Volume 6 Number 10 1979 Nucleic Acids Research

Sequence analysis of cloned cDNA encoding part of an immunoglobulin heavy chain

John Rogers, Patrick Clarke and Winston Salser

Molecular Biology Institute and Department of Biology, University of California, Los Angeles, CA 90024, USA

Received 8 March 1979

ABSTRACT

The recombinant plasmid pH21-1 consists of mouse-derived complementary DNA (cDNA) in the *E. coli* plasmid pMB9. The mouse insertion has been completely sequenced, and encodes the CH3 domain and half the CH2 domain of the immunoglobulin γ1 heavy chain. The predicted amino acid sequence differs at several positions from that previously published for this protein. The pattern of codon usage resembles that in some other eukaryotic messenger RNAs. A computer program has been used to predict the optimum secondary structure for the mRNA encoding the CH3 domain and the inter-domain junction.

INTRODUCTION

With the development of elegant techniques (1-4) for cloning DNA complementary to eukaryotic messengers (cDNA), it has become possible to prepare large quantities of many such cDNAs for use both in sequence analysis, and in locating the same sequences in cellular DNA and RNA. The immunoglobulin heavy chain genes are of particular interest, since not only do they undergo the generation of diversity and joining of variable and constant regions which so far appear to be peculiar to immunoglobulins, but also they are members of a developmentally regulated multigene family, and the ancestral heavy chain gene apparently arose by tandem duplication of a still smaller genetic unit, the immunoglobulin domain.

A recombinant DNA plasmid, pH21-1, has been constructed containing sequences from the heavy chain messenger of the IgG1-producing mouse myeloma MOPC21 (R. Wall, K. Toth, G. Paddock, R. Higuchi, and W. Salser, unpublished). Here we report the complete restriction map and DNA sequence of the mouse-derived insert in this cloned plasmid. It contains 459 nucleotides encoding the C-terminal 1½ domains of the γ1 constant region. Some characteristics of the coding sequence are discussed.

MATERIALS AND METHODS

Construction of pH21-1

Construction of a series of cDNA clones containing κ light chain mRNA sequences has been reported (5). The mRNA was from solid tumors of the IgG1-producing mouse myeloma MOPC21 (6). In these same experiments, cDNA clones containing heavy chain mRNA sequences were constructed by the same methods using the 16-17S fraction of the mRNA, in which the principal species is the heavy chain messenger (7). The cDNA was inserted into the EcoRI site of plasmid pMB9, by means of poly-dA and poly-dT "tails" on the respective 3' ends of insert and plasmid, and the recombinant plasmids were cloned in *E. coli* (R. Wall, K. Toth, G. Paddock, R. Higuchi, and W. Salser, unpublished). One clone gave a distinct peak of hybridization with 16-17 S MOPC21 mRNA (R. Wall and D. DeBorde, personal communication) and was designated pH21-1.

Restriction Analysis

Plasmid DNA was prepared as in refs. (8) and (9). EndoR.TaqI was prepared by S. Hendrich using an unpublished technique of M. Komaromy. It was used at 65°C in 10 mM HEPES pH 8.4, 6 mM $MgCl_2$, 6 mM β-mercaptoethanol , 25 mM $(NH_4)_2SO_4$, 100 μg ml^{-1} gelatin. Enzymes HaeIII and HincII were purchased from New England Biolabs, and AluI, HhaI and HinfI from Bethesda Research Laboratories; they were used as suggested by the suppliers.

Polyacrylamide gel electrophoresis of restriction fragments for preparative or analytical purposes was carried out in 20 x 40 cm gels made with 6% acrylamide, 0.2% methylene bisacrylamide, 12% glycerol, in running buffer. Running buffer was 50 mM Tris borate, 1 mM EDTA (TBE). For strand separation the gel consisted of 4% acrylamide plus 0.14% methylene bisacrylamide in the running buffer, which was 36 mM Tris base, 30 mM NaH_2PO_4, 1 mM EDTA. Samples for 6% gels were loaded in the restriction buffer, diluted if necessary, plus ¼ volume of dye solution (0.03% bromphenol blue plus xylene cyanol in 20% glycerol). Samples for strand separation were prepared in 90 μl of the same dye solution made up to 300 μl 0.3 M NaOH, and heated at 37°C for 3 minutes immediately before loading. After electrophoresis, DNA fragments were visualized by ethidium bromide staining and UV fluorescence, or, if end-labelled, by autoradiography. Elution of DNA fragments was carried out as in ref. (10), except that incubation of the crushed gel in elution solution was for 2 days at 42°C.

DNA Sequence Analysis

Restriction fragments purified from digests of 150-300 µg of the 6-kb plasmid pH21-1 were treated with bacterial alkaline phosphatase (Worthington Biochemical, grade f) in 10 mM Tris-HCl pH 8.0, for 30 minutes at 37°C. This mixture was phenol-extracted thrice, ether-extracted twice, ethanol-precipitated, redissolved in 5 mM Tris pH 9.5, 0.1 mM spermidine, and 0.01 mM EDTA.Na_3, and then denatured by boiling. 5' end labelling with ^{32}P was carried out as in ref. (10), using Tris-HCl rather than sodium glycine as buffer. T4 polynucleotide kinase was purchased from PL Biochemicals, and $\gamma^{32}P$-ATP was made from $^{32}P_i$ (ICN Pharmaceuticals) (10). After ethanol precipitation, the labelled ends of the DNA were separated by strand separation, or by another restriction cleavage and electrophoretic separation of the fragments.

