Independent Effects of Alternative Splicing and Structural Constraint on the Evolution of Mammalian Coding Exons

Feng-Chi Chen,*[1,2,3] Chia-Lin Pan,[1] and Hsuan-Yu Lin[1]

[1]Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan, Republic of China

[2]Department of Life Sciences, National Chiao-Tung University, Hsinchu, Taiwan, Republic of China

[3]Department of Dentistry, Chinese Medical University, Taichung, Taiwan, Republic of China

***Corresponding author:** E-mail: fcchen@nhri.org.tw.

Associate editor: John H McDonald

Abstract

Alternative splicing (AS) is known to significantly affect exon-level protein evolutionary rates in mammals. Particularly, alternatively spliced exons (ASEs) have a higher nonsynonymous-to-synonymous substitution rate (*d*N/*d*S) ratio than constitutively spliced exons (CSEs), possibly because the former are required only occasionally for normal biological functions. Meanwhile, intrinsically disordered regions (IDRs), the protein regions lacking fixed 3D structures, are also reported to have an increased evolutionary rate due to lack of structural constraint. Interestingly, IDRs tend to be located in alternative protein regions. Yet which of these two factors is the major determinant of the increased *d*N/*d*S in mammalian ASEs remains unclear. By comparing human–macaque and human–mouse one-to-one orthologous genes, we demonstrate that AS and protein structural disorder have independent effects on mammalian exon evolution. We performed analyses of covariance to demonstrate that the slopes of the (*d*N/*d*S-percentage of IDR) regression lines differ significantly between CSEs and ASEs. In other words, the *d*N/*d*S ratios of both ASEs and CSEs increase with the proportion of IDR (PIDR), whereas ASEs have higher *d*N/*d*S ratios than CSEs when they have similar PIDRs. Since ASEs and IDRs may less frequently overlap with protein domains (which also affect *d*N/*d*S), we also examined the correlations between *d*N/*d*S ratio and exon type/PIDR by controlling for the density of protein domain. We found that the effects of exon type and PIDR on d*N*/d*S* are both independent of domain density. Our results imply that nature can select for different biological features with regard to ASEs and IDRs, even though the two biological features tend to be localized in the same protein regions.

Key words: alternative splicing, intrinsically disordered region, nonsynonymous to synonymous substitution rate ratio.

Introduction

Alternative splicing (AS) is a major mechanism of increasing transcriptome/proteome diversity and regulatory complexity (Keren et al. 2010). In humans, more than 90% of the genes are alternatively spliced (Hallegger et al. 2010). AS has significant effects on the evolution of coding exons. For genes that undergo AS, constitutively spliced exons (CSEs) are present in all of the transcript isoforms and are likely subject to strong selective constraints. In contrast, alternatively spliced exons (ASEs), due to their intermittent presence in transcripts, may be relatively relaxed from selection pressure (Modrek and Lee 2003; Xing and Lee 2004). Indeed, previous studies have demonstrated that ASEs have higher nonsynonymous substitution rates (*d*N) and nonsynonymous-to-synonymous substitution rate (*d*N/*d*S) ratios than CSEs (Chen, Wang, et al. 2006; Chen and Chuang 2007; Ramensky et al. 2008).

Another important determinant of protein evolutionary rate is structural constraint. In most of the cases, peptides and proteins must be correctly folded to carry out normal biological functions. Yet, certain protein regions lack fixed 3D structures, even though they are located in functional proteins. These protein regions (designated as intrinsically disordered regions or "IDRs") may either have limited functional importance or depend on induced-fit structural changes to perform normal protein functions (Uversky and Dunker 2010). Furthermore, IDRs (and possibly ASEs) may less frequently overlap with functional protein domains because the latter usually have fixed 3D structures (Ponting and Russell 2002). It has been reported that experimentally determined IDRs evolve faster than structurally ordered regions and that disordered proteins have fewer evolutionary constraints than ordered proteins (Brown et al. 2002, 2010; Kahali et al. 2009). Interestingly, IDRs tend to be located in alternative protein regions (Romero et al. 2006; Pentony and Jones 2009). These observations raise an important question regarding exon evolution: are AS and structural constraint independent of each other in affecting exon-level protein evolutionary rate? In other words, does the previously observed CSE–ASE difference in evolutionary rate merely reflect the effects of IDRs? Or alternatively, do IDRs evolve faster simply because they tend to be located in alternative protein regions? In addition, whether IDRs and ASEs evolve faster because they tend to lack protein domains is also a question of interest.

