Selective Constraints in Conserved Folded RNAs of Drosophilid and Hominid Genomes

Robert Piskol* and Wolfgang Stephan

1Department of Biology II, Section of Evolutionary Biology, Ludwig-Maximilian-University, Munich, Germany

*Corresponding author: E-mail: piskol@bio.lmu.de.

Abstract

Small noncoding RNAs as well as folded RNA structures in genic regions are crucial for many cellular processes. They are involved in posttranscriptional gene regulation (microRNAs), RNA modification (small nucleolar RNAs), regulation of splicing, correct localization of proteins, and many other processes. In most cases, a distinct secondary structure of the molecule is necessary for its correct function. Hence, selection should act to retain the structure of the molecule, although the underlying sequence is allowed to vary. Here, we present the first genome-wide estimates of selective constraints in folded RNA molecules in the nuclear genomes of drosophilids and hominids. In comparison to putatively neutrally evolving sites, we observe substantially reduced rates of substitutions at paired and unpaired sites of folded molecules. We estimated evolutionary constraints to be in the ranges of (0.974, 0.991) and (0.895, 1.000) for paired nucleotides in drosophilids and hominids, respectively. These values are significantly higher than for constraints at nonsynonymous sites of protein-coding genes in both genera. Nonetheless, valleys of only moderately reduced fitness (s \( \approx \) 10\(^{-4}\)) are sufficient to generate the observed fraction of nucleotide changes that are removed by purifying selection. In addition, a comparison of selective coefficients between drosophilids and hominids revealed significantly higher constraints in drosophilids, which can be attributed to the difference in long-term effective population size between these two groups of species. This difference is particularly apparent at the independently evolving (unpaired) sites.

Key words: RNA secondary structure, selective constraints, selection coefficients.

Introduction

In recent years, it has become obvious that great portions of the genomes of complex organisms are being transcribed to produce noncoding RNAs (ncRNAs) that are involved in a variety of important processes (Amaral et al. 2008). In fact, ncRNAs emerge to be the key players in many developmental systems (Amaral and Mattick 2008) and regulators of diseases (Taft et al. 2010). A secondary structure of the molecule is often necessary to perform its function (MacDonald 1990; Bullock et al. 2003). This structure is composed of double stranded regions (helices) that arise through the formation of Watson–Crick (WC) pairs between complementary nucleotides. If mutations occur in the primary sequence of the molecule, they lead to a disruption of these paired regions, thus changing the conformation of the molecule and impairing or, in the worst case, disabling its original function. Although the original conformation of the molecule in space may be restored through a second (so called “compensatory”) mutation at the position opposing the first mutation, intermediate variants will suffer from a selective disadvantage that ultimately will result in reduced evolutionary rates in that region. Knowledge of these rates can further our understanding of constraints imposed on RNA molecules and may reveal the importance of their structures. Various studies focused on the distribution of fitness effects of new mutations (Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2007; Keightley and Eyre-Walker 2007, 2010) and evolutionary constraints in nonprotein-coding DNA of different genera (Halligan et al. 2004; Halligan and Keightley 2006; Eöry et al. 2010). Also the process of RNA evolution was studied extensively (Stephan and Kirby 1993; Kirby et al. 1995; Stephan 1996; Chen et al. 1999; Innan and Stephan 2001; Chen and Stephan 2003; Knies et al. 2008; Mimouni et al. 2009). However, an analysis of selective constraints in regions of the nuclear genome that are able to form distinct RNA secondary structures with specific functions has hardly been performed. Previous work on mitochondrially encoded transfer RNAs (mt-tRNAs) in mammals (Meer et al. 2010) suggested that large reductions in fitness have to be expected when mutations in the sequence lead to disruption of the mt-tRNA structure. However, mitochondrial DNA differs from nuclear DNA in several respects (high mutation rates, mode of inheritance, and selection; Parsons et al. 1997; John et al. 2010), which may result in different estimates for evolutionary constraints.

Therefore, this study aims to advance our knowledge of conserved ncRNAs that are encoded in hominid and drosophilid nuclear genomes. The availability of ncRNA data sets that were compiled using the same methodology (Pedersen et al. 2006; Stark et al. 2007) allows us to infer constraints for each of the genera and also permits comparison of constraints between drosophilids and hominids. The main focus of our work is the identification of overall selective constraints in ncRNAs, their variation between different genomic locations, and their differences...
between drosophilids and hominids. Furthermore, we aimed to determine the depth of the valleys of reduced fitness that have to be crossed by RNA molecules in the transition from one WC pair to another one and investigated how selective constraints are related to structural features of the ncRNA molecule.

