensus t--acctgcc -aagtt-tcc atc-c-ag-- acta-a--ctg ----a-ga-
1001 1050
ACHT A A TCC GC C G C AGCA G
T37N
CONT G C CAG AA G A T GAAC A
Consensus -t-cttc--- a--tggg-at t-aggaag-c ttcacc---- a-gctgacct
1051 1100
ACHT G A C GGG G C AGC A C C T
T37N
CONT A T T AAA A A GAG T G T C
Consensus -tc-gggatc a-ag---cca -gaa-ct--- -gt-tc-cag gtggtcca-a
1101 1150
ACHT TC TC T ATT GAG T
T37N
CONT CT CT G GGC ACA G
Consensus agg--gtg-- -gatgt---t gag---ggca cagaagca-c tgctgccaca
1151 1200
ACHT CA C AAA TCAC C C T TCT AT G GGAGACAA GGA CAT
T37N
CONT GG T TTG GTGG A T G AAG CA T A..... ..C AGC
Consensus g--gt-a--- ----c-t-c- t---gc--ta -t-------- ---c---tgt
1201 1250
ACHT G C A GA C TGTC CT GACA C
T37N
CONT A A T TG T CTAT AC AGTG T
Consensus gc-tttcaac aggcc-ttcc tg-t--t-at ----c--aca ----c-caga
1251 1300
ACHT AC T C AG CC T GCA C TAG TG
T37N
CONT GT C T GC AT C AGT A CTA ..
Consensus --atc-tctt -atg--caaa gtca--aa-c cca---ag-c ---agct--c
1301 1350
ACHT T CAG GGG C CTCA GTA AACT A TGC A G G CC
T37N
CONT C TTC CAT G TCTG TCC GTTC G CAA G T C ..
Consensus a-caag---t ---g-t---- ---agg---- tgg-a---a- -ctg-atg--

```
    1351                                                                1400
ACHT            G      G GGC    G TG   C CC GTGC AC GT G      CA G    TG
T37N       .......... .......... .......... .......... ..........
CONT            A      A TAT    T CT   T AT CAAG CT TG T      TG T    AA
Consensus  tgg-tctct- ---aca-c-- g-c--t---- --c--a-tgg c--t-gca--

    1401                                                                1450
ACHT       TG     CCT   T T  ....     T CT GGA A GT A    C GATTCCCTGT POLY(A)
T37N       .......... .......... .......... .......... ..........
CONT       GT     TTA   A A  TACA     A TA CTC T TG T    T  POLY(A) ...
Consensus  --tggc---g -c-gc----t ta-c--t--- -g--g-cag- ..........
```

FIGURE 42

Protein Sequence Homology between Human Alpha$_I$-Antichymotrypsin and Murine Contrapsin

The deduced protein sequence of alpha$_I$-antichymotrypsin, from (Chandra et al, 1983) and contrapsin, from (Figure 40 and Hill et al, 1984) were aligned. Dots indicate gaps introduced to maximize the homology. Capital letters indicate differences to the consensus sequence, depicted by either small case letter or dashes, where there is no homology. The start of mature alpha$_I$-antichymotrypsin is double underlined. The row of x's represents the region between two subclones which was not sequenced. An arrow indicates the position at which the sequence of Chandra et al, (1983) comes into variance with the alpha$_I$-antichymotrypsin deduced protein sequence of Hill et al (1984). The reactive centre regions of the protease inhibitors are boxed (Hill et al, 1984). The consensus sequence is numbered at either end of each line. Position one indicates the first possible amino acid of the alpha$_I$-antichymotrypsin signal peptide.

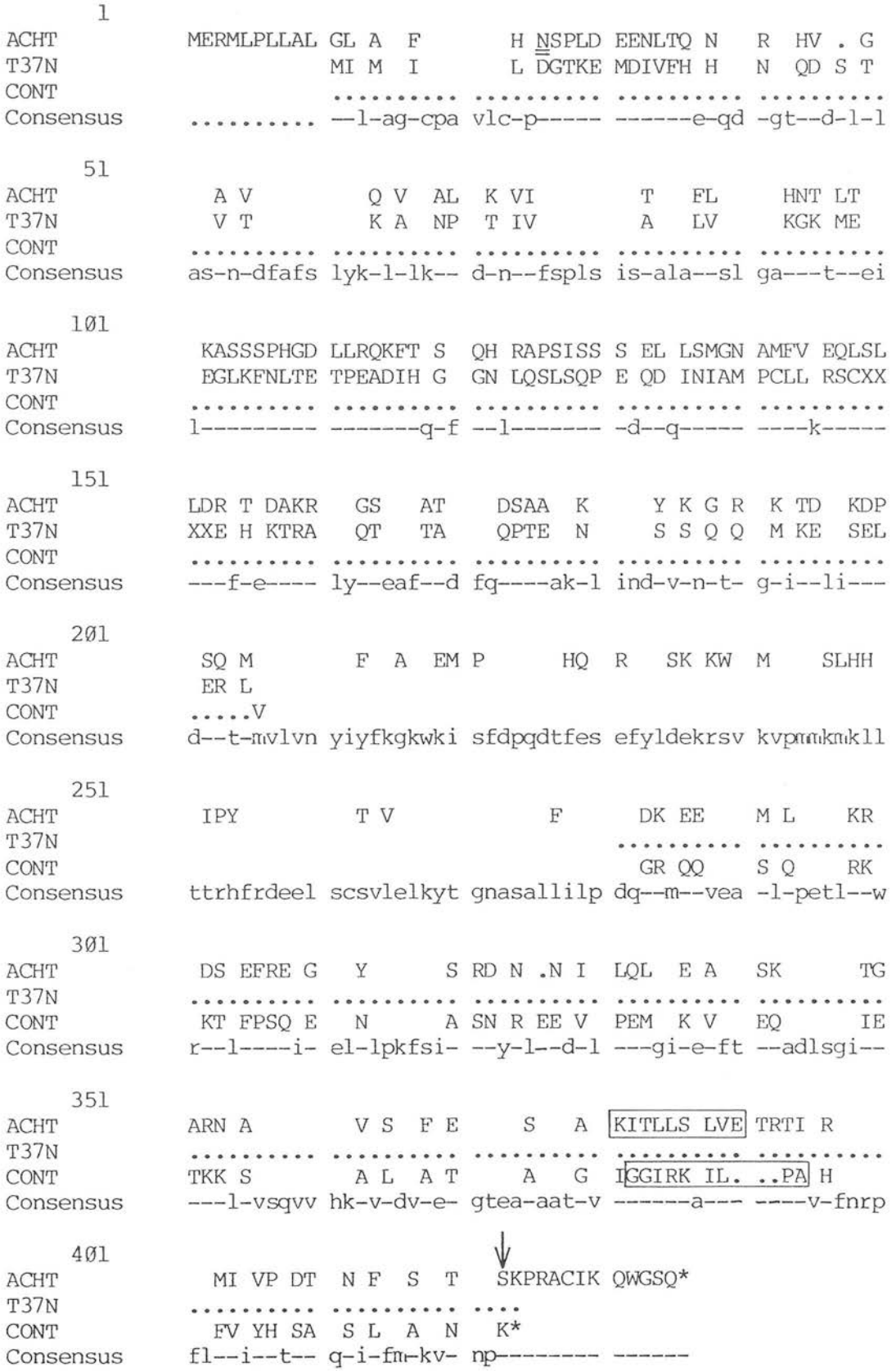

1
50
ACHT MERMLPLLAL GL A F H NSPLD EENLTQ N R HV . G
T37N MI M I L DGTKE MDIVFH H N QD S T
CONT
Consensus —l-ag-cpa vlc-p----- ------e-qd -gt--d-l-l
51
100
ACHT A V Q V AL K VI T FL HNT LT
T37N V T K A NP T IV A LV KGK ME
CONT
Consensus as-n-dfafs lyk-l-lk-- d-n--fspls is-ala--sl ga---t--ei
101
150
ACHT KASSSPHGD LLRQKFT S QH RAPSISS S EL LSMGN AMFV EQLSL
T37N EGLKFNLTE TPEADIH G GN LQSLSQP E QD INIAM PCLL RSCXX
CONT
Consensus l-------- -------q-f --l------- -d--q----- ----k-----
151
200
ACHT LDR T DAKR GS AT DSAA K Y K G R K TD KDP
T37N XXE H KTRA QT TA QPTE N S S Q Q M KE SEL
CONT
Consensus ---f-e---- ly--eaf--d fq----ak-l ind-v-n-t- g-i--li---
201
250
ACHT SQ M F A EM P HQ R SK KW M SLHH
T37N ER L
CONT V
Consensus d--t-mvlvn yiyfkgkwki sfdpqdtfes efyldekrsv kvpmmkmkll
251
300
ACHT IPY T V F DK EE M L KR
T37N GR QQ S Q RK
CONT
Consensus ttrhfrdeel scsvlelkyt gnasallilp dq--m--vea -l-petl--w
301
350
ACHT DS EFRE G Y S RD N .N I LQL E A SK TG
T37N
CONT KT FPSQ E N A SN R EE V PEM K V EQ IE
Consensus r--l----i- el-lpkfsi- --y-l--d-l ---gi-e-ft --adlsgi--
351
400
ACHT ARN A V S F E S A KITLLS LVE TRTI R
T37N
CONT TKK S A L A T A G IGGIRK IL. ..PA H
Consensus ---l-vsqvv hk-v-dv-e- gtea-aat-v ------a--- ----v-fnrp
401
435
ACHT MI VP DT N F S T SKPRACIK QWGSQ*
T37N
CONT FV YH SA S L A N K*
Consensus fl--i--t-- q-i-fm-kv- np-------- ------

FIGURE 43

Sequence Comparison of Protease Inhibitors to Determine Reading Frames

The underlined sequences of the mouse (contrapsin) and human (alpha$_I$-antichymotrypsin) protease inhibitors shown in Figure 41 were aligned with the primate (alpha$_I$-antitrypsin) protease inhibitor sequence (Kurachi et al, 1981) to determine the origin of the sequence length discrepancies. Dots indicate gaps introduced to maximize the homology. Capital letters indicate differences to the consensus sequence, which is depicted either by small case letters or dashes, where there is no homology. The consensus sequence is numbered at either end of the line, position 311 corresponds to position 311 of Figure 41. The row of x's represents the region between two subclones which was not sequenced.

```
        311                                                        360

T37N                              T      A     .CC      A      C
∝1AT  C     G       G  T                       .     G      TC GG
ACHT      .C         G     .                   C  A            A  T
Con   tctagaaggcctcaagttcaacctcacggaga-ttcctgaggcaga-atc


        361                                                        410

T37N     .C          TGGCA          A AG GT     G     . .       A
∝1AT     . G  A           G A       C  T           A   . .        AG
ACHT     C      A         C         G  C          T    TT     .
Con   ca-tcagggcttccag-acctcct-cg-accctcaaccag--ccagacga


        411                                                        460

T37N       GAT    A A A    A.                  A TG      GT
∝1AT         C         C  CC      C G C     CC     C    AG     A
ACHT  TG     G         GT  G  A               G      G   CAA  .
Con   ccagct-cagctga-cat-ggcaatgccatgttt-tcaaaaag--cctgc


        461                                                        510

T37N     XXXXXXXXXXG A    CA T  A GA A G GC           GA T
∝1AT     G. AG     T      TTT          T    A A              A
ACHT     TC GC     C G     AC                 GG      TGG    C
Con   ag--t--tgga-aagttc--ggaggatgccaaga-tctgtaccactc-ga
```

