J Mol Evol (1998) 47:763–771

JOURNAL OF MOLECULAR EVOLUTION

Ancient and Recent Intron Stability in the *Artemia* Hemoglobin Gene

Charles M. Matthews, Clive N.A. Trotman

Biochemistry Department, University of Otago, Box 56, Dunedin, New Zealand

Received: 17 March 1998 / Accepted: 6 May 1998

Abstract. *Artemia* has evolved three distinct hemoglobins formed by the association of two nine-domain globin polymers. Sequence analysis of cDNA clones corresponding to two polymers, named T and C, indicates that their genes are the products of a duplication event some 60 million years ago. The present study indicates the presence of 22 introns in each of the T and C polymer genes. The 22 introns are classified into two groups: 17 correspond to positions within globin domains, and 5 correspond to interdomain linkers (or N- and C-terminal extensions). Intron position and reading frame phase are precisely conserved between T and C polymers for all 22 introns, but within each gene the position and phase are not always consistent from domain to domain or from linker to linker. The discordance of *Artemia* hemoglobin introns is discussed in terms of different model mechanisms and constraints: intron sliding, intron loss or gain, and the exon definition model of primary transcript RNA splicing. The results suggest that constraints of pre-mRNA processing should be considered when considering intron positional changes in homologous genes.

Key words: Protein evolution — *Artemia* — Intron — Exon — Hemoglobin

Introduction

The origin of spliceosomal introns is still a controversial issue despite two decades passing since their discovery. Two main theories have been proposed to account for the presence of introns in the protein coding genes of eukaryotes: the exon theory of genes (Gilbert 1987; Gilbert et al. 1997), an extension of the introns-early view (Doolittle 1978; Gilbert 1978); and the insertion theory of introns, or introns-late view (Rogers 1990; Cavalier-Smith 1991; Stolzfus et al. 1994). The introns-early theory is that most, if not all, extant genes were formed by the linking-together of primordial minigenes which survive today as exons, with introns representing the locations of original noncoding spacers between the minigenes. According to this view, introns facilitated the formation of today's diverse repertoire of genes by events of exon shuffling (Dorit et al. 1990; Long et al. 1996).

The identification of chimeric proteins (Doolittle 1985; Patthy 1985) provides support for the existence of exon shuffling events but does not prove that exon shuffling occurred significantly in the progenote. Studies on intron phase correlations with respect to the reading frame and the finding of an excess of symmetrical exons (Long et al. 1995a,b) have been taken as support for the introns-early hypothesis. However, Hurst and McVean (1996) have pointed out that since most introns appear in phase 0 (i.e., between codons), the finding that most symmetrical exons are in phase 1 (following the first nucleotide of the codon) is in conflict with an exon shuffling explanation.

In the introns-late view it is not disputed that spliceosomal introns once present can aid the evolution of new proteins by exon shuffling but it is doubted that spliceosomal introns were present in primordial genes or that exon shuffling events occurred in primordial ancestors. The introns-late view questions how a eubacterial cell lacking compartmentation of transcription and transla-

Correspondence to: C.N.A. Trotman;
e-mail: clive.trotman@stonebow.otago.ac.nz

tion events could cope with spliceosomal intron interruptions in their protein coding genes. The translation of unspliced or partially spliced pre-mRNAs would result in both chimeric and truncated proteins (Cavalier-Smith 1991).

Both sides have disagreed about the existence of a correlation of intron clustering with the boundaries of subjectively defined modules of protein structure (Go 1981; de Souza et al. 1996; Stoltzfus et al. 1994). One problem in such an analysis is in determining which intron positions are the most ancient, as the present distribution of intron positions in homologous genes is likely to be the result of intron loss and gain and, if it happens, intron sliding. The two theories are not mutually exclusive, however, since it is possible that a large initial complement of introns of early origin has undergone extensive loss, and it is the surviving introns that have been moved and replicated more recently by recombinant means.

Homologous genes coding for a particular protein often have introns in different locations in different organisms. The focus of much of the debate is whether some of the closely but inexactly aligned introns may be considered homologous or not. The introns-late view is that the discrepant intron positions cannot be explained adequately by patterns of inheritance and loss but are more likely to be due to events of loss and gain. The introns-early view is that homologous introns can change position and phase and thereby appear discordant. Inferred movement of an intron over a short distance is commonly referred to as intron sliding. Mechanisms have been proposed to explain how an intron might slide but as yet there is no firm evidence (Stoltzfus et al. 1997). Intron positions in several genes, including the globin genes of *Artemia,* could be interpreted as evidence of sliding (Jellie et al. 1996).

Artemia hemoglobins are large molecules comprising two polymers (identified as T and C), each being a continuous chain of nine globin domains. Individual globin domains within a polymer are 17–35% identical at the amino acid level and are thought to have arisen from a monomer by an ancient series of gene duplication events (Jellie et al. 1996). The amino acid identity of 88% and DNA identity of 84% between the aligned T and C polymers indicate that they arose by duplication of the entire nine domain gene much more recently.

Thus, different time marks can be identified in the history of the gene. After partial correction for coincident mutations using the Poisson formula and taking the commonly cited equation for globins of about 5 MY per 1% amino acid divergence, the multiplication from one to nine domains appears to have occupied a period about 500–700 MYA (Trotman et al. 1994). Given the insufficiency of the Poisson correction for this problem, the flatness and uncertainty of empirical correction curves at higher divergences (Dayhoff et al. 1983), which would raise the 700 MY figure to about 1150 MY, and the uncertain validity of the 5 MY per 1% evolutionary period over such a time scale, we refer to the period of gene lengthening from one to nine domains descriptively as ''a billion years ago.'' The 12% divergence of the T and C variants incurs minimal correction but the applicability of the 5 MY unit evolutionary period to a long polymeric structure with further quaternary constraints is uncertain and the corresponding date of 60 MY similarly is descriptive.

