Volume 13 Number 7 1985 Nucleic Acids Research

# Isolation and characterization of rat and human glyceraldehyde-3-phosphate dehydrogenase cDNAs: genomic complexity and molecular evolution of the gene+

J.Yun Tso, Xiao-Hong Sun, Teh-hui Kao, Kimberly S.Reece and Ray Wu*

Section of Biochemistry, Molecular and Cell Biology, Division of Biological Sciences, Wing Hall, Cornell University, Ithaca, NY 14853, USA

Received 22 January 1985; Accepted 4 March 1985

ABSTRACT

Full length cDNAs encoding the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH) from rat and man have been isolated and sequenced. Many GAPDH gene-related sequences have been found in both genomes based on genomic blot hybridization analysis. Only one functional gene product is known. Results from genomic library screenings suggest that there are 300-400 copies of these sequences in the rat genome and approximately 100 in the human genome. Some of these related sequences have been shown to be processed pseudogenes. We have isolated several rat cDNA clones corresponding to these pseudogenes indicating that some pseudogenes are transcribed. Rat and human cDNAs are 89% homologous in the coding region, and 76% homologous in the first 100 base pairs of the 3'-noncoding region. Comparison of these two cDNA sequences with those of the chicken, Drosophila and yeast genes allows the analysis of the evolution of the GAPDH genes in detail.

## INTRODUCTION

Study of the process of molecular evolution has long been based on amino acid sequences. By comparing the amino acid sequences of the same or related proteins from various organisms, the evolutionary clock and phylogenetic trees can be constructed. Because of the degenerate nature of genetic codes, events leading to substitutions of amino acid residues can only be hypothesized. Furthermore, silent mutations that lead to synonymous substitutions as well as changes in the noncoding sequence in the gene are not detectable in the protein sequence data. With the recent advances in DNA sequencing techniques, we believe that nucleotide sequence data would be a better indicator for providing a more detailed picture of the process of molecular evolution. For this study we chose the gene coding for the enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH).

GAPDH is a glycolytic enzyme which is a tetramer of four identical subunits, each of MW 37,000. Based on x-ray studies, the protein chain has been shown to fold into two domains: one for dinucleotide binding and the other for catalysis. The coenzyme binding sites are conserved in the

tertiary structure of the various dehydrogenases[1]. The protein sequence of GAPDH is highly conserved among six different organisms so far reported[2]. In view of the well characterized structural information of this enzyme at the protein level, studies of the gene structure should be of particular interest.

Our second interest in the GAPDH gene concerns the presence of pseudogenes in the mammalian genome. Our laboratory has previously obvserved that the cytochrome c gene has many related sequences in the mammalian genomes, many of which have been shown to be processed pseudogenes[3]. We wanted to know whether the presence of an unusually large number of pseudogenes is unique to the cytochrome c gene or if it is common to other "housekeeping" genes. In this report, we present data on the isolation and sequencing of two GAPDH cDNAs, one from rat and one from man. By using these cDNAs as probes, we detected a large number of GAPDH gene-related sequences in mammalian genomes. The sequences of these two cDNAs together with the previously determined sequences of GAPDH genes from yeast[4], Drosophila[5] and chicken[6] allow us to analyze the history of the evolution of the GAPDH gene.

## METHODS

cDNA libraries were screened by the methods of Grunstein and Hogness[7] for colony hybridization and Benton and Davis[8] for plaque hybridization. Nick translated probes were prepared according to Rigby et al.[9]. Genomic blot analysis was performed according to Southern[10]. The dideoxynucleotide chain termination method of Sanger[11] was used for DNA sequence determination.

## RESULTS

Isolation and characterization of Rat GAPDH cDNA - We used a cDNA clone containing a partial GAPDH sequence from chicken[12] as a probe to screen the rat cDNA libraries for GAPDH clones. The probe used was a 243 bp HindIII-HindIII fragment containing coding information for amino acid residues 225-306 of GAPDH[6]. Three cDNA libraries were used in these screenings. One was a rat brain cDNA library given to us by Jack Dixon of Purdue University. It was constructed according to the method of Okayama and Berg[13]. The other two were rat liver cDNA libraries constructed by linker additions to the double-stranded cDNA. These fragments were inserted into either the BamH1 site of the plasmid pBR322 (constructed in

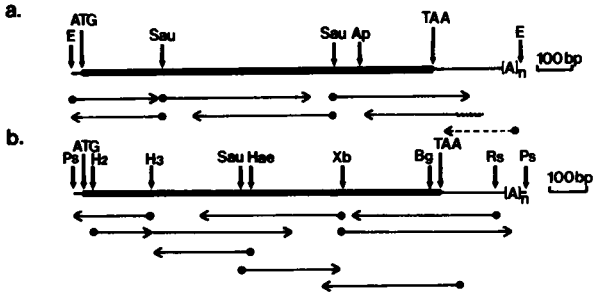


Fig. 1. Restriction maps and sequencing strategy. (a) rat GAPDH cDNA. (b) human GAPDH cDNA. The DNA sequence in each subclone was determined by the dideoxynucleotide chain termination method using a commercial M13 sequencing primer or, where indicated by a "~~~", using an isolated restrictin fragment as a primer. Thick lines represent the coding regions. The broken arrow indicates the region where sequence was determined by the Maxam and Gilbert method in order to bypass the Poly(dA) track. Abbreviations for restriction sites: E = EcoRI; Sau = Sau3AI; Ap = ApaI; Ps = PstI; H2 = HincII; H3 = HindIII; Hae = HaeIII; Xb = XbaI; Bg = BglI; Rs =RsaI.