To answer the above questions, we retrieved human–mouse and human–macaque one-to-one orthologous genes

and controlled for exon type (CSE/ASE) while evaluating the correlation between evolutionary rate and the proportion of IDRs (designated as "PIDR"). Our results indicate that AS and protein structural constraint have independent effects on the protein evolutionary rate in mammals. Furthermore, CSE/ASE exon type and PIDR each can affect protein evolutionary rates regardless of the density of protein domain. Our results suggest that selection may work independently on different biological features (e.g., exon type and PIDR), even though these features tend to occur at the same physical locations.

Materials and Methods

Data Source

The transcripts of human, rhesus macaque, and mouse were retrieved from Ensembl Version 59 (http://www.ensembl.org) and the University of California–San Cruz (UCSC) Genome Browser (http://genome.ucsc.edu) with reference human, macaque, and mouse genomes as hg19, rheMac2, and mm9, respectively. To ensure data quality, we required that the analyzed transcripts be experimentally validated and have known protein products. And the transcripts must begin with the AUG start codon and end with a legitimate stop codon. The gene orthology and Pfam protein domain annotations (Finn et al. 2010) were also retrieved from Ensembl. The "domain density" of each exon was the proportion of the exon that was annotated to be part of a Pfam protein domain. The exons were classified into CSEs and ASEs as described previously (Shabalina et al. 2010).

Prediction of IDR and Calculation of Evolutionary Rates

IDR was predicted by using Disopred (Ward et al. 2004) with default parameters. MUSCLE (Edgar 2004) was used to align the peptide sequences. The proportion of human–macaque or human–mouse alignable sequence must be $\geq$90% of the human exon length. For each exon, we calculated the percentage of amino acids that were predicted to fall within IDRs (designated as PIDR). The CODEML program of the PAML package V4.4 (Yang 1997) was used to calculate dN and dN/dS ratio. The exons shorter than 81 bp were excluded from the analysis to avoid biases in evolutionary rate estimations. It was reported that the accuracy of the dN/dS ratio test could be considerably compromised in the case of short exons (e.g., $<$50 bp) and that the exons that failed the dN/dS ratio test had a median length of 78 bp (Nekrutenko et al. 2002). Therefore, we selected this 81 bp threshold. However, two different thresholds were also tried in later analyses (99 and 120 bp).

Statistical Test

To distinguish the effects of exon type (CSE/ASE) and PIDR on protein evolutionary rate, an analysis of covariance (ANCOVA) was performed using the R package with reference to (McDonald 2009). Considering that other factors (e.g., G+C content and exon length) may also affect the evolutionary rates of coding exons (Lemos et al. 2005; Galtier et al. 2009), we began our analysis by formulating a "full model" that included all of the four variables:

$$\begin{aligned} d\mathrm{N}/d\mathrm{S} = {} & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \\ & + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{14} X_1 X_4 \\ & + \beta_{23} X_2 X_3 + \beta_{24} X_2 X_4 + \beta_{34} X_3 X_4 + \varepsilon, \end{aligned} \quad (1)$$

where X_1, X_2, X_3, and X_4, respectively, represents the CSE/ASE exon type (CSE = 0, ASE = 1), PIDR, G+C content, and exon length (bp). The products of variables represent the interaction terms. ϵ stands for the error term.

We used three model selection approaches (stepwise, forward, and backward) to select the variables that can best explain the variations of dN/dS. As shown in supplementary figure S1, Supplementary Material online, the three approaches yielded the same best models (Models 1 and 2 had very similar Baysian Information Criterion scores). In other words, G+C content and exon length seem to contribute less to the variations of dN/dS than CSE/ASE exon type and PIDR.

Since Model #2 (which included CSE/ASE exon type, PIDR, and the interaction between these two variables) was one of the two top models and was consistent with previous observations, we first selected this model and examined whether the slopes of dN/dS-PIDR regression lines differed significantly between CSEs and ASEs. The regression model was as follows:

$$d\mathrm{N}/d\mathrm{S} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon. \quad (2)$$

Therefore, for CSEs,

$$(d\mathrm{N}/d\mathrm{S}|X_1 = 0) = \beta_0 + \beta_2 X_2 + \varepsilon, \quad (3)$$

Whereas for ASEs,

$$(d\mathrm{N}/d\mathrm{S}|X_1 = 1) = (\beta_0 + \beta_1) + (\beta_2 + \beta_{12}) X_2 + \varepsilon, \quad (4)$$

If $\beta_{12} \neq 0$, the slopes of CSEs and ASEs would be considered as different. Alternatively, if $\beta_{12} = 0$ but $\beta_1 \neq 0$, the intercepts of the two regression lines would be considered as different. In either case, the CSE/ASE exon type and PIDR would be regarded as having independent effects on dN/dS (McDonald 2009).