To calculate the detrimental effect of mutations in a sequence of interest, it is necessary to know the rate at which mutations accumulate in neutrally evolving regions (regions that are free of the constraint to perform a certain function). Sites used as neutral standards vary between studies and also between species (Koonin and Wolf 2010). Originally, 4-fold degenerate sites were used as the reference for neutral sequence evolution. However, recently, these positions were also found to be subject to evolutionary constraints and were replaced in their function as a neutral standard by repetitive sequences and intronic sequences in hominids (Eöry et al. 2010) and drosophilids (Parsch et al. 2010), respectively. Therefore, we used intronic and intergenic ancestral repeats (ARs) as reference for evolution of ncRNAs in genic and intergenic regions of hominid ncRNAs and positions 8–30 of short introns (≤ 65 nt) as reference for evolution in drosophilid ncRNAs. To calculate the selective pressure against point mutations that disrupt the secondary structure of conserved noncoding RNA molecules and mutations at unpaired positions, we compiled data sets of folded molecules according to distinct genomic regions of Drosophila (focusing on the Drosophila melanogaster/ D. simulans comparison) and vertebrates (human/chimp/panzee comparison). We then estimated substitution rates for paired and unpaired nucleotides within the structures as well as for sequences of putatively neutral evolution. Subsequently, we compared the observed numbers of substitutions in the RNA sequences of interest with their expected numbers (which were obtained from the neutrally evolving genomic regions) and determined the selective constraint C and selection coefficient s scaled by the effective population size Ne.

Materials and Methods

Sequence Data

Sequence data for drosophilids and vertebrates were obtained from the University of California Santa Cruz (UCSC) Genome Browser home page (Kent et al. 2002) in form of MULTIZ multiple sequence alignments for D. melanogaster (dm3) and human (hg18) genome assemblies. The Drosophila alignment, which consists of up to 12 Drosophila species, was analyzed for the D. melanogaster group (D. melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta, and D. ananassae). The vertebrate alignment comprised up to 17 species from which we only used eight for our analysis (human, chimpanzee, mouse, rat, dog, chicken, zebrafish, and pufferfish) because these were the basis for the annotation of conserved RNA secondary structure elements by Pedersen et al. (2006). Annotations of functional RNA secondary structures (folds) in Drosophila (Stark et al. 2007) and human (Pedersen et al. 2006) genomes were obtained from the evofold tracks of the Genome Browser and are based on the previously mentioned multiple alignments of 3–12 and 4–8 species, respectively. Scores in the range [0,1] are available for all positions in the folds and describe the confidence of structure annotation. Annotations of genes, repetitive sequences, and nucleotide quality scores were also obtained from the UCSC Genome Browser home page. The position of genes within the D. melanogaster and human assembly were taken from the reference gene (RefGene) tracks (downloaded on 13 March 2010 and 26 January 2010 for Drosophila and human, respectively). ARs, that is, repetitive sequences common to human and chimpanzee, were determined according to the RepeatMasker annotation (table rmskRM327) for the human assembly. We only considered long terminal repeats, DNA transposons, short interspersed elements, long interspersed elements, and other repeats but excluded simple repeats, low complexity regions, and microsatellites as well as RNA-coding genes from our analysis as described by Eöry et al. (2010).