The present study was conducted to impose a time frame over the *Artemia* globin introns, revealing the stability or instability of different loci and classes of introns (intradomain and interdomain or linker) throughout ancient and relatively recent evolutionary history. We conclude that 44 introns have been completely stable in location for 60 MY but that some discordances were generated over the billion-year time frame. Introducing the linker introns into the data set enables the entire complement of introns to be described by a model which is more supportive of deletion and insertion. Constraints imposed by pre-mRNA splicing may in part determine both the distribution and the stability of introns in the protein encoding genes of eukaryotes.

Materials and Methods

Genomic Library Construction. Genomic DNA was isolated from hatched *Artemia* nauplii following incubation of cysts at 22°C for 44 h. Cysts were collected from Lake Grassmere, New Zealand. Isolation of high molecular weight DNA was performed by a modified procedure of Blin and Stafford (1976) as detailed by Sambrook et al. (1989). The DNA was partially digested with *Sau*3AI (Boehringer Mannheim) and size fractioned by sucrose gradient centrifugation (Kaiser et al. 1995). Fractions of DNA between 14 and 20 kb were chosen for the subsequent cloning into precut LambdaGEM-11 (Promega) according to the manufacturer's instructions.

Library Screening. Approximately 800,000 recombinants were plated on Luria broth plates using cell type KW251 and duplicate lifts were taken using Hybond N membranes (Amersham). Hybridization of membranes with both T and C polymer cDNA probes (Matthews et al. 1998) was performed at 60°C for at least 6 h in 4× SSPE, 0.1% tetrasodium pyrophosphate, 0.5 mg/ml heparin, and 0.5% SDS, and membranes were then washed twice for 10 min in 2× SSPE and 0.5% SDS before autoradiography.

DNA Sequencing. Primers, T and C polymer-specific, were used to amplify selected regions of the positive clones using the polymerase chain reaction (PCR) on a PTC-200 thermal cycler (MJ Research). The size of the products was estimated by agarose gel electrophoresis. All sequencing was performed on uncloned PCR products after purification using a Qiaquick PCR purification kit (QIAGEN). Sequencing was performed using dRhodamine terminator chemistry and an ABI Prism 377 DNA sequencer.

Computer-Assisted DNA Sequence Compilation and Analysis. Translation functions were performed using the program NLDNA (Stockwell 1987). The homology of new DNA sequences was deter-

mined using HOMED (Stockwell 1988) or the GCG (Genetics Computer Group Inc) sequence analysis package, Version 7.3.

Results and Discussion

Identification of Introns

Sequence analysis of C polymer genomic DNA clones against the previously determined cDNA sequence (Matthews et al. 1998) resulted in the identification of 22 introns (Table 1). Of these, 17 introns correspond to intradomain inferred locations in the globin (i.e., structural helices A–H), 3 correspond to interdomain linkers, and 2 correspond to amino- and carboxy-terminal extensions. A previous study on the T polymer gene identified 17 introns corresponding to the intradomain regions (Jellie et al. 1996) and noted the existence of at least two interdomain introns (Trotman et al. 1994). The present study completed the search for extradomain introns in the T gene, finding five as in the C gene. The *Artemia* globin intron data set is now complete and each of the two homologous genes contains 22 introns in identical positions and phases; however, each gene is divided into nine repeating domains in which intron positions and phases are not always consistent. Such regions of the introns as have been sequenced, in the vicinity of exon–intron boundaries, do not reveal any evidence of intron sequence similarity between or within genes.

The data set has been analyzed in order to seek clues to the origin of these introns and whether the present distribution is attributable to intron movements relative to the gene, or to independent events of insertion and deletion, or to both. Since logically the introns located between domains acquired their present positions later than the globin gene came into existence, it is convenient for discussion to distinguish between an intradomain class within the A–H structure (‘‘domain’’ introns for brevity) and those located in linkers or terminal extensions, which, being possibly related, will be grouped as ‘‘linker’’ introns.

The time dimension can be divided into at least four phases: (1) the recent phase since nominally 60 MYA, when the T and C genes arose by duplication of a nine-domain gene; (2) the preceding period following gene enlargement from one to nine domains; (3) the period during which gene enlargement from one to nine domains occurred, around 1 BYA; and (4) the period in which monomeric globin evolved, probably considerably more than 1 BYA.

The conclusion relating to the most recent period of 60 MY until the present is a clear and striking positional stability in which 44 introns have remained completely stable over this time (discounting events for which there is no evidence such as an intron moving identically in both genes, an intron in one gene moving then reverting, or deletion and replacement at the same point). Evidently these intron positions have been inherited and faithfully preserved for at least 60 MY, although the introns themselves may have mutated beyond recognition. This absence of displacement, loss, or insertion can be expressed as equivalent to about two and a half billion intron-years of stability.

The same inheritance and stability cannot both have applied throughout the nominal 1 BY period since the gene coded for a monomeric globin because some introns are now in different positions in different domains. Taking the 60-MY period of stability just described, in conjunction with the existence in each gene of eight B helix introns and six G helix introns, all having stable respective phases and domain loci, there is no basis to reject a continuous inheritance of those intron positions since the monomeric gene era. Conversely, other introns in the set are in discordant locations and domain 4 is missing a B helix intron. The question is whether these discordances have come about by either (1) a process in which exonic mutations have resulted in apparent movement of the original intron and in one case its loss or (2) independent events of complete intron removal and displaced insertion, or of gene conversion, giving an appearance of intron movement. Either mechanism admits the possibility of hidden events where the status quo has been restored.

