Signatures of Domain Shuffling in the Human Genome

Henrik Kaessmann,[1,3] Sebastian Zöllner,[2] Anton Nekrutenko,[1] and Wen-Hsiung Li[1]

[1]*Department of Ecology and Evolution, and* [2]*Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA*

To elucidate the role of exon shuffling in shaping the complexity of the human genome/proteome, we have systematically analyzed intron phase distributions in the coding sequence of human protein domains. We found that introns at the boundaries of domains show high excess of symmetrical phase combinations (i.e., 0–0, 1–1, and 2–2), whereas nonboundary introns show no excess symmetry. This suggests that exon shuffling has primarily involved rearrangement of structural and functional domains as a whole. Furthermore, we found that domains flanked by phase 1 introns have dramatically expanded in the human genome due to domain shuffling and that 1–1 symmetrical domains and domain families are nonrandomly distributed with respect to their age. The predominance and extracellular location of 1–1 symmetrical domains among domains specific to metazoans suggests that they are associated with the rise of multicellularity. On the other hand, 0–0 symmetrical domains tend to be over-represented among ancient protein domains that are shared between the eukaryotic and prokaryotic kingdoms, which is compatible with the suggestion of primordial domain shuffling in the progenote. To see whether the human data reflect general genomic patterns of metazoans, similar analyses were done for the nematode *Caenorhabditis elegans*. Although the *C. elegans* data generally concur with the human patterns, we identified fewer intron-bounded domains in this organism, consistent with the lower complexity of *C. elegans* genes.

[The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: Z. Gu and R. Stevens.]

Structural domains or modules are discrete structural folding units of proteins (Go and Nosaka 1987). Importantly, structural domains also often represent the fundamental functional units of proteins (Copley et al. 2002). It has been suggested that much of the process of gene/protein evolution should be analyzed in terms of domains rather than entire proteins, because domains represent primary protein building blocks that recombine to form various combinations (Koonin et al. 2000).

Interestingly, the diversity of domain combinations seems to increase with the organism's complexity (Rubin et al. 2000). For example, the human genome appears to have evolved more complex domain organizations compared with other eukaryotic organisms for which complete genomes are available (Lander et al. 2001; Li et al. 2001; Venter et al. 2001). This observation may help to explain the relatively modest difference in gene number between humans and, for example, fruitfly and worm, because combinatorial diversity of protein domains can provide a strong increase in the ability to mediate protein–protein interactions without dramatically increasing the absolute size of the protein complement (Pawson and Nash 2000). Thus, domain shuffling, which refers to the duplication of a domain or the insertion of a domain from one gene into another (Graur and Li 1999), may have been a major factor in the evolution of human phenotypic complexity.

In general, domain shuffling is often mediated by intronic recombination of exons encoding the domain (Patthy 1999a). In addition, transduction of genome sequences mediated by retrotransposons likely represents a frequent mechanism to shuffle exons (Moran et al. 1999; Goodier et al. 2000; Kazazian Jr. 2000; Pickeral et al. 2000). Phase combinations of flanking introns are useful indicators of exon shuffling. Intron phase is a parameter that determines the intron position relative to the translational reading frame. Introns that interrupt the reading frame between codons are known as phase 0 introns; those that split codons between the first and second nucleotides are known as phase 1 introns; and those that split codons between the second and the third nucleotides are known as phase 2 introns. Successful shuffling requires that the domain in question is bordered by introns that are of the same phase, that is, that the domain is symmetrical in accordance with the phase-compatibility rules of exon shuffling (Patthy 1999b), because shuffling of asymmetrical exons/domains will result in a shift of the reading frame in the downstream exons of recipient genes. The abundance of genomic data from humans now permits a comprehensive study of protein domains and their coding structure.

In this study, we sought to systematically identify domains in the human genome that are bounded by introns. Analysis of intron-phase distributions of these domains provides insights into the extent, pattern, and timing of domain-shuffling events that have left traces in the human genome. A comparative analysis with *C. elegans* provides an evolutionary perspective to the human data.

RESULTS

Domain Classification

There are several possible relationships between domains and exons (Graur and Li 1999). (1) The domain is encoded by one

[3]Corresponding author.
E-MAIL kaessm@uchicago.edu; FAX (773) 702-9740.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.520702.

12:1642–1650 ©2002 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/02 $5.00; www.genome.org

exon (a class 1 domain); (2) the domain is encoded by several exons (a class 2 domain); (3) one exon encodes several domains; (4) there is no exact correlation between domain and exon, for example, the domain is encoded by partial exons.

In this study, we focused on the first two domain classes (Fig. 1), because phase combinations of flanking introns (boundary introns) may provide evidence for domain shuffling. We annotated protein domains according to the Pfam classification (Bateman et al. 2002), which is currently the most comprehensive collection of protein domain signatures (Copley et al. 2002). It is curated manually and domain boundaries are matched to available structural data, ensuring high quality and reliable boundary estimation (Elofsson and Sonnhammer 1999; Bateman et al. 2002).

When screening the genome for class 1 and class 2 domains, we allowed domain boundaries to be flexible, as these are unlikely to be predicted exactly (Copley et al. 2002). However, we required that the sum of the differences between the two boundary intron positions and domain boundaries is <10% of the domain length (see Methods). According to these criteria, we identified 345 class 1 and 924 class 2 domains in 14,728 human genes (Table 1). Thus, 1269 domains (representing 226 different domain families, i.e., types of domains with discrete structural folding units) of the 20,778 Pfam domains (6%) are bounded by introns, and these domains may reveal signatures of domain shuffling.

Intron Phase Combinations

Previous studies have shown that there is a general excess of symmetrical exons (i.e., they have introns of the same phase at both ends) over the random expectation in eukaryotic genomes, including the human one (Long et al. 1995a; Tomita et al. 1996; Fedorov et al. 1998; Sakharkar et al. 2002). This phenomenon has largely been attributed to exon shuffling. However, if exon shuffling is equivalent to domain shuffling, there are two expected consequences for the intron phase distributions in the genome. First, introns at the boundaries of domains (boundary introns, Fig. 1B) are expected to show excess symmetrical phase correlations due to shuffling constraints. Second, introns that do not coincide with domain boundaries (nonboundary introns, Fig. 1B) are expected to show no excess symmetry.