For sequencing we used four of the base-specific cleavage reactions of Maxam and Gilbert (10), entitled G>A, A>C, T+C, and C. A fifth reaction cleaving at A+G was performed as follows (A. Maxam, personal communication). End-labelled DNA and 1 µg carrier DNA were made up to 30 µl in 17 mM sodium citrate pH 4.0 and heated for 10 minutes at 90°C. 2 µl of 1 M NaOH was added and the mixture sealed in a capillary and heated for a further 30 minutes at 90°C. 20 µl of urea-dye mixture (10) was then added and the sample was ready for loading on a ladder gel.

Ladder gels (20% acrylamide, 0.7% methylene bisacrylamide, 7 M urea in TBE) were made, loaded and run as in ref. (10). In our later runs we used thin gels (11), of thickness 0.32 mm instead of the regular 1.6 mm, and found considerably improved resolution of bands. Ladder gels were autoradiographed at -70°C on Cronex 4 X-ray film with Dupont Hi-plus intensification screens.

Secondary Structure Prediction

The most stable secondary structure for the RNA represented by the sequence was predicted using the computer program of Studnicka *et al*. (12). This program will examine a large number of possible regions of base pairing to find that combination of regions which forms the most stable structure. The program begins by cataloguing all possible regions of 2 or more consecutive base pairs. There were 5231 regions in this "primary region catalogue" for the sequence considered here. It would be prohibitively expensive to consider all of these regions in a single computation cycle.

Therefore we rank the regions and carry out the computation in several cycles. The ranking uses a weighting function which is the sum of the energy of the region itself, divided by the square root of its length, and the energy of the best "local structure" which can be obtained by combining the region with all neighboring regions which are separated from it by less than 10 nucleotides on either strand (W. Salser and L. Nagy, in preparation). Where two primary regions would overlap, a "branch migration" procedure is used to determine the most stable non-overlapping combination of parts of the two primary regions.

The 150 regions with the most favorable weighting factors were chosen for the first cycle and the 100 most stable combinations of these regions were computed, the energies being calculated using the rules given in ref. 13. All these structures shared certain features which permitted us to break up the computation into three smaller jobs for the second cycle. In this second cycle all regions down to a weighting factor of 39 were considered (the equivalent of the top 900 of the original 5231 regions). The alternative structures for the second cycle were in turn examined for common features; these allowed us to subdivide the sequence into eight jobs for the final cycle, in which all regions of two or more base pairs were considered. In theory it would be possible to improve the structure slightly by considering single G-C pairs. For example, according to our base-pairing rules (13), the structure $\begin{array}{l}5'\ \mathrm{CUUC\text{-}GU}\\ 3'\ \dot{\mathrm{G}}\dot{\mathrm{A}}\text{-}\dot{\mathrm{G}}\mathrm{U}\dot{\mathrm{C}}\dot{\mathrm{A}}\end{array}$ is more stable than the computed structure $\begin{array}{l}5'\ \mathrm{CUUCGU}\\ 3'\ \dot{\mathrm{G}}\dot{\mathrm{A}}\mathrm{GU}\dot{\mathrm{C}}\dot{\mathrm{A}}\end{array}$ by 2 kcal. This additional refinement of the structure was not performed.

Biosafety Precautions

P3 physical containment was used throughout for growth of transformed bacteria. The initial isolation of pH21-1 had been carried out in *E. coli* χ1849, an EK1 host, in compliance with the Asilomar Guidelines in effect at that time. When the NIH Guidelines (14) were issued, pH21-1 was transferred to *E. coli* χ1776, an EK2 host, and all subsequent experiments were conducted in accordance with those Guidelines.

RESULTS

Restriction Analysis

Various restriction enzymes were tested in parallel digests of pH21-1 and pMB9 DNA. Since the mouse sequence was inserted at the single EcoRI site in pMB9, it was expected that comparison of the digests would show one band unique to pMB9 which was replaced by one or more bands unique to pH21-1. This was found to be the case, and all the enzymes tested indicated an inserted segment of about 560 bp in length.

Mapping was helped by the observation that each pH21-1 digest exhibited one pair of submolar bands not shown by pMB9. Consistent values for the size of the insert could only be deduced if the doublet was considered to represent a single restriction fragment for each enzyme. It probably resulted from an approx. 44-bp deletion at a single site in a minor population of the plasmid DNA, since on preparative gels the larger band was homogeneous, while the smaller band was seen to be only one of up to seven equally spaced bands, the others being faint (Figure 1). Since no such heterogeneity is seen in pMB9, deletions of various sizes probably occurred in the mouse insert or the A.T joints. There is no repetitive sequence in the mouse insert which could account for this (see below), and the restriction fragment affected always contained the left-hand A.T joint, which therefore could be the site of the deletions.