The interpretation of the locations and discordances rests on the inferred translation of DNA sequence into protein structure by reliance on generic globin structural information from other species. The reliability of this translation is enhanced by the availability of two *Artemia* domains sequenced at the protein level (Moens et al. 1988, 1990), which provide the key to the verification of the reading frame, and by the internal consistency of the *Artemia* alignment across nine domains. Further verification is provided by the independent preparation and sequencing of the T and C cDNA sequences by different persons and technology (manual and automated, respectively). The T and C sequences translate exactly in register, are 88% identical at the amino acid level, and have the same idiosyncrasies such as the rare deletion of residue C2 or C3 from domain 9.

Domain Introns

The pattern of domain introns in the C gene (Fig. 1) is the same as published for the T gene (Jellie et al. 1996). Eight domains contain a B helix intron in the position conventional for globins (B12-2), and only domain 4 does not have a B helix intron. In contrast, the G helix intron positions are more variable. Of the nine C polymer domains, only six have an intron at the usual position in the G helix (G6/G7). The G helix intron of domain 4 is discordant by a single base (G6-2). Domains 3 and 6 lack a G helix intron, yet in both cases introns are present in the F helix and are themselves discordant at F3-1 and F2/F3, respectively.

Table 1. C polymer intron/exon boundaries: nine nucleotides of the intron sequence at the 5′ and 3′ splice sites are shown for each of the C polymer domains and linkers[a]

	Exon				Exon			Polymer intron size (kb)[b]	
Domain	B10	B11	B12	Intron	B12	B13	B14	C	T
				B helix					
C1	TTC	CTA	AG	GTAATTCTA…AATTTATAG	T	GTT	TTT	4.2	1.5
	F	L			S	V	F		
C2	TTC	CAG	AG	GTAAAATTG…CTGTTTCAG	A	TTG	ATC	0.7	2.0
	F	Q			R	L	I		
C3	TTT	GGA	AA	GTGAGTAAC…CAATTTTAG	A	CTC	TTT	1.3	1.3
	F	G			K	L	F		
C4	TTC	ATG	AG	(No intron)	G	ATG	TTC	—	—
	F	M			R	M	F		
C5	TTC	GCT	AA	GTAAGTATA…TTTATTTAG	G	CTG	TTC	3.1	1.5
	F	A			K	L	F		
C6	TTC	ACA	AG	GTCATTATG…CTTTTTTAG	A	CTT	TTC	2.0	2.7
	F	T			R	L	F		
C7	TTC	AAA	AG	GTAAGCTCA…ATTTTTTAG	C	CTT	TTC	3.1	2.7
	F	K			S	L	F		
C8	TTT	GGA	GT	GTAAGTAGT…TAATTTCAG	G	GTA	TTA	2.3	2.0
	F	G			V	I	F		
C9	TTC	AGA	CA	GTAAGTAAA…TTTTCTTAG	G	CTA	TTC	1.0	1.0
	F	R			Q	L	F		
				G and F helix					
	G4	G5	G6		G6	G7	G8		
C1	CAC	TTT	GAG	GTGAGAATT…ATCTTTCAG		GCC	TTT	1.5	4.3
	H	F	E			A	F		
C2	CAT	TTC	CAG	GTAATTAAA…TTCCTTTAG		AAT	TTC	2.6	2.3
	H	F	Q			N	F		
C4	CAC	TTC	AG	GTAAGAAAA…TATTTGCAG	A	AGC	TTC	1.0	1.3
	H	F			R	S	F		
C5	CAA	TTC	GAT	GTGAGTTCA…TTTTTCCAG		CAA	TTT	1.2	2.0
	Q	F	D			Q	F		
C7	ATG	TTC	AAG	GTAACTATC…TATATCCAG		AGT	TTT	1.1	1.4
	M	F	K			S	F		
C8	CAT	TTC	CAG	GTAGGGATC…TTATTTTAG		GCC	TTT	1.4	2.3
	H	F	Q			A	F		
C9	CAT	TTC	GAT	GTAAGTATG…TCTTAATAG		GAC	TTC	3.7	3.8
	H	F	D			D	F		
	F1	F2	F3		F3	F4	F5		
C3	ATC	AAA	G	GTGAATATT…TATTTGAAG	AA	CTG	GGA	1.8	1.1
	I	K			E	L	G		
C6	CTC	AAA		GTAAGTCTT…ACTTTTAAG	GAC	CTT	GGT	1.8	0.6
	L	K			D	L	G		
				Linker					
Linker	NA12	NA13	NA14		NA14	NA15	NA16		
S–/1	GCT	GCC	A	GTAAGTAAA…CTTTTTCAG	TT	GCC	TCT	3.5	1.4
	A	A			I	A	S		
	HA2	HA5	HA6		HA7	HA8	HA9		
L3/4	CTT	CGA	CAG	GTAAGAATT…ATTTTCAAG	GCC	AAC	GTC	2.7	1.1
	L	R	Q		A	N	V		
	HA1	HA2	HA3		HA4	HA5	HA6		
L4/5	GGT	CTT	AAG	GTAGGCTGA…AATTTGTAG	GTT	GCA	TCA	1.5	2.3
	G	L	K		V	A	S		
	HA13	HA14	HA15		HA15	HA16	A1		
L6/7	ATA	ACT	G	GTAAGTTCA…TTATTGCAG	GT	CTT	TCT	4.7	1.2
	I	T			G	L	S		
	HC1	HC2	HC3		HC4	HC5	HC6		
L9/–	GGA	CAA	AGG	GTAAGTTTT…TTATTTCAG	GCT	CGC	CTC	1.2	1.6
	G	Q	R		A	R	R		

[a] Exon sequence triplets are shown under their inferred amino acid positions in helix notation. Translations are given below each codon.
[b] Intron sizes, expressed as kilobases (kb), of the C and T polymer genes were estimated by agarose gel electrophoresis.

