Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 8132–8136, November 1986
Biochemistry

Sequence analysis of the cDNA encoding human liver glycogen phosphorylase reveals tissue-specific codon usage

(muscle/G+C content)

CHRISTOPHER B. NEWGARD, KENICHI NAKANO, PETER K. HWANG, AND ROBERT J. FLETTERICK

Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143

Communicated by Joseph L. Goldstein, July 29, 1986

ABSTRACT **We have cloned the cDNA encoding glycogen phosphorylase (1,4-α-D-glucan:orthophosphate α-D-glucosyltransferase, EC 2.4.1.1) from human liver. Blot-hybridization analysis using a large fragment of the cDNA to probe mRNA from rabbit brain, muscle, and liver tissues shows preferential hybridization to liver RNA. Determination of the entire nucleotide sequence of the liver message has allowed a comparison with the previously determined rabbit muscle phosphorylase sequence. Despite an amino acid identity of 80%, the two cDNAs exhibit a remarkable divergence in G+C content. In the muscle phosphorylase sequence, 86% of the nucleotides at the third codon position are either deoxyguanosine or deoxycytidine residues, while in the liver homolog the figure is only 60%, resulting in a strikingly different pattern of codon usage throughout most of the sequence. The liver phosphorylase cDNA appears to represent an evolutionary mosaic; the segment encoding the N-terminal 80 amino acids contains >90% G+C at the third codon position. A survey of other published mammalian cDNA sequences reveals that the data for liver and muscle phosphorylases reflects a bias in codon usage patterns in liver and muscle coding sequences in general.**

Glycogen phosphorylase (1,4-α-D-glucan:orthophosphate α-D-glucosyltransferase, EC 2.4.1.1) isozymes play a vital role in mobilization of stored sugar in a variety of mammalian tissues. Three forms of the enzyme have been described that are distinguished by their electrophoretic mobilities on gels and their immunological properties (1–3). The three isozymes are tissue-specific; the brain type (also known as the fetal type) is predominant in adult brain and embryonic tissues, while the liver and muscle types are predominant in adult liver and skeletal muscle tissues, respectively (reviewed in ref. 4). The muscle form is the best characterized; both the primary sequence and the x-ray structure of rabbit muscle phosphorylase are known (5–8). The enzyme functions in muscle to provide glucose required for the energy of contraction. Its physiological role is distinct from the liver isozyme, which is centrally involved in the maintenance of blood glucose homeostasis, and from the brain form, which is associated primarily with provision of an emergency glucose supply during brief periods of anoxia or hypoglycemia (4, 9).

Comparisons of the protein and DNA sequences of the phosphorylase isozymes are required to understand the evolutionary and functional relationships among them. Further, such comparisons could ultimately provide insight into how the phosphorylase genes, and perhaps other multigene families, are regulated in a developmental and tissue-specific manner. We have described the cloning and sequencing of the rabbit muscle glycogen phosphorylase cDNA and portions of the C-terminal domain from human and rat muscle (6, 7). We report here the cDNA sequence and the deduced amino acid sequence for human liver glycogen phosphorylase. Comparison of liver and muscle phosphorylase sequences reveals extensive amino acid identity between the two isozymes. Remarkably, the nucleotide sequences are found to be less homologous because of a difference in codon usage patterns. Furthermore, a survey of published sequences reveals that the difference between the liver and muscle phosphorylase nucleotide sequences reflects a general tissue-specific codon usage bias in mammalian liver and muscle cDNA sequences.

MATERIALS AND METHODS

Cloning and Nucleotide Sequencing Strategy for the Human Liver Glycogen Phosphorylase cDNA. A summary of the cloning strategy is shown in Fig. 1. A partial rabbit muscle phosphorylase cDNA (6) encoding amino acids 573–742 (which includes the conserved pyridoxal phosphate cofactor binding site) was labeled with ^{32}P by nick-translation and used to screen a phage λgt11 cDNA library prepared from human liver (courtesy of A. DiLella and S. Woo, Baylor College of Medicine, Houston, TX). The initial screening of 50,000 clones, performed at low stringency [30% (vol/vol) formamide containing 5× NaCl/Cit (1× = 0.15 M NaCl/0.015 M sodium citrate), 5 mM NaH_2PO_4, 0.2 mg of salmon sperm DNA per ml, and Denhardt's solution (0.02% polyvinylpyrrolidone/0.02% Ficoll/0.02% bovine serum albumin) at 42°C], yielded a single clone of about 750 base pairs (HL-1). Sequencing by the dideoxynucleotide method (10) showed that this fragment was homologous to rabbit muscle phosphorylase over the region of the cDNA encoding amino acids from position 660 to the C terminus. The HL-1 fragment was amplified, purified, radiolabeled, and used to rescreen 100,000 clones from the same library under conditions of increased stringency (40% formamide at 42°C), yielding the overlapping fragments HL-2, HL-3, and HL-4. The HL-4 fragment was subsequently used to screen an additional 100,000 clones, yielding the overlapping fragments HL-5, HL-6, HL-7, and HL-8. When the HL-8 fragment was excised from λgt11 with the restriction endonuclease *Eco*RI, a second fragment of ≈360 nucleotides (HL-9) was observed because of an internal *Eco*RI site within the cDNA as shown in Fig. 1. HL-9 was used to screen an additional 500,000 clones, yielding only HL-10, which terminates 23 amino acids from the 5′ end of the coding region. To clone the last small coding fragment (HL-11), we used HL-9 to screen 80,000 clones from a second, randomly primed human liver cDNA library (courtesy of Jing-hsiung Ou, Hormone Research Institute, University of California, San Francisco).

Blot-Hybridization Analysis. Poly(A)$^+$ RNA samples, prepared by the method of Ashley and MacDonald (11), were electrophoresed through a formaldehyde/agarose gel, transferred onto a MSI magna nylon 66 membrane filter, and

Abbreviation: kb, kilobase(s).