We also tried to use model #4 according to the stepwise model selection to include another variable (X_4, exon length). The regression model was

$$\begin{aligned} d\mathrm{N}/d\mathrm{S} = {} & \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ & + \beta_4 X_4 + \beta_{12} X_1 X_2 \\ & + \beta_{24} X_2 X_4 + \varepsilon. \end{aligned} \quad (5)$$

For CSEs,

$$(d\mathrm{N}/d\mathrm{S}|X_1 = 0) = \beta_0 + \beta_4 X_4 + (\beta_2 + \beta_{24} X_4) X_2 + \varepsilon. \quad (6)$$

For ASEs,

$$\begin{aligned} (d\mathrm{N}/d\mathrm{S}|X_1 = 1) = {} & (\beta_0 + \beta_1) + \beta_4 X_4 \\ & + (\beta_2 + \beta_{12} + \beta_{24} X_4) X_2 + \varepsilon, \end{aligned} \quad (7)$$

If we fix X_4, the difference in slope between equations (6) and (7) lies in β_{12}. If $\beta_{12} \neq 0$, the two regression lines have different slopes, and the CSE/ASE exon type and PIDR are regarded as having independent effects on dN/dS.

Table 1. Estimates of Regression Coefficients and P values of the ANCOVA[a].

	Model #2[b]			Model #4[c]		
	Estimate	t value	P value	Estimate	t value	P value
β_0	3.58×10^{-2}	32.00	$<2 \times 10^{-16}$	4.10×10^{-2}	22.72	$<2 \times 10^{-16}$
β_1	5.90×10^{-2}	3.17	0.0016	5.90×10^{-3}	3.15	1.65×10^{-3}
β_2	3.40×10^{-3}	13.41	$<2 \times 10^{-16}$	2.46×10^{-3}	6.86	7.14×10^{-12}
β_4	NA	NA	NA	-3.64×10^{-5}	−3.71	2.06×10^{-4}
β_{12}	3.08×10^{-2}	7.50	6.92×10^{-14}	3.09×10^{-2}	7.50	6.79×10^{-14}
β_{24}	NA	NA	NA	6.30×10^{-5}	3.85	1.21×10^{-4}

NOTE.—NA, not available.

[a] The data points with dS = 0 were excluded.

[b] The regression model was : $d\text{N}/d\text{S} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$ (X_1: exon type: CSE = 0, ASE = 1; X_2: PIDR).

[c] The regression model was $d\text{N}/d\text{S} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_{12} X_1 X_2 + \beta_{24} X_2 X_4 + \varepsilon$ (X_1: exon type: CSE = 0, ASE = 1; X_2: PIDR; X_4: exon length [bp]).

Since the dN/dS ratios are not normally distributed ($P = 0$, by the Pearson chi-square normality test, supplementary fig. S2, Supplementary Material online), all of the statistical tests in the comparisons of evolutionary rates between different exon groups were performed by using the Wilcoxon rank-sum test.

Results and Discussion

The Independence between CSE/ASE Exon Type and PIDR in Affecting the dN/dS Ratio

To disentangle the effects of CSE/ASE exon type and PIDR on mammalian protein evolutionary rate, we formulated a complex regression model by including CSE/ASE exon type, PIDR, G+C content, exon length, and six pairwise interaction terms and used three model selection approaches to choose the variables that can best explain the variations of dN/dS (see Materials and Methods). However, G+C content is not selected in the top four models in any of the model selection methods (Materials and Methods). By contrast, CSE/ASE exon type, PIDR, and the interaction term between these two variables were consistently selected by all three model selection approaches as one of the two top models (supplementary fig. S1, Supplementary Material online). We thus performed an ANCOVA for Model #2 to compare the dN/dS-PIDR regression slopes between CSEs and ASEs. As shown in table 1, the coefficient of the interaction term (β_{12}) deviates significantly from zero (P value $= 6.92 \times 10^{-14}$ by t-test, table 1), indicating independence between the two factors in affecting dN/dS.