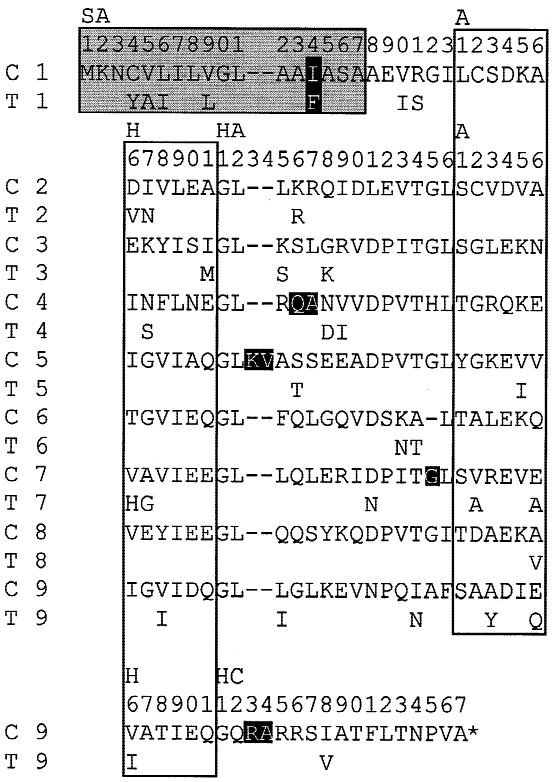

Fig. 1. Alignment of the C and T polymer amino acid sequences around the linker regions. Linkers are designated as those regions joining the H and A helices (both indicated by *open boxes*). The signal peptide is indicated by a *shaded box.* The complete C polymer sequence is shown (single-letter code), but only the differences between the two polymers are shown for the T polymer. Intron positions are indicated by a *block. Small blocks* indicate amino acids whose codon is interrupted by an intron in phase 1; phase 0 introns are indicated by *larger blocks* of two amino acids. The notation SA designates that region joining the signal peptide and A helix of domain 1; similarly HA connects the H and A helices of successive domains and HC connects the H helix of domain 9 with the carboxyl tail. Domain numbers correspond to the A helix. *Dashes* indicate alignment gaps.

Table 2. Summary of intron positions found in *Artemia* C and T polymer genes; the intron phase is shown in parentheses

	Helix position		
Domain	B helix	F helix	G helix
1	B12 (2)	—	G7 (0)
2	B12 (2)	—	G7 (0)
3	B12 (2)	F3 (1)	—
4	—	—	G6 (2)
5	B12 (2)	—	G7 (0)
6	B12 (2)	F3 (0)	—
7	B12 (2)	—	G7 (0)
8	B12 (2)	—	G7 (0)
9	B12 (2)	—	G7 (0)
Linker[a]	Position[b]		
S–/1	SA14 (1)		
L3/4	HA7 (0)		
L4/5	HA4 (0)		
L6/7	HA15 (1)		
L9/–	HC4 (0)		

[a] Nonintradomain introns: linker (L); signal peptide (S).
[b] Positions clarified in Fig 1.

Linker Introns

Covalently joining the nine globin domains are inferred linking peptides of 14–16 residues. An intron was identified within the sequences coding for three of these linkers (L3/4, L4/5, and L6/7, designating the domains they join; Table 2). Two of these are located near the amino ends of linkers L3/4 and L4/5 in phase 0, and the third is near the carboxyl end of linker L6/7 in phase 1. In addition, an intron corresponds to the end of domain 9 (L9/–, phase 0), on the amino terminal side of the carboxy-terminal tail structure, and another intron is in phase 1, four codons upstream of the coding sequence for domain 1 of the mature protein, in the sequence for the putative signal peptide (S–/1).

Superficially the new linker intron data fit the same model of long-term inheritance and occasional slippage as the domain introns. It is possible that gene concatenation as the number of domains increased was accompanied by the preservation of nontranslated sequences between domains, becoming linker introns of which five survive. The intron nearest to the 5′ end of the entire gene occurs between the signal sequence and the first domain, which may provide a clue to the origin of all the linker introns. The retention of only 5 of a possible 10 linker introns may indicate deletion by unknown mechanisms, although the deletion need not have happened as many as 5 times if some gene concatenation events reproduced the deletion.

Two-domain globins with bridging introns in the interdomain coding sequence have been identified in the clam *Barbatia reeveana* (Naito et al. 1991) and in the nematodes *Ascaris suum* (Sherman et al. 1992) and *Pseudoterranova decipiens* (Dixon et al. 1992). The study of the *Barbatia* globin provided experimental evidence for the formation de novo of the bridging intron. Similarity was observed between the bridging intron and regions of both the preceding intron and 3′ untranslated sequences. It was suggested that a crossover event occurred between the 3′ untranslated sequence of a single-domain globin gene and the preceding intron of a similar or identical gene. This mechanism, or a variation, could account for the formation of the two-domain globins in *Ascaris* and *Pseudoterranova.* These nematodes have similar two-domain extracellular globins, and although they both lack preceding introns, both have an intron between the 5′ leader sequence and the mature protein which could serve the same role as proposed for the preceding intron of *Barbatia.* The identical position and phase of both the 5′ intron and the bridging intron in these nematodes is consistent with this proposal.