Mapping of Folds

Functional RNA secondary structures were classified based on their location in the genome. We differentiated between nine sequence types. The main distinction was made between sequences in intergenic regions and sequences that overlap with protein-coding genes. Sequences falling into the latter category were further divided according to the following three criteria: 1) number of splice forms (single transcript: ST and alternatively spliced: AS), 2) inclusion into the mature transcript (exon and intron), and 3) translation of region (UTR and non-UTR). The criterion for an intronic location was met if the sequence did not overlap with any exons of alternative splice forms. Hence, sequences that are annotated as intronic are exclusively located in an intron in all splice forms (if there is more than one), whereas sequences annotated as located in an exon either overlap partially with exons, are located completely within an exon, or are (partially) excluded from some of the mature transcripts but present in others. This categorization leads to a total of nine combinations (one intergenic category, eight categories for sequences overlapping with protein-coding genes resulting from the combination of the three aforementioned criteria). Folds on the complementary strand of annotated genes were not taken into account. We chose to investigate ST and AS genes separately due to evidence for higher selective pressures in the latter sequence type (Eöry et al. 2010), which may be caused by specific factors for splicing (e.g., exonic splice enhancers and silencers, Parmeley et al. 2007 and other highly conserved intronic sequences specific to AS genes, Sorek and Ast 2003). The roles of folded structures also differ substantially depending on their intronic or exonic location. While structured regions in introns predominantly facilitate the correct splicing (Howe and Ares 1997) and efficiency of splicing (Chen and Stephan 2003), folds in coding portions of genes participate in RNA editing (Gott and Emeson 2000; Li et al. 2009) and programmed frameshifting (Farabaugh 1996; Namy et al. 2004).
Furthermore, many folded structures in UTR regions are known to be responsible for correct mRNA localization (Bullock et al. 2003; Irion and Johnston 2007) suggesting the categorization of our data according to this criterion.

Selection Criteria for Folds and Putatively Neutral Sequences

To avoid the wrong estimation of selective constraints due to alignment of nonorthologous sequences or low sequence quality, we applied various selection criteria that were established before (Keightley et al. 2005). We excluded folded regions and ARs from the vertebrate data set if the alignment of human and chimpanzee sequences in that region contained < 50% of aligned nucleotides. We did not allow for more than five mismatches between human and chimpanzee sequences in a window of 25 nucleotides and ten mismatches in a window of 100 nucleotides. In addition, the overall divergence between the two sequences was required to be < 0.1. These criteria for orthology were complemented by requiring > 50% of the chimpanzee sequence to be of high quality (nucleotide quality score > 40).

In addition, folded regions were subjected to criteria similar to those applied in Piskol and Stephan (2008). Thereby, we allowed for the absence of at most one sequence from the alignment of eight vertebrate species, only included folds of low average free energy (as calculated by RNAeval, Hofacker et al. 1994) and discarded folds that overlapped with any kind of repetitve sequences. Apart from removing complete RNA structures and ARs that did not meet the above criteria, we also removed single alignment positions of low quality from fragments that passed the previous checks. These positions were characterized by either a score < 40 in the chimpanzee sequence (applicable to folds and ARs) or a structure confidence score of < 0.9 (only applicable to folds). In case a paired position in a fold had to be removed from the alignment due to its low sequence score, the pairing partner was omitted from the alignment as well. For the Drosophila data set, we relied on the quality of the available sequences, however, we discarded folds and intron sequences from the analysis if two or more species were missing from the alignment of the six species in the D. melanogaster group. Short introns in genes with alternative splicing forms or overlapping gene annotation on the same or opposite strand were not considered. We further omitted sequences according to previous divergence estimates for synonymous sites (Cutter 2008). Thereby, we removed folds and introns if divergence between D. melanogaster and D. simulans was > 0.226. Divergence between D. ananassae and any of the other five species was required to be < 1.324. In addition, positions with low confidence (< 0.9) of the base-specific structure annotations were removed from folded regions.

Substitution Rate Estimation

The estimation of substitution rates in paired and unpaired positions of functional RNAs ($k_{paired}$, $k_{unpaired}$) and neutrally evolving sequences ($k_{neutral}$) of the Drosophila and vertebrate data sets was based on the phylogenies for the D. melanogaster group (Drosophila 12 Genomes Consortium 2007) and the eight vertebrate species (Pedersen et al. 2006), respectively. Estimations were performed using baseml from the PAML package (Yang 2007). The maximum likelihood results for substitution rates are presented based on the REV + $I^r(4) + I^v$ model of sequence evolution (Tavaré 1986), which explains the data best (according to likelihood ratio test: supplementary tables S2 and S3, Supplementary Material online). The estimation of substitution rates for folded regions in each sequence type of drosophilids and hominids, as well as intron regions in drosophilids was performed on a concatenated alignment of all single fragments. Due to the large amount of ARs in the human and chimpanzee genomes, substitution rates for hominid ARs in the nine sequence types were calculated as averages over all 1-Mb windows along the human chromosomes. We did not use dinucleotide substitution models (e.g., RNA7D, Tillier and Collins 1998, RNA16D, Savill et al. 2001) as we were interested in the rate of substitutions at single sites and not in rates for nucleotide pairs. For substitution rates, 95% confidence intervals (CIs) were obtained by bootstrapping 1,000 times by alignment columns (in the case of drosophilid folds and introns as well as hominid folds) or by 1-Mb windows (hominid ARs). When bootstrapping alignments of folded regions by column, we guaranteed the same number of paired and unpaired positions as in the original alignment. The increase of substitution rates through context-dependent substitution processes was taken into account by removing CpG-prone sites from the vertebrate data set (Gaffney and Keightley 2008). Therefore, substitution rate estimates for hominids are provided for non-CpG-prone sites only.