To include one more variable (exon length) in our analysis, we performed an ANCOVA for Model #4 according to stepwise model selection (supplementary fig. S1, Supplementary Material online, Materials and Methods). We actually obtained similar results: β_{12} significantly deviates from zero ($P = 6.79 \times 10^{-14}$; table 1), supporting the independence between CSE/ASE exon type and PIDR in affecting the dN/dS ratio (see Materials and Methods for more details).

Notably, in the above analysis, we classified CSEs and ASEs according to human gene annotations and aligned the human exons against their macaque/mouse counterparts without considering whether the exon types were evolutionarily conserved. It is known that many ASEs are lineage specific (Nurtdinov et al. 2003; Chen, Chen, et al. 2006). Therefore, whether the above observations are also true for CSEs/ASEs in other species remain unknown. To address this issue, we classified mouse exons into CSEs and ASEs according to Ensembl annotations, aligned the exons to their human counterparts, and performed an ANCOVA for Model #2 as mentioned above. We obtained similar results that exon type and PIDR have independent effects on dN/dS (β_{12} deviates significantly from zero, $P = 2.86 \times 10^{-13}$; supplementary table S1, Supplementary Material online). Therefore, our results appear to hold well regardless of interspecies differences in splicing patterns. This is expected because it has been previously shown that conserved ASEs evolve faster than nonconserved ASEs, which in turn evolve faster than conserved CSEs (Chen et al. 2007). Therefore, even if a significant proportion of human ASEs turns out to be CSEs in mouse (or vice versa), they are expected to evolve faster than conserved CSEs, which account for the majority of all CSEs (Chen et al. 2007).

Changes in Parameter Settings Do Not Affect the Overall Results

The independent effects of exon type and PIDR can be easily observed if we divide the CSEs/ASEs into three categories according to PIDR: highly ordered—0–20% IDR, intermediately disordered—20–80% IDR, and highly disordered—80–100% of IDR (table 2) for comparison of dN and dN/dS ratios. As shown in figure 1, ASEs have significantly higher dN and dN/dS ratios than CSEs, which is consistent with previous findings regardless of whether they have high or low PIDRs. Furthermore, for both CSEs and ASEs, dN and dN/dS ratio increase significantly with increasing PIDR. This observation

Table 2. The Numbers of Exons Analyzed in figure 1.

		Human–Mouse	Human–Macaque
Number of transcripts		5,929	9,465
Number of CSEs		8,992	29,877
PIDR	0–20%	5,322	16,539
	20–80%	2,110	7,212
	80–100%	1,560	6,126
Number of ASEs		15,828	51,425
PIDR	0–20%	9,653	28,580
	20–80%	3,856	13,143
	80–100%	2,319	9,702