It was possible to locate some of the restriction sites in the insert

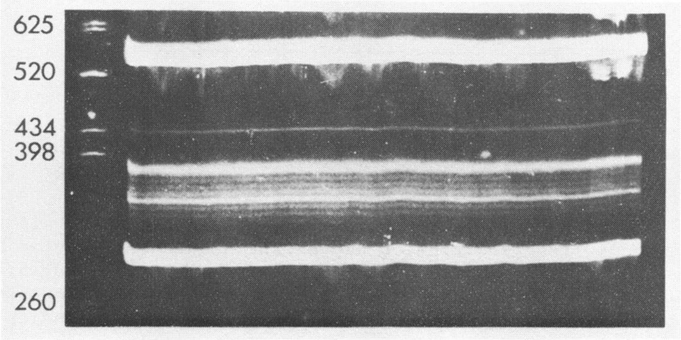

Figure 1. Preparative gel of TaqI-digested pH21-1 (300 µg), showing heterogeneity in one band. The uppermost and brightest member of the set is approximately 372 bp in length. The prominent doublets above and below the set are fragments from the pMB9 parts of the plasmid. The side lane contains HaeIII-digested pMB9, with fragment sizes marked in basepairs.

without further digests using the positions given by Maniatis et al. (2) and I. Cummings (unpublished) for the restriction sites in pMB9 nearest the EcoRI site. Thus of the four pH21-1-specific fragments seen in MboII digests, the 1100-bp fragment must cover the right-hand insert-pMB9 junction, the 460-bp fragment with its 420-bp minor band must cover the left-hand junction, and the 116-bp and 75-bp fragments must be internal. Knowing the positions of some sites and the general location of the deletion-prone region, it was then possible to locate most of the TaqI, HaeIII and AluI sites from the single-enzyme digests. The map was refined by isolating the two pH21-1-specific TaqI fragments and digesting them with HaeIII, AluI, and HaeIII plus MboII. The resulting map was confirmed by subsequent sequencing except for the discovery of an extra MboII site near the middle. The final map is given in Figure 2.

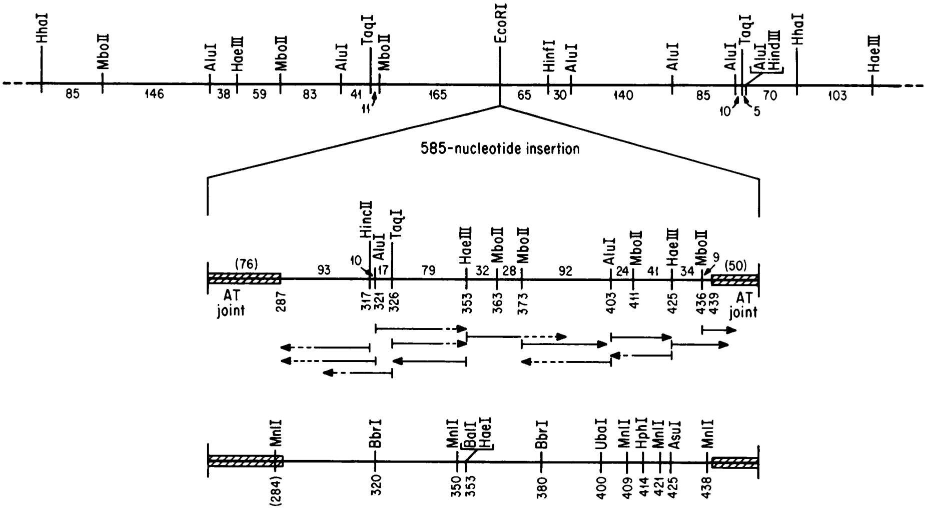

Figure 2. Restriction map of the insert and surrounding region in pH21-1. Top line: Restriction map of the pMB9 sequences around the EcoRI site, as deduced from analysis of the insert-containing HhaI fragment of pH21-1. All sites for the enzymes indicated are shown. Distances between the sites are in nucleotides. Middle line: Restriction map of the sequences inserted into the EcoRI site of pMB9. All sites for the enzymes indicated are shown. Distances between sites are in nucleotides. The numbers below the line denote the codons in or after which the cuts are made. Arrows indicate where sequences were obtained. We did not sequence continuously across all restriction sites, but the continuity of the coding sequence implies that no nucleotides were omitted. Bottom line: Additional sites inferred from the nucleotide sequence, with the numbers of the codons at which the cuts are expected to be made. MnlI and HphI, like MboII, make cuts offset from the recognition sequences by 5-10 nucleotides. There are no sites for EcoRI, BamHI, PstI, HindIII, HhaI, HinfI, HpaII, or HaeII in the insert.

While preparing the HhaI fragment containing the insert for sequencing, we did further restriction analysis in its outer parts, so as to extend the known restriction map for pMB9 (Figure 2). Since parallel digests of whole pH21-1 and pMB9 never showed any differences in bands not covering the EcoRI site, this map is believed to represent the "wild-type" pMB9 structure.

DNA Sequence Analysis

The restriction sites used for Maxam-Gilbert sequence analysis (10) are indicated in Figure 2. Representative 'ladders' are shown in Figure 3, and the complete DNA sequence is in Figure 4. It covers 459 nucleotides between the A.T joints, and encodes amino acids 287 to 439 in the sequence of Adetugbo (15,16). This includes half of the C_H2 domain (amino acids 228-334)

