



Characterization of genomic structure and polymorphisms in the human carbamyl phosphate synthetase I gene[☆]

M.L. Summar^{a,*}, L.D. Hall^a, A.M. Eeds^b, H.B. Hutcheson^a, A.N. Kuo^a, A.S. Willis^b, V. Rubio^c, M.K. Arvin^a, J.P. Schofield^d, E.P. Dawson^e

^aDivision of Medical Genetics, Vanderbilt University, Medical Center North DD2205, Nashville, TN 37232-2578, USA

^bDepartment of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 37232, USA

^cInstituto De Biomedicina, Valencia, Spain

^dClinical Pharmacology, SmithKline Beecham, Harlow, Essex CM19 5AW, UK

^eBioventures, Inc., Murfreesboro, TN, USA

Received 8 November 2002; received in revised form 3 March 2003; accepted 12 March 2003

Received by M. D'Urso

Abstract

Human carbamyl phosphate synthetase I (CPSI) is an essential hepatic enzyme that initiates the urea cycle. Deficiency of this enzyme usually results in lethal hyperammonemia. CPSI is encoded by the *CPSI* gene located on chromosome 2q35. In the present study, we report the coding sequence and define the intron–exon structure of the human *CPSI* gene. These data are compared to the previously defined rat *CPSI* gene structure. This work was generated from direct sequence determination of human genomic DNA (35 introns) and comparison to public domain sequence of anonymous BACs (2 introns). The human *CPSI* gene spans > 120 kb of genomic DNA. *CPSI* has 38 exons and 37 introns, and all adhere to the consensus splicing sequences. Comparison of the human and rat *CPSI* genes reveals that the nucleotide sequences, amino acid sequences, and intron–exon organizations are highly similar. We report the primers and conditions for screening the human CPSI exonic and bordering intronic sequences. We also screened 100 individuals for polymorphisms in the human CPSI gene and identified 14 polymorphisms in the *CPSI* message. The knowledge of the *CPSI* gene structure and the 14 polymorphisms presented in this study will greatly facilitate future molecular studies involving the *CPSI* gene and the enzyme it encodes.

© 2003 Published by Elsevier Science B.V.

Keywords: Carbamyl phosphate synthetase I deficiency; Urea cycle; Genomic organization; Polymorphism

[☆] The full working draft sequence is BAC clone NH0349G04, accession number AC008172.1. The CPSI coding sequence derived from our laboratory is at accession number AF154830. A previously published coding sequence is at accession number NM_001875. Supplementary data includes a list of primer sequences for primers listed in Table 2. Primer sequences are: U1119 (TACTGCTCAGAATCATGGC), U2712 (AGAGTTGTCTGAACCAAGCA), U4295 (CGGAAGCCACATCA GACTGG), U4926 (AATGGTGATCAAGGTAGGAA), L5025 (TGTCTGAGTTGCAGATAG), L5277 (TGGAGAGTGTGACTCC ATCT), U5195 (TGTGACAGAGGCATTTAGAG), L5547 (GGAA TGAACCTTACTTCCAA).

Abbreviations: CPSI, carbamyl phosphate synthetase I; NAG, *n*-acetyl glutamate; CP, carbamyl phosphate; ATP, adenosine triphosphate; CPSID, CPSI deficiency; PCR, polymerase chain reaction; YAC, yeast artificial chromosome, BAC, bacterial artificial chromosome; UTR, untranslated region.

* Corresponding author. Tel.: +1-615-322-7601; fax: +1-615-343-9951.

E-mail address: marshall.summar@vanderbilt.edu (M.L. Summar).

1. Introduction

Human carbamyl phosphate synthetase I (CPSI) (EC 6.3.4.16) is the rate-limiting enzyme that catalyses the first committed step of the hepatic urea cycle. The urea cycle is responsible for the removal of waste nitrogen produced by endogenous and exogenous protein metabolism. CPSI is highly tissue specific, with function and production limited to the liver and a lesser amount in the intestine. The 165 kDa CPSI proenzyme is produced in the cytoplasm and transported into the mitochondria where it is cleaved into its mature 160 kDa form. Mature CPSI enzyme and its cofactor *n*-acetyl glutamate (NAG) catalyse the conversion of ammonia and bicarbonate to carbamyl phosphate (CP) with the expenditure of two ATPs (Rubio and Grisolia, 1981; Rubio et al., 1981).

