Intron/Exon Structure of the Human Gene for the Muscle Isozyme of Glycogen Phosphorylase

J. Burke,[1] P. Hwang,[1] L. Anderson,[2] R. Lebo,[2] F. Gorin,[1] and R. Fletterick[1]
[1]*Department of Biochemistry and Biophysics and* [2]*Howard Hughes Medical Institute, University of California, San Francisco, 94143-0448*

ABSTRACT **The intron/exon organization of the human gene for glycogen phosphorylase has been determined. The segments of the polypeptide chain that corresponds to the 19 exons of the gene are examined for relationships between the three-dimensional structure to the protein and gene structure. Only weak correlations are observed between domains of phosphorylase and exons. The nucleotide binding domains that are found in phosphorylase and other glycolytic enzymes are examined for relationships between exons of the genes and structures of the domains. When mapped to the three-dimensional structures, the intron/exon boundaries are shown to be widely distributed in this family of protein domains.**

Key words: evolution, protein structure, nucleotide binding domain, gene sequence

INTRODUCTION

Glycogen phosphorylase is the largest and most intricately regulated protein in the family of glycolytic enzymes. Phosphorylases play a pivotal role in intracellular energy metabolism by catalyzing the phosphorolysis of glycogen to glucose-1-phosphate. Of the three known mammlian isozymes—muscle, brain, and liver—the muscle isozyme from rabbit is best characterized; its amino acid[1] and cDNA sequences[2] are determined, and X-ray crystallographic structures are refined at high resolution.[3,4] The enzyme functions as a dimer of two identical subunits and is interconverted between two forms, a phosphorylated *a* form and a dephosphorylated *b* form. This phosphorylation, at Ser 14, activates the enzyme, though its activity may be allosterically inhibited by the binding of ligands such as glucose or caffeine.[5,6] The dephosphorylated *b* form is activatable by AMP binding.[7] In all, the enzyme has five allosteric regulatory sites on each subunit that are distributed in three spatially distant regions. Details of the structural and biochemical analysis are presented in recent reviews.[8–10]

Although the protein from muscle has been well characterized, little is known about the family of genes that encodes the glycogen phosphorylases. The structure of the gene for such a complex protein is important to determine since evidence from other systems suggests that the coding regions of eukaryotic genes often correspond to functional domains or structural units of the protein.[11–13]

Comparisons of the many functional or structural domains of phosphorylase with exons of the gene may aid our understanding of the genetic and structural relationships of phosphorylases and of the family of glycolytic enzymes. Clues to the evolutionary development of the interacting regulatory sites of the molecule may be provided by comparing gene and protein structures for the isozymes of phosphorylase.

We report here the isolation and sequence characterization of the gene that encodes glycogen phosphorylase in human muscle. The relationships between the exon structure of the gene and the three-dimensional structure of the protein are discussed and compared with other glycolytic enzymes.

EXPERIMENTAL PROCEDURES

Materials

Enzymes and other materials were obtained from the following sources: Restriction enzymes and DNA polymerase I (Klenow), Boehringer Mannheim; T4 DNA ligase, DNA polymerase I, *Escherichia coli* DNA ligase, New England Biolabs; pUC 8/9 vectors, M13 mp18/19 sequencing vectors, nucleotides, P.L. Biochemicals; α-^{32}P-dCTP, ICN Radiochemicals; nitrocellulose filters, Schleicher and Schuell. Synthetic oligonucleotides were prepared at the Biomolecular Resource Center, UCSF.

Screening of Human Genomic Libraries

Two different genomic libraries were screened for the gene. Approximately 8×10^5 plaques from a Charon 4A human genomic library[14] were screened according to the procedures of Benton and Davis.[15] A 227-base pair cDNA fragment of muscle glycogen phosphorylase from rabbit encoding residues 727–801[16] was used as a hybridization probe. This DNA fragment, as well as all subsequent DNA used as

Received March 9, 1987; accepted June 16, 1987.

Address reprint requests to Dr. Robert Fletterick, Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143-0448.

L. Anderson's present address is Department of Biochemistry, University of California, Berkeley, CA 94720.

F. Gorin's present address is Department of Neurology, University of California, Davis, CA 95616.

hybridization probes, was ^{32}P-labeled by "nick-translation"[17] with α-^{32}P-dCTP to specific activity of 0.5–1.0 × 10^9 cpm/μg. The library filters were incubated in a solution of 50% formamide, 5 × SSC, 1× Denhardt's, 100 μg/ml sonicated salmon sperm DNA, with a probe concentration of 1 × 10^6 cpm/ml, for 16 hours at 42°C. Filters were washed in 2 × SSC at 50°C for 1 hour. Hybridizing phage DNA was identified by autoradiography.

A human genomic cosmid library was constructed by Choo et al. from human leukocyte DNA.[18] DNA from 1 × 10^5 colonies of the library was screened on five nitrocellulose filters. A ^{32}-P-labeled cDNA fragment of human muscle glycogen phosphorylase encoding amino acids 673–820[16] was used as a hybridization probe. Library filters were incubated in a solution of 50% formamide, 5 × SSC, 5 mM NaH_2PO_4, 200 μ/ml sonicated salmon sperm DNA, 1 × Denhardt's, 10% dextran-SO_4, and 2 × 10^6 cpm/ml of probe DNA, for 16 hours at 42°C. Filters were washed in 2 × SSC at 27°C for 30 minutes and then in 0.1% SDS, 0.1 × SSC, at 55°C for 30 minutes. Hybridizing colony DNA was identified by autoradiography.

Restriction Enzyme Analysis

Restriction enzymes were used according to their recommended assay procedures. Digested DNA was electrophoresed through 1% agarose gels, buffered in 1 × TBE (89 mM Tris-borate, 89 mM boric acid, 2 mM EDTA). Restriction fragments of clones L32 and MC1 were transferred to nitrocellose filters as Southern blots.[19]

Hybridization to these blots with ^{32}-P-labeled DNA was carried out under conditions described above. When a filter was screened more than once, previously hybridized probe DNA was removed from the filter by immersion in boiling water for 60 seconds. Successful removal of probe DNA from filter was checked by autoradiography. Restriction maps were determined by digestion with pairs of restriction enzymes and Southern blot analysis.

DNA Sequencing

DNA was sequenced by the chain termination method of Sanger.[20] Figure 1b illustrates the sequencing strategy undertaken. Although regions of the gene were sequenced in one direction only for reasons of economy, each template was sequenced 3–5 times. Potential sequencing errors that could cause frameshift errors were not a problem since we had as a check the complete and highly homologous cDNA sequence of the rabbit muscle enzyme. Sequence data for the introns are less reliable, as these regions were only sequenced once. We report the sequence for the introns only at the junctions of the exons where it is accurate.