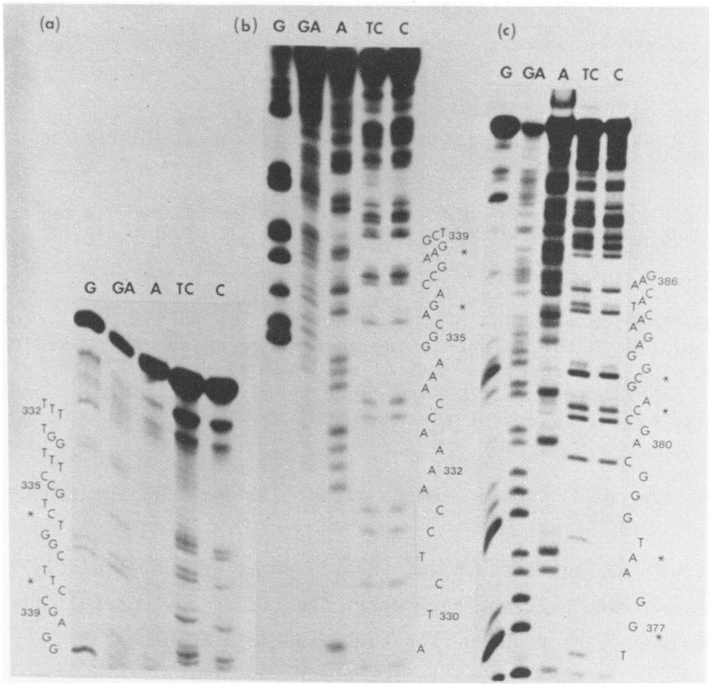

Figure 3. "Ladders" showing sequences from pH21-1. Sequences are read from bottom to top. Codons are numbered. Asterisks indicate codons differing from the reported amino acid sequence (16). (a) Complementary strand with 5' label at HaeIII site at codon 353, 3' cut with TaqI; thin gel. (b) Coding strand covering the same region, with 5' label at TaqI site at codon 326, 3' cut with MboII; regular gel. (c) Coding strand with 5' label at MboII site at codon 373, 3' cut with AluI; thin gel.

```
                        287             291                 *               *
                        Glu Glu Gln Phe Asn Ser Thr Phe Arg Ser Val Ser Glu Leu
                5'-(T)n-GAG GAG CAG TTC AAC AGC ACT TTC CGC TCA GTC AGT GAA CTT
                         1                      20                           40

301                                     311
Pro Ile Met His Gln Asp Trp Leu Asn Gly Lys Glu Phe Lys Cys Arg Val Asn Ser Ala
CCC ATC ATG CAC CAA GAC TGG CTC AAT GGC AAG GAG TTC AAA TGC AGG GTC AAC AGT GCA
                    60                      80              HincII  100

321                                     331                 *       *
Ala Phe Pro Ala Pro Ile Glu Lys Thr Ile Ser Lys Thr Lys Gly Arg Pro Lys Ala Pro
GCT TTC CCT GCC CCC ATC GAG AAA ACC ATC TCC AAA ACC AAA GGC AGA CCG AAG GCT CCA
AluI                TaqI                        140                         160

341                                     351
Gln Val Tyr Thr Ile Pro Pro Pro Lys Glu Gln Met Ala Lys Asp Lys Val Ser Leu Thr
CAG GTG TAC ACC ATT CCA CCT CCC AAG GAG CAG ATG GCC AAG GAT AAA GTC AGT CTG ACC
                    180                       HaeIII                        220

361                                     371                     *   *
Cys Met Ile Thr Asp Phe Phe Pro Glu Asp Ile Thr Val Glu Trp Gln Trp Asn Gly Gln
TGC ATG ATA ACA GAC TTC TTC CCT GAA GAC ATT ACT GTG GAG TGG CAG TGG AAT GGG CAG
      230           MboII       MboII           260                         280

 *   *  383                                 391
Pro Ala Glu Asn Tyr Lys Asn Thr Gln Pro Ile Met Asp Thr Asp Gly Ser Tyr Phe Val
CCA GCG GAG AAC TAC AAG AAC ACT CAG CCC ATC ATG GAC ACA GAT GGC TCT TAC TTC GTC
                    300                         320                         340

401                                     411
Tyr Ser Lys Leu Asn Val Gln Lys Ser Asn Trp Glu Ala Gly Asn Thr Phe Thr Cys Ser
TAC AGC AAG CTC AAT GTG CAG AAG AGC AAC TGG GAG GCA GGA AAT ACT TTC ACC TGC TCT
         AluI       360   MboII                 380                         400

421                                     431                             439
Val Leu His Glu Gly Leu His Asn His His Thr Glu Lys Ser Leu Ser His Ser Pro
GTG TTA CAT GAG GGC CTG CAC AAC CAC CAT ACT GAG AAG AGC CTC TCC CAC TCT CCT-(A)n-3'
                HaeIII               430         MboII                  459
```

Figure 4. Nucleotide sequence of the cDNA insert in pH21-1. Only the coding strand is shown, with the encoded amino acid sequence above. Positions where this differs from the reported amino acid sequence (16) are marked by asterisks. Restriction sites are underlined. Additional underlining indicates less certain parts of the sequence.

and all of the C_H3 domain (amino acids 335-440) except the C-terminal glycine.

The first 30 nucleotides of the sequence, and particularly the first ten, are not entirely certain; they were far from any useful restriction sites and could only be read knowing the amino acid sequence.

The individual nucleotides in the sequences of the A.T joints were only partially resolved, but measurements of the complete poly-A blocks on ladders indicated lengths of 76 ±15 bp on the left and 50 ±15 bp on the right. The left-hand poly-A tail contains at least two T residues. Similar A ⟶ T transversions in A.T joints have been observed previously (2).

DISCUSSION

Amino Acid Substitutions

The DNA sequence implies several substitutions in the published amino acid sequence (15,16), as listed in Table I. In all but the first, the DNA sequence appears certain. There are three possible sources for these discrepancies.