To test that nonboundary introns show no excess symmetry, we used the class 2 domains, which are encoded by several exons (Fig. 1B). We first compared the observed and expected frequencies of symmetrical exons within these domains, whose flanking introns do not coincide with domain boundaries (nonboundary exons) (Fig. 1B). The expected frequency was estimated from the product of the observed frequencies of intron phases 5′ and 3′ of the exons, respectively (see Methods). The results show that there is no significant difference between observed (758) and expected (770.3) numbers of symmetrical nonboundary exons ($P = 0.57$) (Table 2). When testing the three symmetrical combinations (0–0, 1–1, and 2–2) individually, we obtained similar results. Note that the inclusion of class 4 domains (no exact correlation between domain and exons) in the analysis also revealed no significant difference between observed (2554) and expected (2533.3) values of symmetrical nonboundary exons ($P = 0.6$). Thus, nonboundary exons show no excess symmetry.

After elucidating phase correlations of nonboundary introns, we assessed symmetry levels of introns that border domains. Excess symmetry of these boundary introns (Fig. 1B) may provide evidence for domain shuffling. Analysis of boundary intron phases of class 1 domains (domains encoded by single exons; Fig. 1A) reveals that 83% (288/345) of the domains are symmetrical (Table 1). Among the symmetrical domains, 1–1 domains are particularly abundant and constitute about 74% of all class 1 domains. Statistical analysis of intron phase correlations shows that there is a highly significant excess of symmetrical domains ($P < 10^{-7}$; Table 3). Among the three symmetrical combinations, 0–0 and 1–1 domains show significant excess when tested separately; 2–2 domains were not tested, as their frequency is low.

These figures might be driven by symmetrical domains from certain domain families that have successfully expanded by domain shuffling and/or gene duplication. To exclude this possibility, we performed the same analyses, but merged the data into domain families to see whether the excess of symmetry reflects a general trend among class 1 domains. Domains from the same family may be bordered by introns of different phases due to a combination of intron insertion, intron loss, and/or intron sliding events that have occurred since their common origin. Thus, we weighted intron phase combinations of domains belonging to a certain family according to their frequencies. To quantify intron phase combinations of a domain family, we computed the frequency of

Table 1. Observed Numbers of Boundary Intron Phase Combinations of Class 1 Domains (Encoded by One Exon) and Class 2 Domains (Encoded by Several Exons) in Human Genes

	Domain class 1		Domain class 2	
Comb.[a]	Domain	Domain family	Domain	Domain family
0–0	29	8.5	176	51
1–1	257	24.3	298	32.6
2–2	2	0.2	30	8.7
0–1	11	6.9	82	19.1
0–2	9	4.4	74	23.7
1–2	8	0.6	50	13.7
1–0	6	3.1	103	21.9
2–0	2	3	63	17.6
2–1	21	4.9	48	7.6
Total	345	56	924	196

[a]Combination of flanking intron phases.

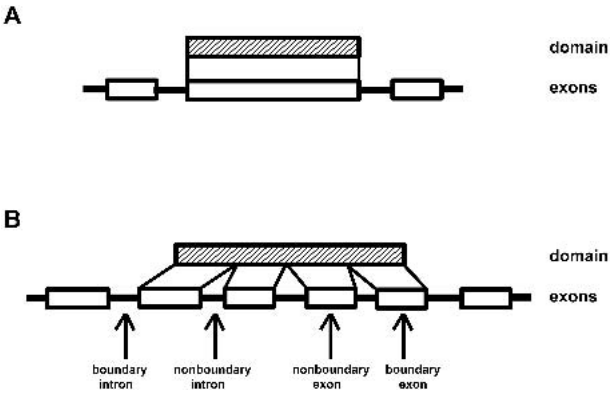

Figure 1 Illustration of a class 1 domain (*A*) and a class 2 domain (*B*) encoded by four exons. The different types of introns and exons in the coding sequence of class 2 domains (as used in the analysis of intron phases) are indicated (see text for details).

Table 2. Observed and Expected Numbers of Symmetrical Nonboundary Exons or Sets of Nonboundary Exons[a]

No. of exons	Comb.[b]	Obs.[c]	Exp.[d]	*P*
1 exon (2007)	0–0	508	533	0.21
	1–1	154	139.7	0.21
	2–2	96	97.6	0.87
	Σ	**758**	**770.3**	**0.57**
2 exons (842)	0–0	243	252.9	0.46
	1–1	59	55.2	0.6
	2–2	27	32.3	0.34
	Σ	**329**	**340.4**	**0.42**
3 exons (454)	0–0	121	130	0.35
	1–1	30	32.2	0.69
	2–2	16	17.1	0.79
	Σ	**167**	**179.3**	**0.24**
4 exons (301)	0–0	95	92.4	0.75
	1–1	26	20.5	0.21
	2–2	5	10	0.11
	Σ	**126**	**122.9**	**0.72**
5 exons (187)	0–0	52	53.9	0.76
	1–1	22	17	0.2
	2–2	4	4.8	n.d.
	Σ	**78**	**75.7**	**0.73**
6 exons (118)	0–0	37	36.8	0.97
	1–1	14	11.6	0.46
	2–2	1	1.9	n.d.
	Σ	**52**	**50.3**	**0.75**

[a]The total number of cases in each group are indicated in parentheses. The expected values were calculated based on the phase frequencies of flanking introns (see Methods). Summed symmetrical combinations are shown in boldface. Some *P*-values were not determined (n.d.) due to the low sample size.
[b]Combination of flanking intron phases.
[c]Observed.
[d]Expected.

each phase class among the different intron phase combinations carried by the members of this family, with the total count of each family equal to one. For example, a domain family with four members, of which two carry introns with the combination 1–1 and the other two carry a 2–1 combination, is counted to have 0.5 class 1–1 and 0.5 class 2–1.

The analysis reveals that class 1 domain families show a significant ($P < 0.05$) excess of symmetry (Table 3). The *P*-value for an excess of 1–1 domain families is suggestive ($P < 0.07$), whereas 0–0 domain families may show nonsignificant excess due to a small sample size. In general, the symmetry signals (as determined by the *P*-values) in this family-based analysis are weaker than those determined without merging the data into families (Table 3). This difference is likely due to lower sample sizes in the family-based analysis but also to a few symmetrical domain families that have particularly expanded (Table 4). In summary, however, boundary introns of class 1 domains and families show signals of symmetry, and therefore suggest domain shuffling, whereas nonboundary exons in the human genome show levels of symmetry according to the random expectation.