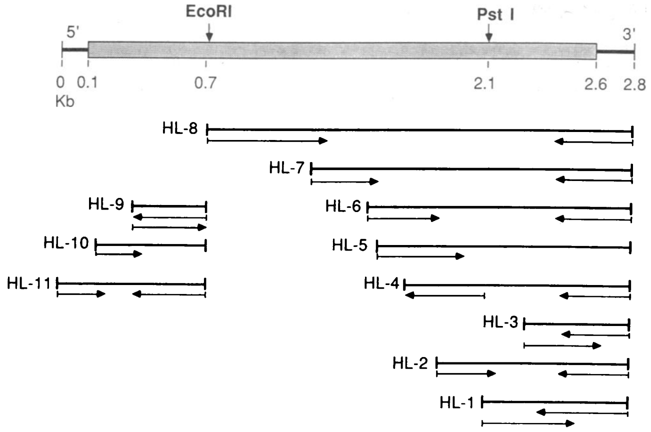

FIG. 1. Cloning and nucleotide sequencing strategy for the full-length human liver glycogen phosphorylase cDNA. The arrows under the HL fragments indicate the direction and distance sequenced. The restriction endonuclease sites *Pst* I and *Eco*RI used for subcloning are shown. kb, Kilobases.

hybridized with radiolabeled downstream rabbit muscle cDNA (encoding amino acids 304–842 and 20 bases of 3' untranslated region), downstream human liver cDNA (encoding amino acids 195–845 and 20 bases of 3' untranslated region), or upstream human liver cDNA (encoding amino acids 1–195 and 115 bases of 5' untranslated region). All panels were hybridized at the same low stringency (35% formamide containing 5× NaCl/Na Cit, 5 mM NaH_2PO_4, 0.2 mg of salmon sperm DNA per ml, and Denhardt's solution at 42°C) and were washed identically (twice for 45 min each at 55°C in 2× NaCl/Na Cit containing 0.1% $NaDodSO_4$).

RESULTS

Description of Human Liver Glycogen Phosphorylase cDNA and Comparison with the Rabbit Muscle Sequence. The nucleotide and amino acid sequences of rabbit muscle and human liver phosphorylases are compared in Fig. 2. Their amino acid sequences were found to be 80.1% identical; the percentage decreased to 73% at the nucleotide level. The liver sequence contains no insertions or deletions relative to that of muscle but extends three amino acid residues past the C terminus of the muscle form.

The sequence divergence of these genes is not random. The coding region of the liver phosphorylase cDNA has an overall G+C content of 48.7% compared with 60.3% in the rabbit muscle homolog. This difference can be largely ascribed to variation at the third codon position, where in liver phosphorylase 60% of the nucleotides are either deoxyguanosine or deoxycytidine residues compared with 85.8% in the muscle isozyme. The difference in G+C content is not due to the species difference because G+C content at the third codon position is similar in human muscle and rabbit muscle phosphorylase cDNA sequences.*

The liver and muscle phosphorylase sequences are completely unrelated at their 3' ends. The divergence begins abruptly at amino acid 830 and continues to the C terminus of the muscle sequence at residue 842 (the liver isozyme is three residues longer) and into the 3' untranslated regions. The human liver 3' untranslated region is 170 residues long and contains only 28% G+C; the rabbit muscle 3' untranslated region, in contrast, is 222 residues long and contains 60% G+C.

In contrast to the 3' end of the message, the 5' end is remarkably conserved between the liver and muscle sequences. A sharp increase in G+C content is observed in the liver sequence beginning at amino acid 80 and continuing upstream through the 5' untranslated region. The increase in G+C content in this region of the liver message is not due to the high amino acid homology, since the segment of nucleotide sequence encoding amino acids 30–80 (an important structural component of the AMP allosteric activation site) has a much higher third-codon-position G+C content (98%) than found in other segments encoding conserved 50-amino acid stretches such as 80–130 (44%; part of the active site) and 630–680 (56%, pyridoxal phosphate cofactor binding site) (8). Amino acid identity is maintained between the liver and muscle phosphorylases in spite of nucleotide divergence in the latter two fragments because the majority of changes in nucleotide sequence involve silent G·C ↔ A·T substitutions at the third codon position.

Blot-Hybridization Analysis of Liver, Muscle, and Brain Poly(A)$^+$ mRNA. Blot-hybridization analysis was undertaken to confirm that the clones isolated from the liver cDNA libraries hybridize preferentially with phosphorylase mRNA expressed in liver and to provide further insight into the structural relationship of the three phosphorylase isozymes. Blots containing brain, liver, and muscle poly(A)$^+$ mRNA, isolated from an adult rabbit, were probed with either downstream rabbit muscle (Fig. 3 *Left*), downstream human liver (3 *Center*), or upstream human liver (3 *Right*) phosphorylase cDNA. The muscle phosphorylase probe hybridized to single bands of ≈3.2 and ≈3.4 kb in the liver and muscle lanes, respectively. The signal was at least 30-fold more intense in the muscle lane than in the liver lane, even though only one-fifth as much muscle RNA was loaded. In addition, the muscle probe hybridized to two bands in the brain lane: the predominant one was ≈4.2 kb in size, and a minor species appeared to comigrate with the muscle form. The liver probe hybridized to the same species as the muscle probe in the muscle and liver lanes, but the signal was ≈3 times more intense in the liver lane than in the muscle lane. Interestingly, no signal was seen in the brain lane when the downstream liver probe was used. The upstream human liver phosphorylase probe (Fig. 3 *Right*) gave a different hybridization profile from that seen with the downstream liver probe. The muscle phosphorylase signal was ≈3 times as intense as that in the liver lane, even though the liver band was of similar intensity to that observed with the downstream liver probe. In addition, the upstream liver probe hybridized to the 4.2-kb band in the brain lane, which was previously recognized by the downstream muscle probe but not the downstream liver probe.