(1) _In vivo_ variation. It is conceivable that since the separation of the MOPC21 tumor line used to make the cDNA from the MOPC21 (P3K) line used for protein sequencing in Cambridge, one or both lines had either accumulated 'somatic' mutations or switched to expression of a different γ chain gene. Somatic mutations have been documented in subclones of the Cambridge line (17-20), and switches in γ2 gene expression have been induced in another myeloma line _in vitro_ (21); serological evidence has been presented for a second γ1 gene in mice (22). However, these events are detected only in single cloned cells, not in the cell population as a whole. Moreover, the decreased homology of pH21-1 with other γ chains (see below) makes such an explanation unlikely.

(2) DNA cloning errors. This is suggested by the fact that 5 of the 8 substitutions would abolish amino acid identities with other γ chains (Table I). In particular, valine-296 is conserved in every γ, α, and ε chain known,

TABLE I. Substitutions in γ1 amino acid sequence

Position	DNA sequence:		AA sequence:		Other AA sequences:			
	Codon	AA	Codon	AA	Mouse γ2a	Human γ1	Rabbit γ	G-pig γ2
296	*(TCA)	(S)	GTN	V	V	V	V	V
299	GAA	E	GCN	A	A	V	T	V
336	AGA	R	AAR	K	S	Q	E	A
338	AAG	K	AGR	R	R	R	L	R
377	TGG	W	TCN/AGY	S	N	S	K	S
378	AAT	N	GAY	D	N	N	D	N
381	CCA	P	GCN	A	T	E	A	P
382	GCG	A	CCN	P	E	P	E	V$^+$-S$^+$-E

Numbering and mouse γ1 amino acid sequence are from ref. (16). Other amino acid sequences are from refs. (15, 25). AA = amino acid; $^+$ = insertion; * = DNA sequence uncertain.

although not in a μ chain (15, 23-28). Our DNA sequence is uncertain at that point but appears not to encode valine.

(3) Protein sequencing errors. 4 of the 8 substitutions involve the interchange of pairs of nearby amino acids; one involves an acid-to-amide change; and one (codon 377) would replace a serine which was not definitely present in the peptide sequenced (15) with a tryptophan which could have been present in greater molarity than estimated (15).

Since the pH21-1 sequence was completed, the corresponding chromosomal γ1 gene has been cloned and partially sequenced (29). The exchange between positions 336 and 338 is confirmed, and additional exchanges are found in the C_H1 domain (29). Both the nature of the "substitutions" in pH21-1, and the presence of the same and similar substitutions in the chromosomal gene, imply that most of them represent errors in protein sequencing. In the analyses that follow, the DNA sequence is taken to be correct unless otherwise stated.

Comparison with Previous Sequence Data

Some sequence data on the MOPC21 γ1 mRNA has already been published in the form of a ribonuclease T1 oligonucleotide catalogue (30). In that work, the T1 oligonucleotides were not sequenced, but secondary digestion products were aligned so that where possible they matched the amino acid sequence. Of the ten oligonucleotides listed in the region we have sequenced, all but three are confirmed. Two of the three (h6 and h29) were located in the right places but the true nucleotide sequence is a permutation of the suggested one; our sequence is equally consistent with the secondary digestion products tabulated by Cowan *et al*. (30). Our data show that the third oligonucleotide, h18, does not derive from the place that was suggested by Cowan *et al*.

Adetugbo and Milstein (20) have also inferred the mRNA sequence from codon 341 through 355 from the amino acid sequence of the MOPC21 frameshift mutant IF3. Their predicted sequence is confirmed by our analysis. They also suggested (17-19) that the premature termination mutant IF1 contained a nonsense mutation at serine-358. Since we find this codon to be AGU, however, the mutation cannot be a single base substitution. It could perhaps be an insertion of U before codon 358, or a 4-base deletion including it, either of which would create a termination codon in the appropriate position.

Codon Usage

As in other eukaryotic messengers, the pattern of codon usage is nonrandom, as shown in Tables II and III. C is clearly preferred in redundant positions, and G is the next most abundant. Codons for glutamine and glutamic acid show particular preferences for G over A.

This distribution can be compared with that in the mouse immunoglobulin C_{κ} domain (31) (Table III), which also shows a preference for C although the other bases are used equally. The coding sequences for hemoglobin α and β in the rabbit, which like immunoglobulin C_{γ} and C_{κ} are believed to have diverged near to the time of origin of the vertebrates (41,42), also share some but not all anomalies in individual codon usage (Table III and (13)), suggesting that such anomalies may be generally conserved over long spans of evolutionary time.

Table III also shows that the most general features of codon usage, preferences for or against given bases in the third position of codons, are shared by large groups of animal genes. The genes for immunoglobulin C regions, hemoglobins, peptide hormones, and histones all fall into Group I, which has high frequency of C, moderate to high G, moderate to low U, and low A. Genes for an immunoglobulin V-region and for ovalbumin fall into Group II, with uniform usage except for a mild deficiency of G. The genes of SV40 are the only known representatives of Group III; they all have high U and low C. The significance of these patterns is unknown, although analysis of hemoglobin β usage (43) suggested a correlation with the relative abundances of tRNAs.

TABLE II. Codon usage.