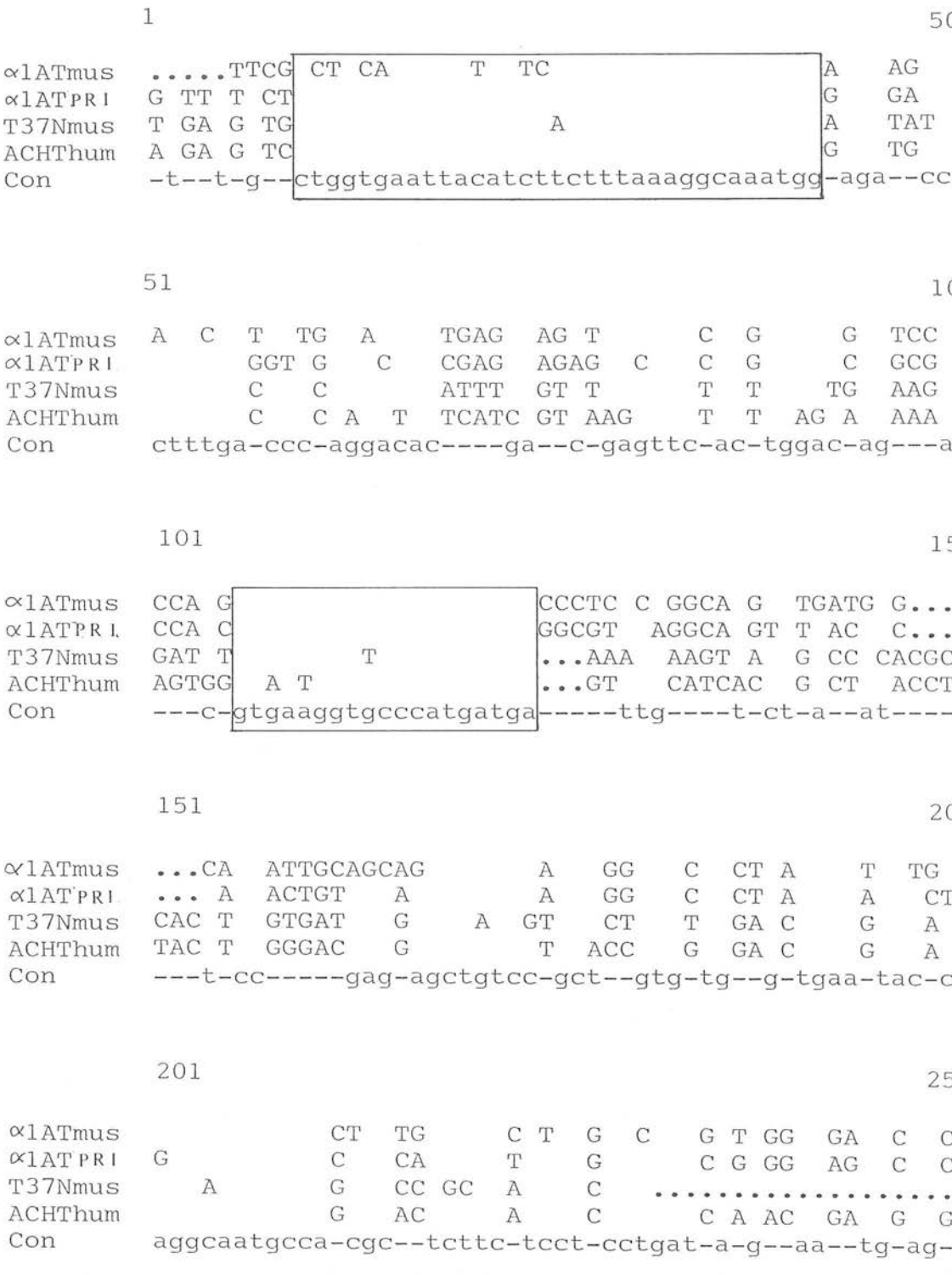
(Figure 41). The gap in the pUC T37N sequence (Figure 43, position 430) adjacent to a very strong cytosine residue in the sequencing gel from which the sequence was read and therefore could have masked a second nucleotide. Additional sequencing of the other strand would clarify the precise number of residues at this position and would determine the structure of the small section of sequence which was omitted from the preliminary sequencing.

In addition to determining the cause of the gap between contrapsin and alpha$_I$-antichymotrypsin the alignment with alpha$_I$-antitrypsin showed two areas of very high sequence homology between the protease inhibitors. Further comparison of these regions with the addition of the mouse alpha$_I$-antitrypsin sequence (Hill et al, 1984) confirmed that those two regions have beeen well conserved within mammalian protease inhibitors (Figure 44 boxed regions). These regions show over 90% homology between all the sequences, which compares with ~70% homology between both the alpha$_I$-antitrypsins (including the reactive site region) and the alpha$_I$-antichymotrypsin/contrapsin comparison (active site region 30%) and a level of 60% homology between contrapsin and primate alpha$_I$-antitrypsin. No other areas of such extensive homology were observed. It would therefore seem possible that the boxed regions of Figure 44 correspond to sequences in the protein important for plasma protease inhibitor control or function.

FIGURE 44

Sequence Comparison of Mammalian Plasma Protease Inhibitors

The sequences of mouse alpha$_1$-antitrypsin (α_1AT mus) (Hill *et al*, 1984), primate alpha$_1$-antitrypsin (α_1AT pri) (Kurachi *et al*, 1981), alpha$_1$-antichymotrypsin (ACHThum) (Chandra *et al*, 1983) and pUC T37N (T37Nmus), which is mouse contrapsin, were compared. Dots indicate gaps introduced to maximize the homology. Capital letters indicate differences to the consensus sequence, which is depicted by small case letters or dashes, where there is no homology. The consensus sequence is numbered at the end of each line, positions 1 and 250 equate to positions 629 and 825 in Figure 41. Regions of outstanding homology between the sequences (>90%, over 15Bp) are boxed.


1
50
∝1ATmus .....TTCG CT CA T TC A AG
∝1ATPRI G TT T CT G GA
T37Nmus T GA G TG A A TAT
ACHThum A GA G TC G TG
Con -t--t-g--ctggtgaattacatcttctttaaaggcaaatgg-aga--cc
51
100
∝1ATmus A C T TG A TGAG AG T C G G TCC
∝1ATPRI GGT G C CGAG AGAG C C G C GCG
T37Nmus C C ATTT GT T T T TG AAG
ACHThum C C A T TCATC GT AAG T T AG A AAA
Con ctttga-ccc-aggacac----ga--c-gagttc-ac-tggac-ag---a
101
150
∝1ATmus CCA G CCCTC C GGCA G TGATG G...
∝1ATPRI CCA C GGCGT AGGCA GT T AC C...
T37Nmus GAT T T ...AAA AAGT A G CC CACGC
ACHThum AGTGG A T ...GT CATCAC G CT ACCT
Con ---c-gtgaaggtgcccatgatga-----ttg----t-ct-a--at----
151
200
∝1ATmus ...CA ATTGCAGCAG A GG C CT A T TG
∝1ATPRI ... A ACTGT A A GG C CT A A CT
T37Nmus CAC T GTGAT G A GT CT T GA C G A
ACHThum TAC T GGGAC G T ACC G GA C G A
Con ---t-cc-----gag-agctgtcc-gct--gtg-tg--g-tgaa-tac-c
201
250
∝1ATmus CT TG C T G C G T GG GA C C
∝1ATPRI G C CA T G C G GG AG C C
T37Nmus A G CC GC A C .....
ACHThum G AC A C C A AC GA G G
Con aggcaatgcca-cgc--tcttc-tcct-cctgat-a-g--aa--tg-ag-

REFERENCES

Aisen, P. and Brown, E.B. (1975) Prog. in Hematology 9 pp 25 - 56. "Structure and Function of Transferrin."

Aisen, P. and Listowsky, I. (1980) Ann. Rev. Biochem. 49 pp 357 - 393. "Iron Transport and Storage Proteins."

Aldred, A.R.; Howlett, G.J.; and Schreiber, G. (1984) Biochem. and Biophys. Res. Comm. 122 pp 960 - 965. "Synthesis of Rat Transferrin in Escherichia coli Containing a Recombinant Bacteriophage."

Alt, F.W.; Rosenberg, N.; Cassanova, R.J.; Thomas, E. and Baltimore, D. (1982) Nature 296 pp 325 - 331 "Immunoglobulin Heavy-Chain Expression and Class Switching in a Murine Leukaemia Cell Line."

Alwine, J.C.; Kemp, D.J. and Stark, G.R. (1977) Proc. Natl. Acad. Sci. USA 74 pp 5350 - 5354. "Method for Detection of Specific RNAs in Agarose Gels by Transfer to Diazobenzyloxymethyl- Paper and Hybridization with DNA Probes."

Aviv, H. and Leder, P. (1972) Proc. Natl. Acad. Sci. USA 69 pp 1408 - 1412 "Purification of Biologically Active Globin Messenger RNA by Chromatography Oligothymidylic Acid-Cellulose."

Barnes, D. and Sato, G. (1980) Cell 22 pp 649 - 655. "Serum-Free Cell Culture: a Unifying Approach."

Barth, R.K.; Gross, K.W.; Gremke, L.C. and Hastie, N.D. (1982) Proc. Natl. Acad. Sci. USA 79 pp 500 - 504. "Developmentally Regulated mRNAs in Mouse Liver."

Beach, R.L.; Popiela, M. and Festoff, B.W. (1983) FEBS Lett. 156 pp 151 - 156. "The Identification of Neurotrophic Factor as Transferrin."

Bennett, K.L.; Lalley, P.A.; Barth, R.K. and Hastie, N.D. (1982) Proc. Natl. Acad. Sci. USA 79 pp 1220 - 1224. "Mapping the Structural Genes coding for the Major Urinary Proteins in the Mouse: Combined use of Recombinant Inbred Strains and Somatic Cell Hybrids."

Berget, S.M. (1984) Nature 309 pp 179 - 182. "Are U4 Small Nuclear Ribonucleoproteins Involved in Polyadenylation ?"

Biggin, M.D.; Gibson, T.J.; and Hong, G.F. (1983) Proc. Natl. Acad. Sci. USA 80 pp 3963 - 3965. "Buffer Gradient Gels and ^{35}S label as an Aid to Rapid DNA Sequence Determination."

Bird, A.P. (1984) Nature. 307 pp 503 - 504. "DNA Methylation - How Important in Gene Control ? "

Birnboim, H.C. and Doly, J. (1979) Nucl. Acid Res. 7 pp 1513 - 1523. "A Rapid Alkaline Extraction Procedure for Screening Recombinant Plasmid DNA."

Bishop, J.O. (1979) J. Mol. Biol. 128 pp 545 - 559 "A DNA Sequence Cleaved by Restriction Endonuclease R Eco Rl in Only One Strand."

Bishop, J.O.; Clark, A.J.; Clissold, P.M.; Hainey, S. and Francke, U. (1982) EMBO J. 1 pp 615 - 620 "Two Main Groups of Mouse Major Urinary Protein Genes, Both Largely Located on Chromosome 4."

Bishop, J.O.; Selman, G.G.; Hickman, J.; Black, L.; Saunders, R.D.P. and Clark, A.J. (1985) Mol. and Cell Biol. 5 (7) pp 1591 - 1600. "The 45 - kb Unit of Major Urinary Protein Gene Organisation Is a Gigantic Imperfect Palindrome."

Bonner, W.M. and Laskey, R.A. (1974) Eur. J. Biochem. 46 pp 83 - 88 "A Film Detection Method for Tritium Labelled Proteins and Nucleic Acids in Polyacrylamide Gels."

Bostain, K.A.; Lemire, J.M.; Cannon, L.C. and Halvorson, H.O. (1980) Proc. Natl. Acad. Sci. USA 77 pp 4504 - 4508. "In vitro Synthesis of Repressible Yeast Acid Phosphatase: Identification of Multiple mRNA and Products."

Boyer, H.W. and Roulland-Dussoix, D. (1969) J. Mol. Biol. 41 pp 459 - 472 "A Complementation Analysis of the Restriction and Modification of DNA in Escherica coli."

Breathnach, R.; Benoist, C.; O'Hare, K.; Gammon, F. and Chambon, P. (1978) Proc. Natl. Acad. Sci. USA 75 pp 4853 - 4857. "Ovalbumin Gene: Evidence for a Leader Sequence in mRNA and DNA Sequences at the Exon-Intron Boundaries."

Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50 pp 349 - 383. "Organization and Expression of Eukaryotic Split Genes Coding for Proteins."

Bridges, C.B. (1936) Science 83 pp 210 - 211 "The Bar "Gene" a Duplication."

Brown, D.D. (1981) Science 211 pp 667 - 674. "Gene Expression in Eukaryotes."

Brown, J.P.; Hewick, R.M.; Hellstrom, I.; Hellstrom, K.E.; Doolittle, R.F. and Dreyer, W.J. (1982) Nature 296 pp 171 - 173. "Human Melanoma-Associated Antigen p 97 is Structurally and Functionally Related to Transferrin."

Cech, T.R. (1983) Cell 34 pp 713 - 716. "RNA Splicing: Three Themes With Variations."

Chan, E.W.; Dale, P.J.; Greco, I.L.; Rose, J.G. and O'Connor, T.E. (1980) Biochem. et Biophys. Acta 606 pp 353 - 361. "Effects of Polyethylene Glycol on Reverse Transcriptase and Other Polymerase Activities."

Chandra, T.; Stackhouse, R.; Kidd, V.J.; Robson, K.J.H. and Woo, S.L.C. (1983) Biochemistry 22 pp 5055 - 5061. "Sequence Homology Between Human $alpha_I$ -Antichymotrypsin, $alpha_I$-Antitrypsin, and Antithrombin III."

Charnay, P.; Treisman, R.; Mellon, P.; Chao, M.; Axel, R. and Maniatis, T. (1984) Cell 38 pp 251 - 263. "Differences in Human alpha- and beta-Globin Gene Expression in Mouse Erythroleukemia Cells: The Role of Intragenic Sequences."

Clark, A.J.; Clissold, P.M. and Bishop, J.O. (1982) Gene pp 221 - 230. "Variation Between Mouse Major Urinary Protein Genes Isolated From a Single Inbred Line."

Clark, A.J.; Clissold, P.M.; Al Shawi, R.; Beattie, P. and Bishop, J.O. (1984a) EMBO J. 3 pp 1045 - 1052. "Structure of Mouse Major Urinary Protein Genes: Different Splicing Configurations in the 3' - Non-Coding Region."

Clark, A.J.; Hickman, J. and Bishop, J.O. (1984b) EMBO J. 3 pp 2055 - 2064. "A 45 - Kb DNA Domain with Two Divergently Orientated Genes is the Unit Organization of the Murine Major Urinary Protein Genes."

Clark, A.J.; Ghazal, P.; Bingham, R.W.; Barrett, D. and Bishop, J.O. (1985a) EMBO J. 4 pp 3159 - 3165. "Sequence Structures of a Mouse Major Urinary Protein Gene and Pseudogene Compared."

Clark, A.J.; Chave-Cox, A.; Ma, X. and Bishop, J.O. (1985b) EMBO J. 4 pp 3167 - 3171. "Analysis of Mouse Major Urinary Protein Genes: Variation Between the Exonic Sequences of Group 1 Genes and a Comparison with an Active Gene outwith Group 1 Both Suggest that Gene Conversion has Occurred Between MUP Genes."

Clissold, P.M. and Bishop, J.O. (1981) Gene 15 pp 225 - 235. "Molecular Cloning of cDNA Sequences Transcribed from Mouse Liver Endoplasmic Reticulum Poly (A) mRNA."

Clissold, P.M.; Mason, P.J. and Bishop, J.O. (1981) Proc. Natl. Acad. Sci. USA 78 pp 3697 - 3701. "Comparison of Poly (A) mRNA Prepared from Membranes and Free Polyribosomes of Mouse Liver."

Clissold, P.M. and Bishop, J.O. (1982) Gene 18 pp 211 - 220 "Variation in Mouse Major Urinary Protein (MUP) Genes and MUP Gene Products Within and Between Inbred Lines."

Clissold, P.M.; Hainey, S. and Bishop, J.O. (1984) Biochemical Genetics 22 pp 379 - 387. "Messenger RNAs Coding for Mouse Major Urinary Proteins are Differentially Induced by Testosterone."

Cochet, M.; Gannon, F.; Hen, R.; Maroteaux, L.; Perrin, F. and Chambon, P. (1979) Nature 282 pp 567 - 574. "Organization and Sequence Studies of the 17-Piece Chicken Conalbumin Gene."

Cohen, B.L. (1960) Genet. Res., Camb. 1 pp 431 - 438 "Genetics of Plasma Transferrins in the Mouse."

Crabtree, G.R. and Kant, J.A. (1982) Cell 31 pp 159 - 166. "Organization of the Rat Gamma-Fibrinogen Gene: Alternative mRNA Slice Patterns Produce the Gamma A and Gamma B Chains of Fibrinogen."

Craik, C.S.; Sprang, S.; Fletterick, R. and Rutter, W.J. (1982) Nature 299 pp 180-182. "Intron-exon splice junctions map at protein surfaces."

Derman, E.; Krauter, K.; Walling, L.; Weinberger, C.; Ray, M. and Darnell, J.E. Jr. (1981) Cell 23 pp 731 - 739. "Transcriptional Control in the Production of Liver - Specific mRNAs."

Diamond, A.; Cooper, G.M.; Ritz, J. and Lane, M.A. (1983) Nature 305 pp 112 - 116. "Identification and Molecular Cloning of the Human Blym Transforming Gene Activated in Burkitt's Lymphomas."

Dolan, K.P.; Unterman, R.; McLaughlin, M.; Nakhasi, H.L.; Lynch, K.R. and Feigelson, P. (1982) J. Biol. Chem. 257 pp 13527 - 13534. "The Structure and Expression of Very Closely Related Members of the Alpha $_{2\mu}$ Globulin Gene Family."

Domdey, H.; Wiebauer, K.; Kazmaier, M.; Müller, V.; Odink, K. and Fey, G. (1982) Proc. Natl. Acad. Sci. USA 79 pp 7619 - 7623. "Characterization of the mRNA and Cloned cDNA Specifying the Third Component of Mouse Complement."

Donis-Keller, H.; Maxam, A.M. and Gilbert, W. (1977) Nucl. Acid Res. 4 pp 2527 - 2538 "Mapping Adenines, Guanines and Pyrimidines in RNA."

Efstratiadis, A.; Kafatos, F.; Maxam, A.M.; and Maniatis, T. (1976) Cell 7 pp 279 - 288. "Enzymatic in vitro Synthesis of Globin Genes."

Efstratiadis, A.; Posakony, J.W.; Maniatis, T.; Laww, R.M.; O'Connell, C.; Spritz, R.A.; De Riel, 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) Cell 21 pp 653 - 668. "The Structure and Evolution of the Human Beta-Globin Gene Family."

Eiferman, F.A.; Young, P.R.; Scott, R.W. and Tilghman, S.M. (1981) Nature 294 pp 713 - 718. "Intragenic Amplification and Divergence in the Mouse Alpha-Fetoprotein Gene."

Ekblom, P.; Thesleff, I.; Saxén, L.; Miettinen, A. and Timple, R. (1983) Proc. Natl. Acad. Sci. USA 80 p 2651 - 2655. "Transferrin as a Fetal Growth Factor: Acquisition of Responsiveness Related to Embryonic Induction."

Esumi, H.; Takahashi, Y.; Sato, S.; Nagase, S. and Sugimura, T. (1983) Proc. Natl. Acad. Sci. USA 80 pp 95 - 99. "A Seven- Base-Pair Deletion in an Intron of the Albumin Gene of Analbuminemic Rats."

Fisher, D.H.; Dodgson, J.B.; Hughes, S. and Engel, J.D. (1984) Proc. Natl. Acad. Sci. USA 81 pp 2733 - 2737. "An Unusual 5' Splice Sequence is Efficiently Utilized in vivo."

Frazier, W.A.; Angeletti, R.H. and Bradshaw, R.A. (1972) Science 176 pp 482 - 488. "Nerve Growth Factor and Insulin."

Fritton, H.P.; Igo-Kemenes, T.; Nowock, J.; Strech-Jurk, U.; Theisen, M. and Sippel, A.E. (1984) Nature 311 pp 163 - 165. "Alternative Sets of DNase I - Hypersensitive Sites Characterize the Various Functional States of the Chicken Lysozyme Gene."

Gergen, J.P.; Stern, R.H. and Wensink, P.C. (1979) Nucl. Acid Res. 7 pp 2115 - 2135. "Filter Replicas and Permanent Collections of Recombinant DNA Plasmids."

Ghazal, P.; Clark, A.J. and Bishop, J.O. (1985) Proc. Natl. Acad. Sci. USA, 82 pp 4182 - 4185. "Evolutionary Amplification of a Pseudogene."

Gil, A. and Proudfoot, N.J. (1984) Nature 312 pp 473 - 474 "A Sequence Downstream of AATAAA is Required for Rabbit Beta-Globin mRNA 3' End Formation."

Gilbert, W. (1978) Nature 271 pp 501. "Why Genes in Pieces."

Gilbert, W. (1981) Science 214 pp 1305 - 1312. "DNA Sequencing and Gene Structure."

Goubin, G.; Goldman, D.S.; Luce, J.; Neiman, P.E. and Cooper, G.M. (1983) Nature 302 pp 114 - 119. "Molecular Cloning and Nucleotide Sequence of a Transforming Gene Detected by Transfection of Chicken B-Cell Lymphoma DNA."

Gubler, U. and Hoffman, B.J. (1983) Gene 25 pp 263 - 269. "A Simple and Very Efficient Method for Generating cDNA Libraries."

Hainey, S. and Bishop, J.O. (1982) Genet. Res., Camb. 39 pp 31 - 42 "Allelic Variation at Several Different Genetic Loci Determines the major Urinary Protein Phenotype of Inbred Mouse Strains."

Hanahan, D. (1983) J. Mol. Biol. 166 pp 557 - 580. "Studies on Transformation of Escherichia coli with Plasmids."

Hastie, N.D.; Held, W.A. and Toole, J.J. (1979) Cell pp 449 - 457 "Multiple Genes coding for the Androgen-Regulated Major Urinary Proteins of the Mouse."

Heilig, R.; Perrin, F.; Gannon, F.; Mandel, J.L. and Chambon, P. (1980) Cell 20 pp 625 - 637. "The Ovalbumin Gene Family: Structure of the X Gene and Evolution of Duplicated Split Genes."

Hill, R.E.; Shaw, P.H.; Boyd, P.A.; Baumann, H. and Hastie, N.D. (1984) Nature 311 pp 175 - 177. "Plasma Protease Inhibitors in Mouse and Man: Divergence within the Reactive Centre Regions."

Hill, R.L.; Brew, K.; Vanaman, T.C.; Trayer, I.P. and Mattock, P. (1969) Brookhaven Symp. Biol. 21 pp 139 - 152. "The Structure, Function and Evolution of Alpha-Lactalbumin."

Hong, G.F. (1982) J. Mol. Biol. 158 pp 539 - 549 "A Systematic DNA Sequencing Strategy."

Huebers, H.; Baner, W.; Huebers, E.; Csisba, E. and Finch, C. (1981) Blood 57 pp 218 - 228. "The Behaviour of Transferrin Iron in the Rat."

Ingolia, T.D. and Craig, E.A. (1982) Proc. Natl. Acad. Sci. USA 79 pp 2360 - 2364 "Four Small Drosophila Heat Shock Proteins are Related to Each Other and to Mammalian Alpha-Crystallin."

Jeltsch, J.M. and Chambon, P. (1982) Eur. J. Biochem. 122 pp 291 - 295. "The Complete Nucleotide Sequence of the Chicken Ovotransferrin mRNA."

Jones, C.W. and Kafatos, F.C. (1980) Cell 22 pp 855 - 867. "Structure, Organization and Evolution of Developmentally Regulated Chorion Genes in a Silkmoth."