These data indicate that the phosphorylase clones isolated from the human liver cDNA library encode the liver isozyme and not the brain or muscle forms and that the upstream probe recognizes a region that is more highly conserved in the three phosphorylase isozymes than that detected by the downstream probe.

Comparison of Codon Usage in Other Liver and Muscle cDNA Sequences. To determine whether the difference in G+C content observed between liver and muscle phosphorylase isozymes indicates a general tissue-specific codon usage bias, we compared the published cDNA sequences of

*The comparison covers the portions of the human muscle phosphorylase coding sequence that have been characterized, including the previously described cDNA fragment encoding amino acids from position 733 to the C terminus (6) and unpublished data on gene fragments encoding amino acids 1–233 and a cDNA clone encoding residues 580–680. Over these regions, the human muscle phosphorylase sequence has a third-codon-position G+C content of 75%, as compared to 83% in the rabbit muscle sequence and 58% in the human liver message. Human muscle coding sequences are not always lower in G+C content than their rabbit muscle counterparts, since human muscle aldolase contains 84% G+C at the third codon position compared to 81% for the rabbit muscle aldolase message (ref. 12; D. Tolan and E. Penhoet, personal communication).

```
                                                                                                                                1
                                                                                                                            MetGly
LIVER     GTTGAAAGCTCCTGGCGCGGCGGGGCGGACTCCACCCCTGCCCGGCAGCCCAGCGCCTCCGGCCGCACTTCCAGCTCTCTGCGCAGCCCGCCGCGCAGCCCGCCGCCC-CAGCCATGGGC
MUSCLE    -------------------------------------- C  T   T GAGGCAC    A    A C G   -  C  TCCTT CA GTCT   C G       T   G        TC
                                                                                                                               Ser
                                                                20                                                          40
          GluProLeuThrAspGlnGluLysArgArgGlnIleSerIleArgGlyIleValGlyValGluAsnValAlaGluLeuLysLysSerPheAsnArgHisLeuHisPheThrLeuValLys
LIVER     GAACCGCTGACAGACCAGGAGAAGCGGCGGCAGATCAGCATCCGCGGCATCGTGGGCGTGGAGAACGTGGCAGAGCTGAAGAAGAGTTTCAACCGGCACCTGCACTTCACGCTGGTCAAG
MUSCLE    CGG  C   T              A AAA   A      G G      C G CC           T   A G        A  A AC        C        T     C     G
          Arg      Ser               Lys         Val      LeuAla               Thr            Asn
                                                                60                                                          80
          AspArgAsnValAlaThrThrArgAspTyrTyrPheAlaLeuAlaHisThrValArgAspHisLeuValGlyArgTrpIleArgThrGlnGlnHisTyrTyrAspLysCysProLysArg
LIVER     GACCGCAACGTGGCCACCACCCGCGACTACTACTTCGCGCTGGCGCACACGGTGCGGGACCACCTGGTGGGGCGCTGGATCCGCACGCAGCAGCACTACTACGACAAGTGCCCCAAGAGG
MUSCLE            T        GC G  A              A     C     T     C        C                          A        T  G   GA
                            Pro                                                                                 Glu   Asp
                                                                100                                                         120
          GluTyrTyrLeuSerLeuGluPheTyrMetGlyArgThrLeuGlnAsnThrMetIleAsnLeuGlyLeuGlnAsnAlaCysAspGluAlaIleTyrGlnLeuGlyLeuAspIleGluGlu
LIVER     GAATATTACCTCTCTCTGGAATTTTACATGGGCCGAACATTACAGAACACCATGATCAACCTCGGTCTGCAAAATGCCTGTGATGAGGCCATTTACCAGCTTGGATTGGATATAGAAGAG
MUSCLE    ATC  C     G   T    G  C           G  GC              G G     G C T  G G        C  C       CC        G  CC     C G  G
          Ile                                                   Val      Ala   Glu                  Thr                  Met
                                                                140                                                         160
          LeuGluGluIleGluGluAspAlaGlyLeuGlyAsnGlyGlyLeuGlyArgLeuAlaAlaCysPheLeuAspSerMetAlaThrLeuGlyLeuAlaAlaTyrGlyTyrGlyIleArgTyr
LIVER     TTAGAAGAAATTGAAGAAGATGCTGGACTTGGCAATGGTGGTCTTGGGAGACTTGCTGCCTGCTTCTTGGATTCCATGGCAACCCTGGGACTTGCAGCCTATGGATACGGCATTCGGTAT
MUSCLE    C G  G     C  G  G  C  C  G  G     C  C  C  G  CC G  G  A     T  TC    C           A  C  G  T           C     G  C  C

                                                                180                                                         200
          GluTyrGlyIlePheAsnGlnLysIleArgAspGlyTrpGlnValGluGluAlaAspAspTrpLeuArgTyrGlyAsnProTrpGluLysSerArgProGluPheMetLeuProValHis
LIVER     GAATATGGGATTTTCAATCAGAAGATCCGAGATGGATGGCAGGTAGAAGAAGCAGATGATTGGCTCAGATATGGAAACCCTTGGGAGAAGTCCCGCCCAGAATTCATGCTGCCTGTGCAC
MUSCLE      G T      C  T  C         T C GG  C      A G     G  T     C     TC C  C  C     C         G    G  T  G    CA
             Phe                     CysGly         Met                                             Ala            Thr
                                                                220                                                         240