Phe	UUU	0	Ser	UCU	3	Tyr	UAU	0	Cys	UGU	0
	UUC	8		UCC	2		UAC	4		UGC	3
Leu	UUA	1		UCA	1	STOP	UAA	0	STOP	UGA	0
	UUG	0		UCG	0		UAG	0	Trp	UGG	4
Leu	CUU	1	Pro	CCU	4	His	CAU	2	Arg	CGU	0
	CUC	3		CCC	4		CAC	4		CGC	1
	CUA	0		CCA	3	Gln	CAA	1		CGA	0
	CUG	2		CCG	1		CAG	7		CGG	0
Ile	AUU	2	Thr	ACU	5	Asn	AAU	4	Ser	AGU	3
	AUC	4		ACC	5		AAC	6		AGC	4
	AUA	1		ACA	2	Lys	AAA	5	Arg	AGA	1
Met	AUG	4		ACG	0		AAG	8		AGG	1
Val	GUU	0	Ala	GCU	2	Asp	GAU	2	Gly	GGU	0
	GUC	4		GCC	2		GAC	4		GGC	4
	GUA	0		GCA	2	Glu	GAA	2		GGA	1
	GUG	4		GCG	1		GAG	10		GGG	1

TABLE III. Frequencies of bases at third-base positions.

(a) Mouse immunoglobulin γ1 (pH21-1)

	U	C	A	G	total
Observed	28	62	20	43	153
Expected for uniform usage	38.09	38.09	35.59	41.26	153.03
Codon usage index	0.74	1.63	0.56	1.04	

(b) Other animal genes

		U	C	A	G	total	ref.
Group I							
Mouse Ig Cγ1	(P)	0.74	1.63	0.56	1.04	153	-
Mouse Ig Cκ		0.81	1.44	0.82	0.86	107	(31)
Rabbit Hb α		0.37	1.99	0.16	1.41	141	(32)
Rabbit Hb β		1.12	1.12	0.25	1.46	148	(33)
Rat insulin	(P)	0.91	1.43	0.39	1.30	98	(34)
Rat GH	(S)	0.77	1.68	0.34	1.21	216	(35)
Human CS	(P)	0.39	1.70	0.45	1.41	168	(36)
Sea urchin histones	(P)	0.84	1.67	0.65	0.96	534	(37)
Group II							
Mouse Ig Vλ	(S)	1.23	1.11	1.10	0.58	128	(38)
Chicken Ov		1.09	1.06	1.10	0.78	386	(39)
Group III							
SV40 all genes		1.52	0.54	1.18	0.78	1514	(40)

The codon usage index is the frequency at which a base appears at the third position in codons, divided by the frequency at which it would appear if all the possible codons for each aminoacid were used uniformly. Stop codons are not included. (P) = partial sequence, (S) = signal ("pre") sequence included. Ig = immunoglobulin, Hb = hemoglobin, GH = growth hormone, CS = chorionic somatomammotropin, Ov = ovalbumin.

Base Composition and Dinucleotide Frequency

The base composition and dinucleotide frequencies in the coding strand are shown in Table IV. There is a severe underabundance of the dinucleotide CG, comparable to that in total eukaryotic DNAs (44,45) and hemoglobin β genes (33,43,46). Demonstration that CG is not deficient in some other coding sequences, such as hemoglobin α genes (32,45), has ruled out earlier models in which it was proposed that ribosomes are unable to translate CG-rich sequences effectively. Subsequently, it has been suggested that CG in eukaryotes is a hotspot for mutation (13). The mechanism may be methylation

TABLE IV. Frequencies of nucleotides and dinucleotides.

A	134 (29%)		AA	35	AT	20	AG	44	AC	35
T	88 (19%)		TA	10	TT	16	TG	29	TC	32
G	105 (23%)		GA	32	GT	17	GG	26	GC	30
C	132 (29%)		CA	57	CT	35	CG	5	CC	35
total	459 (100%)						average			28.625

of the C followed by deamination to yield T (32,45). Coulondre et al. (47) have shown that methylated Cs in E. coli are indeed hotspots for mutation, and suggested the same deamination mechanism.

Therefore, and since no amino acid is required to have CG in its codon, we asked whether the remaining CGs might be maintained by selective pressure on the nucleotide sequence. The five amino acids encoded by these Cs (arg-295, ile-326, pro-337, ala-382, phe-399) are conserved slightly more than average in other γ chains (15,25). One possible pressure for conserving CGs might be selection for an RNA secondary structure; however, these five CGs do not fall preferentially into regions of strong base pairing in our secondary structure prediction (below).

Homologies between C_H2 and C_H3 domains

The DNA sequence covers homologous parts of two immunoglobulin domains, C_H2 and C_H3. This is the first nucleotide sequence to cover regions which are believed to have evolved by tandem duplication of an ancestral gene, so we have examined the sequence for possible nucleotide homology between the domains. Comparison of all known heavy chain sequences (15, 23-28) indicates that the most probable alignment is between codons 287-290 and 392-395, codons 292-324 and 396-428, and codons 325-334 and 430-439, although the positions of the two deletions cannot be defined unequivocally. Within these regions, excluding codons opposite deletions and codons which differ from the sequence of Adetugbo (16), there are 45 pairs of codons of which 9 are identical in the two domains. With such low amino acid homology, reflecting the very ancient divergence of the C_H2 and C_H3 domains, little nucleotide homology would be expected, particularly since the conserved amino acids are probably selected for function. (This is implied by the fact that 15 of the 18 amino acids in conserved positions are also conserved in at least 3 of the 4 heavy chain classes now sequenced (γ, α, ε, μ), compared to only 24% of the other amino acids in the regions compared. In addition, most of those conserved in both C_H2 and C_H3 are also conserved in the C_H1 domain.)