The *CPSI* gene is a highly conserved ancient gene with representatives from each of the well-defined domains of Bacteria, Archaea, and Eukarya (Schofield, 1993). An analysis of the mammalian *CPSI* coding sequence indicates that the *CPSI* gene encodes a protein that has arisen from a fusion of loci from two separate ancestral subunits that are found in yeast and *Escherichia coli* (Nyunoya et al., 1985a; van den Hoff et al., 1995). In bacteria, the smaller enzyme subunit is responsible for the catalytic transfer of the amide nitrogen from glutamine to the catalytic center for CP synthesis situated on the larger enzymatic subunit (Nyunoya et al., 1985a). In contrast, the human and rat *CPSI* enzymes are unable to process glutamine because they lack the cysteine residue that is essential for aminotransferase activity in the yeast and the bacterial enzymes (Nyunoya et al., 1985b; Rubio, 1993). In addition to housing the active sites for CP synthesis and a set of duplicated ATP-binding domains, the carboxy end of the protein on the large subunit is the site of the binding domain for NAG. The binding of NAG to *CPSI* is theorized to cause a conformational change in the enzyme that exposes the ATP-binding domains (Rubio, 1993). Although studies attempting to localize the *CPSI* ATP-binding sites have obtained varied results (Nyunoya et al., 1985a; Powers-Lee and Corina, 1986, 1987), the *CPSI* enzyme does have a sequence bearing a high degree of homology to known ATP/bicarbonate binding domains and other ATP-binding sites. Despite the performance of *CPSI* enzyme functional studies, the exact function of the NH₄-terminal portion of the enzyme is not known.

Deficiencies in *CPSI* enzyme function reflect the severity of the underlying molecular defect (Summar et al., 1995). *CPSI* deficiency (*CPSID*) is inherited in an autosomal recessive mode and presents as either a devastating metabolic disease in neonates or a more insidious late-onset condition. In the present study we have determined the intron–exon organization of the human *CPSI* gene. We also have identified 14 polymorphisms in the gene that encodes the *CPSI* enzyme, one having an implication in environmental toxicity. In addition to presenting an evolutionary understanding of this gene, the information presented will facilitate studies of *CPSI* mutations and their role in the disruption of normal urea cycle function.

2. Materials and methods

2.1. cDNA sequence

RNA was extracted from a normal human liver and reverse transcribed using oligo(dT) and primers derived from rat *CPSI* sequence. Subsequent PCR reactions were done using primers derived from rat sequence and fragments were cloned and sequenced. 5' and 3' sequences were obtained from screening a human hepatic cDNA library and obtaining partial clones spanning these regions. Using this

sequence as a base we have determined the cDNA sequence from over ten individuals in order to arrive at a consensus sequence. Sequence was compared to that which had been previously published. (Haraguchi et al., 1991).

2.2. Genomic sequence

Using the rat intronic sequence locations worked out by researchers at the University of Amsterdam (van den Hoff et al., 1995), we designed a series of exon-based primers flanking the introns. We screened and isolated a YAC containing human *CPSI* (Research Genetics, Huntsville, AL) that served as an enriched template. Amplified fragments were gel purified and subsequently sequenced. The presence of an intron was established by comparison of the genomic fragments with the expected sizes of complementary cDNA fragments. With the exception of introns 1 and 21, all intron locations were determined in this fashion. Sequences for introns 1 and 21 were obtained from an anonymous BAC clone (NH0349G04: accession number AC008172.1) sequenced in GenBank.

2.3. Polymorphisms

As part of our mutation detection program we identified a number of changes shared by both *CPSID* patients and controls. Allele frequencies were established using a group of 100 unrelated samples.

3. Results

Utilizing various experimental protocols, we report for the first time the complete structure of the human *CPSI* gene and identify fourteen polymorphisms that are located in its coding and non-coding regions.

3.1. Organization of the human *CPSI* gene

Our data indicate that *CPSI* is a relatively large gene spanning 122,497 bases. Compilation of the sequencing data, as shown in Table 1, demonstrates that the human *CPSI* gene contains 38 exons (average size of 122 bp) and 37 introns (average size > 3.1 kb). Exon sizes range from 56 to 260 bp while intron sizes range from 415 to 21,160 bp. The coding message is 5746 nucleotides in length and consists of a 123-nucleotide 5' untranslated region, an open reading frame of 4500 nucleotides and a 1123-nucleotide 3' untranslated region. All 38 of the 5' donor sequences adhere to the 'GT' consensus at the start of the introns. All of the intron 3' acceptor sequences at the intron–exon junctions of the *CPSI* gene contain the invariant 'AG' (Shapiro and Senapathy, 1987). Also, 37 of the 38 exons end with a 'G'.