Note that the lack of excess symmetry of nonboundary exons is in contrast to the entire set of exons in the human genome. Analysis of intron phase combinations of 92,052 exons from 14,728 genes (the genomic background) used in this study reveals a highly significant excess of symmetrical exons (41% observed vs. 37% expected; $P < 10^{-11}$), consistent with previous studies (Long et al. 1995b; Tomita et al. 1996; Fedorov et al. 1998). To exclude that this discrepancy is due to sample size differences, we repeatedly (10,000 times) and randomly sampled the same number of exons as found within domains (2007 nonboundary exons) from the genomic background. More than 99% (9939/10,000) of these draws show significant ($P < 0.05$) excess of symmetrical exons. Thus, contrary to nonboundary exons, random exons in the genome show excess symmetry. Together with the observation that class 1 domains tend to be highly symmetrical, this suggests that the excess symmetry in the genome is, in fact, due to exon shuffling, but generally involves exons encoding complete protein domains.

To see whether domains of class 2 also reflect evolution by domain shuffling, we subjected them to similar analyses as described above for single exon domains. Among the 924 domains of this type, we identify 504 (54%) symmetrical domains (Table 1). Again, we find that 1–1 domains are most abundant (~32% of all domains), whereas 0–0 domains (~19%), and especially 2–2 domains (~3%), are less common.

To compare observed and expected levels of symmetry of class 2 domains, we split these domains into groups defined by the number of exons encoding them. We considered five groups of domains encoded by two to six exons, respectively, because each of these groups contains a sufficient number of domains for statistical analysis. The random expectation of symmetrical class 2 domains was calculated according to the frequencies of their boundary intron phases (see Methods). The results reveal significant excess of symmetry for domains encoded by two to four exons, and domain families show significant excess when encoded by two exons (Table 5). The *P*-value of three exon domain families ($P = 0.12$) is suggestive in view of the low sample size. The excess of symmetry is unevenly distributed among the three symmetrical classes (Table 5). The 1–1 domains show significant excess for domains encoded by two to four exons and for families encoded by two exons. The 0–0 domains show significant excess when encoded by two exons, whereas all other comparisons show no significant excess.

To confirm that sets of exons within domains (similarly to single nonboundary exons—see above) show no excess symmetry, we assessed symmetry levels of nonoverlapping sets of two to six exons that are located inside class 2 domains (i.e., introns flanking these sets of exons do not coincide with

Table 3. Observed and Expected Numbers of Symmetrical Class 1 Domains and Domain Families in Humans[a]

	Comb.[b]	Obs.[c]	Exp.[d]	*P*
Domains (345)	0–0	29	5.3	$<10^{-10}$
	1–1	257	226.2	$<10^{-3}$
	2–2	2	1.4	n.d.
	Σ	**288**	**233**	**$<10^{-7}$**
Domain family (56)	0–0	8.5	5.2	0.13
	1–1	24.3	17.9	0.07
	2–2	0.2	0.8	n.d.
	Σ	**33**	**23.8**	**<0.05**

[a]The total number of class 1 domains and families are indicated in parentheses. Summed symmetrical combinations are shown in boldface. The expected values were calculated based on the frequencies of flanking intron phases. Some *P*-values were not determined (n.d.) due to the low sample size.
[b]combination of flanking intron phases.
[c]Observed.
[d]Expected.

Table 4. Common Symmetrical Domain Families in the Human Genome[a]

Domain family	Pfam id[b]	No.[c]	Comb.[d]	Frequ.[e]
Sushi domain/SCR repeat/CCP module	PF00084	144	1–1	0.98
CUB domain	PF00431	65	1–1	0.95
Laminin-type EGF-like (LE) domain	PF00053	34	1–1	0.82
von Willebrand factor type A domain	PF00092	30	1–1	0.83
Myosin head (motor domain)	PF00063	28	0–0	0.79
EGF-like domain	PF00008	22	1–1	0.82
Discoidin domain/Coagulation factor 5/8 type C domain (FA58C)	PF00754	18	1–1	1.00
Kringle	PF00051	18	1–1	1.00
Crystallin	PF00030	16	0–0	1.00
von Willebrand factor, type C repeat	PF00093	16	1–1	0.94
Major histocompatibility complex protein, Class I	PF00129	16	1–1	0.81
PAN domain	PF00024	13	1–1	0.85
Low density lipoprotein-receptor class A (LDLRA) domain	PF00057	13	1–1	0.92
MHC Class II, alpha chain, alpha-1 domain	PF00993	12	1–1	1.00
MAM domain	PF00629	11	1–1	0.91
IPT/TIG domain	PF01833	9	1–1	1.00
PHD-finger	PF00628	8	0–0	0.75
SEA domain	PF01390	8	1–1	1.00
Link domain	PF00193	8	1–1	1.00
Thrombospondin type I domain	PF00090	7	1–1	0.71
Pancreatic trypsin inhibitor (Kunitz)	PF00014	7	1–1	1.00
Band 4.1 family	PF00373	7	0–0	0.71
GPS domain	PF01825	5	0–0	1.00
TRAF-type zinc finger	PF02176	5	0–0	0.80
Thyroglobulin type-1 repeat	PF00086	5	1–1	1.00
PKD domain	PF00801	4	1–1	1.00
WAP-type (Whey Acidic Protein) four-disulfide core domain	PF00095	4	1–1	1.00
Triple function domain (TRIO)	PF00650	4	0–0	1.00
PX (Bem1/NCF1/PI3K) domain	PF00787	4	0–0	1.00
HYR domain	PF02494	4	1–1	1.00

[a]Class 1 and class 2 domains are combined in this overview. Only domain families that occur more than four times in the dataset and for which the symmetrical intron phase combination has a frequency >.7 are shown.
[b]Pfam database identification (http://www.sanger.ac.uk/software/Pfam/index.shtml).
[c]Total number of occurrences.
[d]Flanking intron phase combination.
[e]Frequency of intron phase combination among members of respective domain family.

domain boundaries). None of these sets shows significant excess of symmetry (Table 2). All of the observed values are very close to the random expectation. This corroborates that only introns that border domains show excess symmetry as a consequence of domain shuffling. Thus, domains rather than exons appear to be the primary shuffling units in the genome. Note that sample sizes are larger for all symmetry tests involving nonboundary introns than for tests of boundary introns (Tables 2, 3, and 5). Thus, differences in sample size are unlikely to account for the lack of statistically significant phase correlations of nonboundary introns. Altogether, our results reveal striking signatures of domain shuffling even among domains that are encoded by several exons.

Age Distribution of Shuffling Events

To put the signals of domain shuffling detected in the human genome into an evolutionary perspective, we classified the domains into two major age groups according to their phylogenetic distribution. The first group (old domains) consists of domain families that are shared between eukaryotes and prokaryotes, whereas the second group (new domains) contains domain families that are specific to multicellular animals (metazoans). The remaining domains that occur in humans, other multicellular eukaryotes, as well as unicellular eukaryotes were not considered separately in the following analyses.