Modeling

The computer graphics program Insight[21] was used to model the positions of the amino acid substitutions and evaluate the structural location of the splice junctions. Solvent accessible surface calculations were done using T.J. Richmond's program, which applies the principles developed by Lee and Richards.[22]

RESULTS

Isolation and Characterization of a Human Genomic Clone Containing the C-Terminal Exons of Muscle Glycogen Phosphorylase

A single clone, designated L32, was isolated from the Charon 4A lambda library and restriction mapped. Southern blot analysis of an Eco R1 digestion of L32 revealed a 10-kilobase (kb) fragment that was shown by additional restriction mapping, blotting, and sequence analysis to contain the C-terminal region of the glycogen phosphorylase gene. Since previous analysis of partial CDNA's of the human and rabbit muscle isozymes had shown 96% amino acid identity,[16] exons of the human muscle gene could be identified by comparison of translated DNA sequence to the known amino acid and cDNA sequences of glycogen phosphorylase from rabbit muscle. Intron/exon junctions were further confirmed by the splice junction consensus sequence.[23] Sequence analysis of the 10-kb Eco R1 fragment of L32 revealed three exons encoding residues 725–770, 770–792, and 793-poly A signal of the 3′ untranslated region. No additional sequence data was obtained from clone L32 since restriction mapping showed that the 10-kb fragment was the most upstream of those generated by Eco Rl digestion so most of the gene was not in the clone.

Isolation and Characterization of a Clone Containing the Complete Human Muscle Glycogen Phosphorylase Gene

The human genomic cosmid library was screened and a single clone, termed MC1, was isolated and restriction mapped. In order to identify coding regions, four fragments (encoding amino acids 1–68, 68–158, 158–560, 573–743) of a rabbit muscle phosphorylase cDNA[2] were used as probes of a Southern blot of digestions of MC1. All four probes show specific hybridizations to restriction fragments of MC1. Portions of four hybridizing Bam Hl restrictions fragments and one hybridizing Hind lll fragment of MC1 were subcloned into M13 and sequenced. Sequencing revealed an additional 17 exons in the muscle phosphorylase gene. Therefore, 20 exons completely encode muscle glycogen phosphorylase. The overlapping restriction maps of clones L32 and MC1 allow the construction of a complete map of the human muscle glycogen phosphorylase gene (Fig. 1a). From these data, the size of the gene [from the initiator ATG codon to the poly (A) addition site] is determined at 12.5 kb and is of average size for a 100,000-dalton (D) protein.

3′and 5′ Flanking Regions

The gene was sequenced 241 bases beyond the TGA translation stop codon to the poly (A) signal, AA-

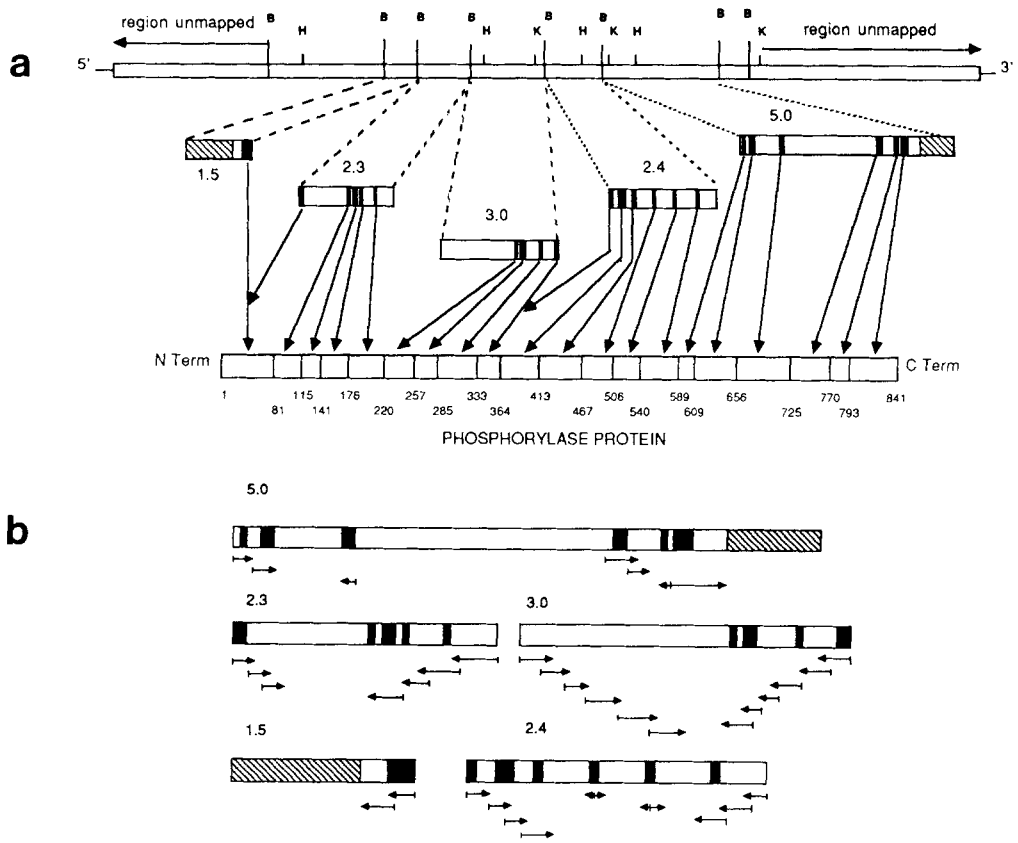

Fig. 1. **a:** Map of the human gene encoding glycogen phosphorylase in muscle. The gene's twenty exons are shown dispersed through five adjacent Bam HI restriction fragments. The sizes of each of these fragments are shown in kilobases. A schematic of the phosphorylase protein is shown below these highlighted fragments. Correlation of exon to encoded region in the protein is shown by arrows. **b:** Sequencing strategy. Each of the five adjacent Bam H1 restriction fragments is shown with arrows below indicating sequenced region and direction. Four of these fragment (excluding the 5.0-kb fragment) were subcloned into M13, mp 18, and mp19, and sequenced 200–300 bases from each end with M13 universal primer. Internal regions were then sequenced by either similar methods after removal of sequenced regions by various restriction digestions and religations in M13 or priming with 17 base oligonucleotides, synthesized specifically to hybridize to the ends of regions previously sequenced. Fifteen such oligonucleotides were used. The six exons in the 5.0-kb BAM H1 restriction fragment were revealed after subcloning two internal PST 1 restriction fragments (1.4 kb and 0.7 kb) into M13 and sequencing directly with the universal primer.