Jongstra, J.; Reudelhuber, T.L.; Oudet, P.; Benoist, C.; Chae, C.B.; Jeltsch, J.M.; Mathis, D.J. and Chambon, P. (1984) Nature 307 pp 708 - 714. "Induction of Altered Chromatin Structures by Simian virus 40 Enhancer and Promoter Elements."

Karin, M. and Mintz, B. (1981) J. Biol. Chem. 256 (7) pp 3245 - 3252. "Receptor-Mediated Endocytosis of Transferrin in Developmentally Totipotent Mouse Teratocarcinoma Stem Cells."

Kessler, S.W. (1975) J. of Immunology 115 pp 1617 - 1624. "Rapid Isolation of Antigens from Cells with a Staphlococcal Protein A - Antibody Adsorbent: Parameters of the Interaction of Antibody - Antigen Complex with Protein A."

Kieny, M.P.; Lathe, R. and Lecocq, J.P. (1983) Gene 26 pp 91 - 99. "New Versatile Cloning and Sequencing Vectors Based on Bacteriophage M13."

King, C.R. and Piatigorsky, J. (1983) Cell 32 pp 707 - 712. "Alternative RNA Splicing of the Murine Alpha A-Crystallin Gene: Protein - Coding Information Within an Intron."

Klein, P.A.; Roop, B.L. and Roop, W.E. (1966) Nature 212 pp 1376 - 1377. "Starch Gel Electrophoretic Patterns of Murine Transferrin."

Knowler, J.T. and Wilks, A.F. (1980) T.I.B.S., October pp 268 - 271. "Ribonucleoprotein Particles and the Maturation of Eukaryote mRNA."

Krämer, A.; Keller, W.; Appel, B. and Lührmann, R. (1984) Cell 38 pp 299 - 307. "The 5' Terminus of the RNA Moiety of U1 Small Nuclear Ribonucleoprotein Particles is Required for the Splicing of Messenger RNA Precursors."

Kuhn, N.J.; Woodworth-Gutai, M.; Gross, K.W. and Held, W.A. (1984) Nucl. Acid Res. 12 pp 6073 - 6090. "Subfamilies of the Mouse Major Urinary Protein (MUP) Multigene Family: Sequence Analysis of cDNA Clones and Differential Regulation in the Liver."

Kurachi, K.; Chandra, T.; Frienzner Degen, S.J.; White, T.T.; Marchioro, T.L.; Woo, S.L.C. and Davie, E.W. (1981) Proc. Natl. Acad. Sci. USA 78 pp 6826 - 6830. "Cloning and Sequence of cDNA Coding for $Alpha_1$-Antitrypsin."

Kurtz, D.T. (1981) J. Mol. and App. Genetics 1 pp 29 - 38 "Rat Alpha $_{2\mu}$ -Globulin is Encoded by a Multigene Family."

Laemmli, U.K. (1970) Nature 227 pp 680 - 685. "Cleavage of Structural Proteins During Assembly of the Head of Bacteriophage T4"

Langford, C.J.; Klinz, F.J.; Donath, C. and Gallwitz, D. (1984) Cell 36 pp 645 - 653. "Point Mutations Identify the Conserved, Intron-Contained TACTAAC Box as an Essential Splicing Signal Sequence in Yeast."

Laperche, Y.; Lynch, K.R.; Dolan, K.P. and Feigelson, P. (1983) Cell 32 pp 453 - 460. "Tissue - Specific Control of Alpha $_{2\mu}$-Globulin Gene Expression: Constitutive Synthesis in the Submaxillary Gland."

Laskey, R.A. and Mills, A.D. (1975) Eur. J. Biochem. 56 pp 335 - 341. " Quantitative Film Detection of ^3H and ^{14}C in Polyacrylamide Gels by Flourography."

Lawn, R.H.; Adelman, J.; Bock, S.C.; Franke, A.E.; Hounk, C.M.; Najarien, R.L.; Seeburg, P.H. and Wion, K.L. (1981) Nucl. Acid Res. 9 pp 6103 - 6115. "The Sequence of Human Serum Albumin cDNA and its Expression in E. coli."

Lawson, G.M.; Knoll, B.J.; March, C.J.; Woo, S.L.C.; Tsai, M.J. and O'Malley, B.W. (1982) J. Biol. Chem. 257 pp 1501 - 1507. "Definition of 5' and 3' Structural Boundaries of the Chromatin Domain Containing the Ovalbumin Multigene Family."

Leicht, M.; Long, G.L.; Chandra, T.; Kurachi, K.; Kidd, V.J.; Mace, Jr.M.; Davie, E.W. and Woo, S.L.C. (1982) Nature 297 pp 655 - 659. "Sequence Homology and Structural Comparison Between the Chromosomal Human $Alpha_I$-Antitrypsin and Chicken Ovalbumin Genes."

Levin, M.J.; Tuil, D.; Uzan, G.; Dreycus, J.C. and Kahn, A. (1984) Biochem. and Biophys. Res. Com. 122 pp 212 - 217. "Expression of the Transferrin Gene during Development of Non-Hepatic Tissues."

Liang, T.J. and Grieninger, G. (1981) Proc. Natl. Acad. Sci. USA 78 pp 6972 - 6976. "Direct Effect of Insulin on the Synthesis of Specific Plasma Proteins: Biphasic Response of Hepatocytes Cultured in Serum- and Hormone-Free Medium."

Lomedico, P.T.; Rosenthal, N.; Efstratiadis, A.; Gilbert, W.; Kolodner, R. and Tizard, R. (1979) Cell 18 pp 545 - 558. "The Structure and Evolution of the Two Non-Allelic Rat Preproinsulin Genes."

Lomedico, P.T. and McAndrew, S.J. (1982) Nature 299 pp 221 - 226. "Eukaryotic Ribosomes Can Recognise Preproinsulin Initiation Codons Irrespective of their Position Relative to the 5' End of mRNA."

MacGillivray, R.T.A.; Mendez, E.; Sinha, S.K.; Sutton, M.R.; Limeback-Zins, J. and Brew, K. (1982) Proc. Natl. Acad. Sci. USA 79 pp 2504 - 2508. "The Complete Amino Acid Sequence of Human Serum Transferrin."

MacGillivray, R.T.A.; Mendez, E.; Shewale, J.G.; Sinha, S.K.; Lineback-Zins, J. and Brew, K. (1983) J. Biol. Chem. 258 pp 3543 - 3553. "The Primary Structure of Human Serum Transferrin."

Maki, R.; Roeder, W.; Traunecker, A.; Sidman, C.; Wabl, M.; Raschke, W. and Tonegawa, S. (1981) Cell 24 pp 353 - 365. "The Role of DNA Rearrangement and Alternative RNA Processing in the Expression of Immunoglobulin Delta Genes."

Mandel, M. and Higa, A. (1970) J. Mol. Biol. 53 pp 159 - 162. "Calcium Dependant Bacteriophage DNA Infection."

Maniatis, T.; Jeffrey, A. and Van de Sande, H. (1975) Biochemistry 14 pp 3787 - 3794. "Chain Length Determination of Small Double and Single Stranded DNA Molecules by Polyacrylamide Gel Electrophoresis."

Maniatis, T.; Hardison, R.C.; Lacy, E.; Lauer, J.; O'Connell, C.; Quon, D.; Sim, G.K. and Efstratiadis, A. (1978) Cell 15 pp 687 - 701. "The Isolation of Structural Genes from Libraries of Eukaryotic DNA."

Maniatis, T.; Fritsch, E.F. and Sambrook, J. (1982) in "Molecular Cloning a Laboratory Manual." "Cold Spring Harbour Laboratory."

de Martynoff, G.; Pays, E. and Vassart, G. (1980) Biochem. Biophys. Res. Com. 93 pp 645 - 653. "Synthesis of a Full Length DNA Complementary to Thyroglobulin 33S Messenger RNA."

Mather, J.P. and Sato, G.H. (1979) Exp. Cell Res. 124 pp 215 - 221. "The Use of Hormone-Supplemented Serum-Free Media in Primary Cultures."

McKnight, G.S.; Lee, D.C.; Hemmaplardh, D.; Finch, C.A. and Palmiter, R.D. (1980) J. Biol. Chem. 255 pp 144 - 147. "Transferrin Gene Expression: Effects of Nutritional Iron Deficiency."

McKnight, S.L. (1982) Cell 31 pp 355 - 365. "Functional Relationships Between Transcriptional Control Signals of the Thymidine Kinase Gene of Herpes Simplex Virus."

Meehan, R.R.; Barlow, D.P.; Hill, R.E.; Hogan, B.L.M. and Hastie, N.D. (1984) EMBO J. 3 pp 1881 - 1885. "Pattern of Serum Protein Gene Expression in Mouse Visceral Yolk Sac and Foetal Liver."

Messing, J. (1979) Recombinant DNA Technical Bulletin, NIH Publication No.79 - 99, 2, No.2 pp 43 - 48. "A Multi-Purpose Cloning System Based on the Single-Stranded DNA Bacteriophage M13."

Messing, J. and Vieira, J. (1982) Gene 19 pp 269 -276. "A New Pair of M13 Vectors for Selecting Either DNA Strand of Double-Digest Restriction Fragments."

Mills, F.C.; Fisher, L.M.; Kuroda, R.; Ford, A.M. and Gould, H.J. (1983) Nature 306 pp 809 - 812. "DNase I Hypersensitive Sites in the Chromatin of Human μ Immunoglobulin Heavy-Chain Genes."

Miyata, T.; Yasunaga, T. and Nishida, T. (1980) Proc. Natl. Acad. Sci. USA 77 pp 7328 - 7332. "Nucleotide Sequence Divergence and Functional Constraint in mRNA Evolution."

Morgan, E.H. (1983) in Plasma Protein Secretion by the Liver (Glaumann, H.; Peters, T.Jr. and Redman, C., eds) pp 331 - 355, Academic Press, London.

Mount, S.M. (1982) Nucl. Acid Res. 10 pp 459 - 472. "A Catalogue of Splice Junction Sequences."

Nawa, H.; Kotani, H. and Nakanishi, S. (1984) Nature 312 pp 729 - 734. "Tissue-Specific Generation of the Two Preprotachykinin mRNAs From One Gene by Alternative RNA Splicing."

North, G. (1984) Nature 312 pp308 - 309 "Multiple Levels of Gene Control in Eukaryotic Cells."

Nudet, U.; Katcoff, D.; Zakut, R.; Shani, M.; Carmon, Y.; Finer, M.; Czosnek, H.; Ginsburg, I. and Yaffe, D. (1982) Proc. Natl. Acad. Sci. USA 79 pp 2763 - 2767. "Isolation and Characterization of Rat Skeletal Muscle and Cytoplasmic Actin Genes."

Okayama, H. and Berg, P. (1982) Molecular and Cellular Biol. 2 pp 161 - 170. "High-Efficiency Cloning of Full Length cDNA."

Padgett, R.A.; Mount, S.M.; Steitz, J.A. and Sharp, P.A. (1983) Cell 35 pp 101 - 107. "Splicing of Messenger RNA Precursors is Inhibited by Antisera to Small Nuclear Ribonucleoprotein."

Palmiter, R.D. and Lee, D.C. (1980) J. Biol. Chem. 255 pp 9693 - 9698. "Regulation of Gene Transcription by Estrogen and Progesterone."

Palmiter, R.D.; Mulvihill, E.R.; Shepard, J.H. and McKnight, G.S. (1981) J. Biol. Chem. 256 pp 7910 - 7916. "Steroid Hormone Regulation of Ovalbumin and Conalbumin Gene Transcription."

Parish, J.H. and Kirby, K.S. (1966) Biochem. Biophys. Acta. 129 pp 554 - 562. Reagents Which Reduce Interactions Between rRNA and Rapidly Labelled RNA from Rat Liver."