Our coding sequence (AF154830) was derived from > 10 templates and was confirmed against the full working draft sequence (BAC clone NH0349G04: accession number

Table 1
Intron organization of human CPS1

Intron	5' exon/intron	3' intron/exon	cDNA position	Intron size (bp)	Preceding exon size (bp)
1	GTC AAGgtaatacccatattg	caactgtttcttcagGCACAG	250	16438	125
2	GGGAGGgtgagtaatgctttt	ggcatgtttaccagGTACC	360	2938	109
3	ATCAAGgtagtagcagtggtg	tgtctgtctttctagGTTTCA	505	930	144
4	GAAAAGgtaagaatgtaata	atTTTTTcttatagGTTCC	595	2203	89
5	GATAAGgtataatcatcatct	TTTTTaatTTgatagGGTACC	652	2846	56
6	ACCAAGgtgaggggttttcc	ggTTTTTaaatggcagGATGTC	743	5348	80
7	GTAAGGtaagtaatttgctt	gtgtctcttttccagCGAGGA	835	1958	101
8	AGAAAgtgcaatgaaccttg	ttctccctgatttagATTTT	964	565	128
9	CAACAGgtgaggtatTTTca	tccttttctctccagAGGGCA	1071	924	106
10	AATGAGgtaaatgatgtcaat	ttcttattccttttagGGGAT	1210	909	138
11	ACTGAGgtacgtcaaaaagat	tgtatTTTTTcttagTACCTG	1288	1551	76
12	GTGAGgtcagtagtggtgct	tccttattgcttatagGTTTCC	1387	880	98
13	ATGAGGtgagagaatatgat	TTTTTgttctttcagGAAGAA	1483	3789	95
14	ACTGTGgtgagttcttataag	ttggTtcttcttttagGAGTGG	1673	993	189
15	GAATCGgtaaggattctttgc	tgttctctgttgacagATTGAG	1831	1489	157
16	ACAAAGgtatgtatTTTTgta	gtcaccatTTctagGCCTTT	1960	2771	128
17	ACACAGgtaggcaagtatct	ttcttggatataatagGTGACT	2105	1483	144
18	GACTGGgtaagaccagaataa	gtcttctTTTTatagCTACCC	2316	1419	210
19	GGAGAGgtgagtccttggtt	aaacatgtattacagGTCATG	2515	3557	198
20	GCCAAGgtaagatgtttacaag	ttctctcttggcagGCCATT	2692	4129	176
21	CAACAGgtaaggcagtgctgc	tctattaaatcctagTGAGTC	2811	21160	260
22	AAACAGgtaaaaggagtttccc	tgttttctcttacagATTGAT	2953	1306	129
23	GGTCAGgtaggaatgggcaaa	ttcttatttctcagGAGCAT	3019	780	65
24	ACATTGgtaaaataataaatt	tcattgtctctgcagGCAGCA	3083	2424	63
25	tcattgtctctgcagGCAGCA	gagtgatTTTTccagGCATGT	3265	5197	181
26	ACTTTGgtaaggagagaaca	gtttctattaattagAATGAA	3460	415	194
27	TTTGAGgtaaacactgtatatt	cccttcttcttccagTGGGTC	3528	1822	67
28	TCTCAGgtagtgccaatttc	tttctctctgtagcagGAGCAC	3604	3586	75
29	GGAAGGgtaagtgtttattc	TTTTTTTTggccagGTTATC	3682	2422	77
30	GAAAAGgtcatcatttataaa	tgctTTTTTaatcagGTGAAG	3790	1966	107
31	GTCTTGgtaagaaatgccaa	atTTTTatctctcagGTGATT	3880	1796	89
32	ATTAAGgtaacattttcaaaa	ttctttttccaacagGCTCCC	4051	2467	170
33	GGAGAGgtaactagtttaata	tttatttctaaacagGTGGCT	4126	4988	74
34	ATCCAGgtaagtggtttgtgg	gaaattccttttccagCAATCA	4225	6617	98
35	TTC AAGgtatgttcattagtt	tcattttaaatgcagCTGTTT	4285	766	59
36	CAGAAAgtaaactaggcat	tgtttatTTTTccagATTGAT	4398	1166	112
37	TTTCAGgtatagtcTTTTct	ccattatatttccagGTGACC	4528	750	129
				116748	4496
				3155	122
					Total Average
	Sequence preceding exon 1		aaagatcgtgtgcaGTCAG		
	Sequence following the stop codon		GCAGCA [TAG] agatgcagaccccc		
	Sequence following the polyadenylation signal		CTATTAAGAGGTaatgcagttgaatctggt		
	Size of exon 38		1219		