TAAA, in the 3′ untranslated region (see Fig. 2). The sequence identically matches that previously determined in this laboratory from a partial cDNA of human muscle glycogen phosphorylase.[16]

The 5′ region of the gene was sequenced 224 bases upstream from the ATG initiation codon. Alignment of this sequence with the published 75 bases of the 5′ untranslated region from rabbit muscle glycogen phosphorylase[2] reveals 74% identity. The putative promoter sequence, TATAA, is not found in this region and further characterization of the 5′ region of the cDNA would be required to identify the mRNA transcription start site since an intron may be found in this region. (A consensus sequence for a start site is found at − 149.) A possible 3′ acceptor consensus sequence is observed to be at position −156 and may mark the downstream end of an intron in the 5′ region.

Coding DNA

The sequence of the coding region shows that the rabbit and human amino acid sequences are 97% identical, differing at only 22 of 842 positions. These differences were modeled by substituting the altered side chains on the three-dimensional model for rabbit muscle glycogen phosphorylase. All changes are conservative and unremarkable suggesting that the biochemical analyses and conclusions that were derived for the enzyme from rabbit are directly applicable to the human enzyme.

As anticipated from the rabbit cDNA sequence, the coding regions are rich in G + C (57%). This is accounted for by the third position of the codons which is a G or a C for 78% of the codons. This observation may be general for genes expressed in muscle tissue.[24] The introns also show this enrichment in G + C content (data not shown). The G + C composition of the 6.1 kb of introns that are sequenced is 57%. The coding DNA of rabbit and human phosphorylase is 91.8% identical.

Table I shows the sequence data for the intron/exon boundaries and the positions of the 19 introns in the amino acid sequence. The DNA sequences at the splices are as expected from the consensus sequences

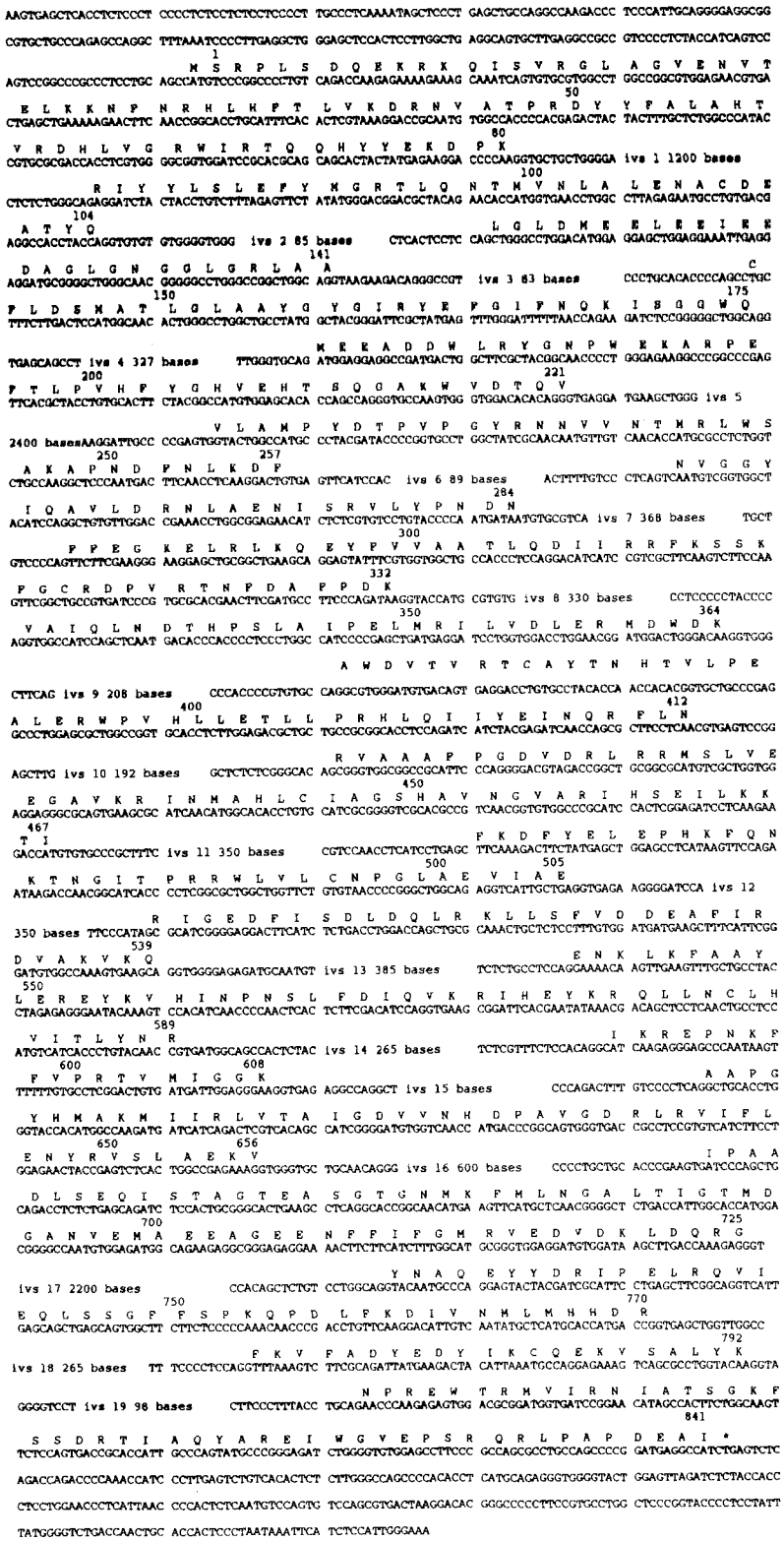

Fig. 2. The sequence of the gene for muscle glycogen phosphorylase and the translated amino acid sequence.

TABLE I. Comparison of Splice Junction Sequences for Phosphorylase*

No.	Amino acids	Position	5′ donor: $^{A}_{C}$CAG/GTG	Length	3′ acceptor: YYXCAG/XX
1	PK/RI	80	AAG/GTG	1,200	GGGCAG/AG
2	YQ/LG	114	CAG/GTG	85	CTCCAG/CT
3	LAACF	141	CAG/GTA	83	CCCCAG/CC
4	WQ/ME	175	CAG/GTG	327	GTGCAG/AT
5	TQVVV	219	CAGG/GTG	2,400	CCCGAG/TG
6	KDFNV	257	ACT/GTG	89	CCTCAG/TC
7	DN/FF	284	AAT/GTG	368	CCCCAG/TT
8	DK/VA	333	AAG/GTA	330	CCCCAG/GT
9	DK/AW	363	AAG/GTG	208	TGCCAG/GC
10	LN/RV	412	AAC/GTG	192	GCACAG/CG
11	KTIKF	467	CAT/GTG	350	CCTGAG/CT
12	AE/RI	505	GAG/GTG	350	CCATAG/CG
13	KQ/EN	539	CAG/GTG	385	CTCCAG/GA
14	YNRIK	589	ACC/GTG	265	CCACAG/GC
15	GK/AA	608	AAG/GTG	125	CCTCAG/GC
16	EKVIP	656	AAG/GTG	600	CCGAAC/TG
17	QRGYN	725	AGG/GT	2,200	TGGCAG/GT
18	HDRFK	770	CCG/GTG	265	CTCCAG/GT
19	YK/NP	792	AAG/GTA	98	CTGCAG/AA