Among the 1269 class 1 and class 2 domains, there are 222 new domains belonging to 36 families (Table 6). It has been suggested that 1–1 domains have expanded in metazoans and may therefore be associated with the origin of animal multicellularity (Patthy 1999b). To see whether domains identified in this study reflect such a pattern, we tested whether 1–1 domains are nonrandomly distributed with respect to their age. On the basis of the overall frequency of 1–1 domains and families among the 1269 domains, we calculated the frequency of new 1–1 domains that is expected to occur by chance among the new domains (Table 6). Intriguingly, there is a significant excess of new 1–1 domains ($P < 10^{-3}$) and domain families ($P < 10^{-3}$) over the random expectation, which lends statistical support to the notion that an expansion of 1–1 domains accompanied the rise of multicellularity in animals (Patthy 1999b). New 0–0 domains, on the other hand, are significantly under-represented. The dramatic expansion of 1–1 domains specific to metazoans is illustrated by the average number of new 1–1 domains per domain family, which is about nine, and thus, more than three times larger than that of new 0–0 domains (~3 domains/family). Notably, all new 1–1 domains identified in this study are extracellular and involved in various processes that are characteristic of multicellular animals, consistent with previous results (Patthy 1999b). For example, the MAM domain occurs in the extracellular region of functionally diverse pro-

Table 5. Observed and Expected Numbers of Symmetrical Class 2 Domains in the Human Genome[a]

Category	Comb.	Obs.	Exp.	*P*
2 exon domains				
Domain (367)	0–0	31	16.1	<0.02
	1–1	194	156.9	$<10^{-4}$
	2–2	10	7.4	0.33
	Σ	**235**	**180.4**	**$<10^{-7}$**
Domain family (74)	0–0	10.6	8.6	0.47
	1–1	21.5	14.6	<0.05
	2–2	2.8	3.4	n.d.
	Σ	**34.9**	**26.6**	**<0.05**
3 exon domains				
Domain (155)	0–0	23	18.1	0.22
	1–1	50	38.5	<0.05
	2–2	6	3.8	n.d.
	Σ	**79**	**60.4**	**<0.01**
Domain family (63)	0–0	12.7	11.7	0.75
	1–1	12.3	8.3	0.14
	2–2	3.5	2.6	n.d.
	Σ	**28.5**	**22.5**	**0.12**
4 exon domains				
Domain (115)	0–0	24	18.9	0.2
	1–1	33	20.9	$<10^{-3}$
	2–2	3	3	n.d.
	Σ	**60**	**42.8**	**$<10^{-3}$**
Domain family (49)	0–0	10.1	9.6	0.86
	1–1	9.5	5.9	0.11
	2–2	1	0.6	n.d.
	Σ	**20.6**	**16.1**	**0.17**
5 exon domains				
Domain (70)	0–0	25	22.8	0.57
	1–1	5	2.7	n.d.
	2–2	0	1.9	n.d.
	Σ	**30**	**27.4**	**0.52**
Domain family (39)	0–0	15	12.4	0.37
	1–1	3.5	1.9	n.d.
	2–2	0	1	n.d.
	Σ	**18.5**	**14.3**	**0.16**
6 exon domains				
Domain (75)	0–0	27	22.8	0.29
	1–1	6	4.3	n.d.
	2–2	3	3.1	n.d.
	Σ	**36**	**30.2**	**0.17**
Domain family (43)	0–0	16.7	14.6	0.5
	1–1	2.7	2	n.d.
	2–2	2.3	1.6	n.d.
	Σ	**21.7**	**18.2**	**0.28**

[a]Total numbers of domains and families in each category are indicated in parentheses. Summed symmetrical combinations are shown in boldface. The expected values were calculated based on the frequencies of flanking intron phases. (Comb.) Combinations; (Obs.) observed; (Exp.) expected.

teins such as meprin (a cell surface glycoprotein), in which it has an adhesive function (Beckmann and Bork 1993).

Analysis of all exons in the 14,728 genes from this study, as well as previous data from multiple species (Long et al. 1995b), reveals the genome-wide dominance of 1–1 exon shuffling. Although 0–0, 1–1, and 2–2 symmetrical exons are all in excess over the random expectation (9%, 23%, and 8% excess, respectively), 1–1 exons show higher excess relative to the other two symmetrical classes. Thus, the extensive expansion of 1–1 exons is likely not restricted to known domains that are bordered by introns as identified in this study, but probably involves also presently uncharacterized domains, or domains for which the ancestral relationship of introns and domain boundaries has been lost.

The introns-early theory suggests that phase 0 introns are associated with domain boundaries in ancient proteins, because these introns mediated the assembly of primordial protein domains by exon shuffling (Doolittle 1978; Gilbert 1987). To test this hypothesis, we compared observed and expected values of old 0–0 domains. We find an excess of 0–0 domains and families among the 401 domains and 95 families that are old, whereas 1–1 domains are significantly underrepresented in this age category (the *P*-value for 1–1 domain families is suggestive, $P = 0.1$; Table 6). However, the excess of old 0–0 domains is not significant. This may be due to more recent intron gain/loss and/or other evolutionary events that have complicated the picture, provided that the pattern is a consequence of ancient exon shuffling.

Domain Shuffling in *C. elegans*

The *C. elegans* genome is more than one order of magnitude smaller than the human genome and has a lower intron density (Deutsch and Long 1999). To see whether signatures of domain shuffling similar to those found in the human genome are detectable in the *C. elegans* genome, we analyzed intron phase combinations of *C. elegans* domains.

The total number of domains in *C. elegans* in which introns coincide with domain boundaries is comparatively low (286 domains representing 110 domain families). The lower number of domains is due to a smaller proportion of genes with known domains—7754/14,728 (~53%) in humans and 4271/19,682 (~29%) in *C. elegans*—and an on average lower number of domains per gene (~2.7 vs. ~1.8). Furthermore, the proportion of domains with flanking introns among all domains is lower than that in humans (6.1% vs. 3.7%), which is likely to be due to the overall lower intron density in the *C. elegans* genome in general (Deutsch and Long 1999), and modular genes in particular (Patthy 1999a).