AC008172.1) that was identified by a BLAST search. In the process of this characterization, we found a number of discrepancies between our consensus *CPSI* coding sequence (AF154830) and the previously published *CPSI* coding sequence (NM_001875). Although the previously published *CPSI* coding sequence shared a 99% similarity with that derived in our characterization, we found 34 discrepant portions between the two sequences including single and double base pair mismatches. In addition we found discrepancies between our intron exon structure and the mRNA-genomic alignment obtained with sequence NT_0005403 published in LocusLink 1373. This entry

includes an extra exon that through our repeated analysis is actually part of intron 21 because it includes a stop codon and the database reports an 'indeterminate' boundary for intron 21/exon 22. NT_0005403 also reports that what we have found from genomic amplification are the first 12 bases of exon 7 as the last bases of exon 6. We have confirmed both these findings and shown our sequence is the correct one.

3.2. Polymorphisms in the human *CPSI* gene

During the characterization of the *CPSI* gene, the cDNA,

and our mutation studies in CPSI deficiency, fourteen polymorphisms were detected in screening a cohort of 100 unrelated individuals (Table 2). Three single base pair polymorphisms were identified in exons 21 (bp 2802), 35 (bp 4249), and 36 (bp 4340). The G/C polymorphism in exon 21 encodes a silent glycine to glycine change (G893G). The G/A polymorphism in exon 35 encodes a glycine to serine substitution (G1376S). The C/A polymorphism in exon 36 encodes a threonine to asparagine substitution (T1405N). The CAC/GCT polymorphism, located in exon 10 starting at base pair 1153 causes a threonine to be changed to an asparagine (T344A). The exonic polymorphisms screened are around 40% heterozygous. Four single base pair polymorphisms were identified in the 3' untranslated portion of the CPSI gene while one insertion/deletion trinucleotide polymorphism was identified in the 5' untranslated region. Statistical analysis performed on these exonic polymorphisms in a group of 100 patients indicates no significant linkage disequilibrium, demonstrating their usefulness in diagnostic purposes. Additionally, we identified six intronic polymorphisms whose heterozygosities range from 20 to 50% in a

small sample group. Two of these intronic polymorphisms have been reported on the SNP database accessed through LocusLink 1373 and their percent heterozygosity is included in Table 2. Our intronic polymorphisms include two variable GT repeats, three single base substitutions, and one single base pair deletion.

3.3. Exon primer pairs

In Table 3 we list the primer pairs we have found to have the greatest success in amplifying all the exons, exon flanking regions, 5' untranslated, 3' untranslated, and cDNA flanking sequences. The fragments generated by these PCR reactions are suitable for mutation screening by a number of methods.

4. Discussion

4.1. Organization of the human CPSI gene

Our data of the intron and exon sequence shows that

Table 2
CPSI polymorphisms

Polymorphism	Name	Location	Heterozygosity	Primer pair
<i>Translated polymorphisms</i>				
<u>ACC</u> ↔ <u>GCT</u>	Thr344Ala	Exon 10	44% ^a	U1119 L(I10) + 37
<u>GGC</u> ↔ <u>GGG</u>	Gly893Gly	Exon 21	43% ^a	U2712 L(I21) + 47
<u>GGT</u> ↔ <u>AGT</u>	Gly1376Ser	Exon 35	40%	U(I34)-81 L(I35) + 47
<u>ACC</u> ↔ <u>AAC</u>	Thr1405Asp	Exon 36	44% ^a	U4295 L(I35) + 47
<i>3' Untranslated polymorphisms</i>				
C ↔ T		4972		U4926 L5025
C ↔ G		5006		U4926 L5277
G ↔ T		5064		U4926 L5277
C ↔ G		5318		U5195 L5547
<i>Intronic polymorphisms</i>				
(GT) ⁿ		Intron 5		
G ↔ T (rs2287600)		Intron 18	44% ^b	
T deletion		Intron 25		
A ↔ T (rs3213784)		Intron 27	50% ^b	
(GT) ⁿ		Intron 33		
A ↔ G		Intron 36		
<i>5' Untranslated polymorphisms</i>				
CTT insertion		– 118		U5' – 74 L175

^a % heterozygosity based on analysis of 100 unrelated samples.