Park, I.; Schaeffer, E.; Sidoli, A.; Baralle, F.E.; Cohen, G.N. and Zakin, M.M. (1985) Proc. Natl. Acad. Sci. USA 82 pp 3149 - 3153. "Organization of the Human Transferrin Gene: Direct Evidence that it Originated by Gene Duplication."

Pelham, H.R.B. and Jackson, R.J. (1976) Eur. J. Biochem. 67 pp 247 - 257. "An Efficient mRNA-Dependant Translation System from Reticulocyte Lysates."

Perler, F.; Efstratiadis, A.; Lomedico, P.; Gilbert, W.; Kolodner, R. and Dodgson, J. (1980) Cell 20 pp 555 - 566. "The Evolution of Genes: The Chicken Preproinsulin Gene."

Picard, D. and Schaffner, W. (1984) Nature 307 pp 80 - 82. "A Lymphocyte-Specific Enhancer in the Mouse Immunoglobulin K Gene."

Pikielny, C.W.; Teem, J.L. and Rosbash, M. (1983) Cell 34 pp 395 - 403. "Evidence for the Biochemical Role of an Internal Sequence in Yeast Nuclear mRNA Introns: Implications for U1 RNA and Metazoan mRNA Splicing."

Plowman, G.D.; Brown, J.P.; Enns, C.A.; Schroder, J,; Nikinmaa, B.; Sussman, H.H., Hellstrom, K.E. and Hellstrom, I. (1983) Nature 303 pp 70 - 72. "Assignment of the Gene for Human Melanoma-Associated Antigen p 97 to Chromosome 3."

Proudfoot, N.J. (1984) Nature 307 pp 412 - 413 "The End of the Message And Beyond."

Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263 pp 211 - 214 "3' Non-coding Region Sequences in Eukaryotic Messenger RNA."

Proudfoot, N.J. and Maniatis, T. (1980) Cell 21 pp 537 - 544. "The Structure of a Human Alpha-Globin Pseudogene and its Relationship to Alpha-Globin Gene Duplication."

Putnam, F.W. (1975) in The Plasma Proteins (Putnam, F.W., Ed.) 2nd ed., Vol I, pp 265 - 316, Academic Press, London.

Rave, N.; CrKvenjakov, R. and Boedtker, H. (1979) Nucl. Acid Res. 6 pp 3559 - 3567. "Identification of Procollagen mRNAs Transferred to Diazobenzyloxymethyl Paper from Formaldehyde Agarose Gels."

Reisner, A.H. (1985) Nature 313 pp 801 - 803. "Similarity Between the Vaccinia Virus 19K Early Protein and Epidermal Growth Factor."

Retzel, E.F.; Collett, M.S. and Faras, A.J. (1980) Biochemistry 19 pp 513 - 518. "Enzymatic Synthesis of Deoxyribonucleic Acid by the Avion Retrovirus Reverse Transcriptase in vitro: Optimum Conditions Required for Transcription of Large Ribonucleic Acid Templates."

Reudelhuber, T. (1984) Nature 312 pp 700 - 701. "Upstream and Downstream Control of Eukaryotic Genes."

Richards, R.I.; Shine, J.; Ullrich, A.; Wells, J.R.E. and Goodman, H.M. (1979) Nucl. Acid Res. 7 pp 1137 - 1146. "Molecular Cloning and Sequence Analysis of Adult Chicken Beta Globin cDNA."

Rigby, P.W.J.; Dieckmann, M.; Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113 pp 237 - 251. "Labelling Deoxyribonucleic Acid to High Specific Activity in vitro by Nick Translation with DNA Polymerase I."

Royal, A.; Garapin, A.; Cami, B.; Perrin, F.; Mandel, J.L.; Le Meur, M.; Brégégère, F.; Gannon, F.; Le Pennec, J.P.; Chambon, P. and Kourilsky, P. (1979) Nature 279 pp 125 - 132. "The Ovalbumin Gene Region: Common Features in the Organization of Three Genes Expressed in Chicken Oviduct Under Hormonal Control."

Ruskin, B.; Krainer, A.R.; Maniatis, T. and Green, M.R. (1984) Cell 38 pp 317 - 331. "Excision of an Intact Intron as a Novel Lariat Structure During pre-mRNA Splicing in vitro

Sanger, F. and Coulson, A.R. (1978) FEBS lett. 87 pp 107 - 110. "The Use of Thin Acrylamide Gels for DNA Sequencing."

Sawatzki, G.; Anselstetter, V. and Kubanek, B. (1981) Biochem. et Biophys. ACTA 667 pp 132 - 138. "Isolation of Mouse Transferrin Using Salting-Out Chromatography on Sepharose CL-6B."

Schibler, U.; Pittet, A.C.; Young, R.A.; Hagenbüchle, O.; Tsoi, M.; Gellman, S. and Wellauer, P.K. (1982) J. Mol. Biol. 155 pp 247 - 266. "The Mouse Alpha-Amylase Gene Family."

Schreiber, G.; Dryburgh, H.; Millership, A.; Matsuda, Y.; Inglis, A.; Phillips, J.; Edwards, K. and Maggs, J. (1979) J. Biol. Chem. 254 pp 12013 - 12019. "The Synthesis and Secretion of Rat Transferrin."

Setzer, D.R.; McGrogan, M. and Schimke, R.J. (1982) J. Biol. Chem. 257 pp 5143 - 5147. "Nucleotide Sequence Surrounding Multiple Polyadenylation Sites in the Mouse Dihydrofoliate Reductase Gene."

Shahan, K. and Derman, E. (1984) Molecular and Cellular Biology 4 pp 2259 - 2265. "Tissue-Specific Expression of Major Urinary Protein (MUP) Genes in Mice: Characterization of MUP mRNAs by Restriction Mapping of cDNA and by in vitro Translation."

Shaw, P.H.; Held, W.A. and Hastie, N.D. (1983) Cell 32 pp 755 - 761. "The Gene Family for Major Urinary Proteins: Expression in Several Secretory Tissues of the Mouse."

Shen, S.H.; Slighton, J.L. and Smithies, O. (1981) Cell 26 pp 191 - 203. "A History of the Human Fetal Globin Gene Duplication."

Slightom, J.L. ; Blechl, A.E. and Smithies, O. (1980) Cell 21 pp 627-638. "Human fetal Ggamma and Agamma globin genes: Complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes."

Shewale, J.G. and Brew, K. (1982) J. Biol. Chem. 257 pp 9406 - 9415. "Effects of Fe^{3+} Binding on the Microenvironments of Individual Amino Groups in Human Serum Transferrin as Determined by Differential Kinetic Labelling."

Shore, G.C. and Tata, J. R. (1977) J. Cell Biol. 72 pp 726 - 743. "Two Fractions of Rough Endoplasmic Reticulum From Rat Liver."

Shreffler, D.C. (1960) Proc. Natl. Acad. Sci. USA 46 pp 1378 - 1384. "Genetic Control of Serum Transferrin Type in Mice."

Shreffler, D.C. (1963) J. of Heredity 54 pp 127 - 129. "Linkage of the Mouse Transferrin Locus."

Skinner, M.K. and Griswold, M.D. (1980) J. Biol. Chem. 255 pp 9523 - 9525. "Sertoili Cells Synthesize and Secrete Transferrin Like Protein."
Slightom et al (1980) see bottom page 233.
Southern, E.M. (1975) J. Mol. Biol. 98 pp 503 - 517. "Detection of Specific Sequences Among DNA Fragments Separated by Gel Electrophoresis."

Speicher, D.W. and Marchesi, V.T. (1984) Nature 311 pp 177 - 180. "Erythrocyte Spectrin is Comprised of Many Homologous Triple Helical Segments."

Stark, G.R. and Williams, J.G. (1979) Nuc. Acid Res. 6 pp 195 - 204. "Quantitative Analysis of Specific Labelled RNAs using DNA Covalently Linked to Diazobenzyloxymethyl-Papers."

Stein, J.P.; Catterall, P.K.; Means, A.R. and O'Malley, B.W. (1980) Cell 21 pp 681 - 687. "Ovomucoid Intervening Sequences Specify Functional Domains and Generate Protein Polymorphism."

Stencher, V.J. and Thornbecke, G.J. (1967) J. Immunology 99 pp 660 - 668. "Sites of Synthesis of Serum Proteins."

Stone, E.M.; Rothblum, K.N. and Schwartz, R.J. (1985) Nature 313 pp 498 - 500. "Intron-Dependent Evolution of Chicken Glyceraldehyde Phosphate Dehydrogenase Gene."

Sturtevant, A.H. (1925) Genetics 10 pp 117 - 147. "The Effects of Unequal Crossing Over at the Bar Locus in Drosophila."

Takahara, H. and Sinohara, H. (1982) J. Biol. Chem. 257 pp 2438 - 2442. "Mouse Plasma Trypsin Inhibitors."

Talmadge, K.; Vamvakopoulos, N.C. and Fiddes, J.C. (1984) Nature 307 pp 37 - 40. "Evolution of the Genes for the Beta Subunit of Human Chorionic Gonadotropin and Lutenizing Hormone."

Thomas, P.S. (1980) Proc. Natl. Acad. Sci. USA 77 pp 5201 - 5205. "Hybridization of Denatured RNA and Small DNA Fragments Transferred to Nitrocellulose."

Tosi, M.; Young, R.A.; Hagenbüchle, O. and Schibler, U. (1981) Nucl. Acid Res. 9 pp 2313 - 2323. "Multiple Polyadenylation Sites in a Mouse Alpha-Amylase Gene."

Treisman, R.; Orkin, S.H. and Maniatis, T. (1983) Nature 302 pp 591 - 596. "Specific Transcription and RNA Splicing Defects in Five Cloned Beta-Thalassaemia Genes."

Ullrich, A.; Bell, J.R.; Chen, E.Y.; Herrera, R.; Petruzzelli, L.M.; Dull, T.J.; Gray, A.; Coussens, L.; Liao, Y.C.; Tsubokawa, M.; Mason, A.; Seeburg, P.H.; Grunfield, C.; Rosen, O.M. and Ramachandran, J. (1985) Nature, 313 pp 756 - 761. "Human Insulin Receptor and its Relationship to the Tyrosine Kinase Family of Oncogenes."

Unterman, R.D.; Lynch, K.R.; Nakhaschi, H.L.; Dolan, K.P.; Hamilton, J.W.; Cohn, D.V. and Feigelson, P. (1981) Proc. Natl. Acad. Sci. USA 78 pp 3478 - 3482. "Cloning and Sequence of Several Alpha $_{2u}$ -Globulin cDNAs."

Uzan, G.; Frain, M.; Park, I.; Besmond, C.; Maessen, G.; Trépat, J.S.; Zakin, M.M. and Kahn, A. (1984) Biochem. and Biophys. Res. Comm. 119 pp 273 - 281. "Molecular Cloning and Sequence Analysis of cDNA for Human Transferrin."

Vandenbergh, J.G. (1976) Biol. Reprod. 15 pp 260 - 265. "Chromatographic Separation of Puberty Accelerating Pheromone From Male Mouse Urine."

Vannice, J.L.; Taylor, J.M. and Ringold, G.M. (1984) Proc. Natl. Acad. Sci. USA 81 pp 4241 - 4245. "Glucocorticoid-Mediated Induction of Alpha$_{I}$-Acid Glycoprotein: Evidence for Hormone-Regulated RNA Processing."

Velcich, A. and Ziff, E. (1984) Nature 312 pp 594 - 595. "Repression of Activators."

Vieira, J. and Messing, J. (1982) Gene 19 pp 259 - 268. "The pUC Plasmids, an M13 mp7-Derived System for Insertion Mutagenesis and Sequencing with Synthetic Universal Primers."

Von der Ahe, D.; Janich, S.; Scheidereit, C. Renkawitz, R.; Schutz, G. and Beato, M. (1985) Nature 313 pp 706 - 709. "Glucocorticoid and Progesterone Receptors Bind to the Same Sites in Two Hormonally Regulated Promoters."