this laboratory) or the EcoRI site of the phage vector λgt 11 (constructed and given to us by Schwarzbauer et al.[14]). Although many positive clones were obtained from all three libraries, most of them contained cDNA inserts too short (0.5-0.8 kb) to code for the full-length GAPDH protein. This enzyme from various organisms has a molecular weight of about 37,000 dalton, so the minimum length for the full-length cDNA should be about 1 kb. Out of the 50 positive clones isolated, only one clone from the rat liver cDNA library constructed in λgt11 contained an insert of sufficient length (1.3 kb) for a full-length GAPDH cDNA. The insert from this clone, λRcGAP123, was excised by EcoRI digestion and subcloned into the plasmid pUC13 (designated pRcGAP123) for restriction mapping (Fig. 1a) and into phage M13 vectors, mp10 and mp11, for sequencing. Based on DNA sequence analysis (Fig. 2 and 3) the insert in λRcGAP123 is 1241 bp long and contains a 30 bp 5'-noncoding region, a 1002 bp coding region and a 207 bp 3'-noncoding region. The consensus polyadenylation signal, AATAAA, is located 15 bp preceding the poly(dA) track. The coding region encodes a protein of 333 amino acid residues whose sequence is highly homologous to the published GAPDH protein sequences from various organisms[2]. No nucleotide insertion, deletion or substitution of a sequence that codes for an invariant amino acid (unacceptable substitution) is observed. This cDNA clone is thus likely to encode the functional GAPDH in rat. We used this

```
                                                1                                  30
a          GGTCAGTGCCGGCCTCGTCTCATAGACAAG ATG GTG AAG GTC GGT GTC AAC GGA TTT GGC CGT
b GTCGCCAGCCGAGCCACATCGCTCAGACACC ATG GGG  A                A                   T
a                                              Met Val Lys Val Gly Val Asn Gly Phe Gly Arg
b                                  Met Gly Lys
                                                 1                                 10
                                  60                                  90
a ATT GGC CGC CTG GTC ACC AGG GCT GCC TTC TCT TGT GAC AAA GTG GAC ATT GTT GCC ATC AAC
b       G                           T   T AAC  C   GT           T                   T
a Ile Gly Arg Leu Val Thr Arg Ala Ala Phe Ser Cys Asp Lys Val Asp Ile Val Ala Ile Asn
b                                       Asn Ser Gly
                                  20                                  30
                            120                                 150
a GAC CCC TTC ATT GAC CTC AAC TAC ATG GTC TAC ATG TTC CAG TAT GAT TCT ACC CAC GGC AAG
b                                   T                A         C       T       A
a Asp Pro Phe Ile Asp Leu Asn Tyr Met Val Tyr Met Phe Gln Tyr Asp Ser Thr His Gly Lys
b
                               40                                  50
                         180                                 210
a TTC AAT GGC ACA GTC AAG GCT GAG AAT GGG AAG CTG GTC ATC AAT GGG AAA CCC ATC ACC ATC
b     C         C                 C           T   G          A   T
a Phe Asn Gly Thr Val Lys Ala Glu Asn Gly Lys Leu Val Ile Asn Gly Lys Pro Ile Thr Ile
b     His                                                     Asn
                           60                                  70
                    240                                 270
a TTC CAG GAG CGC GAT CCC GCT AAC ATC AAA TGG GGT GAT GCT GGT GCT GAG TAT GTC GTG GAG
b               A         T C   A       G       C          C          C
a Phe Gln Glu Arg Asp Pro Ala Asn Ile Lys Trp Gly Asp Ala Gly Ala Glu Tyr Val VAl Glu
b                         Ser Lys
                      80                                  90
                300                                 330
a TCT ACT GGC GTC TTC ACC ACC ATG GAG AAG GCT GGG GCT CAC CTG AAG GGT GGT GCC AAA AGG
b   C                                               T T   C     G   A
a Ser Thr Gly Val Phe Thr Thr Met Glu Lys Ala Gly Ala His Leu Lys Gly Gly Ala Lys Arg
b                                                       Gln
                100                                 110
             360                                 390
a GTC ATC ATC TCC GCC CCT TCC GCT GAT GCC CCC ATG TTT GTG ATG GGT GTG AAC CAC GAG AAA
b               T       C   T                       C   C                   T       G
a Val Ile Ile Ser Ala Pro Ser Ala Asp Ala Pro Met Phe Val Met Gly Val Asn His Glu Lys
b
             120                                 130
          420                                 450
a TAT GAC AAC TCC CTC AAG ATT GTC AGC AAT GCA TCC TGC ACC ACC AAC TGC TTA GCC CCC CTG
b             AG            C A           C                                   A
a Tyr Asp Asn Ser Leu Lys Ile Val Ser Asn Ala Ser Cys Thr Thr Asn Cys Leu Ala Pro Leu
b                         Ile
          140                                 150
       480                                 510
a GCC AAG GTC ATC CAT GAC AAC TTT GGC ATC GTG GAA GGG CTC ATG ACC ACA GTC CAT GCC ATC
b                                   T               A
a Ala Lys Val Ile His Asp Asn Phe Gly Ile Val Glu Gly Leu Met Thr Thr Val His Ala Ile
b
       160                                 170
   540                                 570                                 600
a ACT GCC ACT CAG AAG ACT GTG GAT GGC CCC TCT GGA AAG CTG TGG CGT GAT GGC CGT GGG GCA
b         C                               C   G   A                       C       T
a Thr Ala Thr Gln Lys Thr Val Asp Gly Pro Ser Gly Lys Leu Trp Arg Asp Gly Arg Gly Ala
b
   180                                 190                                 200
```

```
                                        630                                 660
a GCC CAG AAC ATC ATC CCT GCA TCC ACT GGT GCT GCC AAG GCT GTG GGC AAG GTC ATC CCA GAG
b CT                        C   T       C                                   T
a Ala Gln Asn Ile Ile Pro Ala Ser Thr Gly Ala Ala Lys Ala Val Gly Lys Val Ile Pro Glu
b Leu
                                        210                                 220
                                    690                                 720
a CTG AAC GGG AAG CTC ACT GGC ATG GCC TTC CGT GTT CCT ACC CCC AAT GTA TCC GTT GTG GAT
b                                               C   C   T G     C   G   A   G       C
a Leu Asn Gly Lys Leu Thr Gly Met Ala Phe Arg Val Pro Thr Pro Asn Val Ser Val Val Asp
b                                                       Ala
                                    230                                 240
                                750                                 780
a CTG ACA TGC CGC CTG GAG AAA CCT GCC AAG TAT GAT GAC ATC AAG AAG GTG GTG AAG CAG GCG
b       C       T   A   A             A
a Leu Thr Cys Arg Leu Glu Lys Pro Ala Lys Tyr Asp Asp Ile Lys Lys Val Val Lys Gln Ala
b
                                250                                 260
                            810                                 840
a GCC GAG GGC CCA CTA AAG GGC ATC CTG GGC TAC ACT GAG GAC CAG GTT GTC TCC TGT GAC TTC
b T G           C   C                                   C       G       C
a Ala Glu Gly Pro Leu Lys Gly Ile Leu Gly Tyr Thr Glu Asp Gln Val Val Ser Cys Asp Phe
b Ser                                                   His             Ser
                            270                                 280
                        870                                 900
a AAC AGC AAC TCC CAT TCC TCC ACC TTT GAT GCT GGG GCT GGC ATT GCT CTC AAT GAC AAC TTT
b         G   A     C                   C                       C       C     C
a Asn Ser Asn Ser His Ser Ser Thr Phe Asp Ala Gly Ala Gly Ile Ala Leu Asn Asp Asn Phe
b         Asp Thr                                                           His
                        290
                    930                                 960
a GTG AAG CTC ATT TCC TGG TAT GAC AAT GAA TAT GGC TAC AGC AAC AGG GTG GTG GAC CTC ATG
b   C                               C       T
a Val Lys Leu Ile Ser Trp Tyr Asp Asn Glu Tyr Gly Tyr Ser Asn Arg Val Val Asp Leu Met
b                                       Phe
                    310                                 320
                990
a GCC TAC ATG GCC TCC AAG GAG TAA
b     C
a Ala Tyr Met Ala Ser Lys Glu
b     His
                330
```

Fig. 2. Coding sequences and deduced amino acid sequences of (a) rat and (b) human GAPDH cDNA including portion of 5'-untranslated regions. Both nucleotide and amino acid sequences are numbered according to the rat sequences. Only those nucleotides and amino acids which are different from the rat are shown for the human sequences.

cDNA as a probe to analyze mRNA from brain, muscle, and liver by mRNA hybridization analysis (Northern blot). In all cases only one species of mRNA of 1.4 kb hybridized (data not shown).