Indeed, little nucleotide sequence homology can be found between C_H2 and C_H3. In the 36 pairs of codons for nonidentical residues, 36/108 nucleotides are identical; the value expected by chance is 27/108. In the 9 pairs of codons for identical residues, omitting nucleotides uniquely specified by the coding requirement, 6/11 nucleotides are identical; the value expected by chance is 5/11. The slight excess over chance is attributable to selection, for conservation of chemical (and thus coding) similarities in amino acid replacements, and for C in third-base positions.

Secondary structure

There is usually little point in attempting to predict the secondary structure of a fragment of an RNA, because long-range interactions may make the most stable folding of the complete RNA very different from that of the fragment (13). However, it seemed of interest to predict the secondary structure of the pH21-1 sequence, in order to find out whether the domain structure of the protein might be reflected in the structure of the RNA. One might anticipate either that the C_H3 domain would fold up into a structure independent of C_H2, or that there might be prominent local structure around the junction between them.

The computed secondary structure is presented in Figure 5. The overall stability is 0.345 kcal/nucleotide. This is more than the 0.331 kcal/nucleotide computed for the complete rabbit β globin mRNA using a less powerful computer program (13) and less than the 0.407 kcal/nucleotide found for the rabbit α globin mRNA using some of the same computer programs used here (32). Examination of each part of the structure separately did not turn up any regions of local low stability of the sort which might indicate that the local sequence paired with other portions of this mRNA.

The most striking feature of the structure is the prominent stem formed from nucleotides 46-65 and 210-230. The C_H2-C_H3 junction is approximately in the center of the loop formed by this stem. The junction itself has now been defined by an RNA splice point in codon 335 (29) as shown in Figure 5. If this portion of the structure is correct, the fact that the splice point is in a region of little secondary structure bounded by the very strong stem would limit the possible roles of secondary structure in directing splicing. That secondary structure does have an important role is suggested by the fact that the loosely conserved primary sequence common to all splice points (48, 49) is too small to give the specificity needed. One can imagine that secondary structure sequesters some potential splicing sequences so that they

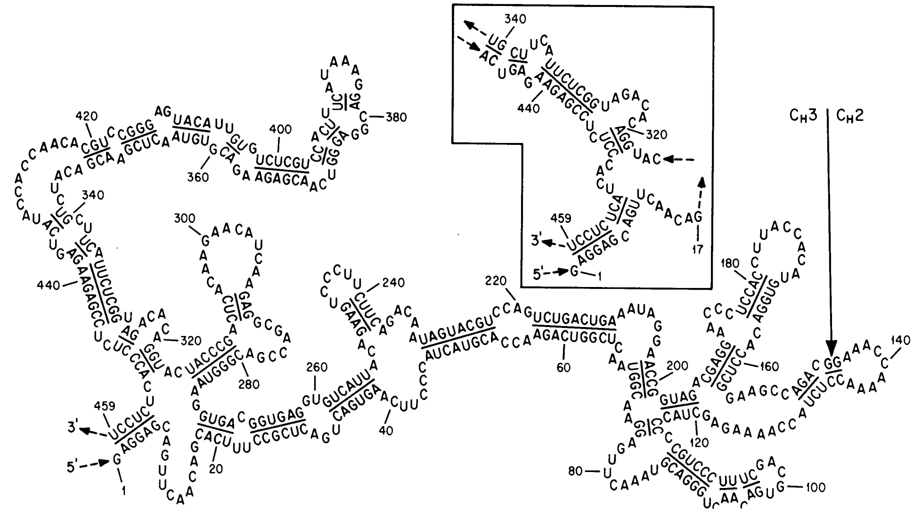

Figure 5. Computed secondary structure of the γl messenger RNA fragment , codons 287-439. The inset shows an equally stable structure for the beginning and end of the sequence. The arrow marks the position of the RNA splice between the CH2 and CH3 domains (29). The total energy of the structure (13) is -158.5 kcal.

are unavailable for splicing, and brings others into reasonably close proximity in a way that facilitates the correct splicing. Since we have not examined the intervening sequence itself, it also remains possible that the secondary structure of the actual precursor molecule contributes in a more definite way to the splicing specificity. Clearly more work will be required to determine the exact role of secondary structure in RNA splicing.

ACKNOWLEDGEMENTS

We are grateful to R. Wall for communicating data on the identification of pH21-1 by hybridization. D.B. DeBorde kindly gave us pH21-1 DNA, and R. Hammen prepared additional supplies. We particularly thank L. Nagy for running the secondary structure cycles on the computer. J.R. acknowledges a Regents' Fellowship of the University of California and an assistantship at the Molecular Biology Institute. The project was supported by the American Cancer Society grant VC240 and NIH grants GM18586 and HL21831 to Winston Salser.