If domain duplication in *Artemia* involved a crossover event between the intron within the leader peptide coding

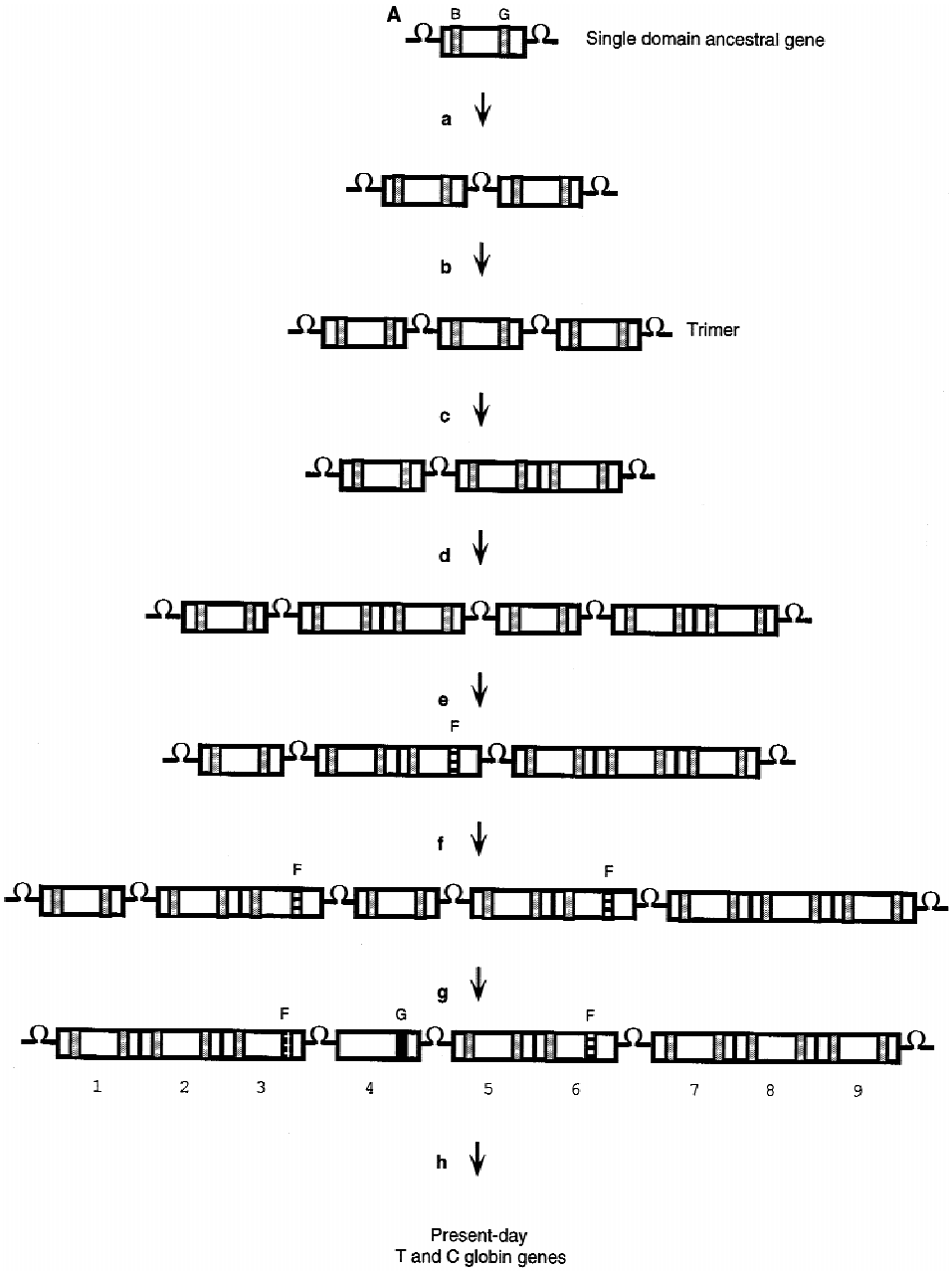

Fig. 2. Stages in a possible origin of the nine-domain hemoglobin polymers of *Artemia.* **A** The ancestral single-domain globin gene. The *open box* represents the globin domain and positions of intradomain introns are indicated by *smaller shaded boxes* (helix position indicated *above*); potential interdomain introns are represented by loops. **a, b** Domain duplication events, facilitated by the presence of interdomain (linker) introns. **c** Loss of a linker intron. **d** Duplication of the trimer. **e** Linker intron loss and movement (presumably by loss/gain) of the G helix intron in the third domain to an F helix position. **f** Duplication of the amino-terminal three domains to yield a nine-domain globin gene with two domains containing an F helix intron. A single base movement (by loss/gain) of one of the F helix introns. **g** Loss and gain of an intron in the G helix of the fourth domain and then loss of the B helix intron in the same domain. Movement (by loss/gain) of the amino terminal intron and the linker intron between domain 6 and domain 7 (L6/7). **h** Duplication of the entire nine-domain gene to produce two genes for T and C polymers.

sequences and a 3′ untranslated sequence, then all linker introns should be in the same position and phase 1 as the leader intron, which they are not (Fig. 1). This could mean that interdomain linkage was not necessarily facilitated by introns, which were gained later, or that the introns have moved. If domain linkage were facilitated by introns, then which, if any, of the five linker introns preserve the position and phase of the ancestral gene? The three introns in phase 0 at the amino terminal end of the linkers L3/4, L4/5, and L9/– would be consistent with the proposed trimer model (Fig. 2). This, however, suggests that the remaining two introns either have been

acquired recently or have moved to their current locations.