Isolation and characterization of human GAPDH cDNA

The human liver cDNA library was screened using a fragment of rat GAPDH coding sequence as a probe. The cDNA library used was constructed by Chander *et al.*[15] using the GC tailing method at the *Pst*I site of pBR322. Many positive clones were obtained. 80% of them contained an insert long

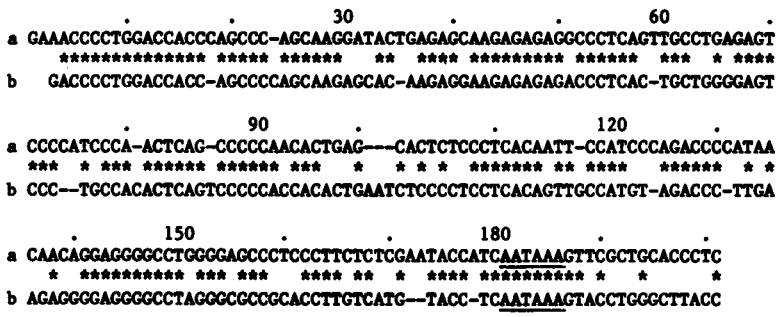


Fig. 3. 3'-noncoding sequence of (a) rat and (b) human GAPDH cDNA. DNA sequences start immediately after the stop codon of the coding region. Gaps "-" are introduced to maximize homology. Identical bases are indicated by "*". Polyadenylations follow immediately at the end of the sequences. The polyadenylation signal AATAAA is underlined. The nucleotides in the rat sequence are numbered, disregarding the gaps.

enough to be nearly full-length GAPDH cDNA. The insert of one of these clones, pHcGAP, is about 1.2 kb long and was subcloned into M13 vectors for DNA sequence analysis. The sequence strategy is summarized in Fig. 1b. 90% of the sequence was obtained by sequencing from both strands, and each clone was sequenced more than once. Sequence data (Fig. 2 and 3) indicate that this cDNA contains a coding region of 1008 bp for GAPDH and has a 31 bp sequence for the 5'-noncoding region and a 216 bp sequence for the 3'-noncoding region. The consensus sequence AATAAA is also present 15 bp upstream from the polyadenylation site. Although the amino acid sequence deduced from this cDNA contains no unacceptable substitution when compared to other GAPDH sequences, it differs from the published human muscle GAPDH amino acid sequence[16] at 27 positions. This result contradicts the previous observation that there is only one human GAPDH in various tissues[17]. To eliminate the possibility of the occurrence of GAPDH isoenzymes in human, we screened a human muscle cDNA library[18] at low stringency using pHcGAP as a probe. The partial sequences of four individual muscle cDNA clones are identical to that of the liver cDNA clone. There is thus no evidence for a second functional GAPDH gene in muscle. The discrepancy in the amino acid sequence of human GAPDH derived from protein sequencing and from nucleotide sequencing is probably due to errors in protein sequencing.

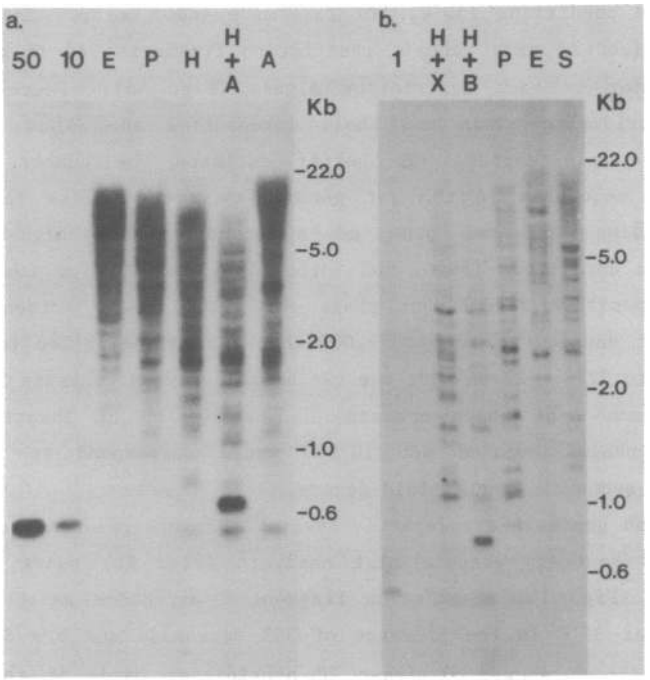


Fig. 4. (a) Genomic Southern analysis of the rat liver DNA (10µg)digested with ApaI (A), HindIII + ApaI (H+A), HindIII (H), PvuII (P) and EcoRI (E). A short subcloned GAPDH cDNA fragment was cleaved from the vector and loaded in amounts to represent 10 and 50 genomic copies. DNA fragments were electrophoresed on a 0.8% agarose gel and transferred to a nitrocellulose filter. The filter was hybridized to a nick-translated Sau3A-ApaI fragment from rat GAPDH cDNA.

(b) Genomic Southern Analysis of human placenta DNA (20µg) digested with SstI (S), EcoRI (E), PvuII (P), HindIII + BglI (H+B) and HindIII + XbaI (H+X). Cloned GAPDH cDNA was cleaved with HindIII + XbaI and loaded in amounts to represent one genomic copy. DNA fragments were electrophoresed on a 0.8% agarose gel and transferred to a nitrocellulose filter. The filter was hybridized to a nick translated HindIII + XbaI fragment from human GAPDH cDNA.