*If the splice falls with a codon, two amino acids on either side are shown. The lengths of the introns are approximate.

that have been characterized at intron boundaries.[23] The amino acids at the 5′ side of the splice are frequently those with polar side chains. A codon for a polar amino acid (Q, N, E, or K) is observed 5′ for all of the 11 splices between codons. This bias has been previously noted.[25]

Figure 2 shows that the coding region of the gene is divided into 20 exons that code for an average of 41 amino acids with a standard deviation of 15. The longest is exon 1, which codes for the N-terminal 80 amino acids. The shortest is exon 16, which codes for 19 amino acids. The 19 introns account for about 10 kb of the gene. The shortest, number 3, is 83 bases; and the longest, intron 5 is 2,400 bases. Three are longer than 1 kb and four are shorter than 100 bases. Ten are approximately 300 bases long.

Secondary and Tertiary Structural Relationships With Intron Positions

The tertiary structure of phosphorylase is well described by defining two domains. The N-terminal 482 residues which includes a subdomain of 56 residues composing the glycogen binding structure is the largest domain defined by x-ray crystallography. The domain has a core of nine mixed parallel and antiparallel strands of β-sheet surrounded by an irregular array of α-helices. Five strands of this β-structure form many of the intersubunit contacts. The C-terminal domain, residues 483–842, is the second largest domain so far described and is also a β-sheet core (of parallel strands) surrounded by αhelices. Within this domain is a typical nucleotide binding subdomain[26] commonly found in dehydrogenases[27] and other glycolytic enzymes.[28] This subdomain in phosphorylase contributes the majority of the side chains found in the active site.

Domain and Subdomain Boundaries

The N-terminal domain ends at amino acid 482. This point can be precisely defined in the folded protein since the preceding amino acids fold tightly with the N-terminal residues and the proceeding amino acids form very tight contacts with the residues of the C-terminal domain. The closet splice point to amino acid 482 is intron number 11 at amino acid 467. The N- and C-terminal domain boundary is therefore not coincident with an intron position in the gene.

The substructure of the N-terminal domain that binds the enzymes to glycogen extends from residues 384 through 440 in the N-terminus. These 56 amino acids (Fig. 3a) form a compact subdomain unique to phosphorylase which is set off from the surface of the rest of the N-terminal domain.[29] Neither the beginning nor the end of this subdomain is marked by intron boundaries. Intron 10 maps to amino acid 412 which is nearly at the midpoint of this structure. The amino acid sequence of potato phosphorylase[30] shows a 74 amino acid insertion at intron position 10. A second substructure also unique to phosphorylase is the all-β structure that is formed by a segment of 100 amino acids from 153 to 250 (Fig. 3b). These form five long meandering β-strands that form twisted and self-associating two-strand ribbons. This structure forms a large part of the subunit interface (dimer contacts ll) in Table III and is presumably involved in allosteric switching between subunits.[31] This structural unit terminates eight amino acids before intron position 6. It is not initiated at an intron position.

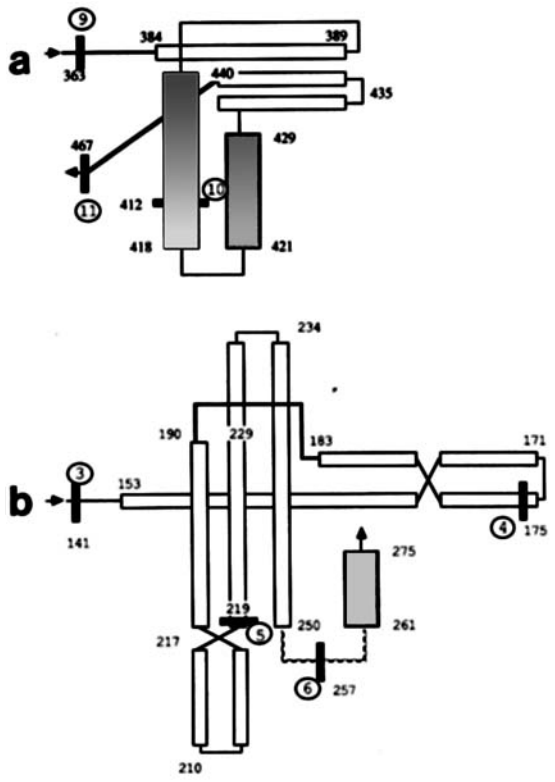

Fig. 3. **a:** The two α-helices (shaded) and three β strands (open) that form the 56 amino acid glycogen binding subdomain in the N-terminal domain of phosphorylase. Splice 9 is well before the first amino acid of this structure and splice 11 is well beyond the end. Splice 10 fragments the structure in asymmetrical halves. Intron/exon boundaries are marked by bards. **b:** The β-ribbon structure that forms the subunit contacts at the control face. The tertiary structure corresponding to exon 3 is similar in structure to that of exon 4, although the latter is 10 amino acids longer. Intron 5 is in a different relative position from intron 4 in the tertiary structure. Intron 6 occurs in a position corresponding to disordered chain in the phophorylase *a*-three-dimensional structure.

Within the C-terminal half of the protein, a domain frequently observed in glycolytic enzymes is found. This is the nucleotide binding domain (NBD) which shares topological and tertiary structural homology with those found in the dehydrogenases and other glycolytic enzymes.[32] This structure is a duplication of three parallel β-strands covered on one face by two interconnecting α-helices—a so-called mononucleotide binding fold (MNBF). Two of these units are joined by a crossover which can be a segment of polypeptide chain with irregular structure or can contain an α-helix. The boundaries of this domain, 562–712, are not coincident with intron positions.

Secondary Structural Elements

Table II shows the relationships of the intron/exon boundaries to the secondary structural elements. Only four of the 19 are further than three residues before the ends of α-helices and only one (No. 5) is internal to a β-strand. Thus 14 of 19 are near ends of secondary structural elements or within turns or loops[33] linking secondary structural elements. A remarkably varied assortment of α-helices and β-strands, alternating singly or in pairs correspond to exon-encoded polypeptide chains.