          PheTyrGlyLysValGluHisThrAsnThrGlyThrLysTrpIleAspThrGlnValValLeuAlaLeuProTyrAspThrProGluProGlyTyrMetAsnAsnThrValAsnThrMet
LIVER     TTCTATGGAAAAGTAGAACACACCAACACCGGGACCAAGTGGATTGACACTCAAGTGGTCCTGGCTCTGCCATATGACACCCCCGAGCCCGGCTACATGAATAACACTGTCAACACCATG
MUSCLE            CCG   G  G       G CAG   G        G G     A  G     G     CA    C  C     A  G T   T      CGC  C   GTC
                   Arg            SerGln   Ala      Val                     Met               Val         Arg      Val
                                                                260                                                         280
          ArgLeuTrpSerAlaArgAlaProAsnAspPheAsnLeuArgAspPheAsnValGlyAspTyrIleGlnAlaValLeuAspArgAsnLeuAlaGluAsnIleSerArgValLeuTyrPro
LIVER     CGCCTCTGGTCTGCTCGGGCACCAAATGACTTTAACCTCAGAGACTTTAATGTTGGAGACTACATTCAGGCTGTGCTGGACCGAAACCTGGCCGAGAACATCTCCCGGGTCCTCTATCCC
MUSCLE               G  CAA   G  C        C       AG     C  C  C  T G      C     C           C        T           T        G  C
                         Lys                     Lys               Gly
                                                                300                                                         320
          AsnAspAsnPhePheGluGlyLysGluLeuArgLeuLysGlnGluTyrPheValValAlaAlaThrLeuGlnAspIleIleArgArgPheLysAlaSerLysPheGlySerThrArgGly
LIVER     AATGACAATTTTTTTGAAGGGAAGGAGCTAAGATTGAAGCAGGAATACTTTGTGGTGGCTGCAACCTTGCAAGATATCATCCGCCGTTTCAAAGCCTCCAAGTTTGGCTCCACCCGTGGT
MUSCLE      C  T  C  C  C  G           GC GC          G     C        G  C   C    G  C           C    GT        A  C    G CG GACCCC
                                                                                                    Ser               CysArgAspPro
                                                                340                                                         360
          GlnGlyThrValPheAspAlaPheProAspGlnValAlaIleGlnLeuAsnAspThrHisProArgIleAlaIleProGluLeuMetArgIlePheValAspIleGluLysLeuProTrp
LIVER     CAAGGAACTGTGTTTGATGCCTTCCCGGATCAGGTGGCCATCCAGCTGAATGATACTCACCCTCGCATCGCGATCCCTGAGCTGATGAGGATTTTTGTGGATATTGAAAAACTGCCCTGG
MUSCLE    GTGC C  CAAC  C               A    A           C     C  C     CTCGC G  C     C     C      G GC G     CC G  GCGG   GA
          ValArg   Asn                  Lys                              SerLeu                     ValLeu      Leu   Arg   Asp

                                                                380                                                         400
          SerLysAlaTrpGluLeuAsnGlnLysThrPheAlaTyrThrAsnHisThrValLeuProGluAlaLeuGluArgTrpProValAspLeuValGluLysLeuLeuProArgHisLeuGlu
LIVER     TCCAAGGCATGGGAGCTCAACCAGAAGACCTTCGCCTACACCAACCACACAGTGCTCCCGGAAGCCCTGGAGCGCTGGCCCGTGGACCTGGTGGAGAAGCTGCTCCCTCGACATTTGGAA
MUSCLE    GA      C     AG G CGGTA     G G                  G     G  C  G  G                  C    CC      C      G  G  C  CC CC G
          Asp            ValThrVal      Cys                                                   His   Leu   Thr                  Gln
                                                                420                                                         440
          IleIleTyrGluIleAsnGlnLysHisLeuAspArgIleValAlaLeuPheProLysAspValAspProLeuArgArgMetSerLeuIleGluGluGluGlySerLysArgIleAsnMet
LIVER     ATCATTTATGAGATAAATCAGAAGCATTTAGATAGAATTGTGGCCTTGTTTCCTAAAGATGTGGACCCTCTGAGAAGGATGTCTCTGATAGAAGAGGAAGGAAGCAAAAGGATCAACATG
MUSCLE         C  C     C  C   CGCTTCC GA CC CG G C   TGC   C  GGGG  C  C    GG   C GC C     G  G  G  G    GC CCGTG   C C
                               ArgPhe   Asn   ValAla   Ala      Gly         Arg                  Val      GlyAlaVal
                                                                460                                                         480
          AlaHisLeuCysIleValGlySerHisAlaValAsnGlyValAlaLysIleHisSerAspIleValLysThrLysValPheLysAspPheSerGluLeuGluProAspLysPheGlnAsn
LIVER     GCCCATCTCTGCATTGTCGGTTCCCATGCTGTGAATGGCGTGGCTAAAATCCACTCAGACATCGTGAAGACTAAAGTATTCAAGGACTTCAGTGAGCTAGAACCTGACAAGTTTCAGAAT
MUSCLE      G  C  G     C CG  C  G  C  C  C  C  T      CGC        G  G   C C    AG CCA C            TAC     G  G   C T     C     C
                         Ala                           Arg         Glu   Leu   LysThrIle            Tyr            His
                                                                500                                                         520
          LysThrAsnGlyIleThrProArgArgTrpLeuLeuLeuCysAsnProGlyLeuAlaGluLeuIleAlaGluLysIleGlyGluAspTyrValLysAspLeuSerGlnLeuThrLysLeu
LIVER     AAAACCAATGGGATCACTCCAAGGCGCTGGCTCCTACTCTGCAACCCAGGACTTGCAGAGCTCATAGCAGAGAAAATTGGAGAAGACTATGTGAAAGACCTGAGCCAGCTGACGAAGCTC
MUSCLE      G     C  C     C  TC          GG T  G  T     T  G  G  C   A    T  T   CGC  C  G  G  G  CA CTC       GA       CGC     G
                                           Val                        Ile         Arg         Glu   IleSer      Asp      Arg
                                                                540                                                         560
          HisSerPheLeuGlyAspAspValPheLeuArgGluLeuAlaLysValLysGlnGluAsnLysLeuLysPheSerGlnPheLeuGluThrGluTyrLysValLysIleAsnProSerSer
LIVER     CACAGCTTCCTGGGTGATGATGTCTTCCTCCGGGAACTCGCCAAGGTGAAGCAGGAGAATAAGCTGAAGTTTTCTCAGTTCCTGGAGACGGAGTACAAAGTGAAGATCAACCCATCCTCC