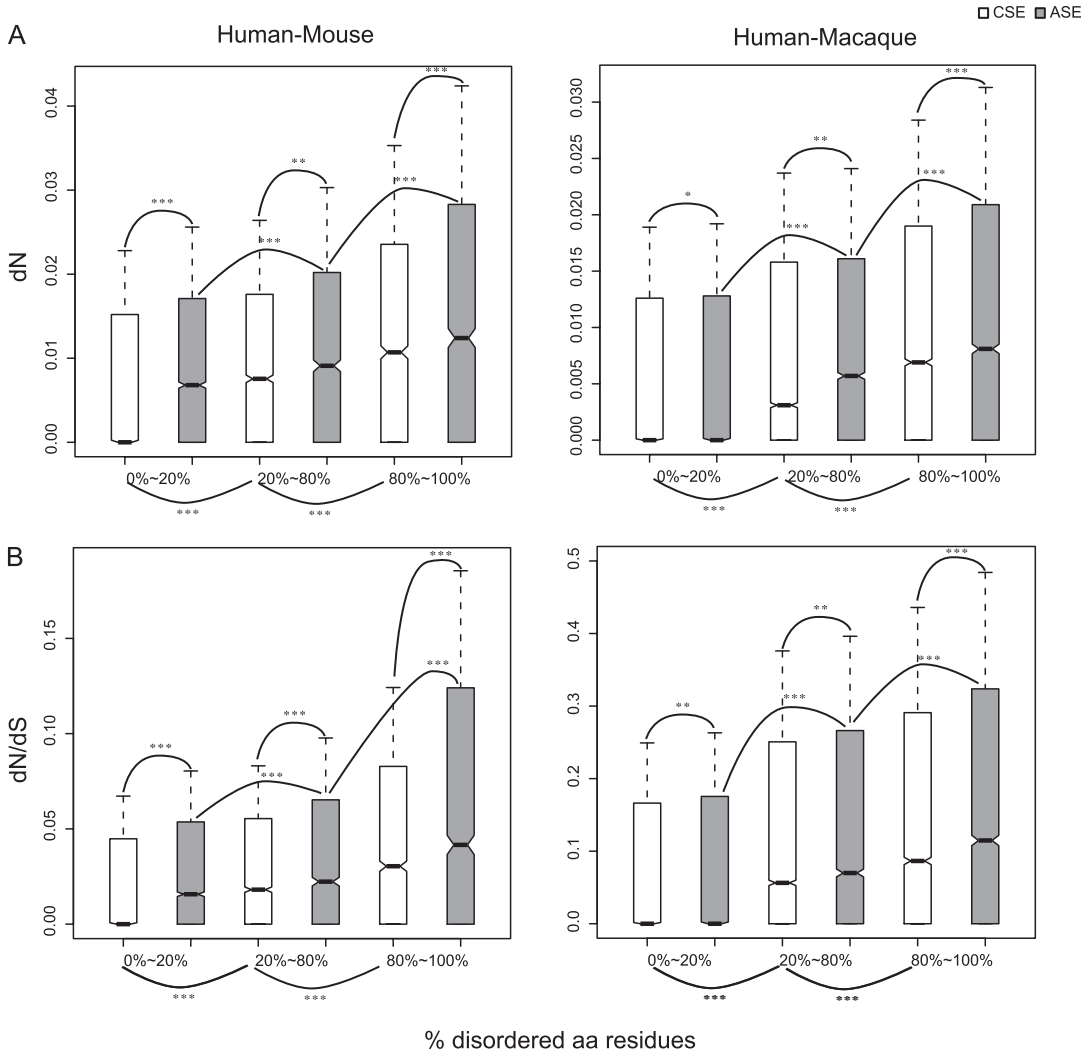

FIG. 1. The distributions of *d*N and *d*N/*d*S ratios of human–mouse and human–macaque pairwise comparisons. The curves with stars indicate statistical significance in the comparison of different exon groups. *$P < 0.05$; **$P < 0.01$; and ***$P < 0.001$.

is consistent with the ANCOVA results that the CSE/ASE exon type and PIDR affect the evolutionary rates of coding exons independently.

To examine whether different parameter settings may affect our results for the human–mouse comparison, we also used 1) two different PIDR binning schemes (0–10%, 10–90%, and 90–100%; 0–30%, 30–70%, and 70–100%; supplementary fig. S3, Supplementary Material online); 2) a different criterion for selecting orthologous exon pairs for analysis (the alignable length of the macaque or mouse exons must be $\geq$85% of the corresponding human exon; supplementary fig. S4, Supplementary Material online); 3) a different alignment tool (ClustalW, Larkin et al. 2007; supplementary fig. S5, Supplementary Material online); and 4) a different tool to predict IDR (PreDisorder, Deng et al. 2009; supplementary fig. S6, Supplementary Material online). We also used a different exon classification scheme according to how frequent an exon is included in the alternative transcript isoforms (high frequency: $>$50%; low frequency: $\leq$50%) with two different data sources (Ensembl and UCSC; supplementary fig. S7, Supplementary Material online).

We obtained similar results for all of these different parameter settings. In addition, we tried to use only "pure coding sequences" (the exons that do not overlap with any untranslated regions of other transcript isoforms) for the analysis and again obtained similar results (supplementary fig. S8, Supplementary Material online). Finally, in the above analyses, we filtered out exons shorter than 81 bp to avoid potential biases in evolutionary rate estimates (Materials and Methods). Changing this threshold to 99 or 120 bp does not affect the overall trend (supplementary fig. S9, Supplementary Material online) (The ANCOVA results for the data sets used in supplementary figs. S4–S9, Supplementary Material online can be found in supplementary table S2, Supplementary Material online). Therefore, our results appear to hold well regardless of changes in a variety of parameter settings.