REFERENCES

1 Higuchi, R., Paddock, G.V., Wall, R. and Salser, W. (1976) Proc. Nat. Acad. Sci. USA 73, 3146-3150
2 Maniatis, T., Kee, S.G., Efstratiadis, A. and Kafatos, F.C. (1976) Cell 8, 163-182
3 Rabbitts, T.H. (1976) Nature 260, 221-225
4 Rougeon, F., Kourilsky, P. and Mach, B. (1975) Nucl. Acids Res. 2, 2365-2378
5 Wall, R., Gilmore-Hebert, M., Higuchi, R., Komaromy, M., Paddock, G.V., Strommer, J. and Salser, W. (1978) Nucl. Acids Res. 5, 3113-3128
6 Potter, M. and Lieberman, R. (1967) Adv. Immunology 7, 91-145
7 Cowan, N.J. and Milstein, C. (1973) Eur. J. Biochem. 36, 1-7
8 Curtiss, R., personal communication
9 Padayatty, J., Cummings, I., Manske, C., Higuchi, R., Woo, S. and Salser, W. (1979) in preparation
10 Maxam, A.M. and Gilbert, W. (1977) Proc. Nat. Acad. Sci. USA 74, 560-564
11 Sanger, F. and Coulson, A.R. (1978) FEBS Lett. 87, 107-110
12 Studnicka, G.M., Rahn, G.M., Cummings, I. and Salser, W. (1978) Nucl. Acids Res. 5, 3365-3387
13 Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biol. 42, 987-1004
14 NIH Guidelines for Research Involving Recombinant DNA Molecules (1976) Federal Register 41, 27902
15 Adetugbo, K., Poskus, E., Svasti, J. and Milstein, C. (1975) Eur. J. Biochem. 56,503-519
16 Adetugbo, K. (1978) J. Biol. Chem. 253, 6068-6075
17 Secher, D.S., Milstein, C. and Adetugbo, K. (1977) Immunological Rev. 36, 51-72
18 Adetugbo, K., Milstein, C. and Secher, D.S. (1977) Nature 265, 299-304
19 Adetugbo, K. (1978) J. Biol. Chem. 253, 6076-6080
20 Adetugbo, K. and Milstein, C. (1978) J. Mol. Biol. 121, 239-254
21 Preud'homme, J.-L., Birshtein, B.K. and Scharff, M.D. (1975) Proc. Nat. Acad, Sci. USA 72, 1427-1430
22 Stanislawski, M. and Mitard, M. (1976) Immunochemistry 13, 979-984
23 Edelman, G.M., Cunningham, B.A., Gall, W.E., Gottlieb, P.D., Rutishauser, U. and Waxdal, M.J. (1969) Proc. Nat. Acad. Sci. USA 63, 78-85
24 Wolfenstein-Todel, C., Frangione, B., Prelli, F. and Franklin, E.C. (1976) Biochem. Biophys. Res. Commun. 71, 907-914
25 Bourgois, A., Fougereau, M. and Rocca-Serra, J. (1974) Eur. J. Biochem. 43, 423-435
26 Fougereau, M., Bourgois, A., de Preval, C., Rocca-Serra, J. and Schiff , C. (1976) Annales d'Immunologie 127c, 607-631
27 Low, T.L.K., Liu, Y.-S.V. and Putnam, F.W. (1976) Science 191, 390-392
28 Torano, A. and Putnam, F.W. (1978) Proc. Nat. Acad. Sci. USA 75, 966-970
29 Sakano, H., Rogers, J.H., Hüppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. and Tonegawa, S. (1979) Nature 277, 627-633
30 Cowan, N.J., Secher, D.S. and Milstein, C. (1976) Eur. J. Biochem. 61, 355-368
31 Hamlyn, P.H., Brownlee, G.G., Cheng, C.C., Gait, M.J. and Milstein, C. (1978) Cell 15, 1067-1075
32 Heindell, H.C., Liu, A., Paddock, G.V., Studnicka, G.M. and Salser, W. (1978) Cell 15, 43-54
33 Efstratiadis, A., Kafatos, F.C. and Maniatis, T. (1977) Cell 10, 571-585
34 Ullrich, A., Shine, J., Chirgwin, J., Pictet, R., Tischer, E., Rutter, W.J. and Goodman, H.M. (1977) Science 196, 1313-1319

35 Seeburg, P.H., Shine, J., Martial, J.A., Baxter, J.D. and Goodman, H.M. (1977) Nature 270, 486-494

36 Shine, J., Seeburg, P.H., Martial, J.A., Baxter, J.D. and Goodman, H.M. (1977) Nature 270, 494-499

37 Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H.O. and Birnstiel, M.L. (1978) Cell 14, 655-671

38 Bernard, O., Hozumi, N. and Tonegawa, S. (1978) Cell 15, 1133-1144

39 McReynolds, L., O'Malley, B.W., Nisbet, A.D., Fothergill, J.E., Givol, D., Fields, S., Robertson, M. and Brownlee, G.G. (1978) Nature 273, 723-728

40 Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K.N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L. and Weissman, S.M. (1978) Science 200, 494-502

41 Goodman, M., Moore, G.W. and Matsuda, G. (1975) Nature 253, 603-608

42 Marchalonis, J.J. (1977) Immunity in Evolution. Harvard University Press, Cambridge, Mass.

43 Kafatos, F.C., Efstratiadis, A., Forget, B.G. and Weissman, S.M. (1977) Proc. Nat. Acad. Sci. USA 74, 5618-5622

44 Russell, G.J., Walker, P.M.B., Elton, R.A. and Subak-Sharpe, J.H. (1976) J. Mol. Biol. 108, 1-23

45 Salser, W., Liu, A., Cummings, I., Strommer, J., Padayatty, J. and Clarke, P. (1978) in: Cellular and Molecular Regulation of Hemoglobin Switching, Stamatoyannopoulos, G. and Nienhuis, A., eds., Grune and Stratton, in press

46 Konkel, D.A., Tilghman, S.M. and Leder, P. (1978) Cell 15, 1125-1132

47 Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978) Nature 274, 775-780

48 Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978) Proc. Nat. Acad. Sci. USA 75, 4853-4857

49 Catterall, J.F., O'Malley, B.W., Robertson, M.A., Staden, R., Tanaka, Y. and Brownlee, G.G. (1978) Nature 275, 510-513.