The following patterns emerge from the data. First, there is a significant excess of symmetrical domains (Table 7) (class 2 domains encoded by more than two exons were not analyzed separately, as their frequency is low). Furthermore, class 2 domain families show a significant excess of symmetry, and the *P*-value for class 1 domain families is suggestive ($P = 0.06$). Among the symmetrical classes, 1–1 domains and families reveal significant excess for class 1 domains, whereas there is no significant excess of class 2 domains bordered by phase 1 introns. On the other hand, the excess of 0–0 domains and families tends to be larger for class 2 than for class 1 domains. Second, 1–1 domains and families are in excess among domains specific to metazoans (0–0 domains are significantly underrepresented here) (Table 8); all of these new domains are extracellular or suggested to be extracellular (data not shown). Third, 0–0 domains show slight excess among old domains that are shared between eukaryotes and prokaryotes, whereas 1–1 domains are under-represented in this category (Table 8). In summary, the *C. elegans* genome reveals signatures of domain shuffling similar to those from the human genome.

Examples of Symmetrical Domain Families

Some domain families are particularly illustrative with regard to the association of domain shuffling and boundary intron phases. The discoidin domain has been described previously from a variety of different genes (Patthy 1999b). It has weak homologs in bacteria and is predominant in multicellular animals, in which it has substantially expanded. Our data reveal that it occurs 18 times in 16 different human genes representing 7 different gene families, and that it is encoded by two to

Table 6. Observed and Expected Numbers of Old and New Domains and Families in Humans That Are Bounded Symmetrically by Phase 0, Phase 1, or Phase 2 Introns[a]

	New						Old					
	Domain (222)			Family (36)			Domain (401)			Family (95)		
	Obs.	Exp.	*P*	Obs.	Exp.	*P*	Obs.	Exp.	*P*	Obs.	Exp.	*P*
0–0	13	32.6	$<10^{-3}$	4.6	8.9	0.1	80	71	0.19	27.5	23.4	0.33
1–1	138	109.5	$<10^{-3}$	15.4	7	$<10^{-3}$	125	148.7	<0.02	11.5	18.5	0.07
2–2	3	4.8	n.d.	1.2	1.4	n.d.	13	11.8	0.72	3.9	3.7	n.d.

[a]Total numbers in each category are indicated in parentheses. The expected numbers are based on the fraction of 0–0, 1–1, and 2–2 domains in the entire data set (1269 domains). Some *P*-values were not determined (n.d.) due to the low sample size. (Obs.) Observed; (Exp.) expected.

four exons. The domain has spread to 13 different chromosomes. Despite the fact that this domain shows five different intron phase combinations and occurs together with other domains in modular proteins, it is always bounded symmetrically by introns of phase 1 (Fig. 2; Table 4). This indicates that although this domain accumulated internal introns of different phases over time, the conserved phase symmetry of its boundary introns strikingly reflects its evolution by domain shuffling. Interestingly, the combinations of intron phases in general correspond to the different types of gene families carrying the domain. This is in line with recent data, which show that the exon–intron structure of protein domains may be a good measure of evolutionary relatedness of genes (Betts et al. 2001).

We also identify novel 1–1 modules that have hitherto not been described. For example, the PKD module is an extracellular domain that has an immunoglobulin-like fold and is involved in protein–protein interactions (Gluecksmann-Kuis et al. 1995). It occurs as a 1–1 domain (in four of four cases; Table 4) in three genes from two different gene families, and is encoded by two exons. The domain was first identified in the human polycystic kidney disease protein, in which it occurs together with other 1–1 modules such as the LDL domain. In contrast to the majority of extracellular 1–1 modules that have expanded in the metazoan lineage, this domain also occurs in prokaryotes (mainly in Archaea). Like most other modules, it is also found tandemly repeated in a gene, which may be a prerequisite for successful expansion of domains by domain shuffling (Patthy 1999b).

The TIG domain is an atypical 1–1 module, because it is found in extracellular cell-surface receptors, as well as intracellularly, in transcription factors, and because it occurs in a modular fashion in both eukaryotes and a variety of bacteria (http://www.sanger.ac.uk/software/Pfam/index.shtml). The TIG domains identified in this study are encoded by 2–3 exons, occur in three different gene families, and show the typical 1–1 pattern for the boundary introns in eight of eight cases (Table 4).

The PAN domain is structurally homologous to domains from the plasminogen protein family, apple domains, and various nematode domains, and it mediates protein–protein or protein–carbohydrate interactions. Our data confirms recent predictions (on the basis of the distant homology to the apple domain; Tordai et al. 1999) that the PAN domain, which is found in diverse proteins of eukaryotes, is a typical 1–1 module. Interestingly, it consistently shows this 1–1 pattern in both human (11/13 times; Table 4) and *C. elegans* genes (4/4 times). The seven human genes (from two gene families) carrying the PAN domain show an intriguing intron phase distribution. All of the 13 copies of this type of domain carry an internal phase 2 intron. Most of the domains—including those that occur tandemly repeated—carry these introns at approximately the same position (data not shown). This indicates that a phase 2 intron was present in the PAN domain before it was duplicated and/or shuffled and illustrates that domains encoded by several exons can expand by multiple-exon shuffling.

DISCUSSION

Phase Symmetries and Domain Shuffling

Comparative analyses of complete genomes/proteomes from humans and other eukaryotes have suggested that the complexity of the human genome may be largely due to elaborate domain architectures encoded by human genes. In this study, we have systematically screened for traces of shuffling events in the human genome sequence that have created this combinatorial diversity. We provide a comprehensive survey of protein domains that are bordered by introns.

Table 7. Observed and Expected Numbers of 0–0 and 1–1 Symmetrical Class 1 and Class 2 Domains in *C. elegans*[a]

Category	Comb.	Obs.	Exp.	*P*
Class 1 domains				
Domain (83)	0–0	13	8.7	0.13
	1–1	36	25.5	<0.05
	2–2	2	1.2	n.d.
	Σ	**51**	**35.4**	**$<10^{-3}$**
Domain family (41)	0–0	9.9	8.5	0.59
	1–1	10.7	6.1	<0.05
	2–2	0.8	0.9	n.d.
	Σ	**21.4**	**15.5**	**0.06**
Class 2 domains				
Domain (81)	0–0	13	7.1	<0.05
	1–1	21	17.5	<0.34
	2–2	5	4.3	n.d.
	Σ	**39**	**28.9**	**<0.05**
Domain family (47)	0–0	8.3	4.7	0.11
	1–1	10.7	8.4	0.38
	2–2	4.3	3.2	n.d.
	Σ	**23.3**	**16.3**	**<0.05**

[a]Total numbers of domains in each category are indicated in parentheses. Summed combinations of symmetrical domains are shown in boldface. Note that only class 2 domains encoded by 2 exons are considered here (see text for details). (Comb.) Combinations; (Obs.) observed; (Exp.) expected.