Notably, here, we require that the percentages of alignable length between orthologous exonic sequences to be larger than 90% to ensure the quality of the sequence alignments and to avoid confusion in PIDR-based exon groupings (Materials and Methods). This requirement may bias the analyzed sequences toward highly conserved

exons. However, we observed similar trends in both human–mouse and human–macaque comparisons, which represent different levels of genetic divergences. Furthermore, reducing the threshold to 85% does not affect the overall results (supplementary fig. S4 and table S2, Supplementary Material online). Therefore, our results may have reflected the biological truth.

The Effects of CSE/ASE Exon Type and PIDR on *d*N/*d*S Ratio Are Independent of Exon Order and Protein Domain Density

Next, we ask whether the increase of protein evolutionary rates in ASEs and high-PIDR exons are associated with other factors. It has been reported, for example, that exons encoding signal peptides (usually at the N-terminal) evolve faster than other exons (Li et al. 2009). Similarly, C-terminal exons are also suggested to evolve rapidly (Ermakova et al. 2006). Therefore, we compared the proportions of first and last coding exons of different exon groups. Our results indicate that CSEs tend to have a larger proportion of first coding exons (fig. 2*A*) but a smaller proportion of last coding exons than ASEs (fig. 2*B*). Meanwhile, for both CSE and ASE, the proportion of first coding exons is highest in the medium-PIDR exon group, whereas the proportion of last coding exons tends to increase with PIDR. However, the difference in the proportion of last coding exons is statistically insignificant between medium- and high-PIDR exons. Therefore, the influences of first/last coding exons on the evolutionary rate differences between different exon groups remain ambiguous. We thus excluded all of the first/last coding exons and compared the *d*N/*d*S again. In fact, excluding these exons does not affect our overall results (supplementary fig. S10 and table S2, Supplementary Material online).

We also tried to clarify whether the effects of PIDR on *d*N/*d*S ratio actually depended on other evolutionary rate determinants (e.g., the density of functional protein domain, G+C content, and exon length). In fact, domain density tends to decrease as PIDR increases for both CSEs and ASEs (supplementary fig. S11, Supplementary Material online). PIDR is also correlated with both G+C content and exon length (supplementary table S3, Supplementary Material online). Furthermore, *d*N/*d*S ratio is correlated separately with protein domain density, G+C content, and exon length (supplementary table S3, Supplementary Material online). Therefore, we performed partial correlation analyses with reference to Kim and Yi (2007) to examine the correlations between *d*N/*d*S ratio and PIDR by simultaneously controlling three potential confounding factors

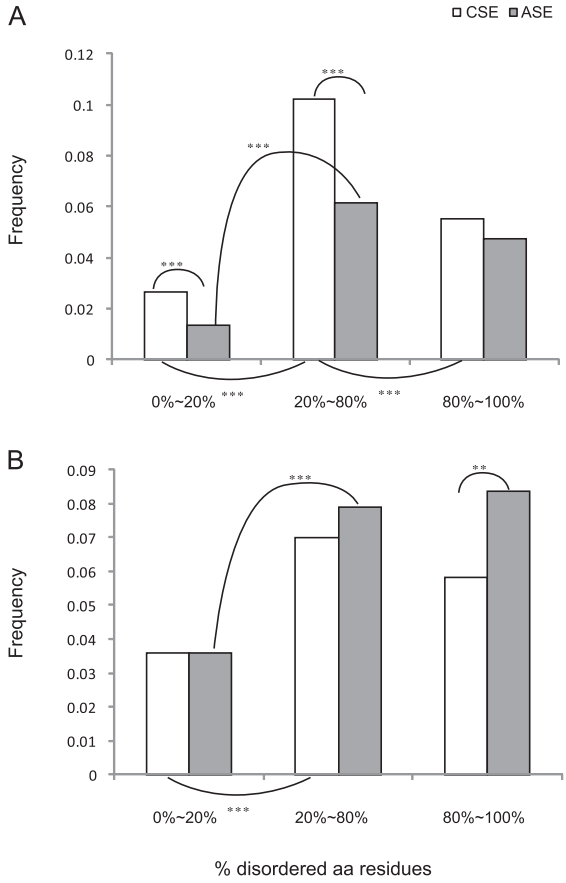

Fig. 2. The proportions of (*A*) first and (*B*) last coding exons in different PIDR categories. The curves with stars indicate statistical significance in the comparison of different exon groups. *$P < 0.05$; **$P < 0.01$; and ***$P < 0.001$.