^b % heterozygosity listed on LocusLink.

Table 3
Primers used for exon amplification

Exon amplified and fragment size	Primer	Sequence	Melting temp. (°C)	Exon amplified and fragment size	Primer	Sequence	Melting temp. (°C)
1 (206)	U0129	GAGGATTTTGACAGCTTTCA	66	23 (244)	U(122) – 147	AATTCCACGGGTCATACATT	67
	L(11) + 85	TATTTGCCTTGTAAGGGACA	66		L(123) + 32	CTTCTGGATAGGCCAATTT	67
2 (208)	U(11) – 28	GCTAAGATTCATGCAACTGTT	65	24 (289)	U(123) – 35	ATGATTTCTGCATCTTCCTT	64
	L(12) + 69	CCATGAGTGAGATCTCCAG	65		L(124) + 191	TTGAGTCCCACCTACCACCT	64
3 (244)	U(12) – 77	TCAGAGCATGTATGCAGATT	66	25 (345)	U(124) – 23	TGCCAATCTCATGTCTCTG	67
	L(13) + 30	AATGGCCACCACTGGTACTA	68		L(125) + 141	TAAGACAATGGATGCCACTG	67
4 (297)	U(13) – 114	CCTAAGTCTCATGTCACTGGA	65	26 (289)	U(125)32	CCAATGGCTGATATTGTGAG	67
	L(14) + 94	CATGATCTTCCCTGTCTGCG	65		L(126) + 63	AACATTCAGGTTGCAGCTCT	68
5 (269)	U(14) – 91	GAAGATAGATGAGGCCAAGT	64	27 (375)	U(126) – 101	ATAAAGAAGCTGGGCTACCA	64
	L(15) + 119	GGCCAGATTCACCTCCTTTAC	67		L(127) + 205	AGTCTTCTATGGCCACTCC	68
6 (296)	U(15) – 97	TAAGTCTTGACAGCAGCTGTT	67	28 (249)	U(127) – 77	GGAAGTCTTCTGAGAACCAG	64
	L(16) + 119	CTGGTGTCTTCTGGAAGATA	66		L(128) + 99	TTGCATTAAGAGATAAATGTCA	64
7 (277)	U(16) – 141	TGTTAACTGCCAGTCACTCC	66	29 (233)	U(128) – 123	TTCAGCCTTGAGTATTTTGC	66
	L(17) + 45	CAATATGACAAACCCTCAC	65		L(129) + 33	CAGGAATGAAGAGATGAGA	65
8 (290)	U(17) – 26	TCATCATTTCTGTCTCTTTT	65	30 (262)	U(129) – 103	GCTCAGTGCATCTGTTAGGA	66
	L(18) + 135	CCTGCCTCAGAAGTCACAA	67		L(130) + 48	TTTCTGAGTGTTCCTTTTC	66
9 (211)	U(18) – 67	TGTTCAATTCAGAAGCAACAT	64	31 (290)	U(130) – 162	AAACCTAATAGTCCCTCCAT	68
	L(19) + 38	GAGCTGGTTTACTGTAAGCA	64		L(131) + 39	CAGTGGTGGTGTCTCTCAG	68
10 (213)	U(19) – 165	ACTGGGATTAACAATCTGTGAC	66	32 (329)	U(131) – 43	GCTCTCAATGCTCCTTTCT	64
	L(110) + 40	TTCTCATCACCAACTGAACAG	66		L(132) + 116	CCTAGTCCAACCTTTTAGTT	64
11 (247)	U(110) – 57	GCCACACTTGACATTCATTG	67	33 (342)	U(132) – 162	ATCTTTCAAGTCGGATGCTT	67