Table 8. Observed and Expected Numbers of Old and New Domains and Families in *C. elegans* Genes That Are Bounded Symmetrically by Phase 0, Phase 1, or Phase 2 Introns[a]

	New						Old					
	Domain (40)			Family (19)			Domain (134)			Family (52)		
	Obs.	Exp.	*P*	Obs.	Exp.	*P*	Obs.	Exp.	*P*	Obs.	Exp.	*P*
0–0	1	7	0.01	1	3.9	n.d.	29	24.7	0.34	12.1	10.7	0.63
1–1	23	11.4	$<10^{-4}$	6.1	2.6	n.d.	8	27.3	$<10^{-4}$	3.4	7.2	0.13
2–2	0	2.1	n.d.	0	1.3	n.d.	7	7	1	3.3	3.5	n.d.

[a]Total numbers of domains in each category are indicated in parentheses. Some *P*-values were not determined (n.d.) due to the low sample size. (Obs.) Observed; (Exp.) Expected.

Comparative statistical analyses of the phases of these boundary introns and of nonboundary introns located within the domains reveal clear signatures of domain shuffling. Whereas phases of nonboundary introns are not correlated, boundary introns show excess phase symmetries. This pattern is observed for domains encoded by as many as four exons. As a corollary, although some domains may have evolved from smaller units such as repeats (Andrade et al. 2001), this finding provides strong evidence that exon shuffling usually affects entire structural and/or functional units of genes. It also suggests that the overall excess of symmetrical exons and sets of exons in the genome (Long et al. 1995a; Tomita et al. 1996; Fedorov et al. 1998; this study) is a consequence of domain shuffling.

Our results, therefore, do not support the suggestion that weak selection may act against asymmetry of exons, because erroneous splicing of asymmetrical exons results in a frameshift of the mRNA (Lynch 2002). Such selection is unlikely to have played a major role in shaping symmetry levels of exons in the genome, because nonboundary exons are also expected to show excess symmetry according to this hypothesis.

However, selection acting on one form of alternative splicing may have played some evolutionary role in shaping symmetry levels of exons in the genome; exon-skipping without frameshifts requires that the skipped exons or sets of exons are symmetrical. Our finding that only introns bordering domains show correlated phases suggests that selection affecting this form of alternative splicing would mainly involve exons or sets of exons encoding structural and functional domains. It is notable, however, that not in all exon-skipping events the reading frame seems to be maintained (Hide et al. 2001). Furthermore, alternative splicing events that affect exons at the beginning or ends of genes do not restrict phase combinations of exons. Thus, it will be interesting to estimate the extent to which selection associated with in-frame domain splicing may have affected and/or maintained intron phase correlations. However, such analyses will have to await more reliable information on alternative transcripts present in the human genome.

Age Distribution of Domain-Shuffling Events

The majority of domains that are flanked symmetrically by introns of the same phase are of the 1–1 type. We show that this is due to a dramatic expansion of 1–1 modules that is likely to have commenced around the time of the origin of multicellular animals. Thus, we provide statistical support for the hypothesis of a burst of domain shuffling that coincides with the Big bang of metazoan radiation in the Cambrian period (Patthy 1999b). The preference for phase 1 introns flanking these domains may be a result of intron insertion into proto-splice sites (Patthy 1999b). According to this hypothesis, proto-splice site sequences that favor phase 1 intron insertion occur more frequently in domain boundary regions because of their biased amino acid composition.

The nonrandom distribution of symmetrical phase combinations of ancient domain families—an excess of 0–0 domains and a paucity of 1–1 domains—is compatible with the introns-early hypothesis, according to which exons encoding primordial domains shuffled via phase 0 introns to generate more complex proteins (Doolittle 1978; Gilbert 1987). However, although 1–1 phase combinations are statistically significantly under-represented among old domains, 0–0 combinations show nonsignificant excess. Several studies have indicated that ancient proteins show an overall excess of phase 0 introns in boundary regions of domains, consistent with our results (de Souza et al. 1996, 1997; Roy et al. 1999; Fedorov et al. 2001). If the excess of old domains with a symmetrical arrangement of phase 0 introns in our survey reflects primordial shuffling events in the progenote, it may not be surprising that this excess is not large, because intron gain, intron loss, and/or intron sliding may have obscured signals from these ancient events.

In the introns-late view, introns were not present in the progenote, but inserted later, mediating exon shuffling in genes for the complex eukaryotes (Orgel and Crick 1980; Cavalier-Smith 1985; Sharp 1985; Hickey and Benkel 1986). The introns-early and introns-late hypotheses have been the subject of intense debate. Together, our results are compatible with aspects of both hypotheses (de Souza et al. 1998). Whereas introns may have inserted into domain boundary regions (predominantly in phase 1) late in evolution—providing the foundation for a dramatic expansion of 1–1 modules starting around 600 million years ago—0–0 domains may have been the predominant primordial shuffling units of ancient proteins.

Domain Shuffling in *C. elegans*

Our comparative analysis of domains bounded by introns in the nematode *C. elegans* sheds light on the intron-phase patterns identified in humans. In general, the *C. elegans* data seem to reflect the same overall waves of domain shuffling, such as the expansion of 1–1 domains specific to metazoans. However, the lower intron density (Sakharkar et al. 2002), the lower proportion of multidomain proteins, and the less complex domain structure (Lander et al. 2001; Li et al. 2001; Venter et al. 2001) of *C. elegans* genes leave fewer domains that are bordered by introns.