The linker intron data neither reinforce nor negate the previous conclusion of possible slippage arising out of the overall stability of the domain introns (Jellie et al. 1996). Taken as a whole, however, the pattern of all 44 introns can be analyzed in terms of typical exon sizes in a way that lends greater support to independent loss and gain.

Constraints on Intron Position

It has been suggested (Maroni 1996; Robberson et al. 1990; Talerico and Berget 1994) that natural selection may act in a species specific manner to control the size of coding and noncoding elements. Exon size would be determined by the properties of the gene expression machinery. In vertebrate systems the efficiency of premRNA splicing in a multiintron gene decreases when internal exons (exons other than those at the 5′ or 3′ ends of a gene) exceed 300 bases (Robberson et al. 1990). An early stage in spliceosomal assembly is the identification of the intron 3′ splice site and its associated polypyrimidine tract. The intron 5′ splice site is also defined and both splice sites are paired. Robberson et al. (1990) suggest that in genes containing large introns and small exons, the 3′ splice site of an intron is paired to a downstream 5′ splice site across the interposing exon. This pairing of splice sites defines the borders of all internal exons and has been termed ‘‘exon definition.’’ Exon definition accounts for the phenotype of splice site mutants and provides a reason why cryptic donor and acceptor splice sites within intron sequences are not normally utilized. The pairing of splice sites across an exon thus accounts for the observed upper limit on exon size in a multiexon gene with large introns. A lower size limit to vertebrate exons of approximately 51 bases has also been observed (Dominski and Kole 1991). This is thought to be due to steric problems arising from splice sites being located too close together, interfering with the assembly of the appropriate splicing factors. A few extremely small vertebrate exons exist, but it is likely that these exons require special mechanisms of splicing for inclusion in the spliced transcript (Carlo et al. 1996).

The 300-base pair upper size limit for exons appears to be less rigid in lower eukaryotes such as *Drosophila.* The genes of lower eukaryotes often contain small introns that would be predicted to be too short to be spliced efficiently if present in the genes of vertebrates (Wieringa et al. 1984), whereas their exons often exceed 550 bases (Talerico and Berget 1994), which is predicted to be too large to function in a vertebrate. Talerico and Berget (1994) suggest that in multiintron genes with small introns and large exons, initial pairing of splice sites across intron sequences may occur, a process of ‘‘intron definition,’’ as distinct from ‘‘exon definition.’’

Sliding or Loss and Gain of Introns in Artemia *Globin Genes*

The relative sizes of introns and exons in the *Artemia* globin genes appear to be similar to the vertebrate pattern with short exons flanked by large introns. In vertebrate genes, internal introns average 1127 bp and internal exons average 137 bp (Hawkins 1988). Average figures for *Artemia* globin genes are comparable: 2000-bp introns and 200-bp exons. As these introns are much larger than exons, splice site pairing more likely occurs by the exon definition model, implying an upper size limit for internal exons.

The three discordant *Artemia* globin intradomain introns (i.e., the G or F introns in domains 3, 4, and 6) are all located in domains which have a linker intron at their 3′ end. Domain 4 is unique in being flanked by linker introns on both sides and this domain has undergone two apparent intron changes: loss of the B helix intron and a change of position of the G helix intron. Only five linker introns exist, and four of them either have apparently moved or are located close to a discordant intron.

If size limits are imposed on exons, this should provide a test of whether intron sliding or loss and gain best explains the observed pattern. If the loss and gain hypothesis is correct, then we should expect to find discordant introns within relatively intron-rich regions; otherwise the intermediate state in which a gene has lost an intron would result in an oversized exon. Such large exons are likely to be selected against, as they are more likely to result in an exon-skipping phenotype (Robberson et al. 1990). Intron gain before loss could occur only provided a lower size limit was not violated. Too short an exon is also likely to result in aberrant splicing because of steric problems arising during assembly of spliceosomal components (Black 1991). Conversely, if introns can move by some unknown sliding process, then the positions of discordant introns need not be in intron-rich regions, as no transient state with amalgamated exons would exist.

Domain 4 is the only domain missing a B helix intron, resulting in the formation of the largest internal exon, 347 bp. This is markedly larger than any of the other internal exons, the next largest being 245 bp. The upper size limit for internal exons in the genes of *Artemia* has not been established and it is not known whether the 347-bp exon results in a small degree of exon skipping or whether accessory sequences within neighboring introns or exons aid in the accurate determination of splice sites flanking this exon. A small proportion of vertebrate exons do have a size exceeding 300 bp (Hawkins 1988). Considering the current intron positions, if domain 2, 3, 6, 8, or 9 were to lose a B helix intron, this would generate exon sizes ranging from 402 bp (if domain 6 lost its B helix intron) to 479 bp (for domain 8). Assuming that the largest tolerated exon size observed

(347 bp) is close to the upper limit (possibly about 400 bases) for correct splicing to occur, then only four of the nine domains could lose a B helix intron.

The G helix intron pattern more strongly supports the existence of this upper exon size limit. Of the nine domains, only four could potentially lose a G helix intron and maintain an exon size less than 400 bp, and in three of these a G helix intron either is missing or has apparently shifted.