## Genomic complexity of the GAPDH gene-related sequences in the rat and human chromosomes

We have used both cDNAs to analyze the complexity of the GAPDH gene-related sequences in the corresponding genomes. The rat cDNA probe was a 561 bp Sau3A-ApaI partial restriction fragment derived from λRcGAP123 (nucleotide numbers 235-796, Fig. 2). The same restriction fragment was loaded on the gel to give the equivalent of 10 and 50 gene copies as control in the Southern genomic blot analysis. Even under high-stringency

hybridization conditions (50°C, 50% formamide and 5 x SSC) the rat cDNA probe hybridized to many genomic restriction fragments and in most cases formed a smear in each restriction digest (Fig. 4a). There are many distinct hybridizing bands and their intensities show that they must represent multiple copies. To better estimate the number of GAPDH gene-related sequences in the rat genome, we screened the rat genomic library[19] using the rat cDNA as a probe. Under high-stringency hybridization conditions (50°C, 50% formamide, 5 x SSC), we found on the average 10 positive recombinant phage per 10,000 phage screened. This number almost doubled to 17 per 10,000 when the hybridization temperature was lowered to 37°C. Given that the rat haploid genome contains $3 \times 10^9$ bp and each recombinant phage contains $1.5 \times 10^4$ bp of insert DNA, 17 recombinant phage detected per 10,000 would correspond to 340 GAPDH gene-related sequences per haploid genome.

The human genome has relatively fewer GAPDH gene-related sequences as indicated by Southern genomic blot analysis (Fig. 4b) using a 550 bp HindIII-XbaI (Fig. 1b) restriction fragment from pHcGAP as a probe and hybridizing at 37°C in the presence of 50% formamide and 5 x SSC. Each restriction digest showed at least 30 hybridizing bands in the genomic blot. Some of the bands correspond to more than one gene copy based on the relative intensity of the bands compared to the copy number controls. The screening of the human genomic library, at the same conditions as the genomic blot, established that there are approximately 100 copies of GAPDH gene-related sequence per genome. We believe that library screening is a more accurate estimate of gene copy number than genomic blot analysis since many hybridizing bands of lower sequence homology may not be detectable on the genomic blot due to limited amounts of immobilized DNA.

Many GAPDH gene-related sequences are processed pseudogenes.

With the hope of better understanding the nature of these GAPDH gene-related sequences and identifying expressed functional genes, we isolated 60 recombinant phage from a limited screening of the rat genomic library. Based on the restriction analysis each of these phage is unique. Under very stringent hybridization conditions (60°C, 50% formamide 5 x SSC) only 5 of the 60 isolates hybridized to rat cDNA. The GAPDH gene-related sequences of these 5 recombinant phage, as well as the sequence of one phage with less homology, have been partially determined. Preliminary data indicate that all 6 are processed GAPDH pseudogenes lacking introns and containing mutations which would render them functionless. Furthermore, a

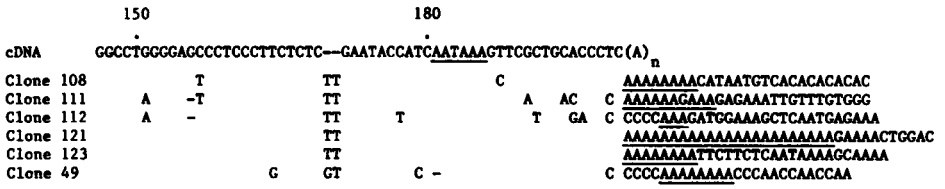


Fig. 5. The 3'-noncoding sequences of six rat GAPDH pseudogenes. The sequences were obtained by primer extension from M13 subclones of each pseudogene using a 79 bp HinfI-HaeIII fragment (nucleotide numbers 68-147, Fig. 3) from the rat cDNA as primer. All pseudogenes except clone 49 are highly homologous to the cDNA based on dot hybridizations. "-" indicates gaps introduced for maximum homology. The polyadenylation signal AATAAA and poly(dA) tracks for each clone are underlined.

track of poly(dA) was found at a position closely corresponding to the location of the poly(A) track of GAPDH mRNA. The position of the poly(A) track of all six clones is shown in Fig. 5 next to that of the cDNA. The sequences near the 3' end of the GAPDH gene transcription unit of these clones were obtained by the primer extension sequencing method using an 79 bp HinfI-HaeIII fragment (nucleotide numbers 68-147, Fig. 3) as the primer and M13 subclones from various phage as templates. High sequence homology is found at the 3'-noncoding region up to the poly(A) track, but no homology is observed outside the region corresponding to the processed transcription unit of GAPDH mRNA in these phage. These characteristics are all in agreement with those described for processed pseudogenes[20]. Curiously, all six processed pseudogenes contain a HindIII site (AAGCTT) at nucleotide numbers 190-195 (Fig. 2) instead of the sequence AAGCTG which is found in the functional rat GAPDH cDNA. This observation explains the occurrence of an intense hybridizing band of about 640 bp(about 50 or more gene copies) in the genomic blot of HindIII and ApaI double digest. The latter enzyme cut the cDNA at a unique site (nucleotide number 796). We reasoned that the synonymous mutation leading to the loss of the HindIII site was a recent event in the functional gene and many of these pseudogenes probably were generated much earlier so that they still retain this site in their coding region. Because many of the processed pseudogenes are intronless, cutting them at two unique sites within the coding region would generate a common fragment of nearly identical size from each pseudogene. The other hybridizing bands could be processed pseudogenes that lost either one of these unique sites, or alternatively they might be related sequences containing introns in their coding region.