MUSCLE     T TCG ATG   AC  C  A C    A       TG G  G  A           A  C           CG GGCC A        G   A     G  CC C        CAA   G
          Leu   TyrValAsp   GluAla   Ile   AspVal                                 AlaAlaTyr      Arg            His         Asn
                                                                580                                                         600
          MetPheAspValGlnValLysArgIleHisGluTyrLysArgGlnLeuLeuAsnCysLeuHisValIleThrMetTyrAsnArgIleLysLysAspProLysLysLeuPheValProArg
LIVER     ATGTTTGATGTCCAGGTGAAGAGGATACATGAGTACAAGCGACAGCTCTTGAACTGTCTGCATGTGATCACGATGTACAACCGCATTAAGAAAGACCCTAAGAAGTTATTCGTGCCAAGG
MUSCLE    C    C  C     A     AC    C     A        G     GC C     C     C        CC             C     G  G  C  T     TG T     TC
          Leu                                                                     Leu                  Glu   Asn   PheVal
                                                                620                                                         640
          ThrValIleIleGlyGlyLysAlaAlaProGlyTyrHisMetAlaLysMetIleIleLysLeuIleThrSerValAlaAspValValAsnAsnAspProMetValGlySerLysLeuLys
LIVER     ACAGTTATCATTGGTGGTAAAGCTGCCCCAGGATATCACATGGCCAAAATGATCATAAAGCTGATCACTTCAGTGGCAGATGTGGTGAACAATGACCCTATGGTTGGAAGCAAGTTGAAA
MUSCLE      C  C  G     A  G  G  C  A  T  G  C           G        C  A  T     CG CA T GG        C   C C     GG        GA CGCC CCGT
                Met                                                                  AlaIleGly            His      Val      AspArg   Arg
                                                                660                                                         680
          ValIlePheLeuGluAsnTyrArgValSerLeuAlaGluLysValIleProAlaThrAspLeuSerGluGlnIleSerThrAlaGlyThrGluAlaSerGlyThrGlyAsnMetLysPhe
LIVER     GTCATCTTCTTGGAGAACTACAGAGTATCTCTTGCTGAAAAAGTCATTCCAGCCACAGATCTGTCAGAGCAGATTTCCACTGCAGGCACCGAAGCCTCGGGGACAGGCAATATGAAGTTC
MUSCLE             C           C G  C  A  G     G     G  C  G  TG C  C  C  G        C     C  G        G     C  C  C     C
                                                                Ala
                                                                700                                                         720
          MetLeuAsnGlyAlaLeuThrIleGlyThrMetAspGlyAlaAsnValGluMetAlaGluGluAlaGlyGluGluAsnLeuPheIlePheGlyMetSerIleAspAspValAlaAlaLeu
LIVER     ATGCTAAATGGGGCCCTAACTATCGGGACCATGGATGGGGCCAATGTGGAAATGGCAGAAGAAGCTGGGGAAGAGAACCTGTTCATCTTTGGCATGAGCATAGATGATGTGGCTGCTTTG
MUSCLE         C     A     G  C  T  C  G     C  C     C     G     C  G  G  G  A  G      T C               C GG G  A       ACAGAC C
                                                                                    Phe               ArgValGlu      AspArg
                                                                740                                                         760
          AspLysLysGlyTyrGluAlaLysGluTyrTyrGluAlaLeuProGluLeuLysLeuValIleAspGlnIleAspAsnGlyPhePheSerProLysGlnProAspLeuPheLysAspIle
LIVER     GACAAGAAAGGGTACGAGGCAAAAGAATACTATGAGGCACTTCCAGAGCTGAAGCTGGTCATTGATCAAATTGACAATGGCTTTTTTTCTCCCAAGCAGCCTGACCTCTTCAAAGATATC
MUSCLE       C   G       A C  CC G  G     C  CCGCA C  G     TCG  A A    C  G  GC GAGT GC     C  C  C  G        G     G     G  C  T
             GlnArg      Asn   Gln         AspArgIle         ArgGlnIle   Glu   LeuSerSer
                                                                780                                                         800
          IleAsnMetLeuPheTyrHisAspArgPheLysValPheAlaAspTyrGluAlaTyrValLysCysGlnAspLysValSerGlnLeuTyrMetAsnProLysAlaTrpAsnThrMetVal
LIVER     ATCAACATGCTATTTTATCATGACAGGTTTAAAGTCTTTGCAGACTACGAAGCCTATGTCAAGTGTCAAGATAAAGTGAGTCAGCTGTACATGAATCCAAAGGCCTGGAACACAATGGTA
MUSCLE    G          CA GC C      C    C              T  T    AG  C        C  G  GCG   C  CGCCT      A   C  C GA AG    CGCGG     G
          Val         MetHis                                 Glu               GluArg      Ala      Lys      ArgGlu   ThrArg

                                                                820                                                         840
          LeuLysAsnIleAlaAlaSerGlyLysPheSerSerAspArgThrIleLysGluTyrAlaGlnAsnIleTrpAsnValGluProSerAspLeuLysIleSerLeuSerAsnGluSerAsn
LIVER     CTCAAAAACATAGCTGCCTCGGGGAAATTCTCCAGTGACCGAACAATTAAAGAATATGCCCAAAACATCTGGAACGTGGAACCTTCAGATCTAAAGATTTCTCTATCCAATGAATCTAAC
MUSCLE    A  CGG        CA       C  G        C     C  C   GCCC G  C  G GGG G      GGT     G  C  GCGG AGCG C GC AGCCC  G C  GAAG TA
          IleArg         Thr                              AlaGln      ArgGlu      Gly            ArgGlnArgLeuProAlaProAsp   LysIle
          LysValAsnGlyAsn *
LIVER     AAAGTCAATGGAAATTGAACTCTACAATGTCTCTAGAAAACATAGCTTCTTACTGAACTTGAACATTTTTACAACATTCACTGGTTTTTGTTTTGTTAGCTAATAATCTATAATAGTTGA
MUSCLE    CCCTAG------------G CAG  CCCAA G GGCCCTG  GTCTGCAAGCT  GGGCCAGCGCCAGC C CCGTCCAGAGT  GGGG CC G AGTCAG CCTCCAAG CCCCTCCTG
          Pro *
LIVER     GTATCTCTGGAATGGGGAGGGAAATTATATGTAATAGAGCTTAAAAATAAAGTGTCAATTTCCAAGGA----------------------------------------------------
MUSCLE    AA C C CATTCCCCCCAGAA C  GGTCCCAGTGCCC   G GCCTCAGG CAC GGGCCC  TCCT TTTATGGGGTCCGACCAACTGCGCCCACTCCCCAATAAACTCTCCCTCCTT
```