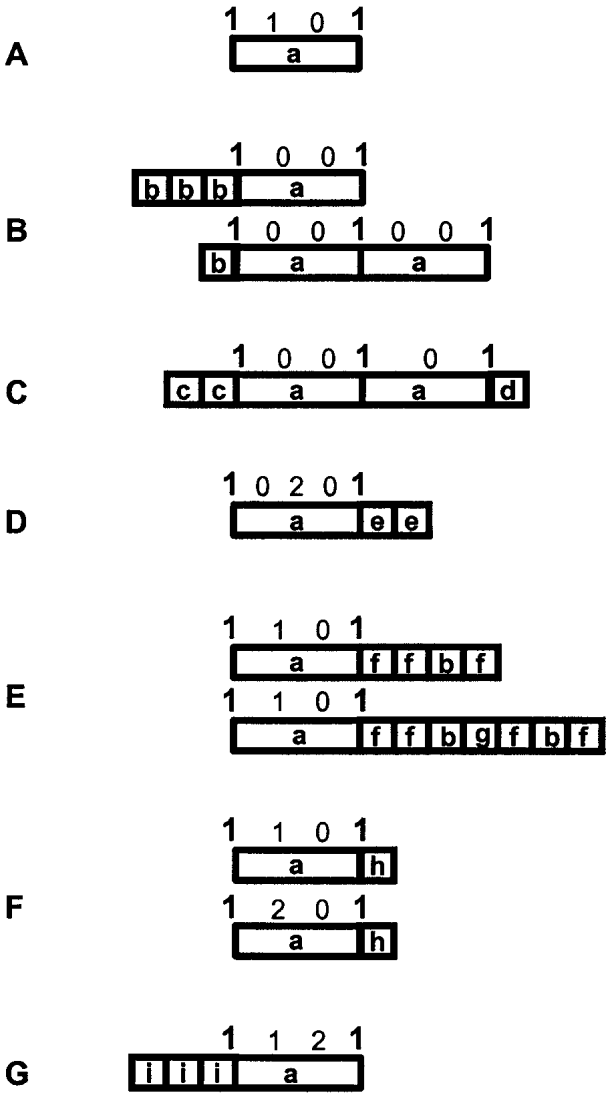

Figure 2 Intron-phase combinations of the discoidin domain in 10 human genes representing 7 gene families (*A*–*G*). Boxes represent the discoidin domain (a) and its neighboring domains (b–i) (not drawn to scale). Numbers indicate the phase class of the introns. Phase 1 introns found at the boundaries of the discoidin domain are shown in boldface. The seven different Ensembl gene families (http://www.ensembl.org) are as follows: (*A*) endothelial and muscle cell-derived neuropilin-like protein, (*B*) lactadherin milk fat globule EGF factor, (*C*) neuropilin precursor vascular endothelial cell growth factor, (*D*) carboxypeptidase H, (*E*) contactin-associated protein like, (*F*) discoidin domain receptor, and (*G*) coagulation factor VIII precursor. The letters within the boxes refer to the following Pfam signatures: discoidin domain (a), EGF-like domain (b), CUB domain (c), MAM domain (d), zinc carboxypeptidase (e), laminin G type domain (f), fibrinogen carboxy-terminal globular domain (g), protein kinase domain (h), and multicopper oxidase (i).

METHODS

Data Preparation

The human data was retrieved from the Ensembl 4.28 database (Hubbard et al. 2002). Genomic exon positions, intron phase information, and gene description information were downloaded (ftp://ftp.ensembl.org/pub/current_human/data/mysql/) and stored in a MySQL relational database system. Ensembl also provides coordinates of Pfam domains (Bateman et al. 2002) relative to each protein sequence that have been calculated using the Interpro tool (ftp://ftp.ebi.ac.uk/pub/databases/interpro/) (Zdobnov and Apweiler 2001); this information was downloaded as well. For genes with several alternative transcripts, we kept the longest coding structure to retain the maximum number of domains. Furthermore, we excluded genes without Pfam signatures. This database (14,728 genes) was used for further analysis. Exon and protein domain coordinates relative to the coordinates of each transcript were calculated using a combination of PERL scripts (PERL DBI) and MySQL database queries.

The *C. elegans* data was retrieved from WormBase (release 30) (ftp://ftp.sanger.ac.uk/pub/C.elegans_sequences). Protein sequences and genomic coordinates of exons and genes were downloaded and stored in the database. All but one isoform of each gene were removed using a previously published method (Gu et al. 2002). Intron phases were calculated on the basis of genomic exon coordinates using PERL scripts. We annotated Pfam domains of *C. elegans* genes using the Interpro tool (ftp://ftp.ebi.ac.uk/pub/databases/interpro/) (Zdobnov and Apweiler 2001). The program was run on a LINUX cluster at the Argonne National Laboratory.

Domain Classification

To identify domains that are bounded by introns, we screened our database (see above) using a combination of PERL scripts and MySQL queries. The relationship of flanking intron positions and domain boundaries had to fulfill the following criteria as follows: $|dn - in| + |dc - ic| < 0.1 \cdot (dc - dn)$; in which dn is the coordinate of the amino terminus of the domain, dc the carboxy-terminal coordinate of the domain in the intron position at the amino terminus of the domain, and ic the position of the intron at the carboxyl terminus of the domain. In cases in which more than one pair of introns satisfied this definition, we chose introns closest to the domain boundary.

We classified domains according to their phylogenetic distribution using the taxonomy information in the Pfam database (http://www.sanger.ac.uk/Software/Pfam/ftp.shtml).

Analysis of Intron Phase Combinations

The random expectation of intron-phase combinations at exon or domain boundaries was calculated on the basis of the observed frequencies of flanking intron phases. The expectation is P_iP_jN, in which P_i and P_j are the frequencies of introns with phases i and j (5′ and 3′), respectively, and N is the total number of cases.

To quantify intron-phase combinations of a domain family, we computed the frequency of each phase class among the different intron-phase combinations carried by the members of this family, setting the total count of each family equal to one.

Statistical Analyses

The χ^2 test was used to assess significance levels in the analyses. For the analysis of observed and expected phase correlations, we used a two-way χ^2 test comparing the summed combinations of symmetrical (0–0, 1–1, and 2–2) and asymmetrical (0–1, 0–2, 1–2, 1–0, 2–0, 2–1) classes (df = 1). Symmetrical classes were also compared individually to the summed combinations of the remaining classes by use of a two-way test procedure. For iteratively assessing phase symmetries of random exons drawn from the genomic background, we implemented this test in a PERL script.

The prior expectation of old and new domains with a specific symmetrical phase combination of boundary introns (0–0 or 1–1) was calculated on the basis of the proportion of that combination among all domains. Observed and expected values were analyzed in a two-way test of the symmetrical

combination and the summed combinations of the remaining classes.

ACKNOWLEDGMENTS

We thank Z. Gu for constructive discussions and for providing *C. elegans* data; M. Long for comments on the manuscript; R. Blocker for UNIX/Linux system maintenance; and R. Stevens for granting access to Unix clusters at the Argonne National Laboratory. This study was supported by NIH grants GM30998 and GM65499. H.K. was supported by a fellowship from the European Molecular Biology Organization and an Emmy Noether fellowship from the Deutsche Forschungsgemeinschaft. S.Z. was supported by the Katz fellowship.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Andrade, M.A., Perez-Iratxeta, C., and Ponting, C.P. 2001. Protein repeats: Structures, functions, and evolution. *J. Struct. Biol.* **134:** 117–131.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30:** 276–280.