Calculation of Selective Constraints and Selection Coefficients

The calculation of selective constraint for each sequence type was performed according to the two-lineage approach using the formula $C = 1 - \left( \frac{N_{obs}}{N_{exp}} \right)$ (Halligan et al. 2004). $N_{obs}$ is the observed number of differences between two sequences of a certain length. $N_{exp}$ denotes the number of expected substitutions in a neutrally evolving genomic region of the same size. Hence, $C$ describes the fraction of mutations that were removed by selection (due to their selective disadvantage). To avoid over/underestimation of constraint due to differences in GC content between the sequence of interest and the neutrally evolving sequences, Halligan et al.’s (2004) method corrects $N_{exp}$ by adjusting the substitution rates of nucleotide changes that alter the GC content according to the equilibrium GC content ($GC^*$). Thereby, we assumed GC* in Drosophila and humans to be 0.37 (Halligan et al. 2004; Duret and Arndt 2008). This correction allows us to compute estimates of $C$ that are free of the GC compositional bias and makes a comparison of $C$ for sequence types of varying GC content possible.

Furthermore, we computed the scaled selection coefficient $N_{s}$ using the symmetrical bidirectional model of sequence evolution (Innan and Stephan 2001; eqs. 5a and 6).
This model is based on Kimura's (1985) idea of compensatory neutral mutations which states that individual mutations are deleterious but harmless in certain combinations. It assumes that the two intermediate (deleterious) allelic states of a compensatory mutation are subject to selective constraints of the same strength \((s_1 = s_2)\) and mutation rates to and from the intermediate states are equal \((\mu_1 = \mu_2)\). The model parameters are effective population size \((N_e)\), mutation rate \((\mu)\), and selection coefficient \(s\) (the amount by which fitness of the intermediate is reduced). The ratio between the expected waiting time for the appearance of a double mutant that will successfully reach fixation in the population \((T)\) and the expected waiting time for two independent neutral substitutions \((T_{1\text{neu}})\) is a function of these parameters. Or, vice versa, after observing the ratio \(N_e \mu\) in the data (which corresponds to \(\frac{TT}{T_{1\text{neu}}}\)), we are able to find a numerical solution for the scaled selective constraint \(N_e \mu\) given that we know the scaled mutation rate \(\theta (= 4N_e \mu)\). To obtain \(N_{\text{exp}}\), it has to be assumed that the two taxa are separated by only a short phylogenetic distance such that no multiple hits have occurred.

### Results and Discussion

**Composition of Data Sets**

We applied the selection criteria outlined in Materials and Methods to the complete set of conserved drosophilid RNA secondary structures, thus reducing it from 22,682 to 16,575 folds. The majority of these folds falls into intergenic regions (11,647), whereas the remaining 4,928 folds are located in positions within protein-coding genes. Among the latter, we distinguished between folds in regions that are annotated as ST genes (2,137 folds) and AS genes (2,791 folds). The position of the RNA secondary structure within the gene was characterized by the inclusion of the region into the mature transcript (intron/exon) and the structure's location in the untranslated or translated portion of the gene (UTR/nUTR), which resulted in four combinations (intron/UTR, intron/nUTR, exon/UTR, and exon/nUTR). From the 47,510 folds in the vertebrate data, we selected 10,412 structures and partitioned them according to the same methodology. Positions of the selected regions are available from the authors upon request. Table 1 shows the number of folds, number of sites, and GC content for each category in D. melanogaster and human, respectively. For D. melanogaster, the numbers are given for all sites, whereas in humans, only non-CpG-prone sites were considered. The most striking difference between the two data sets is the elevated number of folds in intergenic regions of drosophilids (almost 70.3%). Even though, this sequence type contains the highest percentage of folds in hominids as well (29.1%), their fraction is much lower than in drosophilids and a remarkable part of structures can also be found in coding parts of the genome. The GC content in drosophilids and hominids is significantly lower at paired than unpaired positions. This difference can be attributed to the adjustment of the secondary structure algorithm to favor pairings in AT-rich regions (Pedersen et al. 2006). Furthermore, a clear difference between the GC content in folds and in neutrally evolving sequences (tables 1 and 2) can be observed. The GC content of neutrally evolving positions (table 2) lies between the elevated content at unpaired and reduced
Table 2. Nucleotide Composition of Sequences Used as Neutral Standards.