The secondary structural elements that form the domains are often adjacent in three-dimensional space and pack together. For the entries in Table II, the exon is defined to be compact if all secondary structural elements in a chain segment are in intimate contact and exclusively form a compact and space filling structure. Compactness was determined by visual inspection of the atomic model of glycogen phosphorylase. Table II shows that only seven of the twenty exons correspond to compact structures in the folded protein.

Surface Positions of Splice Junctions

The solvent accessibility of rabbit muscle phosphorylase was calculated to define buried or solvent-inaccessible amino acid side chains (25 Å^2) or less solvent accessible surface per side chain). By this criterion, 49% of the side chains in phosphorylase are buried, an unusually large fraction.[34] (With 20 Å^2 angstroms as the threshold value for accessible surface, 43% of the side chains are solvent inaccessible.) Except for Ser and Thr, which are less solvent accessible than other proteins.[34] the usual hydrophobic residues compose the majority of this group of buried side chains. This calculation and visual inspection shows that only two of the 19 residues at the splice points are not at the surface of the protein. The splice point after exon three maps 15A from the surface of the protein. The intron after exon 15 maps 5A from the surface.

A survey of families of bacterial and eukaryotic proteins with known tertiary structures and sequenced genes suggested that splice points often correspond to positions of length-variation of the polypeptide chain.[35] Aside from the mammalian phosphorylase sequences which show no length-variations except at the C-terminus, sequences are determined for phosphorylases from potato[30] *Escherichia coli*[36] and yeast.[37] These four sequences display length variations, with yeast phosphorylase having 13 positions in the polypeptide chain with insertions or deletions in comparison with mammalian phosphorylases. Nine of the 19 intron positions fall within three amino acids of this selected set of length-variations. This fraction is expected to occur by chance, assuming a normal distribution, with a probability of 0.25, and therefore the correlation of splice points with length variations may not be significant for the family of phosphorylases.

Functional Units of Structure

Numerous functions can be ascribed to chain segments of phosphorylase by examination of the three-dimensional structures of the enzyme complexed with ligands. As the mechanisms of allosteric switching and the kinase and phosphatase binding sites on the

TABLE II. Intron/Exon Structure of the Human Glycogen Phosphorylase Gene and Correlations With Tertiary Structure of the Protein*

Exon	No. amino acids	Secondary structure	Intron position	Surface	Length variable	Compact
1	80	ααα	Start β	Y	N	N
2	34	βαα	Turn	Y	Y	N
3	26	αβα/2	Mid α	N	N	N
4	34	α/2β	End β	Y	N	N
5	44	βββ/2	Mid β	Y	Y	N
6	38	ββ	Turn	Y	N	Y
7	27	αβ	Turn	Y	N	N
8	49	βαα	Turn	Y	N	N
9	30	βα	Turn	Y	N	Y
10	49	αββαα/2	Mid α	Y	Y	N
11	55	α/2αββαβα	Turn	Y	Y	Y
12	38	αβαα	Turn	Y	Y	N
13	34	αα/2	Mid α	Y	N	N
14	50	α/2ββα	Turn	Y	Y	N
15	19	β	End β	N	N	Y
16	48	αβα/2	Mid α	Y	N	Y
17	69	α/2βαβαβα	Turn	Y	Y	Y
18	50	ααα	Turn	Y	Y	Y
19	22	α	Turn	Y	Y	—
20	50	αβ		—	—	N

*The notation of α/2 or β/2 denotes a splice that maps to near the midpoint of a secondary structural element.

surface of phosphorylase are poorly defined, we can be certain of only some functional assignments such as subunit contacts or ligand binding residues. The activation locus where AMP, ATP, Glc-6-P, and Ser 14 Pi bind is complex. Some side chains such as those of Arg 309 and Arg 310 which bind phosphate are multifunctional. Table III lists the exons of phosphorylase and the functional properties defined by examination of the tertiary structures.

This compilation shows that most functions are distributed among multiple exons and that many functions can be ascribed to some exons. Thus exon 1 codes for a chain segment that participates in seven functions. The residues which form the dimer contacts and link the active sites and allosteric effector sites are divided into two sets, dimer 1 and dimer II in Table III. These correspond to the noncovalent links of the helical structures at the control face[38] and the β-strands at the catalytic face[38] of the dimer. The dimer contacts are distributed among seven exons. The amino acids that form the active site are distributed among nine exons. The AMP binding site is split among three exons as is the G_n binding site. There is no case where a unique function can be wholly identified with polypeptide chain encoded by a single exon.

DISCUSSION

As is typical of eukaryotic genes, the gene for muscle glycogen phosphorylase shows segmentation between coding and noncoding DNA. The coding regions of the gene are of usual length found for eukaryotic genes.[39] Much has been written about the fragmentation of eukaryotic genes and correlations of the exons with the structure and function of the protein product.[40] The relationships between these patterns has been the subject of rich speculation but defensible correlations are few.

It is commonly believed that domains (about 150 amino acids) or compact folding units of about 20–40 amino acids[41] are encoded by exons in the gene. This relationship is not always apparent since introns which fragment the gene may be inserted or deleted in the course of evolution so that the gene segmentation pattern varies with species or isozyme. By examining a single present-day gene, it is therefore problematic to discern whether a particular set of introns was invasive (itinerant introns) or associative and used to construct the gene from component exons. A description of the evolution of the chicken glyceraldehyde-3-phosphate dehydrogenase (GPD) gene[40] discusses these issues. In this case and in other proteins of the α/β class,[42] it is plausible that some introns mark prototype units of structure such as an αβ-unit of two secondary structural elements. Four of the repeating αβ-units of secondary structure that form the triose phosphate isomerase structure, $(\alpha\beta)_8$, are exon-encoded in the chicken or maize genes.[43] Three of these have an intron position interrupting the α helices, making the correlation difficult to see. The same $(\alpha\beta)_8$ structure occurs in pyrivate kinase, where exons 2 and 3 are strictly defined αβ-units, and exons 7 and 8 are again cases where the helices are interrupted by intron positions.[44] The evidence that these pseudosymmetric structures were

TABLE III. Functional Assignments of Exons in Phosphorylase

Exon	Amino acids	Binding site for ligands	Allosteric control
1	1–80	Adenine Ribose Glucose of Glc-6-P Ser-Pi	Dimer I contacts Kinase binding Phosphatase binding
2	81–114	—	Dimer I contacts
3	115–142	Divalent metal, active site Pi	Dimer I contacts
4	143–176	—	Dimer II contacts
5	177–219	—	Dimer II contacts
6	220–258	Pi of AMP	Dimer II contacts (active site links)
7	258–284	Glucose	Dimmer II active site gate
8	285–335	Pi of AMP	—
9	336–363	Glycogen	—
10	363–413	Glycogen	Active site activation
11	413–467	Glycogen	Active site activation
12	468–505	Glucose pyridoxal phosphate	—
13	506–540	—	—
14	541–589	Glucose	—
15	590–608	Purine inhibitor	—
16	609–656	Purine inhibitor pyridoxal phosphate	—
17	657–725	Glucose pyridoxal phosphate	—
18	726–769	—	—
19	770–792	—	—
20	793–842	—	Dimer I contacts

formed from intron-mediated fusion of $\alpha\beta$-units is tantalizing but inconclusive from the small set of available data.