Wahl, G.M.; Stern, M. and Stark, G.R. (1979) Proc. Natl. Acad. Sci. USA 76 pp 3683 - 3687. "Efficient Transfer of Large DNA Fragments from Agarose Gels to Diazobenzyloxymethyl-Paper and Rapid Hybridization by using Dextran Sulfate."

Wahli, W.; Dawid, I.B.; Wyler, T.; Weber, R. and Ryffel, G.U. (1980) Cell 20 pp 107 - 117. "Comparative Analysis of the Structural Organization of Two Closely Related Vitellogenin Genes in X laevis."

Weissman, C. (1984) Nature 311 pp 103 - 104. "Excision of Introns in Lariat Form."

Wiebauer, K.; Domdey, H.; Diggelmann, H. and Fey, G. (1982) Proc. Natl. Acad. Sci. USA 79 pp 7077 - 7081. "Isolation and Analysis of Genomic DNA Clones Encoding the Third Component of Mouse Complement."

Wieringa, B.; Meyer, F.; Reiser, J. and Weissman, C. (1983) Nature 301 pp 38 - 43. "Unusual Splice Sites Revealed by Mutagenic Inactivation of an Authentic Splice Site of the Rabbit Beta-Globin Gene."

Wieringa, B.; Hofer, E. and Weissmann, C. (1984) Cell 37 pp 915 - 925. "A Minimal Intron Length but No Specific Internal Sequence is Required for Splicing the Large Rabbit Beta-Globin Intron."

Williams, J.; Chasteen, D. and Moreton, K. (1982) Biochemistry J. 201 pp 527 - 531. "The Effect of Salt Concentration on the Iron Binding Properties of Human Transferrin."

Williams, J.; Elleman, T.C.; Kingston, I.B.; Wilkins, A.G. and Kuhn, K.A. (1982) Eur. J. Biochem. 122 pp 297 - 303. "The Primary Structure of Hen Ovotransferrin."

Winter, G.; Fields, S.; Gait, M.J. and Brownlee, G.G. (1981) Nucl. Acid Res. 9 pp 237 - 245. "The Use of Synthetic Oligodeoxynucleotide Primers in Cloning and Sequencing Segment 8 of Influenza Virus (A/PR/8/34)."

Winter, G. and Coulson, A.R. (1982) in "M13 Cloning Manual" M.R.C. Laboratory of Molecular Biology, Hills Road, Cambridge.

Wittels, S.; Berger, L. and Birkenmeier, C.S. (1979) Biochemistry 18 pp 5143 - 5149. "Inhibition of Intractable Nucleases with Ribonucleoside-Vanadyl Complexes: Isolation of Messenger Ribonucleic Acid from Resting Lymphocytes."

Womack, J.E. (1979) Genetics 92 pp 5 - 12 (Supplement). "Single Gene Differences controlling Enzyme Properties in the Mouse."

Woo, S.R.C.; Chandra, T.; Means, A.R. and O'Malley, B.W. (1977) Biochemistry 16 pp 5670 - 5676. "Ovalbumin Gene: Purification of the Coding Strand."

Yamada, Y.; Avvedimento, V.E.; Mudryj, M.; Ohkubo, H.; Vogeli, G.; Irani, M.; Pastan, I. and de Crombrugghe, B. (1980) Cell 22 pp 887 - 892. "The Collagen Gene: Evidence for its Evolutionary Assembly by Amplification of a DNA Segment Containing an Exon of 54 Bp."

Yang, F.; Lum, J.B.; McGill, J.R.; Moore, C.M.; Naylor, S.L.; Van Bragt, P.H.; Baldwin, W.D. and Bowman, B.H. (1984) Proc. Natl. Acad. Sci. USA 81 pp 2752 - 2756. "Human Transferrin: cDNA Characterization and Chromosomal Location."

Young, R.A.; Hagenbüchle, O. and Schibler, U. (1981) Cell 23 pp 451 - 458. "A Single Mouse Alpha-Amylase Gene Specifies Two Different Tissue-Specific mRNAs."

Youvan, D.C. and Hearst, J.E. (1979) Proc. Natl. Acad. Sci. USA 76 pp 3751 - 3754. "Reverse Transcriptase Pauses at N^2-Methylguanine During *in vitro* Transcription of *Escherichia coli* 16S Ribosomal RNA."

ADDENDUM

The EMBO Journal vol.4 no.12 pp.3167–3171, 1985

Analysis of mouse major urinary protein genes: variation between the exonic sequences of Group 1 genes and a comparison with an active gene outwith Group 1 both suggest that gene conversion has occurred between MUP genes

A.J.Clark[1], A.Chave-Cox, X.Ma and J.O.Bishop

Department of Genetics, University of Edinburgh, West Mains Road, Edinburgh EH9 3JN, UK

[1]Present address: A.F.R.C. Animal Breeding Research Organisation, West Mains Road, Edinburgh, UK

Communicated by J.O.Bishop

Here we compare the exonic sequences of four Group 1 mouse major urinary protein (MUP) genes and four Group 1 cDNA sequences. These define seven different nucleotide sequences which differ from each other by 0.35% of bases on average, and which would code for seven different MUP proteins that could probably be resolved physically into at least five classes. The sequences differ at 13 nucleotide positions and at six codons, and although they are closely related their descent cannot be described by a simple series of duplications. We also describe the sequence of another liver cDNA (pMUP15) which has diverged from the Group 1 consensus sequence in 14.6% of bases. The divergence is much greater over exons 1–3 than over exons 4–6, suggesting that an ancestral gene conversion event has occurred. pMUP15 also differs from the Group 1 genes in having a longer signal peptide sequence and a different splice configuration between exons 6 and 7. Unlike the Group 1 sequences, pMUP15 contains a potential N-linked glycosylation site. Other published work has shown that a shorter cDNA clone which is identical over their common sequence to pMUP15 codes for MUP proteins that are unusually large in size and acidic in pI. We show here that mouse urine does indeed contain a glycosylated MUP protein with those properties, presumably the product of the gene that corresponds to pMUP15.

Key words: mouse/major urinary protein genes/variation/gene conversion

Introduction

Major urinary protein (MUP) is the most abundant product of male mouse liver and MUP mRNA makes up ~5% of liver mRNA on a weight basis (Hastie and Held, 1978; Clissold and Bishop, 1981). MUP mRNA is also found in the submaxillary, sublingual, mammary and lachrymal glands (Hastie *et al.*, 1979; Shaw *et al.*, 1983). The protein is made up of a number of different components, many of which can be resolved on one- or two-dimensional gels. MUPs are under multihormonal control and different components display different patterns of hormonal regulation. For example, testosterone stimulates the expression of some genes more than others in female liver (Clissold *et al.*, 1984) and thyroxine and growth hormone induce the expression of different subsets of MUP genes (Knopf *et al.*, 1983). The mouse genome contains ~35 MUP genes (Bishop *et al.*, 1982), most of which can be classified into two groups (Group 1 and Group 2), each with ~15 members. The Group 1 genes are active, while the Group 2 genes are pseudogenes (Ghazal *et al.*, 1985). In the previous paper (Clark *et al.*, 1985) we presented the full sequence of the transcription unit of a Group 1 gene and a Group 2 gene. Here we present the exonic sequences of three more different Group 1 genes and two Group 1 cDNA clones and compare these with each other and with two published cDNA sequences (Kuhn *et al.*, 1984). The mRNA corresponding to each of these sequences could in principle be translated to give a MUP protein. Seven out of the eight code for slightly different MUP proteins. The nucleotide sequences are closely related, but the pattern of their relationships is complex, suggesting that gene conversion may have played a part in their origin. We also describe a liver cDNA clone, pMUP15, which belongs neither to Group 1 nor to Group 2 and show that in intron 6 the corresponding mRNA is spliced differently from the Group 1 mRNAs. The pMUP15 sequence differs from the Group 1 and Group 2 sequences by ~15 and 17%, respectively. The distribution pattern of the differences along the sequences suggests that an ancient gene conversion event may have occurred between them.

Results

Comparison of the sequences of different Group 1 MUP genes

In the previous paper (Clark *et al.*, 1985) we described the sequence of a complete Group 1 gene and a complete Group 2 gene. We have also determined the exonic sequences of three other MUP genes (isolated as genomic clones from BALB/c DNA libraries, Clark *et al*, 1982) and of three MUP cDNA clones isolated from a BALB/c liver cDNA library (Figure 1). The four genes had all been classified as belonging to Group 1 by DNA-RNA hybridisation methods. The sequencing data confirm their close similarity, and show that each one codes for a different MUP protein (Tables I and III). Two of the three cDNA clones (pMUP8 and pMUP11) correspond to transcripts of Group 1 genes. Both are incomplete copies of the mRNA, with a deficiency at the 5′ end. pMUP8 is 670 and pMUP11 714 nucleotides long while the length of the long form of MUP mRNA is 879 nucleotides (Clark *et al.*, 1984a). Gene BL1 and pMUP8 are identical over their common length. pMUP8 may therefore correspond to the mRNA transcribed from BL1. The coding sequence of pMUP11 corresponds to a fifth MUP protein. All of the Group 1 sequences are very closely related. The average divergence from the Group 1 consensus sequence is 0.35%. The five Group 1 sequences which are different at the DNA level also differ from each other at the protein level by at least one amino acid.

Table I shows the differences between the regions of seven MUP genes that code for mRNA. Five of these are described above. Two more are cDNA clones isolated and sequenced by Kuhn *et al.* (1984). Where possible, the part of intron 6 that is present in short Group 1 mRNA is included, with intron residue numbers. Variation is observed at a total of 19 different sites, of which 14 are present in long-form mRNA. At 18 sites only two alternative nucleotides are found. At 13 of these sites the minority nucleotide is present in only one of the seven genes, at four sites it is present in two genes and at one site it is present in three. The nucleotide differences listed in Table I affect 10 codons (Table II). Four are silent differences while six are re-