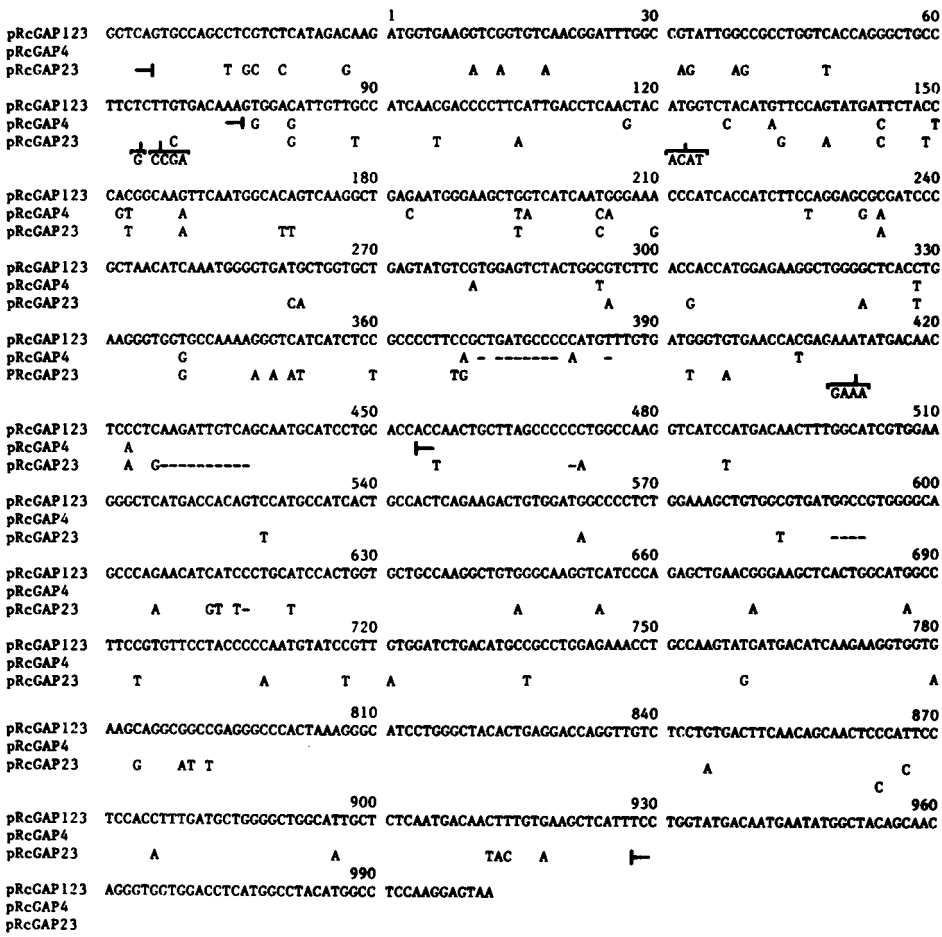

Fig. 6. DNA sequences of two nonfunctional GAPDH cDNA clones, pRcGAP4 and pRcGAP23. The sequences are shown in comparison with that of the authentic GAPDH cDNA clone pRcGAP123. Only those nucleotides that are different from pRcGAP123 are shown. "-" indicates gaps introduced for maximum homology. Sequences bracketed by the symbol "⊣⊢" are insertions. The symbols "⊣" and "⊢" indicate the beginning and the end of the pRcGAP4 and pRcGAP23 clones.

In the human genome, at least a portion of GAPDH gene-related sequences can be accounted for by the presence of intronless processed pseudogenes. Their presence is indicated by the intense 800 bp hybridizing band in the lane containing genomic DNA digested with both HindIII and BglI, two enzymes which cut uniquely in the coding region of the cDNA (Fig.

1b). However, no intense hybridizing band was observed when the genomic DNA was digested with HindIII and XbaI. XbaI also cuts uniquely in the coding sequence. Again, we reasoned that the acquisition of the XbaI site in GAPDH gene may have been a recent event and that many processed pseudogenes were generated prior to this occurrence. Partial sequence analysis of one random human genomic clone also indicates that it is a processed pseudogene (data not shown).