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Sequence Type</th>
<th>Number of Sites</th>
<th>GC Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drosophilids</td>
<td>Introns</td>
<td>82,678</td>
<td>0.3402</td>
</tr>
<tr>
<td>Hominids</td>
<td>Intergenic AR</td>
<td>262,033,297</td>
<td>0.4606</td>
</tr>
<tr>
<td>Hominids</td>
<td>Intron AR</td>
<td>35,563,484</td>
<td>0.4848</td>
</tr>
</tbody>
</table>

Note.—AR, ancestral repeats.

content at paired sites in both drosophilids and hominids. In hominids, however, the unpaired positions are more similar to neutral sequences in terms of GC content, whereas in drosophilids, the neutral standard matches more closely the GC content of paired nucleotides. In general, this pattern holds for all sequence types except for folds in the coding portion of genes (sequence types: exon nUTR ST and exon nUTR AS). These contain higher overall numbers of GC nucleotides than the designated neutral standards which is in accordance with the generally higher number of GC nucleotides in coding portions of the genome (Pozzoli et al. 2008). However, nearly all sequences show a nucleotide composition that deviates from the equilibrium GC content of 0.37 in drosophilids and hominids and thus require a correction of constraint estimates as described.

Substitution Rate Variation

Based on the alignments of D. melanogaster and D. simulans as well as human and chimpanzee, we calculated substitution rates $k_{\text{paired}}$, $k_{\text{unpaired}}$ at paired and unpaired sites in conserved noncoding RNAs and $k_{\text{neut}}$ for sequences that evolve under putatively neutral conditions (drosophilids: positions 8–30 of short introns, hominids: ARs). Even though further inference of sequence constraints from uncorrected substitution rates should be performed with caution due to differences in GC contents between test and neutral sequences, substitution rates may serve as a first indicator of restrictive conditions for sequence evolution. To account for the impact of a context dependent increase of the mutation rate CpG prone nucleotides were removed from the human/chimpanzee alignment. Figure 1 and supplementary table S1, Supplementary Material online, show substitution rates for the neutral standards as well as paired and unpaired regions of noncoding RNAs of different sequence types.
Type. The values of \( k_{\text{paired}} \) and \( k_{\text{unpaired}} \) represent the divergence between \( D. \text{melanogaster} / D. \text{simulans} \) (fig. 1a) and human/chimpanzee (fig. 1b), respectively. In comparison to \( k_{\text{neut}} \), it is evident that folded RNA molecules evolve under strongly constrained circumstances. This is not only true for paired portions of the folded RNA but also for unpaired regions. Substitution rates are in general lowest at paired positions due to their primary purpose to form and maintain the secondary structure of the RNA molecule \( (k_{\text{paired}} \ll k_{\text{neut}}) \) which agrees with the expectation that nucleotide sites evolve slower under dependence \( (\text{Nasrallah et al. 2011}) \). However, also unpaired sites show considerably reduced rates of molecular evolution \( (k_{\text{unpaired}} \ll k_{\text{neut}}) \). Even though they are predicted not to be involved in the formation of the secondary structure, a nucleotide change at an unpaired position may result in a previously not present pairing that leads to a different (for instance energetically) more favorable but functionally defective structure that should be selected against. Furthermore, unpaired sites may still participate in tertiary interactions with other nucleotides of the RNA molecule through canonical WC and other noncanonical base pairs \( (\text{Leontis and Westhof 2001}) \). Hence, they are also subject to evolutionary constraints and show considerably lower divergence between species than neutrally evolving sites do. While mutations in paired positions inevitably lead to nonisosteric pairs and thus to a strong reduction of the molecule’s fitness (if we assume that the fitness is directly associated with the structure), non-WC pairs can be replaced through other isosteric pairs along neutral evolutionary paths \( (\text{Dutheil et al. 2010}) \). Therefore, selection is stronger on canonical WC pairs that are involved in the formation of the secondary structure than on regions denoted as unpaired, which is reflected by \( k_{\text{paired}} < k_{\text{unpaired}} \).