Nucleotide Binding Domain

Arguments have been made that NBD's are related based on tertiary structure[27] and gene structure.[45] Evidence for modular construction of these genes has also been presented.[47] Figure 4 shows the topology diagrams for the NBD's that are found in proteins with determined three-dimensional structure and gene structures. The set includes glycogen phosphorylase, chicken glyceraldehyde-3-phosphate dehydrogenase (GPD), two forms of alcohol dehydrogenase from maize (MADH) and human (HADH), lactate dehydrogenase (LDH) from human, and phosphoglycerate kinase (PGK) from human. An MNBF (not shown) is found in cat pyruvate kinase.[48] A single intron interrupts this structure and has been positioned by modeling the cat pyruvate kinase structure to the chicken sequence.[44]

This compilation shows four examples where the boundaries of the NBD within the protein are near intron positons. These positions are not precisely at the domain boundaries. In the case of alcohol dehydrogenase an intron position is five amino acids before the start and a second is three amino acids after the end of the domain. Whereas in glycogen phosphorylase, an intron position is 13 amino acids beyond the end (712) and in GPD an intron breaks the last β-strand. This figure also shows that there is little evidence of correlation of modular construction of these domains with exons of the genes. The domains are constructed from $\beta\alpha$-units that repeat three times (with the last α-helix crossover being variable structure). Simple units of $\beta\alpha$ or $\alpha\beta$ or more complex motifs such as $\beta\alpha\beta\alpha$ that are exon-encoded are not common in these domains either in phosphorylase or in the other proteins.[49] This may be expected since the subdomains and α-helices which are duplicated in these structures are of different size. For example, MNBF 1 in the five enzymes represented varies in length from 49 to 84 amino acids, and MNBF 2 varies from 55 to 72 amino acids. The crossover chain varies in length from five amino acids in LDH to 49 amino acids in PGK. Similarly the secondary structural ele-

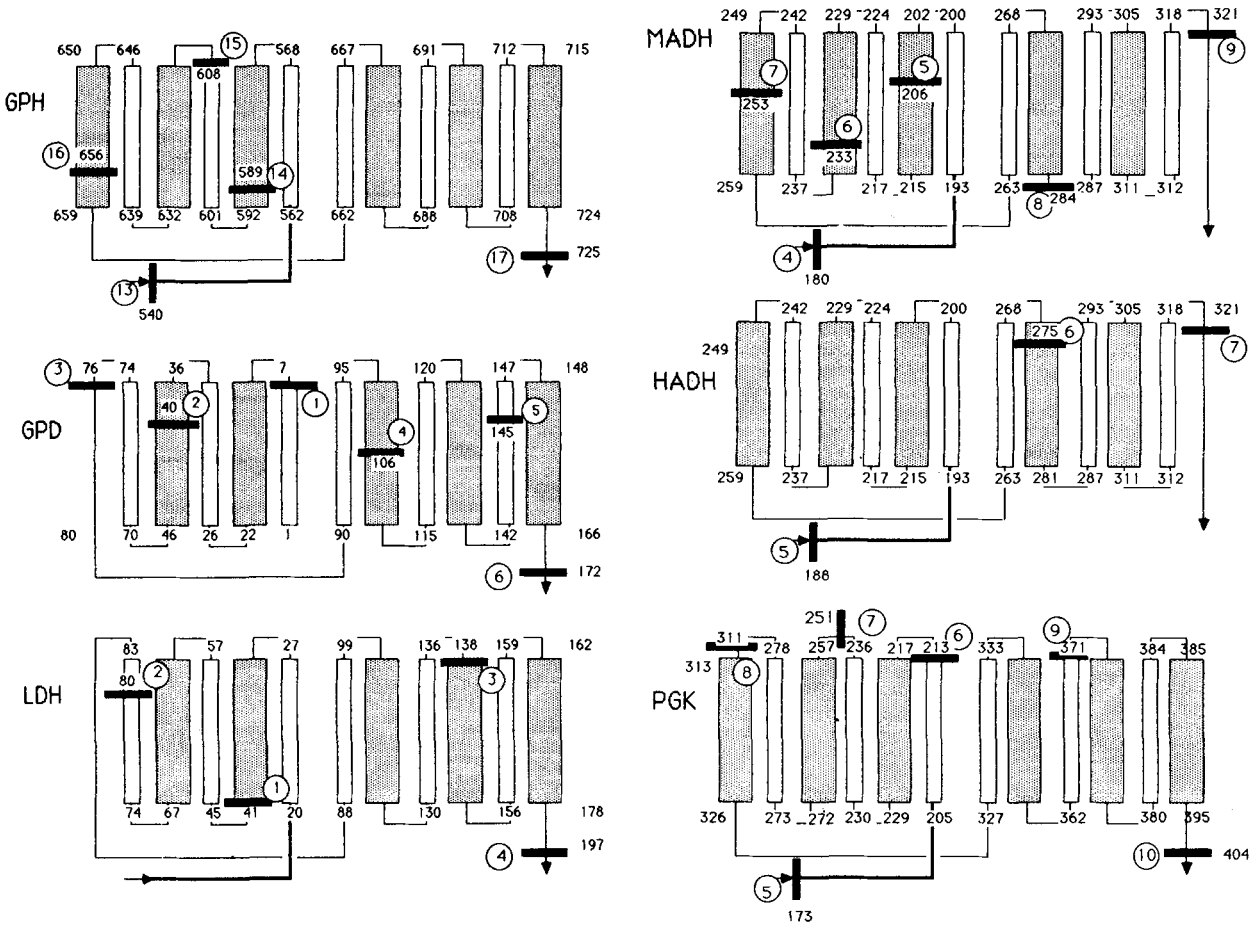

Fig. 4. Schematic diagrams for the twofold symmetrical nucleotide binding domains found in human lactate dehydrogenase (LDH) human phosphorylase (GPH) maize alcohol dehydrogenase (MADH), human alcohol dehydrogenase (HADH), and chicken glyceraldehyde 3-phosphate dehydrogenase (GPD), and human phosphoglycerate kinase (PGK). The α-helices are shaded and the β-strands are open bars. Connecting chain is shown as straight lines. The positions of the splice junctions for each gene are mapped onto the structures as numbered horizontal bards. The amino acid position at the splice is given alongside the bar. The crossover is at the extreme left edge of each schematic. The helix that is found in three of the domains seen at right-hand side is not structurally associated with the nucleotide binding domain but is included in order to define the intron position at the end of these structures.

ments are of different length in each of the proteins. For example, the second α-helix varies from 7 amino acids in MADH to 22 amino acids in glycogen phosphorylase. It may be that the large variation in length observed for these secondary structural elements and the crossover is related to antiquity of the structures and the long time since duplication and divergence.