Some rat GAPDH pseudogenes are transcribed

Among the GAPDH cDNA clones isolated from rat cDNA libraries, two clones were found to be derived from nonfunctional GAPDH mRNAs. Both cDNA clones contain inserts too short to have included the entire coding sequence. The first cDNA clone, designated pRcGAP4, contains an insert, only 370 bp long, which was sequenced. The sequence corresponds to nucleotide numbers 76-453 of pRcGAP123 (Fig.2) but contains 24 nucleotide substitutions, one of which creates an in-frame termination codon. In addition, this clone has 9 nucleotide deletions that would render it nonfunctional. The second cDNA clone, pRcGAP23, failed to hybridize to the authentic rat GAPDH cDNA under stringent conditions (60°C, 50% formamide and 5 x SSC). It has a 948 bp insert which was also sequenced. The sequence corresponds to nucleotide numbers -25 to 927 of pRcGAP123 (Fig.2) but contains 82 nucleotide substitutions, 14 insertions and 18 deletions. The sequences of these two clones are shown together with pRcGAP123 in Fig.6. Interestingly, both clones contain the same internal HindIII site at nucleotide numbers 190-195 which is present in many processed pseudogenes. The mechanism that leads to the transcription of these two nonfunctional genes is not clear. A similar case has been reported for the T-cell receptor cDNA[21]. It is possible that transcription of these genes results from insertion of processed pseudogenes into the transcriptional unit of other genes. Alternatively the transcripts could arise from the transcription of pseudogenes which are not related to the processed type. Since the 5' start point of pRcGAP23 is very close to that of pRcGAP123, it is likely that its transcription start point is similar to that of the authentic GAPDH gene. If this is the case, pRcGAP23 would be the product of a pseudogene that retains the GAPDH gene promoter, a feature which would distinguish it from pseudogenes of the processed type. The number of GAPDH pseudogenes that are transcribed has not been systematically determined.

```
                1         2         3         4         5         6
0123456789012345678901234567890123456789012345678901234567890123456789

AVKVGINGFGRIGRNVFRAALKNPDIEVVAVND-LTNADGLAHLLKYDSVHGRLDAEVVVN 1
MKVGINGFGRIGRQVFRILHSRG-VEVALIND-LTNDKTLAHLLKYDSIYHRFPGEVAYD 2
VRVAINGFGRIGRLVMRIALSRPNVEVVALNDPFITNDYAAYMFKYDSTHGRYAGEVSHD 3
SKIGINGFGRIGRLVLRAAIDKG-ASVVAVNDPFIDVNYMVYLFKFDSTHGRFKGTVAAE 4
SKIGIDGFGRIGRLVLRAALSCG-AQVVAVNDPFIALEYMVYMFKYDSTHGVFKGEVKME 5
VKVGVNGFGRIGRLVTRAAVLSGKVQVVAINDPFIDLNYMVYMFKYDSTHGHFKGTVKAE 6
VKVGVDGFGRIGRLVTRAAFNSGKVDIVAINDPFIDLHYMVYMFQYDSTHGKFHGTVKAE 7
VKVGVNGFGRIGRLVTRAAFSCDKVDIVAINDPFIDLNYMVYMFQYDSTHGKFNGTVKAE 8
GKVKVGVNGFGRIGRLVTRAAFNSGKVDIVAINDPFIDLNYMVYMFQYDSTHGKFHGTVKAE 9

          GFGRIGR V R             ND              DS        V

        7         8         9         10        11        12        13
12345678901234567890123456789012345678901234567890123456789012345678901

DGDVSVNGKEIIVKAERNPENLAWGEIGVDIVVESTGRFTKREDAAKHLEAGAKKVIISAPAKVENITVV 1
DQYLYVDGKAIRATAVKDPKEIPWAEAGVGVVIESTGVFTDADKAKAHLEGGAKKVIITAPAKGEDITLV 2
DKHIIVDGKKIATYQERDPANLPWGSSNVDIAIDSTGVFKELDTAQKHIDAGAKKVVITAPS-STAPMFV 3
GGFLVVNGQKITVFSERDPANINWASAGAEYVVESTGVFTTIDKASTHLKGGAKKVIISAPS-ADAPMFV 4
DGALVVDGKKITVFNEMKPENIPWSKAGAEYIVESTGVFTTIEKASAHFKGGAKKVVISAPS-ADAPMFV 5
NGKLVINGHAITIFQERDPSNIKWADAGAEYVVESTGVFTTMGKAGAHLKGGAKRVIISAPS-ADAPMFV 6
DGKLVIDGKAITIFQERDPANIKWGDAGTAYVVESTGVFTTMEKAGAHLKGGAKRVIISAPS-ADAPMFV 7
NGKLVINGKPITIFQERDPANIKWGDAGAEYVVESTGVFTTMEKAGAHLKGGALRVIISAPS-ADAPMFV 8
NGKLVINGNPITIFQERDPSKIKWGDAGAEYVVESTGVFTTMEKAGAHLQGGAKRVIISAPS-ADAPMFV 9

       G  I        P    W           STG F         H   GAK V I AP         V

       14        15        16        17        18        19        20
12345678901234567890123456789012345678901234567890123456789012345678901

MGVNQDKYDPKAHHVISNASCTTNCLAPFAKVLHQEFGIVRGMMTTVHSYTNNQRILDLP-HKDLRGARA 1
MGVNHEAYDPSRHHIISNASCTTNSLAPVMKVLEEAFGVEKALMTTVHSYTBBZRLLDLP-HKDLRRARA 2
MGVNEVKYT-SDLKIVSNASCTTNCLAPLAKVINDAFGIEEGLMTTVHSLTATQKTVDGPSHKDWRGGRT 3
CGVNLDAYSP-DMKVVSNASCTTNCLAPLAKVINDNFEIVEGLMTTVHATTATQKTVDGPSGKLWRDGRG 4
CGVNLEKYS-KDMTVVSNASCTTNCLAPVAKVLHENFEIVEGLMTTVHAVTATQKTVDGPSAKDWRGGRG 5
MGVNHEKYD-KSLKIVSNASCTTNCLAPLAKVIHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDDRG 6
MGVNHEKYD-NSLKIVSNASCTTNCLAPLAKVIHDHFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRG 7
MGVNHEKYD-NSLKIVSNASCTTNCLAPLAKVIHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRG 8
MGVNHEKYD-NSLKIISNASCTTNCLAPLAKVIHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRG 9

 GVN            VSNASCTTN LAP  KV    F       MTTVH    T    D    K  R  R

       21        22        23        24        25        26        27
12345678901234567890123456789012345678901234567890123456789012345678901

AAESIIPTTTGAAKAVALVLPELKGKLNGMAMRVPTPNVSVVDLVAELEKEVTVEEVNAALKAAAEGELK 1
AAINIIPTTTGAAKATALVLPSLKGRFDGMALRVPTATGSISDITALLKREVTAEEVNAALKAAAEGPLK 2
ASGNIIPSSTGAAKAVGKVLPELQGKLTGMAFRVPTVDVSVVDLTVKLDKETTYDEIKKVVKAAAEGKLK 3
AAQNIIPAATGAAKAVGKVIPALNGKLTGMAFRVPTPNVSVVDLTVRLGKGATYDEIKAKVEEASKGPLK 4
AAQNIIPSSTGAAKAVGKVIPELDGKLTGMAFRVPTPDVSVVDLTVRLGKECSYDDIKAAMKTASEGPLQ 5
AAQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPNVSVVDLTCRLEKPAKYDDIKRVVKAAADGPLK 6
AAQNIIPASTGAAKAVGKVIPELDGKLTGMAFRVPTPNVSVVDLTCRLEKPAKYDDIKKVVKQASEGPLK 7
AAGNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTPNASVVDLTCRLEKPAKYDDIKKVVKQAAEGPLK 8
ALQNIIPASTGAAKAVGKVIPELNGKLTGMAFRVPTANVSVVDLTCRLEKPAKYDDIKKVVKQASEGPLK 9

A   IIP  TGAAKA   V P L G   GMA RVDT   S  D    L                   G L
```

```
        28        29        30        31        32        33
12345678901234567890123456789012345678901234567890123456789012345678901234567890123456

GILAYSEEPLVSRNYNGSTVSSTIDALSTMVIDGKMVKVVSWYDNETGYSHRVVDLAAYINAKGL-   1
GILAYTEDEIVLZBIVMDPHSSIVDAKLTKALGNMXXKVFAWYDNEWGYANRVADLVELVLRKGV-   2
GVLGYTEDAVVSSDFLGDSHSSIFDASAGIQLSPKFVKLVSWYDNEYGYSTRVVDLVEHIAKA---   3
GILGYTDEEVVSTDFLSDTHSSVFDAKAGISLNDKFVKLISWYDNEFGYSNRVIDLIKYMQSKD--   4
GFLGYTEDDVVSSDFIGDNRSSIFDAKAGIQLSKTFVKVVSWYDNEFGYSQRVIDLLKHMQKVDSA   5
GILGYTEDQVVSCDFNGDSHSSTFDAGAGIALNDHFVKLVSWYDNEFGYSNRVVDLMVHMASKE--   6
GILGYTEDQVVSCDFNSNSHSSTFDAGAGIALNDHFVKLISWYDNEFGYSNRVVDLMVHMASKE--   7
GILGYTEDQAASCDFNSNSHSSTFDAGAGIALNDNFVKLISWYDNEYGYSNRVVDLMAYMASKE--   8
GILGYTEHQVVSSDFNSDTHSSTFDAGAGIALNDHFVKLISWYDNEFGYSNRVVDLMAHMASKE--   9

G L Y E   V         SS  DA          K   WYDNE GY  RV DL
```

Fig. 7. Amino acid sequence alignment of GAPDH proteins from nine species. Amino acid sequences are aligned to maximize homology. The nine species are: 1. Bacillus stearothermophilus, 2. Thermus aquaticus, 3. Baker's yeast, 4. Drosophila melanogaster (DmGAP-1), 5. lobster, 6. chicken, 7. pig, 8. rat, 9. human. The invariant amino acid residues are shown in a separate line under the human sequence. The sources of sequence information are cited in the text.

## DISCUSSION

### Isolation and characterization of rat and human GAPDH cDNAs

We have isolated and sequenced full-length GAPDH cDNAs from two mammalian species in this study. Human and rat GAPDH cDNA are highly homologous, having 89% homology on the nucleotide level in their coding regions. Although no homology is observed in their short 5'-noncoding regions, their 3'-noncoding regions are 76% homologous in a stretch of about 100 nucleotides immediately following the stop codon. Conservation of the 3'-noncoding region in mammals has also been reported for actin and tubulin isotypic genes[22,23]. It was suggested that these regions may play some important roles in gene expression and therefore may be under selective pressure. When we compared the 3' region of rat or human with that of chicken and Drosophila GAPDH genes, we did not find any significant homology.