FIG. 2. *(Legend appears at the bottom of the opposite page.)*

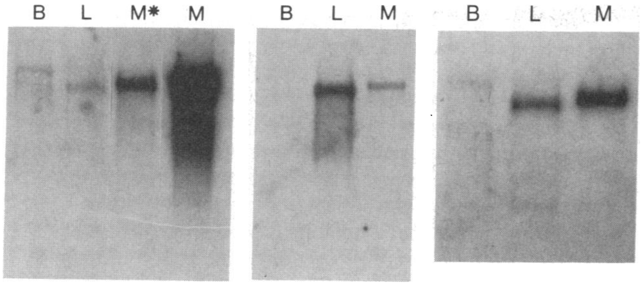

FIG. 3. Tissue distribution of glycogen phosphorylase mRNA in adult rabbit tissues, assessed with liver and muscle phosphorylase probes. Filters containing poly(A)$^+$ RNA were hybridized with radiolabeled downstream rabbit muscle cDNA (*Left*), downstream human liver cDNA (*Center*), or upstream human liver cDNA (*Right*) as described. All panels were hybridized with the same amount of probe (6×10^7 cpm; all probes had specific activities of $\approx 5 \times 10^8$ cpm/μg). Lanes: B, 5 μg of poly(A)$^+$ RNA isolated from brain; L, 5 μg of poly(A)$^+$ RNA isolated from liver; M, 1 μg of poly(A)$^+$ RNA isolated from hind-leg skeletal muscle. All lanes were autoradiographed for 14 hr at −80°C, except for lane M*, which was exposed for only 1.5 hr.

24 liver and 13 muscle proteins from human, rat, and rabbit sources. Only protein-encoding regions of the cDNAs were considered in this analysis. Liver sequences were found to contain an average of 51 ± 6% G+C overall and 59 ± 12% G+C in third codon positions, while muscle sequences contained 58 ± 4% and 80 ± 10%, respectively. Fig. 4 shows the G+C percentage at the third codon position of these sequences plotted against the G+C percentage overall. Muscle sequences as a group were significantly higher in G+C percentage than were liver sequences, both overall and at the third codon position. Fitting this pattern are the muscle and liver phosphorylases, as discussed above, as well as the aldolases, the only other protein for which the cDNA sequences of liver and muscle isozymes are known (12, 13). The grouping of sequences for tissue comparison from three different species in Fig. 4 is justified on the following grounds. First, we included examples of three sequences expressed in skeletal muscle (phosphorylase, aldolase, and creatine kinase) and one in liver (phenobarbital-inducible cytochrome P-450) from more than one species that show minimal variation in G+C content. Second, a plot of the 13 human liver and 5 human muscle sequences shows that the points continue to group in a tissue-specific manner, and a nearly identical slope (0.41 vs. 0.37) and correlation coefficient (0.92 vs. 0.88) were obtained as for the three species together (data not shown).

DISCUSSION

In this paper we have compared the DNA and amino acid sequences of liver and muscle glycogen phosphorylases. The striking disparity that emerges in codon usage patterns has led us to consider whether there is a relationship between tissue-specific genes and G+C content. Our findings suggest that tissue-dependent factors influence codon usage and nucleotide composition in liver and muscle coding sequences, specifically by increasing or decreasing the frequency of deoxyguanosine or deoxycytidine residues at third codon positions.