© IRL Press Limited, Oxford, England

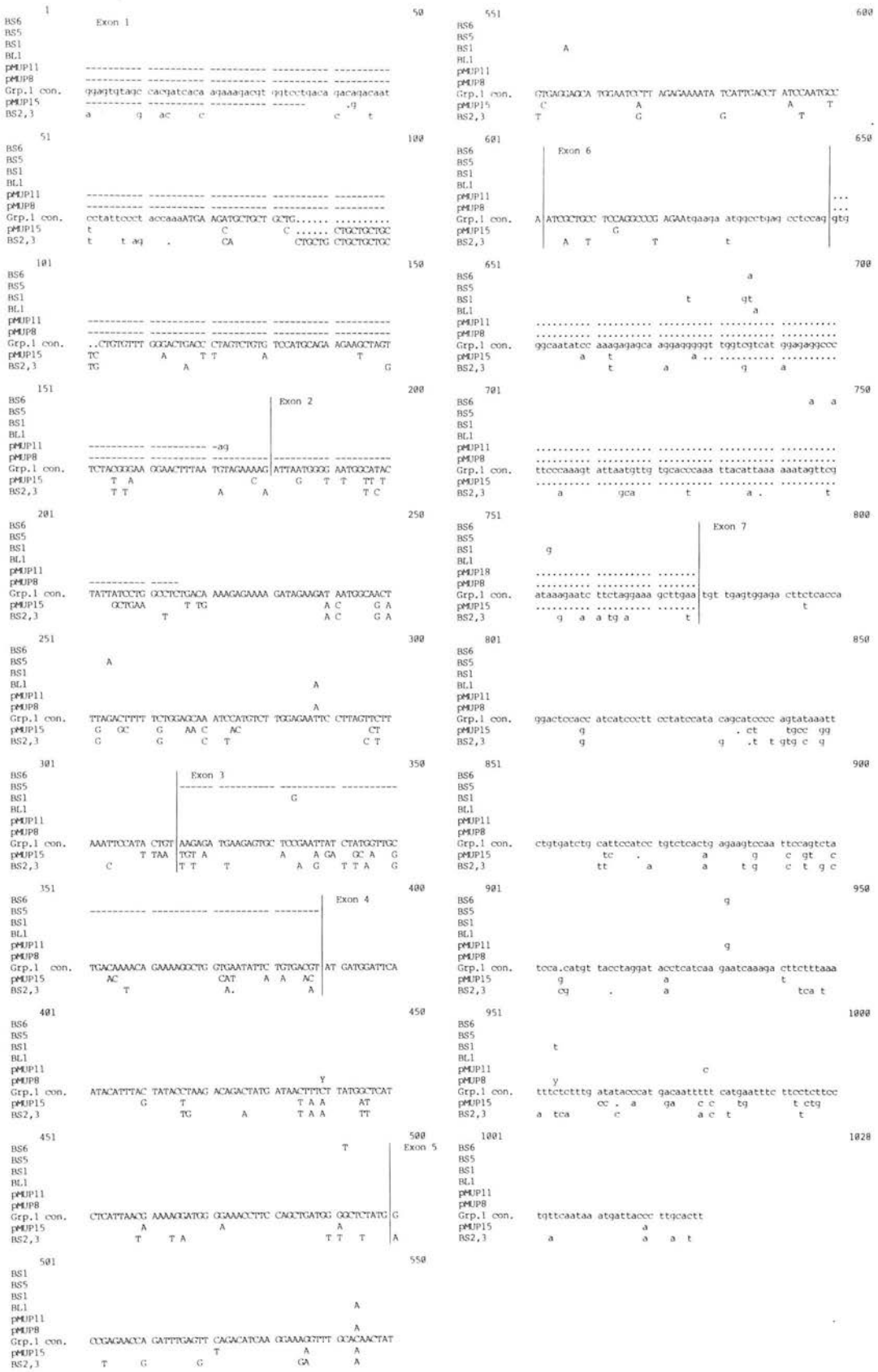


Fig. 1. Sequence comparison of exonic MUP sequences. The exonic sequences of MUP genes cloned in λ phages (BS6, etc.) and MUP cDNA clones isolated from a BALB/c liver cDNA library (pMUP11, etc.) were aligned. The first six of these are Group 1 genes and their consensus sequence (Group 1-Con) is shown immediately underneath and is compard with pMUP15 and BS2,3. The vertical lines delineate the exons.

Table I. Differences between Group 1 MUP genes

	Coding region												3′-non-coding			5′ part of intron 6				
	1	1	1	1	2	2	2	3	4	5	5		7	7	8					1
	0	5	5	9	3	7	8	0	7	2	3		2	7	3	2	3	3	4	0
	8	4	5	2	6	0	6	5	4	6	8		3	5	1	9	8	9	0	6
BS6	C	G	T	G	G	T	T	C	T	C	G		G	C	T	G	C	A	T	A
BS5	C	G	T	G	A	T	T	/	G	C	G		A	C	T	G	C	G	T	A
BS1	C	G	T	G	G	T	T	G	G	C	A		A	T	T	T	G	T	T	G
BL1	C	G	T	G	G	A	T	C	G	A	G		A	C	T	G	C	G	A	A
MUP11	/	A	G	G	G	T	T	C	G	C	G		G	C	C	/	/	/	/	/
1057	G	G	T	A	G	T	G	C	G	C	A		A	T	T	/	/	/	/	/
499	C	G	T	A	G	A	T	C	G	C	G		A	T	T	/	/	/	/	/
				*		*				*			*	*						

The numbers at the top are the residue numbers of the long form of Group 1 mRNA (coding region and 3′-non-coding region), or residue numbers within intron 6 (5′ part of intron 6). 1057 and 499 are from Kuhn *et al.* (1984). / signifies not known. The positions marked * are those at which a single minority nucleotide is present in more than one sequence.

Table II. Codon and amino acid changes at the 10 positions at which Group 1 MUP genes are known to vary

Nucleotide	Amino acid	Consensus codon		Rarer codon		
108	−5	CTG	Leu	CTA	Leu	Silent
154/5	11	GTA	Val	AGA	Arg	
192	24	GTC	Val	GTG	Val	Silent
236	39	AGA	Arg	AAA	Lys	
270	50	AAT	Asn	AAA	Lys	
286	56	TTC	Phe	GTC	Val	
305	62	TCC	Ser	TCG	Ser	Silent
474	118	GGG	Gly	GGT	Gly	Silent
526	136	CAA	Gln	AAA	Lys	
538	140	GAG	Glu	AAG	Lys	

placement differences. Of these, five could be expected to produce a significant difference in the charge of the protein.

Clone pMUP15

The third MUP cDNA clone, pMUP15, is longer than the two Group 1 cDNA clones. It extends 30 bp into the 5′-untranslated region of exon 1 while terminating at the same 3′ position. It also contains two insertions relative to the Group 1 sequences (see below). At the nucleotide level pMUP15 is considerably diverged from both the Group 1 consensus sequence (14.6%) and from the Group 2 gene BS2,3 (17.4%) (Table IV). It has an open reading frame 186 amino acids long which begins at the ATG homologous to the Group 1 and Group 2 genes and terminates at the homologous TGA stop codon. The signal peptide region is 66 nucleotides long, 12 nucleotides [four CTG (Leu) codons] longer than that of the Group 1 genes. The region corresponding to the mature protein is the same length as in the Group 1 genes. The divergence of pMUP15 and the Group 1 consensus is reflected at the protein level. The two sequences differ by 56/182 amino acids (30.7%).

The sequences of both pMUP8 and pMUP11 show that corresponding RNA transcripts were spliced to yield a 45-bp exon 6. The same pattern of splicing is seen in three other Group 1 MUP cDNA clones (Clark *et al.*, 1984a; Kuhn *et al.*, 1984). S1 nuclease protection experiments and Northern blot analyses

Table III. Amino acid patterns at the six known variant sites of Group 1 MUP genes

	1	1				
	1	3	5	5	3	4
	1	9	0	6	6	0
BS1 }	V	R	N	F	Q	K
1057 }	V	R	N	V	Q	K
BL1	V	R	K	F	K	E
MUP11	R	R	N	F	Q	E
499	V	R	K	F	Q	E
BS6	V	R	N	F	Q	R
BS5	V	K	N	/	Q	E

The patterns are arranged in order of charge, with those likely to produce a more positive protein at the top. The numbers at the top are the residue numbers of mature Group 1 MUP protein. Amino acid 56 of BS5 is not known.

Table IV. Divergence between 5′ and 3′ regions of various MUP genes

Comparison	% Divergence		
	Total sequence	Exons 1−3	Exons 4−7
Group 1-Con. versus pMUP15	14.6	20.4	11.1
BS2,3 versus pMUP15	17.4	22.2	14.4
Group 1-Con. versus BS2,3	13.0	12.2	13.5
Consensus versus Group 1-Con.	4.5	4.9	4.3
Consensus versus pMUP15	8.9	15.2	5.2
Consensus versus BS2,3	7.2	6.8	7.7

Group 1-Con. refers to the consensus sequence of the six Group 1 genes sequenced. Consensus refers to the consensus sequence of the comparison between Group 1-Con., MUP15 and BS2,3.

confirm that this is the prevalent mode of splicing of the Group 1 MUP gene transcripts (Clark *et al.*, 1984a). In contrast the transcript of pMUP15 has been spliced to yield a 76-bp exon 6. In principle these different splicing patterns could be due to differences in the region of the Group 1 donor site, in the MUP15 donor site (within intron 6 of the Group 1 gene), or in the common acceptor site, or it could be due to some combination of such differences (Busslinger *et al.*, 1981; Fukumaki *et al.*, 1982; Felber *et al.*, 1982; Treisman *et al.*, 1982, 1983). The gene corresponding to MUP15 has not yet been isolated, so that a complete comparison of the introns is not possible. However, examination of the homologous sequences in the Group 1 genes shows that the splice point that generated pMUP15 occurred 31 bp downstream to that which generated the two Group 1 cDNA clones. In the Group 1 consensus sequence (Figure 1) the sequence immediately 3′ to this point is GTTGGT, which is quite close to the donor site consensus GTARGY. Presumably this sequence, or a closely related counterpart, exists in the MUP15 gene and acts as a functional donor site. However, since there is no divergence between pMUP15 and the Group 1 consensus sequence in the region of the donor site which generates the shorter (45 bp) exon 6, the precursor of pMUP15 mRNA may itself be spliced in two configurations.

Minor MUP proteins are glycosylated

Szoka and Paigen (1978) failed to detect glycosylation of MUP proteins, and indeed the known sequences of the Group 1 genes, which probably code for the bulk of MUP, do not contain an N-linked glycosylation site. However, Kuhn *et al.* (1984) found a potential glycosylation site in the sequence of p199, a cDNA clone that is less complete than MUP15, but identical to it over

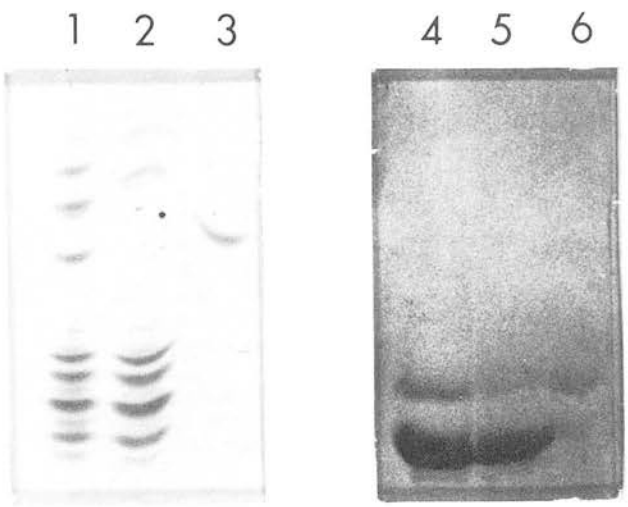


Fig. 2. Glycosylation of a minor MUP component. Purified MUP was fractionated on a column of Con-A-Sepharose. The fractions were resolved by IEF (1–3) and SDS-PAGE (4–6) and stained with Coomassie blue. **Tracks 1** and **4**, fractionated; **2** and **5**, unbound, **3** and **6**, bound to Con-A-Sepharose.

their common length. By hybrid-selection followed by cell-free translation of the hybridised mRNA in the presence of the dog pancreas membrane fraction to effect post-translational modification, Kuhn *et al.* (1984) showed that p199 is homologous in sequence to mRNA that codes for a group of four proteins that migrate more slowly through an SDS gel than most of the MUPs and are also more negative in charge. That is, they show the characteristics expected of glycosylated MUPs.

Our interest in MUP15 led us to ask directly whether mouse urine contains glycosylated MUP proteins, and whether these are the proteins identified by Kuhn *et al.* (1984) as coded for by a gene or a small gene family which is homologous to p199 = MUP15. The MUP proteins were isolated from the urine of both BALB/c and C57BL mice and fractionated on Con-A-Sepharose columns. Both the fractions and the unfractionated MUP proteins were resolved on SDS gels and also separately by isoelectric focusing, and the gels were stained with a general protein dye. Urine from both strains contained a minor fraction of MUP protein that binds specifically to Con-A-Sepharose, and which migrates more slowly on SDS gels and is more negatively charged than the bulk of the MUP protein (Figure 2, only BALB/c shown). Only the band that was retained by the Con-A-Sepharose was stained with a thymol-H_2SO_4 stain specific for glycosylated proteins (data not shown).