### Genomic complexity of the GAPDH gene-related pseudogenes

Since a processed pseudogene has been proposed to be the product of mRNA, reverse transcribed and inserted back into the genome[20], genes that are expressed highly in the germ line cells would have a higher probability of returning to the genome as processed pseudogenes. Our findings of a large number of GAPDH gene-related sequences, as well as those reported for other "house-keeping" enzymes[24] in the mammalian genome, are consistent with the proposition concerning the generation of processed pseudogene.

While this manuscript was in preparation, four independent studies by Hanauer and Mandel[25], Benham et al.[26], Piechaczyk et al.[27], and Arcari et al.[28] also showed the complexity of GAPDH gene-related sequences in mammalian genomes. Two of the studies[25,27] argued that the higher multiplicity of the rat genome is due to amplification of a basal number (as in the human genome) of pseudogenes. We do not believe this explanation is adequate based on Southern blot analysis of the 60 recombinant phage isolated from the rat genomic library. We found that most phage contain only one copy of GAPDH gene-related sequences in an insert (average length 15 kb), and show a different restriction digestion pattern. Furthermore, there is no sequence homology outside of the transcription unit in the limited number of pseudogenes we have partially sequenced. Moreover, when the flanking sequences were taken from these clones and used as probes for genomic blot, they hybridized to only a few bands. There is thus no evidence for tandem duplications or amplification of a limited number of pseudogenes based on the data that we have collected. The higher multiplicity of GAPDH gene-related sequences in the rat genome remains a question for further investigation.

Molecular evolution of GAPDH protein

With the published sequence data of chicken cDNA, three yeast genes, two Drosophila genes[2], and the cDNAs from rat and man reported here, we have a sufficiently large data base for the study of molecular evolution of GAPDH on the protein level. The deduced amino acid sequences from these genes were aligned together with other known GAPDH protein sequences, including those from Bacillus stearothermophilus, Thermus aquaticus, lobster, and pig[2] (Fig. 7).

Based on a pairwise comparison of the amino acid substitutions, and the time since divergence between them, we plotted the divergence time versus the percentage of amino acid substitution which was corrected for multiple hits[29] (data not shown). We found that amino acid residue substitutions occurred at a relatively constant rate over time. The unit evolutionary period (UEP) defined as the time (in million years) required for the fixation of 1% amino acid sequence difference, was calculated to be 24 from the slope of this plot. This value is close to that of cytochrome c (UEP=20), but higher than that of hemoglobin (UEP=5.8).

From the alignment of GAPDH protein sequences shown in Fig. 7, we have found that 107 residues (32%) are invariant in this enzyme over more than 2 billion years of divergence. Of the invariant resdiues, 68% are located in

Table 1. Evolutionary distances of GAPDH genes based on pairwise comparison of nucleotide sequences

| Comparison | T* (MY) | The fraction of nucleotide substitution $K_1$ | $K_2$ | $K_3$ | $K_s$ | $k'_s$ |
|---|---|---|---|---|---|---|
| DmGapdh-1 vs. DmGapdh-2 | 60 | 0.021 | 0.006 | 0.38 | 0.31 | $2.5 \times 10^{-9}$ |
| Human vs. Rat | 75 | 0.040 | 0.021 | 0.31 | 0.27 | $1.8 \times 10^{-9}$ |
| Chicken vs. Human | 270 | 0.090 | 0.034 | 0.51 | 0.39 | $0.7 \times 10^{-9}$ |
| Chicken vs. Rat | 270 | 0.072 | 0.034 | 0.46 | 0.34 | $0.6 \times 10^{-9}$ |
| DmGapdh-1 vs. Human | 600 | 0.230 | 0.130 | 0.82 | 0.61 | $0.5 \times 10^{-9}$ |
| DmGapdh-1 vs. Rat | 600 | 0.230 | 0.130 | 0.73 | 0.53 | $0.4 \times 10^{-9}$ |
| DmGapdh-1 vs.Chicken | 600 | 0.220 | 0.150 | 0.73 | 0.57 | $0.5 \times 10^{-9}$ |
| Yeast vs. Human | 1000 | 0.440 | 0.250 | 2.59 | 2.35 | $1.2 \times 10^{-9}$ |
| Yeast vs. Rat | 1000 | 0.420 | 0.250 | 2.78 | 2.53 | $1.3 \times 10^{-9}$ |
| Yeast vs. Chicken | 1000 | 0.420 | 0.270 | 0.92 | 0.74 | $0.4 \times 10^{-9}$ |
| Yeast vs. DmGapdh-1 | 1000 | 0.450 | 0.250 | 1.28 | 1.03 | $0.5 \times 10^{-9}$ |

* Time of divergence (in million years) for the Drosophila (Dm) gene pair is based on amino acid substitutions in the protein, all the others are based on fossil record.

the second half of the molecule which forms the catalytic domain. This observation implies that there is more functional constraint on the evolution of the catalytic domain than on the coenzyme binding domain. Many dehydrogenases have coenzyme binding sites conserved only in tertiary structure rather than in primary structure. The amino acid sequences in these domains are probably more susceptible to amino acid substitution. Many invariant residues determined to be functionally essential are located in blocks of homology. The most conserved one is located around the essential $Cys_{151}$ in which 12 invariant residues are found in all organisms with only one exception. Other blocks include those containing $His_{178}$ and $Tyr_{313}$, both of which are thought to participate in catalysis[1].

Molecular evolution of GAPDH gene

DNA sequences of GAPDH genes offer us a wide set of data to analyze the evolutionary process. Nucleotide sequence data from man, rat, chicken, Drosophila, and yeast were analyzed according to the method of Kimura[30]. We determined the percentage of divergence for nucleotide substitution at

position 1, 2, and 3 of a codon based on pairwise comparison of two GAPDH coding sequences. The percentage of nucleotide change, or evolutionary distance, for each position, after correction for multiple hit is expressed as $K_1$, $K_2$, and $K_3$ for divergence at codon position 1, 2 and 3, respectively. The synonymous component of $K_3$ is calculated separately and expressed as $K'_s$. Since most of the synonymous nucleotide substitutions occur at the third position, $K'_s$ is a good approximation of the evolutionary distance for silent mutation. The tabulated data based on comparison of nucleotide sequences are shown in Table 1.