Beckmann, G. and Bork, P. 1993. An adhesive domain detected in functionally diverse receptors. *Trends Biochem. Sci.* **18:** 40–41.

Betts, M.J., Guigo, R., Agarwal, P., and Russell, R.B. 2001. Exon structure conservation despite low sequence similarity: A relic of dramatic events in evolution? *EMBO J.* **20:** 5354–5360.

Cavalier-Smith, T. 1985. Selfish DNA and the origin of introns. *Nature* **315:** 283–284.

Copley, R.R., Doerks, T., Letunic, I., and Bork, P. 2002. Protein domain analysis in the era of complete genomes. *FEBS Lett.* **513:** 129–134.

de Souza, S.J., Long, M., Schoenbach, L., Roy, S.W., and Gilbert, W. 1996. Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl. Acad. Sci.* **93:** 14632–14636.

de Souza, S.J., Long, M., Schoenbach, L., Roy, S.W., and Gilbert, W. 1997. The correlation between introns and the three-dimensional structure of proteins. *Gene* **205:** 141–144.

de Souza, S.J., Long, M., Klein, R.J., Roy, S., Lin, S., and Gilbert, W. 1998. Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci.* **95:** 5094–5099.

Deutsch, M. and Long, M. 1999. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27:** 3219–3228.

Doolittle, W.F. 1978. Genes in pieces: Were they ever together? *Nature* **272:** 581–582.

Elofsson, A. and Sonnhammer, E.L. 1999. A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics* **15:** 480–500.

Fedorov, A., Fedorova, L., Starshenko, V., Filatov, V., and Grigor'ev, E. 1998. Influence of exon duplication on intron and exon phase distribution. *J. Mol. Evol.* **46:** 263–271.

Fedorov, A., Cao, X., Saxonov, S., de Souza, S.J., Roy, S.W., and Gilbert, W. 2001. Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. *Proc. Natl. Acad. Sci.* **98:** 13177–13182.

Gilbert, W. 1987. The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.* **52:** 901–905.

Gluecksmann-Kuis, M.A., Tayber, O., Woolf, E.A., Bougueleret, L., Deng, N., Alperin, G.D., Iris, F., Hawkins, F., Munro, C., Lakey, N., et al. 1995. Polycystic kidney disease: The complete structure of the PKD1 gene and its protein. The International Polycystic Kidney Disease Consortium. *Cell* **81:** 289–298.

Go, M. and Nosaka, M. 1987. Protein architecture and the origin of introns. *Cold Spring Harb. Symp. Quant. Biol.* **52:** 915–924.

Goodier, J.L., Ostertag, E.M., and Kazazian, H.H. 2000. Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9:** 653–657.

Graur, D. and Li, W.H. 1999. *Fundamentals of molecular evolution.* Sinauer Associates, Sunderland, MA.

Gu, Z., Cavalcanti, A., Chen, F.C., Bouman, P., and Li, W.H. 2002. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol. Biol. Evol.* **19:** 256–262.

Hickey, D.A. and Benkel, B. 1986. Introns as relict retrotransposons: Implications for the evolutionary origin of eukaryotic mRNA splicing mechanisms. *J. Theor. Biol.* **121:** 283–291.

Hide, W.A., Babenko, V.N., van Heusden, P.A., Seoighe, C., and Kelso, J.F. 2001. The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res.* **11:** 1848–1853.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.

Kazazian Jr., H.H., 2000. Genetics. L1 retrotransposons shape the mammalian genome. *Science* **289:** 1152–1153.

Koonin, E.V., Aravind, L., and Kondrashov, A.S. 2000. The impact of comparative genomics on our understanding of evolution. *Cell* **101:** 573–576.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Li, W.H., Gu, Z., Wang, H., and Nekrutenko, A. 2001. Evolutionary analyses of the human genome. *Nature* **409:** 847–849.

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

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

Lynch, M. 2002. Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci.* **99:** 6118–6123.

Moran, J.V., DeBerardinis, R.J., and Kazazian Jr., H.H., 1999. Exon shuffling by L1 retrotransposition. *Science* **283:** 1530–1534.

Orgel, L.E. and Crick, F.H. 1980. Selfish DNA: The ultimate parasite. *Nature* **284:** 604–607.

Patthy, L. 1999a. Genome evolution and the evolution of exon-shuffling—a review. *Gene* **238:** 103–114.

———. 1999b. *Protein evolution*, p. 228. Blackwell Science, Oxford, UK.

Pawson, T. and Nash, P. 2000. Protein-protein interactions define specificity in signal transduction. *Genes & Dev.* **14:** 1027–1047.

Pickeral, O.K., Makalowski, W., Boguski, M.S., and Boeke, J.D. 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10:** 411–415.

Roy, S.W., Nosaka, M., de Souza, S.J., and Gilbert, W. 1999. Centripetal modules and ancient introns. *Gene* **238:** 85–91.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., et al. 2000. Comparative genomics of the eukaryotes. *Science* **287:** 2204–2215.

Sakharkar, M., Passetti, F., de Souza, J.E., Long, M., and de Souza, S.J. 2002. ExInt: An exon intron database. *Nucleic Acids Res.* **30:** 191–194.

Sharp, P.A. 1985. On the origin of RNA splicing and introns. *Cell* **42:** 397–400.

Tomita, M., Shimizu, N., and Brutlag, D.L. 1996. Introns and reading frames: Correlation between splicing sites and their codon positions. *Mol. Biol. Evol.* **13:** 1219–1223.

Tordai, H., Banyai, L., and Patthy, L. 1999. The PAN module: The N-terminal domains of plasminogen and hepatocyte growth factor are homologous with the apple domains of the prekallikrein family and with a novel domain found in numerous nematode proteins. *FEBS Lett.* **461:** 63–67.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., and Yandell, M. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Zdobnov, E.M. and Apweiler, R. 2001. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17:** 847–848.

WEB SITE REFERENCES

ftp://ftp.ebi.ac.uk/pub/databases/interpro/; Interpro.

ftp://ftp.ensembl.org/pub/current_human/data/mysql/; Ensembl database (human).

ftp://ftp.sanger.ac.uk/pub/; C.elegans_sequences.

ftp://ftp.wormbase.org/pub/wormbase; WormBase.

http://www.sanger.ac.uk/software/Pfam/index.shtml; Pfam database.

Received June 10, 2002; accepted in revised form September 13, 2002.