	L(111) + 113	TTCTGTCTAAGATCATCTGG	66		L(133) + 106	GAAAGCTTCCATGGATTATT	64
12 (258)	U(111) – 112	TGTTGTCTTATCTGAAGTCAA	64	34 (230)	U(133) – 82	TGGTTGACTATTTCTTGCATC	66
	L(112) + 56	TCATGTGATAGGGACAATCC	65		L(134) + 78	CCACTCACCATGTATGACCA	67
13 (339)	U(112) – 82	GGTTGTCTTCTCCAATCTT	64	35 (187)	U(134) – 81	CATCCGGCAACATGTAACA	68
	L(113) + 162	TCGTGATCAAGCTTCTTTAA	64		L(135) + 47	GACAGATTCATGCTGACAGA	66
14 (272)	U(113) – 31	TTCATGTACTGGATCTTTTGT	65	36 (273)	U(135) – 132	AAGTCTCTATCCATGGCACT	64
	L(114) + 49	CACAGCAGTGCATATGTCAA	66		L(136) + 29	TTTCAGAAAACAGTATGCCTA	64
15 (273)	U(114) – 52	AAGGCACAAAATTCAGATT	65	37 (272)	U(136) – 109	ACTCAGCATGGCATTGACTT	68
	L(115) + 62	TCATTTCTCCCTACAACAACAG	65		L(137) + 33	CCATCCAGTCTATATCCAAGG	67
16 (250)	U(115) – 57	GCTGTCTTCTGCCAGTGTGCT	68	38 (UI37)	U(137) – 29	GGTGTGATTCCTACCAFTA	65
	L(116) + 62	TGAAAAGGTCTATCAAGCCAG	68		L4887	CATAGGCAAGCCTAGTCCAC	68
17 (263)	U(116) – 60	TATCCTTGTGGAAGCCAGTG	66	5' – UTR 1 (249)	U74	GGTTAAGAGAAGGAGGAGCTG	68
	L(117) + 56	CCATCAAAGACATGAACTGAA	66		L175	AACCAGTCTTCAGTGTCTCA	67
18 (276)	U(117) – 33	GCCACCAAGTAATGAGTGCT	68	5' – UTR 2 (143)	U118	TTACATGCCCATGGAACATC	68
	L(118) + 31	GCAAACCCATGGTCAATTAT	68		L25	GTGCACCTCAGTGTTAAGGC	66
19 (277)	U(118) – 55	TCTGTTTCAATAATTCTCG	67	3' – UTR 1 (366)	U4661	CCTGAGCCACATGTTATCTA	64
	L(119) + 24	AAGCGTAATAAACAAGGAC	65		L5025	TGTCTGAGTTTGCAGATAG	64
20 (296)	U(119) – 81	TGGAGAAAGTGAGAGAGGAA	65	3' – UTR 2 (351)	U4926	AATGGTGTCAAGGTAGGAA	65
	L(120) + 43	AAGCAACTAGTAGCTGTGGG	65		L5227	TGAGAGTGTGACTCCATCT	65
21 (239)	U(120) – 87	CAGAGGCATGACTGCTATGT	66	3' – UTR 3 (352)	U5195	TGTGACAGAGGCATTTAGAG	64
	L(121) + 34	GTGCTGGCATGAATGAGAG	66		L5547	GGAATGAACCTTACTTCCAA	65
22 (251)	U(121) – 61	TGTCACCGCTTTTATTTGTT	66	3' – UTR 4 (438)	U5406	TCAGCAGATGGTAGACAGTG	64
	L(122) + 49	CACAGAAATCCTGTCACTGG	66		L + 97	AATAAGCAATGCAATCTG	68