The presently reported discordant introns appear to be nonrandomly located, intron positional changes having occurred only where an upper exon size limit was avoided. This suggests that these discordances could well have occurred by the loss and gain mechanism, where an intermediate stage of loss would not breach an upper exon size limit. While the nonrandom distribution of discordant introns is consistent with a loss and gain mechanism, it is not inconsistent with intron sliding. It remains a possibility that introns moved by sliding and that this, in all cases, happened to occur in an intron-rich region of the gene. Given the existence of exon size limits, this points to the sequential order of some intron positional changes. For example, domain 4 has incurred two apparent changes: loss of a B helix intron and change in position of the G helix intron. Loss and gain of the G helix intron would need to happen before loss of the B helix intron to maintain exon sizes below 400 bp.

Conclusions

If the close proximity of discordant intron positions is not the result of intron sliding but due to a loss and gain mechanism, this suggests that intron insertion following intron loss tends to place an intron at or near the original intron position. What selective forces could drive this nonrandom pattern of intron insertion? The lack of a strong consensus of spliceosomal intron splice sites together with the presence of many 5′ and 3′ cryptic splice sites within an intron suggests that other sequences are used by the splicing machinery in determining the authentic splice sites. Enhancer sequences within both introns (Carlo et al. 1996) and exons (Achsel and Shimura 1996; Humphrey et al. 1995) have indeed been identified. According to the exon definition model, splice sites across exons are identified first. The implication is that other sequences are present within exons that aid in identification of exon boundaries.

Using modified antisense oligoribonucleotides, Dominski and Kole (1994) have identified both intron and exon sequences required for splicing of human β-globin pre-mRNAs. They found that at least 25 nucleotides of exon sequence at both the 3′ and the 5′ ends of an intron are required for splicing. Thus, following the loss of an intron it is likely that the presence of these exonic sequences determines to some degree where a new intron can be inserted. In their analysis of the intron–exon boundaries of numerous genes from six species, Long et al. (1998) did not find strong evidence for the existence of a general consensus sequence in exons at splice site boundaries. They concluded that the information content of flanking exonic sequence is low and does not support the introns-late argument for the presence of proto-splice sites. However, exon sequences required for splicing may not easily be detected, as they may not conform to a strong consensus or may in many situations be optional. Dependence upon these accessory sequences for efficient correct splicing may be due to the relative strength of other sequences such as intron 5′ and 3′ splice sites together with the branch point consensus and polypyrimidine tract.

Adherents to the introns-early and introns-late theories offer different explanations to account for the observed present-day gene structures of eukaryotes. In the introns-early view observed intron positions are the result of extensive intron loss and possibly intron sliding. While extensive intron loss presumably by gene conversion events involving processed pre-mRNAs may be true for the yeast, the positions of introns in the genes of many higher organisms would have required extensive intron sliding (Rzhetsky et al. 1997). The exon definition model is restricted to internal exons and therefore is not applicable to genes with one intron such as could possibly have been present in the progenote. As the gene structure of the progenote is unknown, inferences based on these models should be restricted to eukaryotic genes of defined structure. However, reevaluation of other discordant introns in the light of the limitations imposed by splice site selection may aid in resolving which scenario best accounts for both the number and the distribution of introns in homologous genes.

The positions of introns in extant genes may be not so much the product of preferential insertion but rather the result of selection against intron insertion into sites deleterious to posttranscriptional processing events. The dependence of posttranscriptional processing on both intron spatial distribution and intron internal sequence may determine to a large extent both intron numbers and the likelihood of intron retention through evolutionary time. Many pre-mRNAs require an intron for efficient posttranscriptional processing but not all introns are able to provide this function. For example, Liu and Mertz (1996) have found that while the first intron of the human β-globin gene is optional, the presence of the second intron is necessary for efficient downstream processing events. These findings reinforce the importance of considering the downstream events involving pre-mRNAs when comparing introns of homologous genes.

Acknowledgment. We are grateful for generous financial support

from the Marsden Fund, administered by the Royal Society of New Zealand.

References

Achsel T, Shimura Y (1996) Factors involved in the activation of pre-mRNA splicing from downstream splicing enhancers. J Biochem 120:53–60

Black DL (1991) Does steric interference between splice sites block the splicing of a short c-*src* neuron-specific exon in non-neuronal cells? Genes Dev 5:389–402

Blin N, Stafford DW (1976) A general method for isolation of high molecular weight DNA from eukaryotes. Nucleic Acids Res 3: 2303–2308

Carlo T, Sterner DA, Berget SM (1996) An intron splicing enhancer containing a G-rich repeat facilitates inclusion of a vertebrate micro-exon. RNA 2:342–353

Cavalier-Smith T (1991) Intron phylogeny: a new hypothesis. Trends Genet 7:145–148

Dayhoff MO, Barker WC, Hunt LT (1983) Establishing homologies in protein sequences. Methods Enzymol 91:524–545

De Souza SJ, Long M, Schoenbach L, Roy SW, Gilbert W (1996) Intron positions correlate with module boundaries in ancient proteins. Proc Natl Acad Sci USA 93:14632–14636

Dixon B, Walker B, Kimmins W, Pohajdak B (1992) A nematode hemoglobin gene contains an intron previously thought to be unique to plants. J Mol Evol 35:131–136

Dominski Z, Kole R (1991) Selection of splice sites in pre-mRNAs with short internal exons. Mol Cell Biol 11:6075–6083

Dominski Z, Kole R (1994) Identification and characterization by antisense oligonucleotides of exon and intron sequences required for splicing. Mol Cell Biol 14:7445–7454