(domain density, G+C content, and exon length). Interestingly, as shown in table 3, the effects of PIDR on *d*N and *d*N/*d*S ratio are independent of the three confounding factors.

Meanwhile, we were also interested to know whether the effects of CSE/ASE exon type on *d*N/*d*S were dependent on the three abovementioned factors. Since the CSE/ASE exon type is a discrete variable and the partial correlation analysis is inapplicable in this regard, we performed a series of ANCOVA by substituting PIDR (X_2) with domain density, G+C content, or exon length in Model #2 (Materials and Methods). Although none of the β_{12} estimates significantly deviates from zero, β_1 actually is significantly larger than zero in all of the three cases (table 4). Our results indicate that the effects of CSE/ASE exon type are independent of these three confounding factors.

Table 3. Partial Correlation between Evolutionary Rate (*d*N and *d*N/*d*S ratio) and PIDR When Protein Domain Density, G+C content, and Exon Length are Controlled.

					Spearman's Correlation	
Factor 1	**Factor 2**	**Control 1**	**Control 2 (%)**	**Control 3**	**Rho**	***P* value**
***d*N/*d*S ratio**	**PIDR**	**Domain density**	**G+C**	**Exon length**	**0.13**	**$<2 \times 10^{-16}$**
***d*N**	**PIDR**	**Domain density**	**G+C**	**Exon length**	**0.12**	**$<2 \times 10^{-16}$**

Table 4. ANCOVA[a] Results When Domain Density, G+C content, or Exon Length Are Included in the Regression Model.

Coefficient \ X_2	Protein Domain Density			G+C Content			Exon Length		
	Estimate	*t* value	*P* value	Estimate	*t* value	*P* value	Estimate	*t* value	*P* value
β_0	4.88×10^{-2}	16.06	$<2 \times 10^{-16}$	2.93×10^{-2}	5.93	3.05×10^{-9}	4.43×10^{-2}	30.71	$<2 \times 10^{-16}$
β_1	1.15×10^{-2}	2.18	2.90×10^{-2}	2.59×10^{-2}	3.17	1.51×10^{-3}	1.32×10^{-2}	5.53	3.29×10^{-8}
β_2	-1.49×10^{-2}	−4.23	2.41×10^{-5}	3.14×10^{-2}	3.17	1.51×10^{-3}	2.98×10^{-6}	0.41	6.80×10^{-1}
β_{12}	-5.26×10^{-3}	−0.87	3.86×10^{-1}	-2.16×10^{-2}	−1.32	1.88×10^{-1}	1.40×10^{-5}	1.17	2.41×10^{-1}

[a] The regression model was : $dN/dS = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$ (X_1: exon type: CSE = 0, ASE = 1; X_2: protein domain density, G+C content, or exon length).

Potential Caveats

One potential problem in this study is the classification of ASEs and CSEs, for the "CSEs" defined here may actually be absent in some yet unidentified transcript isoforms. However, as we demonstrate in supplementary figure S7 and table S2, Supplementary Material online, the results hold well, even if we classify exons according to a different grouping method (the frequency of being included in different transcript isoforms). In other words, the CSEs and "ASEs" defined here actually have evolutionary properties similar to high- and low-frequency exons, respectively. Therefore, our results can be considered, in a broad view, as demonstrating the differences in protein evolutionary rate between commonly and less frequently used exons with different levels of structural constraint. This study clearly demonstrates that protein structural constraint and AS have independent effects on the evolution of mammalian coding exons.

Concluding Remarks

The independence between CSE/ASE exon type and structural disorderness in affecting protein evolutionary rate is intriguing. ASEs can be viewed as "occasionally unnecessary" exons. In contrast, CSEs can be regarded as "essential" exons. Therefore, ASEs may understandably evolve rapidly due to relaxed selection pressure. Meanwhile, IDRs are relatively free to change because they usually are not bound by the structural constraint important for normal protein functions. Our results appear to suggest that "exon essentiality" and "structural requirement" can be decoupled in view of natural selection. An interesting inference from these observations is that natural selection can work separately on different biological features, although it appears to favor the physical colocalization of these features. Our study thus demonstrates the intricate actions of natural selection in affecting protein evolution.