CPSI adheres to the consensus sequence of a 'GT' at the 5' end of each intron and an 'AG' at the 3' end. Additionally, because 37 out of 38 exons end in a 'G', this is greater than the average consensus of a last 'G' in mammalian exons. Perhaps this observation is not too unusual in light of the fact that the *CPSI* gene has a highly conserved nature that extends between its structure and that of its known non-mammalian bacterial and yeast counterparts (Schofield, 1993). The intron and exon structure is consistent with the previously hypothesized theory that *CPSI* arose from both an internal duplication and from a gene fusion event because

in most genes, larger introns are confined to the first part of the gene and do not occur near the end of the gene (Schofield, 1993). Our finding that introns 1 and 21 are the largest introns of the gene lends credence to the previous theory that the *CPSI* gene arose from an internal duplication in its structure/sequence (Schofield, 1993). The average size of internal vertebrate exons is 137 bp (Hawkins, 1988). In comparison of *CPSI* with this average size, we find that the *CPSI* exons are indeed of average size. The average vertebrate intron size is 1.1 kb (Hawkins, 1988). Even with the exclusion of the two largest introns (intron 1 and

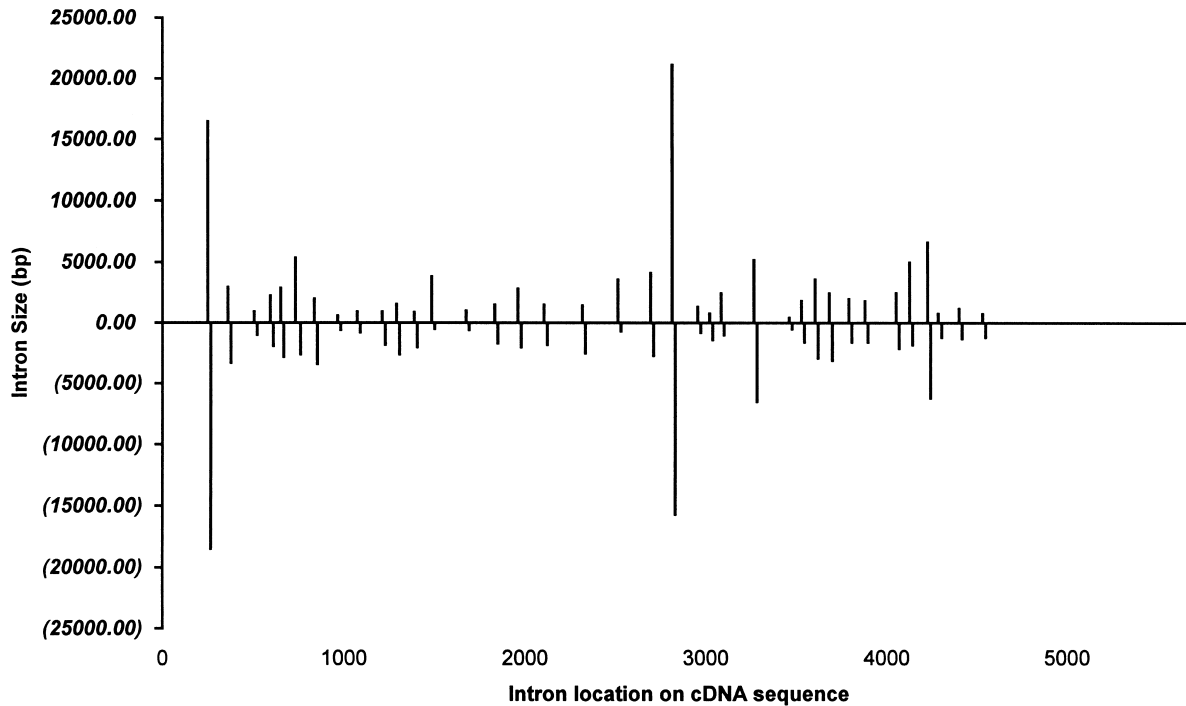


Fig. 1. Comparison of rat and human *CPSI* introns.

intron 21), the average size of the *CPSI* introns (>2.2 kb) is substantially larger than the average intron size.

4.2. Polymorphisms in the human *CPSI* gene

These 14 identified polymorphisms in both coding and non-coding regions will provide a useful tool for prenatal linkage analysis and other genetic studies that involve the *CPSI* gene. Currently, functional studies are underway to determine whether or not these polymorphisms have any effect on normal *CPSI* enzyme function. Preliminary data (not shown) indicates a 30–40% difference in enzyme activity for the T1405N polymorphism. We have recently published data that implicates the T1405N polymorphism as a risk factor for persistent pulmonary hypertension of the newborn (Pearson et al., 2001).