Discussion

Group 1 genes are active

All the available evidence indicates that BS6 is a true gene in that it appears to have all the DNA sequences necessary for correct transcription, processing and translation. The BS6 promoter region is active in fibroblasts in association with the SV40 enhancer (P.Ghazal and J.O.Bishop, unpublished experiments). The translation of BS6 and three other Group 1 genes, BS1, BS5 and BL1, would yield acidic proteins of the size of MUP. A number of lines of evidence show that Group 1 genes are the most abundantly expressed in both BABL/c and C57 livers. (i) 6/8 independently isolated cDNA clones correspond very closely (<0.5% divergence) to the Group 1 consensus sequence (Clark *et al.*, 1984a; Kuhn *et al.*, 1984; and this paper). (ii) Under stringent washing conditions Group 1 genes hybridise considerably more strongly to end-labelled liver mRNA than do other isolated MUP genes (Clark *et al.*, 1982). (iii) From MUP mRNA hybrid-selection-translation experiments and signal intensities in Northern blot hybridisations, Kuhn *et al.* (1984) estimated that the ratio of Group 1 sequences to p199-type (MUP15-type) sequences in total C57BL liver mRNA is ~10. Similarly, Clark *et al.* (1984a) found the ratio of Group 1 to Group 2 mRNA in BALB/c liver mRNA to be ~10.

About half of the 35 MUP genes are pseudogenes (Ghazal *et al.*, 1985). Of the active genes, one or two are identical or very similar in sequence to MUP15 (Kuhn *et al.*, 1984). Thus out of the total of ~18 active genes we have extensive data on the coding sequences of eight, including about half of the Group 1 genes.

Group 1 genes may be related through gene conversion

The seven Group 1 sequences contain a non-random distribution of differences at 19 sites. At 13 sites one sequence differs from all the others. We have argued that the 45-kb units that contain the Group 1 genes are closely related to a common ancestor (Clark *et al.*, 1984b; Ghazal *et al.*, 1985). If so, these single-incidence differences may be explained as relatively recent mutation events. The three different bases present at another site are also consistent with independent mutational changes. At the remaining five sites two different nucleotides are each present in more than one gene (see Table I). The distribution of changes at these five sites may be used in an attempt to establish a pedigree of relationships between the genes. The inter-relationship of the entire set can be explained only by supposing that multiple mutations to the same base have occurred at several of the five sites. However, there is no evident reason why this should have occurred. Two of the sites are in the 3′ non-coding region, and the others correspond to Leu→Leu, Asn→Lys and Glu→Lys, respectively. Furthermore, no simple series of recombination events between any two postulated ancestral combinations of the two variants at each site can explain the observed sequences. The most likely explanation of these relationships would seem to involve gene conversion superimposed on gene duplication.

Group 1 MUP genes code for different protein products

Five of the six Group 1 sequences described here code for proteins that differ from one another by at least one amino acid. Two of the sequences described by Knopf *et al.* (1983) code for two additional Group 1 proteins. The greatest pair-wise difference is three amino acids. The pattern of differences present in the seven protein variants is shown in Table III, where the proteins are listed in order of probable net charge. It is likely that these proteins could be resolved by sensitive methods into at least five components. In fact the two-dimensional gels of Knopf *et al.* (1983) and Kuhn *et al.* (1984) show that the main size class of MUP protein resolves into five charge components.

pMUP15 may have been generated by an ancient gene conversion event

In contrast to the differences between the Group 1 consensus and BS2,3, those between pMUP15 and the Group 1 consensus and those between pMUP15 and BS2,3 are distributed non-uniformly along the sequence. In the latter comparisons the first three exons are considerably more diverged than the last four (Table IV). One possible explanation for this is that the 3′ regions of MUP genes are under stronger selective constraints than the 5′ regions and are not as free to diverge. If this were the case, the ratio

of replacemnet site to silent site differences would be greater in exons 1–3 than in exons 4–6. The corrected frequencies of replacement site and silent site differences (Perler *et al.*, 1980) are 14.4% and 11.6%, respectively, a ratio of 1:2. The frequencies over exons 1–3 are 27.5% and 22.4%, a ratio of 1.2 and over exons 4–6 they are 9.8% and 8.9%, a ratio of 1.1. Thus the disproportionate conservation of exons 4–6 of pMUP15 and the Group 1 consensus seems not to be due to selection. Another possible explanation for the non-uniform divergence between pMUP15 and other MUP genes is gene conversion. The fact that pMUP15 is equally divergent from the Group 1 consensus and BS2,3 suggests that the 3′ region of a gene ancestral to both the Group 1 and Group 2 genes may have converted the homologous regions of MUP15 (or *vice versa*) at about the same time as the onset of divergence of the Group 1 and Group 2 genes. Alternatively, the 5′ regions of MUP15 or the Group 1/Group 2 ancestor may have been converted by a more distantly related MUP gene. The two different regions of divergence define a notional junction between exons 1–3 and exons 4–7, with twice as much divergence to the 5′ side as to the 3′ side of the junction (Table IV). Genomic clones corresponding to pMUP15 are not yet available so that it is not possible to determine precisely the positions of the boundaries of the proposed converted region. However, given that the difference in degree of divergence occurs between exons 3 and 4, a boundary must occur in intron 3. In both BS6 and BS2,3 this intron contains a long stretch of the repeated dinucleotide GT (Clark *et al.*, 1985), which has been found at the boundaries of putative regions of gene conversion (Proudfoot and Maniatis, 1980; Shen *et al.*, 1981).

Materials and methods

Cloned DNA

The isolation of MUP genomic clones and subclones is described in Clark *et al.* (1982, 1984b) and Bishop *et al.* (1982). pMUP11 and pMUP15 were isolated from a cDNA library prepared by using poly(A)$^+$ RNA from female mouse (BALB/c) liver as a template for oligo(dT)-primed DNA synthesis by reverse transcriptase. Single-stranded cDNA >1 kb in size was isolated by polyacrylamide gel electrophoresis and made double-stranded with DNA polymerase I and reverse transcriptase. After treatment with nuclease S1, fragments >1 kb were again isolated by polyacrylamide gel electrophoresis, treated with DNA polymerase I and cloned into *Sma*I-cleaved pUC8. White colonies were isolated into microtitre plates and screened with mouse DNA isolated from the genomic subclones pBS6-5-5 and pBS2-2-2 (Bishop *et al.*, 1982) by the method of Gergen *et al.* (1979). The methods used differed only in minor ways from standard procedures (e.g., Maniatis *et al.*, 1982) and will be published in detail elsewhere (A.Chave-Cox, unpublished data). The propagation of bacteriophage and plasmid clones and the isolation of DNA were carried out as described (Clissold and Bishop, 1982; Clark *et al.*, 1982; Bishop *et al.*, 1982).

DNA sequencing

The exonic sequences of BS1, BS5 and BL1 were obtained by sequencing from nearby restriction sites using subclones generated in M13mp8 and M13mp9 and M13tg130 and M13tg131 (Kieny *et al.*, 1984). The cDNA inserts were excised from pUC8 at the polylinker, recloned into M13mp8 and sequenced by the progressive method of Hong (1982).

Fractionation of MUP with Con-A-Sepharose

Urine was collected from 8- to 10-week-old mice by bladder massage, dialysed overnight against distilled water, and fractionated on a column of Sephadex G-100 developed with 0.2 M NH_4HCO_3. Fractions containing MUP were lyophilised and dissolved in 50 mM Tris-HCl, 500 mM NaCl, 1 mM $MgCl_2$, 1 mM $MnCl_2$, 1 mM $CaCl_2$, pH 6.0. 2 ml (20 mg) of protein was passed over a 7 ml column of Con-A-Sepharose-4B (flow-rate 3.6 ml/h) which was thoroughly washed with the same buffer, and the bound fraction was eluted with the same, containing 0.17 M Na_2-tetraborate. The agarose IEF gel (170 mm running length × 110 × 3 mm) contained 0.74% agarose, 8.75% sorbitol, 3.1% pH 4–6 ampholines and 0.8% pH 3–10 ampholines (both Pharmacia). The electrode solutions were 0.5 M H_2SO_4 and 1 N NaOH. Focusing was for 90 min at 1000 V and 30 min at 1500 V. Acrylamide IEF was performed on 5% preformed plates (LKB Pagplate 1804-111) containing 2% pH 4–5 ampholines. The electrode solutions were 1 M H_3PO_4 and 1 M glycine. Focusing was for 3 h at 1400 V. The SDS-PAGE separating gels contained 15% acrylamide, and the spacers 6%, and the gels were run for 6–8 h at 0.12 mA per mm^2 cross-sectional area. Glycosylated proteins were stained by washing polyacrylamide gels twice (2 h each) in 10 volumes of 25% isopropanol, 10% acetic acid, for 2 h in the same solution containing 0.2% thymol, and for 2.5 h at 35°C in 80% H_2SO_4, 20% ethanol.

Acknowledgements

We are grateful to Morag Robertson and Melville Richardson for technical assistance and to the MRC and the Cancer Research Campaign for financial support.

References

Bishop,J.O., Clark,A.J., Clissold,P.M., Hainey,S. and Francke,U. (1982) *EMBO J.*, **1**, 615-620.
Bishop,J.O., Selman,G.G., Hickman,J., Black,L., Saunders,R.D.P. and Clark, A.J. (1985) *Mol. Cell Biol.*, **5**, 1591-1600.
Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349-383.
Busslinger,M., Moschonas,N. and Flavell,R.A. (1981) *Cell*, **27**, 289-298.
Clark,A.J., Clissold,P.M. and Bishop,J.O. (1982) *Gene*, **18**, 221-230.
Clark,A.J., Clissold,P.M., Al-Shawi,R., Beattie,P. and Bishop,J.O. (1984a) *EMBO J.*, **3**, 1045-1052.
Clark,A.J., Hickman,J. and Bishop,J.O. (1984b) *EMBO J.*, **3**, 2055-2064.
Clark,A.J., Ghazal,P., Bingham,R., Barrett,D. and Bishop,J.O. (1985) *EMBO J.*, **4**, 3159-3165.
Clissold,P.M. and Bishop,J.O. (1981) *Gene*, **15**, 225-235.
Clissold,P.M. and Bishop,J.O. (1982) *Gene*, **18**, 211-220.
Clissold,P.M., Hainey,S. and Bishop,J.O. (1984) *Biochem. Genet.*, **22**, 379-387.
Felber,B.K., Orkin,S.H. and Hamer,D.H. (1982) *Cell*, **29**, 895-902.
Fukumaki,Y., Ghosh,P.K., Benz,E.J., Jr., Reddy,V.B., Lebowitz,P., Forget,B. and Weissman,S. (1982) *Cell*, **28**, 585-593.
Gergen,J.P., Stern,R.H. and Wensink,P.C. (1979) *Nucleic Acids Res.*, **7**, 2115-2136.
Ghazal,P., Clark,A.J. and Bishop,J.O. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4182-4185.
Hastie,N.D. and Held,W.A. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 414-417.
Hastie,N.D., Held,W.A. and Toole,J.J. (1979) *Cell*, **17**, 449-457.
Hong,G.F. (1982) *J. Mol. Biol.*, **158**, 539-549.
Knopf,J.L., Gallagher,J.A. and Held,W.A. (1983) *Mol. Cell. Biol.*, **3**, 2232-2240.
Kuhn,N.J., Woodworth-Gutai,M., Gross,K.W. and Held,W.A. (1984) *Nucleic Acids Res.*, **12**, 6073-6090.
Maniatias,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning, A Laboratory Manual*, published by Cold Spring Harbor Laboratory Press, NY.
Perler,F., Efstratiadis,A., Lomedico,P., Gilbert,W., Kolodner,R. and Dodgson, J. (1980) *Cell*, **20**, 555-565.
Proudfoot,N.J. and Maniatis,T. (1980) *Cell*, **21**, 537-544.
Shaw,P.H., Held,W.A. and Hastie,N.D. (1983) *Cell*, **32**, 755-761.
Shen,S., Slightom,J.L. and Smithies,O. (1981) *Cell*, **26**, 191-203.
Szoka,P.R. and Paigen,K. (1978) *Genetics*, **90**, 597-612.
Treisman,R., Proudfoot,N.J., Shander,M. and Maniatis,T. (1982) *Cell*, **29**, 903-911.
Treisman,R., Orkin,S.H. and Maniatis,T. (1983) *Nature*, **302**, 591-596.

Received on 27 June 1985; revised on 16 September 1985