The positions of the intron within these domains can be described as their positions within a MNBF unit defined by three β-strands, two α-helices, and a crossover as shown in Figure 5. There are seven cases where helix $\alpha 1$ is interrupted by intron positions, four where helix $\alpha 2$ is interrupted by intron positions, and five cases where the crossover chain is interrupted by an intron position. The strand $\beta 3$ is interrupted twice by introns in this set. The positions of the introns mapped onto the secondary structural elements shows the majority correspond to the C-terminal edge of the β-sheet of the N-terminals of the associated α-helices.[50] In these four proteins only 5 of the 21 intron

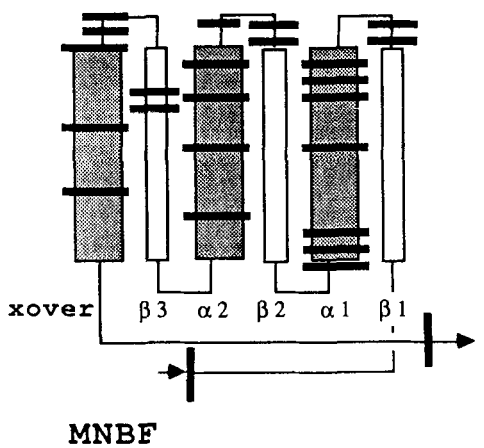

Fig. 5. A compilation of the intron positions for the domains shown in Figure 4. These positions are mapped on an MBNF half-domain. The "Xover" structure in MBNF 1 corresponds to the protein chain after $\beta 3$ ends in MBNF2. The intron positions appear to be distributed everywhere in this structure with the exceptions of elements $\beta 1$, $\beta 2$, and the turn between $\alpha 2$ and $\beta 3$.

positions are at the N-terminal side of the β-structure and strand, therefore away from the active site. Thus the only positions where introns are not found in this set of structures is the turn between α2 and β3 and within β-strands 1 and 2.

CONCLUSIONS

The data taken together for glycogen phosphorylase suggests that intron positions do not mark structurally compact units or functional subdomains. The surface positions of the intron boundaries[51] likely correlate with the biased distribution of polar amino acids observed at exon-intron boundaries. This in turn is a consequence of the consensus nucleotide sequence found for the codon, or fractional codon upstream of the 5′ acceptor. Whether the consensus sequence at the termination of exons is consequential to the mechanism of mRNA splicing or a result of selection is not known.

The widely variant positions of the introns in the nucleotide binding domains for six members of the glycolytic enzymes does not support the argument that intron positions can be used to demonstrate a common genetic ancestor for these domains.

ACKNOWLEDGMENTS

We wish to thank Prudence Bothen for preparation of this manuscript. This work was supported in part by the Muscular Dystrophy Association (R. Lebo) and by NIH grants AM26081 and DK32822 (R. Fletterick). R. Lebo is a consultant to the Howard Hughes Institute. We thank Jean Lockyer for reading the manuscript.