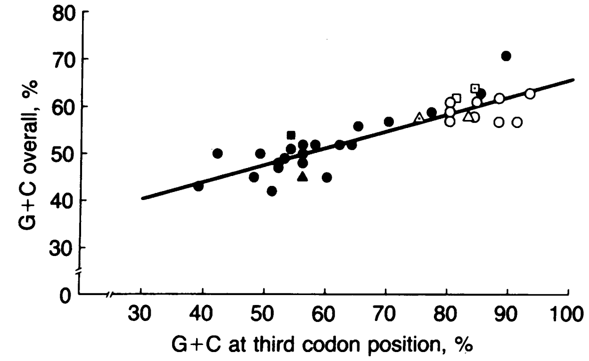

FIG. 4. G+C percentage at the third codon position vs. the overall G+C percentage in liver and muscle cDNA sequences. Human liver phosphorylase (▲), rabbit muscle phosphorylase (△), human muscle phosphorylase (◬), human liver aldolase (■), rabbit muscle aldolase (□), human muscle aldolase (⊡), other liver proteins (●), and other muscle proteins (○) are included. The other liver proteins are phosphoglycerate kinase, serum albumin, triose phosphate isomerase, haptoglobin-2, phenylalanine hydroxylase, alcohol dehydrogenase, apolipoproteins A-I, A-II, B, C-II, and C-III from human; bifunctional peroxisomal dehydrogenase, phospho*enol*pyruvate carboxykinase, ornithine aminotransferase, cathepsin B, α-casein, β-casein, γ-casein, apolipoprotein E, and phenobarbital-inducible cytochromes P-450 from rat and rabbit. The other muscle proteins include actin, acetylcholine receptor, and myoglobin from human; creatine kinase from rat; troponin C and I, beta-tropomyosin, Ca^{2+}/Mg^{2+}-ATPase (fast twitch), and creatine kinase from rabbit. Most of the sequences contain the entire coding region of the protein; a few from each tissue are fragments of at least 200 nucleotides in length. The muscle myoglobin and acetylcholine receptor sequences were derived from their respective gene sequences. All sequences were obtained from the Nucleic Acid Sequence Database and the Nucleic Acid Query Program, National Biomedical Research Foundation (Georgetown University, Washington, DC). The slope of the line = 0.37; the correlation coefficient = 0.88.

Liver and skeletal muscle genes may represent the two extremes of the evolutionary spectrum with regard to tissue-specific codon usage patterns (59% and 80% third-codon-position G+C content in liver and muscle sequences, respectively). Mammalian coding sequences are generally high in third-codon-position G+C content (average of 65%; 72% with immunoglobulins removed from the sample), suggesting that mammalian tissues other than skeletal muscle will have coding sequences enriched in G+C (14–17). A preliminary survey of cDNA sequences from mammalian pancreas indicates an intermediate G+C content [68% at the third codon position, an average of five proteins (18–21)]. More sequence information will be required to determine codon usage patterns in tissues other than liver and muscle.

The reasons for the development of tissue-specific patterns of G+C content are not known. It is of interest, however, that organisms such as the thermophilic bacteria and the proto-

FIG. 2. Comparison of the cDNA and deduced amino acid sequences of human liver and rabbit muscle (7) glycogen phosphorylases, including the 5′ and 3′ untranslated regions of the messages. The numbers above the liver sequence indicate the amino acid number of the liver and muscle isozymes. The residues immediately after the initiator methionine (glycine in liver, serine in muscle) are designated number 1. The liver nucleotide and amino acid sequences are presented in their entirety on the upper two lines of each row, while only nucleotide and amino acid residues that are nonidentical in the muscle sequence are shown on the two bottom lines. The single dashed lines in the 5′ untranslated region of each of the messages indicate single base-pair gaps introduced to allow maximal sequence alignment within the 5′ regions. The stop codons of the two messages are indicated by an asterisk (the dashed lines after the muscle stop codon indicate that the liver message continues for three amino acid residues before encountering a stop codon of its own). The conserved AATAAA polyadenylylation signals in the two messages are underlined.

zoan *Leischmania*, which are exposed to the environmental stresses of high temperature and low pH, respectively, have high G+C content in their coding sequences (22–24), presumably because the greater stability of G·C base pairs aids the processes of gene replication, transcription, and, to a lesser extent, translation.[†] It is possible that skeletal muscle, which undergoes a fall in pH and a rise in temperature during exercise (26, 27), represents a similarly stressful environment that selectively maintains high G+C content in expressed genes.

Alternatively, the difference in G+C content of liver and muscle genes may play a role in regulating tissue-specific expression at the transcriptional or translational level. Although this study has focused on the nucleotide composition of protein-encoding portions of cDNA sequences, G+C bias can extend into 5′ and 3′ flanking portions of genes (15, 16, 28). Since these regions are important in gene regulation, divergence in G+C content may result in the generation of consensus sequences that are recognized in a tissue-specific manner by the proteins involved in expression. When flanking regions of genes mirror the G+C content of their coding regions, methylation patterns also may be affected. Most *in vitro* and some *in vivo* experiments suggest that genes that have heavily methylated flanking regions are transcriptionally inhibited (29, 30).

At a practical level, the strong bias for deoxyguanosine or deoxycytidine residues at the third codon position in muscle genes can be useful in designing cloning strategies. For example, synthetic oligonucleotides based on muscle protein sequence can be made less degenerate by eliminating the codons ending in A or T for amino acids such as glutamic acid and asparagine. Such an approach was recently used to clone rabbit muscle glycogen synthetase in our laboratory (K.N. and R.J.F., unpublished data). In addition, codon bias at the third codon position can be used to identify frame shift errors when sequencing muscle cDNAs (31).

Tissue-specific codon usage patterns also can provide a measure of the evolutionary "distance" between tissues and specific genes within tissues. When viewed in this context, the mosaic G+C content of the liver phosphorylase gene is particularly intriguing. We propose that the high G+C content in the N-terminal region of the liver message indicates that this segment was spliced onto the liver gene from the muscle gene long after the divergence of liver and muscle tissues. Support for this contention includes the fact that a higher level of nucleotide conservation exists between muscle and liver phosphorylases within this region than in other regions of their cDNA. Furthermore, calculation of the evolutionary distance between liver and muscle phosphorylases over the amino acid stretches 30–80, 80–130, and 630–680 by the method of Kimura (32) using only the third codon position for analysis indicates that the N-terminal fragment diverged from a common sequence some 220 million years ago, whereas the other two segments (which are representative of the rest of the protein) diverged from a common ancestor about 440 million years ago. Conservation of the 5′ untranslated regions between the tissue-specific phosphorylase messages also reflects a more recent common ancestor. In contrast, neither the 3′ untranslated regions of liver and muscle phosphorylases nor the 5′ or 3′ untranslated regions of liver and muscle aldolases (33) exhibit significant nucleotide identity. Finally, sequence analysis of the human gene for muscle glycogen phosphorylase shows that the 5′ fragment, stretching from −70 to the codon corresponding to amino acid 80 in the coding region, is encoded by a single exon (J. Burke, P.K.H., and R.J.F., unpublished data). Thus, one possible mechanism for the construction of the mosaic human liver gene is exon shuffling (34), a process that appears to be involved in the evolution of another mammalian gene—that for the receptor of low density lipoprotein (35).