As expected the relationship $K_3 > K_1 > K_2$ holds in every comparison. More interesting is the comparison for silent mutations. Such mutations in theory should be without functional constraint and thus should occur more frequently compared to mutations which result in amino acid replacement. One of the most important problems for molecular evolution today is to ascertain whether silent mutations occur at a maximum and constant rate during the evolution of a molecule. Here we calculated the rate of synonymous substitution by using the formula: $k'_s = K'_s/2T$, where T is time since divergence. It is apparent from the data in Table 1 that synonymous substitutions do not occur at a constant rate. The rate of substitutions is high soon after divergence but slows down with longer divergence time. Only the Drosophila gene pair and the human versus rat gene pair have $k'_s$ values approaching the rate determined by Miyata to be uniform among different genes[31]. Kimura's explanation for this saturation effect is the lower detectability of synonymous substitutions for remote comparison. The alternative interpretation is that only a fraction of synonymous sites in GAPDH gene are neutral and are saturated by mutations in less than 100 million years. The other fraction of synonymous sites accumulates mutations more slowly, at a rate similar to that of the replacement sites. This latter fraction identifies sites which are neutral with regard to the protein, but are under selective pressure for other reasons. It has been proposed that codon usage and mRNA secondary structure are important constraints on this fraction. At any rate, the nonlinear and scattering nature of synonymous substitutions makes it an unreliable time clock except when used for calculating short divergence times. Similar results with regard to synonymous substitutions have been reported in the evolution of the globin gene[32].

Codon usage

We have tabulated the pattern of codon usage in the GAPDH gene from

various organisms (data not shown). The biased nature of codon usage for each gene was estimated by counting the number of codons which are omitted. For each of the following species, out of 61 possible codons for amino acid residues, the numbers of codons not used is given: 27 in yeast, 14 in Drosophila Gapdh-1, 15 in Drosophila Gapdh-2, 7 in chicken, 11 in rat, and 8 in man. The codon usage in GAPDH gene is thus not random, being biased in yeast and Drosophila, but relaxed in chicken, rat and man. The highly biased codon usage in yeast GAPDH gene was proposed to correlate with an abundance of isoacceptor tRNAs for maximum expression[33,34]. In view of the relaxed nature of GAPDH gene codon usage in other organisms, the strategy of codon usage for yeast may not apply to higher eucaryotes.

ACKNOWLEDGEMENT

We thank Drs. Jack Dixon, Richard Hynes, Savio Woo and Larry Kedes for sending us various cDNA libraries. We also thank Dr. Bruce Paterson for sending us the partial chicken GAPDH cDNA clone. The excellent technical assistance of Bruce Blakely and valuable criticism of the manuscript by Dr. Betty Keller are appreciated. This research was supported by research grant GM29179 from the National Institutes of Health, U.S. Public Health Service.

+A preliminary report has appeared in "Proceedings of the 1984 symposium on genetic engineering and biotechnology", June 18-July 6, 1984, pp. 37-49, Beijing, China.

*To whom all correspondence should be addressed.

REFERENCES

1. Liljas, A., and Rossman, M.G. (1974) Ann. Rev. Biochem. 43, 465-507.
2. Dayhoff, M.O. (1978) Altas of Protein Sequence and Structure Vol. 5, Suppl. 3, pp. 62-64, National Biomedical Research Foundation.
3. Scarpulla, R.C., and Wu, R. (1983) Cell 32, 473-482.
4. Holland, J.P., Labieniec, L., Swimmer, C., and Holland, M.J. (1983) J. Biol. Chem. 258, 5291-5299.
5. Tso, J.Y., Sun, X.-H., and Wu, R. (1985) J. Biol. Chem. (Submitted).
6. Dugaiczyk, A., Haron, J.A., Stone, E.M., Dennison, O.E., Rothblum, K.N., and Schwartz, R.J. (1983) Biochemistry 22, 1605-1613.
7. Grunstein, M., and Hogness, D.S. (1975) Proc. Natl. Acad. Sci. USA 72, 3961-3965.
8. Benton, W.D., and Davis, R.W. (1977) Science 196, 180-182.
9. Rigby, P.W.J., Dieckmann, M., Rhodes, C., and Berg, P. (1977) J. Mol. Biol. 113, 237-251.
10. Southern, E.M.,(1975) J. Mol. Biol. 98, 503-517.
11. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H., and Roe,

B.A. (1980) J. Mol. Biol. 143, 161-178.
12. Musti, P.W.J., Zehner, Z., Bostian, K.A., Paterson, B.M.,and Kramer, R.A. (1983) Gene 25, 133-143.
13. Okayama, H., and Berg, P. (1982) Mol. Cell. Biol. 2, 280-289.
14. Schwarzbauer, J.E., Tamkun, J.W., Lemischka, I.R., and Hynes, R.O. (1983) Cell 35, 421-431.
15. Chander, T., Stackhouse, R., Kidd, V.J., and Woo, S.L.C. (1983) Proc. Natl. Acad. Sci. USA 80, 1854-1848.
16. Nowak, K., Wolny, M., and Banas, T. (1981) FEBS Lett. 134, 143-146.
17. Lebherz, H., and Rutter, W.J. (1967) Science 157, 1198-1199.
18. Gunning, P., Ponte, p., Okayama, H., Engel, J., Blau, H., and Kedes, L. (1984) Nucleic Acid Res. 12, 1687-1696.
19. Sargent, T.D., Wu, J.R., Sala-Trepat, J.M., Wallace, R.B., Reyes, A.A., and Bonner, J. (1979) Proc. Natl. Acad. Sci. USA 76, 3256-3260.
20. Lemischka, I., and Sharp, P.A. (1982) Nature 300, 330-335.
21. Hedrick, S.M., Nielsen, E.A., Kavaler, J., Cohen, D.I., and Davis, M.M. (1984) Nature 308, 153-158.
22. Ponte, P., NG., S.-Y., Engle, J., Gunning, P., and Kedes, L. (1984) Nucleic Acid Res. 12, 1687-1696.
23. Cowan, N.J., Bobner, P.R., Fuchs, E.V., and Cleveland, D.W. (1983) Mol. Cell. Biol. 3, 1738-1745.
24. D'Eustachio, P., and Ruddle, F.H. (1983) Science 220, 919-924.
25. Hanauer, A., and Mandel, J.L. (1984) EMBO. J. 3, 2627-2633.
26. Benham, F.J., Hodgkinson, S., and Davies, K.E. (1984) EMBO J. 3, 2635-2640.
27. Piechaczyk, M., Blanchard, J.M., Riaad-El Sabouty, S., Dani, C., Marty, L., and Jeanteur, P. (1984) Nature 312, 469-471.
28. Arcari, P., Martinelli, R., and Salvatore, F. (1984) Nucleic Acids Res. 12, 9179-9189.
29. Dickerson, R.E. (1971) J. Mol. Evol. 1, 26-45
30. Kimura, M. (1980) J. Mol. Evol. 16, 111-120.
31. Miyata, T., Yasunaga, T., and Nishida, T. (1980) Proc. Natl. Acad. Sci. USA 77, 7328-7332.
32. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., and Dodgson, J. (1980) Cell 20, 555-566.
33. Bennetzen, J.L., and Hall, B.D. (1982) J. Biol. Chem. 257, 3026-3031.
34. Ikemura, T. (1982) J. Mol. Biol. 158, 573-597.