Doolittle RF (1985) The genealogy of some recently evolved vertebrate proteins. Trends Biochem Sci 10:233–237

Doolittle WF (1978) Genes in pieces: were they ever together? Nature 272:581–582

Dorit R, Schoenbach L, Gilbert W (1990) How big is the universe of exons? Science 250:1377–1382

Gilbert W (1978) Why genes in pieces? Nature 271:501

Gilbert W (1987) The exon theory of genes. Cold Spring Harbor Symp Quant Biol 52:901–905

Gilbert W, de Souza SJ, Long M (1997) Origin of genes. Proc Natl Acad Sci USA 94:7698–7703

Go M (1981) Correlation of DNA exonic regions with protein structural units in haemoglobin. Nature 293:90–92

Hawkins JD (1988) A survey on intron and exon lengths. Nucleic Acids Res 16:9893–9908

Humphrey MB, Bryan J, Cooper TA, Berget SM (1995) A 32-nucleotide exon-splicing enhancer regulates usage of competing 5′ splice sites in a differential internal exon. Mol Cell Biol 15:3979–3988

Hurst LD, McVean GT (1996) Molecular evolution: A difficult phase for introns-early. Curr Biol 6:533–536

Jellie AM, Tate WP, Trotman CNA (1996) Evolutionary history of introns in a multidomain globin gene. J Mol Evol 42:641–647

Kaiser K, Murray NE, Whittaker PA (1995) Construction of representative genomic DNA libraries using phage lambda replacement vectors. In: Glover DM, Hames BD (eds) DNA cloning 1: A practical approach. IRL Press, Oxford, New York, pp 37–84

Liu X, Mertz JE (1996) Sequence of the polypyrimidine tract of the 3′-terminal 3′ splicing signal can affect intron-dependent pre-mRNA processing *in vivo*. Nucleic Acids Res 24:1765–1773

Long M, de Souza SJ, Gilbert W (1995a) Evolution of the intron-exon structure of eukaryotic genes. Curr Opin Genet Dev 5:774–778

Long M, Rosenberg C, Gilbert W (1995b) Intron phase correlations and the evolution of the intron/exon structure of genes. Proc Natl Acad Sci USA 92:12495–12499

Long M, de Souza SJ, Rosenburg C, Gilbert W (1996) Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c_1 precursors. Proc Natl Acad Sci USA 93:7727–7731

Long M, de Souza SJ, Rosenburg C, Gilbert W (1998) Relationship between ''proto-splice sites'' and intron phases: Evidence from dicodon analysis. Proc Natl Acad Sci USA 95:219–223

Maroni G (1996) The organization of eukaryotic genes. In: Hecht MK, MacIntyre RJ, Clegg MT (eds) Evolutionary biology, Vol 2. Plenum Press, New York, pp 1–19

Matthews C, Vandenberg CJ, Trotman CNA (1988) Variable substitution rates of the eighteen domain sequences in *Artemia* hemoglobin. J Mol Evol 46:729–733

Moens L, Van Hauwaert M-L, De Smett K, Geelen D, Verpooten G, Van Beeuman J, Wodak S, Alard P, Trotman C (1988) A structural domain of the covalent polymer globin chains of *Artemia*. J Biol Chem 263:4679–4685

Moens L, Van Hauwaert M-L, De Smett K, Ver Donck K, Van De Peer Y, Van Beeuman J, Wodak S, Alard P, Trotman C (1990) Structural interpretation of the amino acid sequence of a second domain from the *Artemia* covalent polymer globin. J Biol Chem 265:14285–14291

Naito Y, Riggs CK, Vandergon TL, Riggs AF (1991) Origin of a ''bridge'' intron in the gene for a two-domain globin. Proc Natl Acad Sci USA 88:6672–6676

Patthy L (1985) Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. Cell 41:657–663

Robberson BL, Cote GJ, Berget SM (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol 10:84–94

Rogers JH (1990) The role of introns in evolution. FEBS Lett 268: 339–343

Rzhetsky A, Ayala FJ, Hsu LC, Chang C, Yoshida A (1997) Exon/intron structure of aldehyde dehydrogenase genes support the ''intron-late'' theory. Proc Natl Acad Sci USA 94:6820–6825

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: A laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Sherman DR, Kloek AP, Krishnan BR, Guinn B (1992) *Ascaris* hemoglobin gene: Plant-like structure reflects the ancestral globin gene. Proc Natl Acad Sci USA 89:11696–11700

Stockwell PA (1987) DNA sequences analysis software. In: Rawlings CT, Bishop MJ (eds) Nucleic acid and protein sequence analysis. a practical approach. IRL Press, Oxford, Washington DC, pp 19–45

Stockwell PA (1988) HOMED: a homologous sequence editor. Trends Biochem Sci 13:322–324

Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolittle WF (1994) Testing the exon theory of genes: The evidence from protein structure. Science 265:202–207

Stoltzfus A, Logsdon JM Jr, Palmer JD, Doolittle WF (1997) Intron ''sliding'' and the diversity of intron positions. Proc Natl Acad Sci USA 94:10739–10744

Talerico M, Berget SM (1994) Intron definition in splicing of small *Drosophila* introns. Mol Cell Biol 14:3434–3445

Trotman C, Manning AM, Bray JA, Jellie AM, Moens L, Tate WP (1994) Interdomain linkage in the polymeric hemoglobin molecule of *Artemia*. J Mol Evol 38:628–636

Wieringa B, Hofer E, Weissmann C (1984) A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. Cell 37:915–925