The authors thank C. Craik, P. Hoben-Carter, S. Sprang, J. Ou, R. Edwards, N. Stahl, T. Standing, and M. Selby for helpful discussions. This work was supported by National Institutes of Health Grant AM 32822 and a postdoctoral grant to C.B.N. from the Juvenile Diabetes Foundation.

[†] The stability of codon–anticodon interactions is less affected by the choice of G·C vs. A·T pairs at the third codon position than is expected from their binding energies because of compensatory covalent modifications of residues within and surrounding the anticodon (25).

1. Wosilait, W. D. & Sutherland, E. W. (1956) *J. Biol. Chem.* **218**, 469–481.
2. Schane, H. P. (1965) *Anal. Biochem.* **11**, 371–394.
3. Henion, W. F. & Sutherland, E. W. (1957) *J. Biol. Chem.* **224**, 477–488.
4. David, E. & Crerar, M. M. (1986) *Biochim. Biophys. Acta* **880**, 78–90.
5. Titani, K., Kaide, A., Hermann, J., Ericsson, L. H., Kumar, S., Wade, R. D., Walsh, K. A., Neurath, H. & Fischer, E. H. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 4762–4766.
6. Hwang, P. K., See, Y. P., Vincentini, A. M., Powers, M. A., Fletterick, R. J. & Crerar, M. M. (1985) *Eur. J. Biochem.* **152**, 267–274.
7. Nakano, K., Hwang, P. K. & Fletterick, R. J. (1986) *FEBS Lett.*, in press.
8. Fletterick, R. J. & Madsen, N. B. (1980) *Annu. Rev. Biochem.* **49**, 31–61.
9. Hers, H.-G. (1976) *Annu. Rev. Biochem.* **45**, 167–189.
10. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
11. Ashley, P. L. & MacDonald, R. J. (1985) *Biochemistry* **24**, 4512–4520.
12. Tolan, D. R., Amsden, A. B., Putney, S. D., Urdea, M. & Penhoet, E. E. (1984) *J. Biol. Chem.* **259**, 1127–1131.
13. Paolella, G., Santamaria, R., Izzo, P., Constanzo, P. & Salvatore, F. (1984) *Nucleic Acids Res.* **12**, 7401–7410.
14. Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pave, A. (1980) *Nucleic Acids Res.* **8**, r49–r62.
15. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. & Rodier, F. (1985) *Science* **228**, 953–958.
16. Ikemura, T. (1985) *Mol. Biol. Evol.* **2**, 13–34.
17. Salser, W. (1977) *Cold Spring Harbor Symp. Quant. Biol.* **412**, 985–1002.
18. MacDonald, R. J., Stary, S. J. & Swift, G. H. (1982) *J. Biol. Chem.* **257**, 9724–9732.
19. MacDonald, R. J., Swift, G. H., Quinto, C., Swain, W., Pictet, R. L., Nickovits, W. & Rutter, W. J. (1982) *Biochemistry* **21**, 1453–1463.
20. Swift, G. H., Dagorn, J.-C., Ashley, P. L., Cummings, S. W. & MacDonald, R. J. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 7263–7267.
21. Quinto, C., Quiroga, M., Swain, W. F., Nickovits, W. C., Standring, D. N., Pictet, R. L., Valenzuela, P. & Rutter, W. J. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 31–35.
22. Bibb, M. J., Findlay, P. R. & Johnson, M. W. (1984) *Gene* **30**, 157–165.
23. Winter, G., Koch, G. L. E., Hartley, B. S. & Barker, D. G. (1983) *Eur. J. Biochem.* **132**, 383–387.
24. Kagawa, Y., Nojima, H., Nukiwa, N., Ishizuka, M., Nakajima, T., Yasuhara, T., Tanaka, T. & Oshima, T. (1984) *J. Biol. Chem.* **259**, 2956–2960.
25. Grosjean, H. & Chantrenne, H. (1980) in *Molecular Biology, Biochemistry and Biophysics, Chemical Recognition in Biology*, ed. Chapeville, F. & Henni, A.-L. (Springer, Berlin, New York), Vol. 32, pp. 347–367.
26. Kruk, B., Kaciuba-Uscilko, H., Nazar, K., Greenleaf, J. E. & Kozlowski, S. J. (1985) *Appl. Physiol.* **58**, 1444–1448.
27. Booth, F. W. & Watson, P. A. (1985) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **44**, 2293–2299.
28. Tsutsumi, K., Mukai, T., Tsutsumi, R., Hidaka, S., Arai, Y., Hori, K. & Ishikawa, K. (1985) *J. Mol. Biol.* **101**, 153–160.
29. Max, E. E. (1984) *Nature (London)* **310**, 100.
30. Bird, A. (1986) *Nature (London)* **321**, 209–213.
31. Bibb, M. J., Findlay, P. R. & Johnson, M. W. (1984) *Gene* **30**, 157–166.
32. Kimura, M. (1980) *J. Mol. Evol.* **16**, 111–120.
33. Joh, K., Mukai, T., Yatsuki, H. & Hori, K. (1985) *Gene* **39**, 17–24.
34. Gilbert, W. (1978) *Nature (London)* **271**, 501–502.
35. Sudhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. (1985) *Science* **228**, 815–822.