4.3. Similarity analysis of rat and human *CPSI* gene structures

The rat and human *CPSI* genes both have the same number of exons and share a high degree of size similarity. For comparison, the rat gene contains 38 exons (average size of 125 bp) and 37 introns (average size of 2.8 kb). With the exception of exons 1 and 38, the 38 rat and human exons have the same number of base pairs. In size comparisons between the rat and human, introns vary to a greater extent than exons, but still follow a similar sizing trend between the two species (Fig. 1). For example, introns 1 and 21 are very large in both the human and the rat, whereas intron 26

is the smallest intron in both species. Despite the fact that the human *CPSI* gene coding sequence is larger than that of the rat gene, they share a strong degree of identity (~88%) and their introns are located at the same relative position in their cDNAs. Furthermore, an amino acid comparison reveals over a 97% degree of identity between the two species. A schematic comparison of the alignment of the rat and human *CPSI* genes is shown (Fig. 1). The high degree of homology between the rat and human for their respective *CPSI* cDNA sequences, amino acid compositions, and complete gene structures demonstrates that the *CPSI* gene has been strongly conserved. The larger than average intron size in both species is interesting in light of the known antiquity of this gene and also suggests an increase in size over the evolutionary history of the gene.

Since *CPSI* has such an important function in the human body and because deficiencies of this enzyme lead to extremely serious medical problems, it is vital to gain a better understanding of this enzyme and the gene that encodes it. The genetic characterization and polymorphic data presented in this study will offer researchers a molecular means for studying this enzyme.

Acknowledgements

The authors wish to thank the NIH (ES09915) and the National Urea Cycle Disorders Foundation for their support of this project.

References

- Haraguchi, Y., Takako, U., Takiguchi, M., Endo, F., Mori, M., Matsuda, I., 1991. Cloning and sequence of cDNA encoding human carbamyl phosphate synthetase I. *Gene*. 107, 335–340.
- Hawkins, J.D., 1988. A survey on intron and exon lengths. *Nucleic Acids Res.* 16, 9893–9905.
- Nyunoya, H., Broglie, K.E., Widgren, E.E., Lusty, C.J., 1985a. The gene coding for carbamyl phosphate synthetase I was formed by fusion of an ancestral glutaminase gene and a synthetase gene. *Proc. Natl. Acad. Sci. USA* 82, 2244–2246.
- Nyunoya, H., Broglie, K.E., Lusty, C.J., 1985b. Characterization and derivation of the gene coding for mitochondrial carbamyl phosphate synthetase I of rat. *J. Biol. Chem.* 260, 9346–9356.
- Pearson, D.L., Dawling, S., Walsh, W.F., Haines, J.L., Christman, B.W., Bazyk, A., Scott, N., Summar, M.L., 2001. Neonatal pulmonary hypertension–urea-cycle intermediates, nitric oxide production, and carbamyl phosphate synthetase function. *N Engl. J. Med.* 344, 1832–1838.
- Powers-Lee, S.G., Corina, K., 1986. Domain structure of rat liver carbamyl phosphate synthetase I. *J. Biol. Chem.* 261, 15349–15352.
- Powers-Lee, S., Corina, K., 1987. Photoaffinity labeling of rat liver carbamyl phosphate synthetase I by 8-azido-ATP. *J. Biol. Chem.* 262, 9052–9056.
- Rubio, V., 1993. Structure–function studies in carbamyl phosphate synthetases. *Biochem. Soc. Trans.* 21, 198–202.
- Rubio, V., Grisolia, S., 1981. Human carbamoylphosphate synthetase I. *Enzyme*. 26, 233–239.
- Rubio, V., Ramponi, G., Grisolia, S., 1981. Carbamyl phosphate synthetase I of human liver. Purification, some properties and immunological cross-reactivity with the rat liver enzyme. *Biochim. Biophys. Acta* 659, 150–160.
- Schofield, J.P., 1993. Molecular studies on an ancient gene encoding for carbomyl-phosphate synthetase. *Clin. Sci.* 84, 119–128.
- Shapiro, M.B., Senapathy, P., 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* 15, 7155–7174.
- Summar, M.L., Dasouki, M.J., Schofield, P.J., Krishnamani, M.R., Vnencak-Jones, C., Tuchman, M., Mao, J., Phillips, J.A. III, 1995. Physical and linkage mapping of human carbamyl phosphate synthetase (CPS1) and reassignment from 2p to 2q35. *Cytogenet. Cell Genet.* 71, 266–267.
- van den Hoff, M.J., van de Zande, L.P., Dingenmanse, M.A., Das, A.T., Labruyere, W., Moorman, A.F., Charles, R., Lamers, W.H., 1995. Isolation and characterization of the rat gene for carbamoylphosphate synthetase I. *Eur. J. Biochem.* 228, 351–361.