Supplementary Material

Supplementary figures S1–S11 and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

F.-C.C. was supported by the National Science Council, Taiwan, Republic of China under contract number NSC-98-2311-B-400-002-MY3, and the intramural funding of National Health Research Institutes, Taiwan, Republic of China. We thank Dr Ben-Yang Liao for constructive comments and technical assistance. We also thank Dr Yen-Feng Chiu, Dr Wen-Chang Wang, Ms. Hui-I Gao, and Mr. Tsung-Kai Chang for assistance in statistical tests.

References

Brown CJ, Johnson AK, Daughdrill GW. 2010. Comparing models of evolution for ordered and disordered proteins. *Mol Biol Evol*. 27:609–621.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol*. 55:104–110.

Chen FC, Chaw SM, Tzeng YH, Wang SS, Chuang TJ. 2007. Opposite evolutionary effects between different alternative splicing patterns. *Mol Biol Evol*. 24:1443–1446.

Chen FC, Chen CJ, Ho JY, Chuang TJ. 2006. Identification and evolutionary analysis of novel exons and alternative splicing events using cross-species EST-to-genome comparisons in human, mouse and rat. *BMC Bioinformatics*. 7:136.

Chen FC, Chuang TJ. 2007. Different alternative splicing patterns are subject to opposite selection pressure for protein reading frame preservation. *BMC Evol Biol*. 7:179.

Chen FC, Wang SS, Chen CJ, Li WH, Chuang TJ. 2006. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Mol Biol Evol*. 23:675–682.

Deng X, Eickholt J, Cheng J. 2009. PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics*. 10:436.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

Ermakova EO, Nurtdinov RN, Gelfand MS. 2006. Fast rate of evolution in alternatively spliced coding regions of mammalian genes. *BMC Genomics*. 7:84.

Finn RD, Mistry J, Tate J, et al. (14 co-authors). 2010. The Pfam protein families database. *Nucleic Acids Res*. 38:D211–D222.

Galtier N, Duret L, Glemin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet*. 25:1–5.

Hallegger M, Llorian M, Smith CW. 2010. Alternative splicing: global insights. *FEBS J*. 277:856–866.

Kahali B, Ahmad S, Ghosh TC. 2009. Exploring the evolutionary rate differences of party hub and date hub proteins in Saccharomyces cerevisiae protein-protein interaction network. *Gene* 429:18–22.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*. 11:345–355.

Kim SH, Yi SV. 2007. Understanding relationship between sequence and functional evolution in yeast proteins. *Genetica* 131:151–156.

Larkin MA, Blackshields G, Brown NP, et al. (13 co-authors). 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol*. 22:1345–1354.

Li YD, Xie ZY, Du YL, Zhou Z, Mao XM, Lv LX, Li YQ. 2009. The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. *Gene* 436:8–11.

McDonald JH. 2009. Handbook of biological statistics. Baltimore (MD): Sparky House Publishing. p. 232–237.

Modrek B, Lee CJ. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet*. 34:177–180.

Nekrutenko A, Makova KD , Li WH. 2002. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res*. 12: 198–202.

Nurtdinov RN, Artamonova II, Mironov AA, Gelfand MS. 2003. Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum Mol Genet*. 12:1313–1320.

Pentony MM, Jones DT. 2009. Modularity of intrinsic disorder in the human proteome. *Proteins*. 78:212–221.

Ponting CP, Russell RR. 2002. The natural history of protein domains. *Annu Rev Biophys Biomol Struct*. 31:45–71.

Ramensky VE, Nurtdinov RN, Neverov AD, Mironov AA, Gelfand MS. 2008. Positive selection in alternatively spliced exons of human genes. *Am J Hum Genet*. 83:94–98.

Romero PR, Zaidi S, Fang YY, et al. (11 co-authors). 2006. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A*. 103:8390–8395.

Shabalina SA, Spiridonov AN, Spiridonov NA, Koonin EV. 2010. Connections between alternative transcription and alternative splicing in mammals. *Genome Biol Evol*. 2:791–799.

Uversky VN, Dunker AK. 2010. Understanding protein non-folding. *Biochim Biophys Acta*. 1804:1231–1264.

Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. 2004. The DISOPRED server for the prediction of protein disorder. *Bioinformatics*. 20:2138–2139.

Xing Y, Lee CJ. 2004. Negative selection pressure against premature protein truncation is reduced by alternative splicing and diploidy. *Trends Genet*. 20:472–475.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 13:555–556.