REFERENCES

1. Titani, K., Koide, A., Hermann, J., Ericsson, I., Kumar, S., Wade, R., Walsh, K., Neurath, H., Fischer, E. Complete amino-acid sequence of rabbit muscle glycogen phosphorylase. Proc. Natl. Acad. Sci. U.S.A. 74:4762–4766, 1977.
2. Nakano, K., Hwang, P.K., Fletterick, R.J. Complete cDNA sequence for rabbit muscle glycogen phosphorylase. FEBS Lett. 204:283–287, 1986.
3. Sprang, S., Fletterick, R.J. The structure of glycogen phosphorylase *a* at 2.5 A resolution. J. Mol. Biol. 131:523–551, 1979.
4. Jenkins, J.A., Johnson, L.N., Wilson, K.S. Assignment of the amino acid sequence to the crystal structure of glycogen phosphorylase *b*. Biochemistry 17:5694–5695, 1978.
5. Sprang, S.R., Goldsmith, E.J., Fletterick, R.J., Withers, S.G., Madsen, N.B. Catalytic site of glycogen phosphorylase: structure of the T state and specificity for alpha-D-glucose. Biochemistry 21:5364–5371, 1982.
6. Kasvinsky, P.J., Shechosky, S., Fletterick, R.J., Synergistic regulation of phosphorylase *a* by glucose and caffeine. J. Biol. Chem. 253:9102–9106, 1978.
7. Stura, E.A., Zanotti, G., Babu, Y.S., Sansom, M.S., Stuart, D.I., Wilson, K.S., Johnson, L.N., Van-de-Werve, G. Comparison of AMP and NADH binding to glycogen phosphorylases b. J. Mol. Biol. 170:529–565, 1983.
8. Dombradi, V. Structural aspects of the catalytic and regulatory function of glycogen phosphorylase EC 2.4.1.1. Int. J. Biochem. 13:125–140, 1981.
9. Fletterick, R.J., Madsen, N.B. The structures and related functions of phosphorylase *a*. Annu. Rev. Biochem. 49:31–61, 1980.
10. Fletterick, R.J., Sprang, S.R. Glycogen phosphorylase structures and function. Acc. Chem. Res. 15:361–369, 1982.
11. Go, M. Correlation of DNA exonic regions with protein structural units in hemoglobin. Nature 291:90–92, 1981.
12. Gilbert, W. Why genes in pieces? Nature 271–501, 1978.
13. Blake, C.C.F. Exons and the evolution of proteins. Trends Biochem. Sci. 8:11–13, 1983.
14. Lawn, M., Fritsch, E.R., Parker, G., Manniatis, T. The isolation and characterization of linked α-and β-globin genes from a cloned libary of human DNA. Cell 15:1157–1174, 1978.
15. Benton, W.D., Davis, R.W. Screening λgt recombinant clones by hybridization to single plaques in situ. Science 196:180–182, 1977.
16. Hwang, P.K., See, Y.P., Vincentini, A.M., Powers, M.A., Fletterick, R.J., Crerar, M.M. Comparative sequence analysis of rat, rabbit, and human muscle glycogen phosphorylase cDNAS. Eur. J. Biochem. 152:267–274, 1985.
17. Rigby, P.W.J., Dieckmann, M., Rhodes, C., Berg, P. Labelling deoxyribonucleic acid to high specific activity in vitro by Nick Translation with DNA polymerase I. J. Mol. Biol. 113:237–251, 1977.
18. Choo, K.H., Filby, G., Greco, S., Lau, Y.F., Kan, Y.W. Cosmid vectors for high efficiency DNA-mediated transformation and gene application in mammalian ceus: Studies with human growth hormones gene. Gene 46:277–286, 1986.
19. Southern, E.M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503–517, 1975.
20. Sanger, F., Nicklen, S., Coulson, A.R. DNA sequencing with chain-terminating inhibitors. Pro. Natl. Acad. Sci. U.S.A. 74:5463–5467, 1977.
21. Dayringer, H., Tramantono, A., Sprang, S., Fletterick, R.J. Interactive program for visualization and modelling of protein, nucleic acids and small molecules. J. Mol. Graphics 4:82–87, 1986.
22. Lee, B., Richards, F.M. Solvent accessibility. J. Mol. Biol. 55:379–400, 1971.
23. Breathnach, R., Chambon, P. Organization and expression of eukaryotic split genes coding for proteins. Annu. Rev. Biochem. 50:349–383, 1981.
24. Newgard, C.B., Nakano, K., Hwang, P.K., Fletterick, R.J. Sequence analysis of the cDNA encoding human liver glycogen phosphorylase reveals tissue-specific condon usage. Proc. Natl. Acad. Sci. U.S.A. 83:8132–8136, 1986.
25. Craik, C.S., Laub, O., Bell, G.I., Sprang, S., Fletterick, R., Rutter, W.J. The relationship of gene structure to protein structure. UCLA Symp. Mol. Cell. Biol. 26:35–54, 1982.
26. Rossmann, M.G., Moras, D., Olsen, K.W. Chemical and biological evolution of a nucleotide-binding protein. Nature 250:194–199, 1977.
27. Branden, C.J., Eklund, H., Cambillau, C., Pryor, A.J. Correlation of Exons with structural domains in alcohol dehydrogenase. EMBO J. 3:1307–1310.
28. Banks, R.D., Blake, C.C.F., Evans, P.R., Haser, R., Rice, D.W. Hardy, G.W. Merrett, M. and Phillips, D.W. Sequence, structure and activity of phosphoglycerate kinase. Nature 279:773–777, 1979.
29. Goldsmith, E., Sprang, S., Fletterick, R. Structure of maltoheptaose by difference Fourier methods and a model for glycogen. J. Mol. Biol. 1982 156:411–417, 1982.
30. Nakano, K., Fukui, T. The complete amino acid sequence of potatoe alpha-glucan phosphorylase. J. Biol. Chem. 261:8230–8236, 1986.
31. Sprang, S., Fletterick, R.J. Subunit interactions and the allosteric response in phosphorylase. Biophys. J. 32:175–192, 1980.
32. Fletterick, R.J., Sygusch, J., Semple, H., Madsen, N.B. Structure of glycogen phosphorylase a at A resolution and its ligand binding sites at 6 A. J. Biol. Chem. 251:6142–6146, 1976.
33. Leszczynski, J.F., Rose, G.D. Loops in globular proteins: A novel category of secondary structure. Science 234:849–855, 1986.
34. Sprang, S., Yang, D., Fletterick, R.J. Solvent accessibility properties of complex proteins. Nature 280:1125–1129, 1983.
35. Craik, C.S., Rutter, W.J., Fletterick, R. Splice junctions: Association with variation in protein structure. Science 220:1125–1129, 1983.

36. Palm. D., Goerl, R., Burger, K.J. Evolution of catalytic and regulatory sites in phosphorylases. Nature 316:500–502, 1985.
37. Hwang, P.K., Fletterick. R.J. Convergent and divergent evolution of regulatory sites in eukaryotic phosphorylases. Nature 324:80–84.
38. Fletterick. R.J., Sprang, S., Madsen, N.B. Analysis of the surface topography of glycogen phosphorylase *a*: Implicatons for metabolic interconversion and regulatory mechanisms. Can. J. Biochem. Cell Biol. 57:789–97, 1979.
39. Naora, H., Deacon, N.F. Relationship between the total size of exons and introns in protein coding genes of higher eukaryotes. Proc. Natl. Acad. Sci. U.S.A. 79:6196–6200, 1982.
40. Stone, E.M., Rothblum, K.N., Schwartz, R.J. Intron-dependent evolution of chicken glyceraldehyde phosphate dehydrogenase gene. Nature 313:498–500, 1985.
41. Zehfus, M.H., Rose, G.D. Compact units in proteins. Biochemistry 25:5759–5765, 1986.
42. Levitt, M., Chothia, C. Definition of protein classes. Nature 261:552–558, 1976.
43. Marchionni, M., Gilbert, W. The triosephosphate isomerase gene from maize: introns antedate the plant-animal divergence. Cell 46:133–141, 1986.
44. Lonberg, N., Gilbert, W. Intron-exon structure of the chicken pyruvate kinase gene. Cell 40:81–90, 1985.
45. Duester, G., Jornvall, H., Hatfield, G.W. Intron-dependent evolution of the nucleotide-binding domains within alcohol dehydrogenase and related enzymes. Nucleic Acids Res. 14:1931–1941, 1986.
46. Michelson, A.M., Blake, C.C.F., Evans, S.T., Orkin, S.H. Structure of the human phosphoglycerate kinase gene and the intron-mediated evolution and dispersal of the nucleotide-binding domain. Proc. Natl. Acad. Sci. U.S.A., 82:6965–6969, 1985.
47. Li, S.S.L., Tiano, H.F., Fukasawa, K.M., Yagi, K. Shimizu, M. Sharief, F.S., Nakashima, Y. Pan, Y.C.E. Protein structure and gene organization of mouse lactate dehydrogenase-a isozyme EC 1.1.1.27. Eur. J. Biochem. 149:221–226, 1985.
48. Stuart, D.I., Levine, M., Muirhead, H., Stammers, D.K. Crystal structure of cat muscle pyruvate kinase at a resolution of 2.6 A. J. Mol. Biol. 134:109–142, 1979.
49. Place, A.R., Anderson, S.M., Sofer, W. Introns and domain-coding regions in the dehydrogenase genes. In: "Multidomain Proteins: Structure and Evolution." Hardic, D.C., Coggins, J.R., eds. Amsterdam: Elsevier Science Publishers. 1986:175–193.
50. Branden, C.I. Anatomy of α/β Proteins. Current Communications in Molecular Biology. In: Fletterick, R., Zoller, M. eds. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory. 1986:45–51.
51. Craik, C.S., Sprang, S., Fletterick, R., Rutter, W.J. Intron-exon splice junctions map at protein surfaces. Nature 299:180–182, 1982.