Variation in divergence between sequence types can also be observed; especially with notably higher values for folds in coding regions of the drosophilid genome. However, these differences may stem from the heterogeneous GC content between sequence types that deviates from GC* to various extents and was previously shown to affect the rates of substitution \( (\text{Eöry et al. 2010}) \). In general, bootstrap CIs are larger for the hominid data and thus suggest greater variation of C within each sequence type along the hominid genome. The obtained estimates for the neutral standard sequences match closely with previous results. Based on 82,678 nucleotides that belong to positions 8–30 of short drosophilid introns, we calculated \( k_{\text{neut}} \) to be 0.1100 (95% CI = [0.1074, 0.1125]). This value is similar to a recently reported estimate of 0.123 \( (\text{Parsch et al. 2010}) \).

The discrepancy between these two estimates may stem from a different source of alignments and varying sizes of the data sets. Our substitution rate estimates in intronic ARs \( (0.01035 [0.0102, 0.0105]) \) and intergenic ARs \( (0.01175 [0.0116, 0.0119]) \) in hominids confirm values of a recent study \( (0.0115 [0.0114, 0.0117]) \) for the divergence between human and chimpanzee intergenic ARs \( (\text{Eöry et al. 2010}) \).

Hence, the drastic difference between \( k_{\text{paired}} \) and \( k_{\text{unpaired}} \) in comparison to \( k_{\text{neut}} \) is not caused by excessively high values of the latter but rather very small substitution rates of the former two. This effect is more pronounced in drosophilids and suggests higher constraints on drosophilid folds than on human folds.

**Variation of Selective Constraints**

In order to determine differences in selective pressures on noncoding RNA molecules of different sequence type in the drosophilid and hominid genomes, we calculated selective constraints \( (C) \) after grouping the molecules according to their genomic location as described in Materials and Methods. Thereby, not only the level of constraint on the pairing portion of the folded structures was of interest but also the degree by which unpaired regions are subject to evolutionary restrictions. Therefore, estimates for constraints in paired \( (C_{\text{paired}}) \) and unpaired regions \( (C_{\text{unpaired}}) \) were obtained by using positions 8–30 of short introns as a neutral standard for sequence evolution in drosophilids. Intergenic and intronic ARs served as neutral standards for sequence evolution in intergenic and genic hominid regions, respectively. In addition, we obtained estimates for constraint at paired sites by contrasting their evolution with rates at unpaired sites of the structures \( (C_{\text{paired*}}) \). Due to considerable restrictions on sequence evolution at unpaired sites \( (\text{see previous section on substitution rates}) \), we expected to obtain downwardly biased estimates for \( C_{\text{paired*}} \). Nevertheless, it was worthwhile to consider \( C_{\text{paired*}} \) for comparison with previous studies \( (\text{Innan and Stephan 2001}) \) and to observe how constraints relate to each other depending on the use of different neutral standards. Estimates for \( C_{\text{paired}}, \ C_{\text{paired*}}, \ \text{and \ } C_{\text{unpaired}} \) (table 3) show that evolution in folded regions is subject to strong constraints. These constraints reach levels higher than at nonsynonymous sites in protein-coding genes and even exceed 0.99 (i.e., >99% of mutations are removed due to purifying selection). From the comparison of paired and unpaired sites, we obtained \( C_{\text{paired}}(\text{dros}) \in (0.681, 0.912) \) for drosophilids and \( C_{\text{paired}}(\text{hom}) \in (0.764, 1.000) \) for hominids. These values approach or even surpass constraints at nonsynonymous sites in these genera \( (C_{\text{nonsyn}}(\text{dros}) = 0.86; C_{\text{fold/AS}}(\text{hom}) = 0.759) \) \( (\text{Eöry et al. 2010; Parsch et al. 2010}) \). However, unpaired sites are under selective constraints themselves as evidenced from their substantially lower rates of substitution than putatively neutral sites \( (\text{fig. 1}) \), which suggests that \( C_{\text{paired*}} \) is most likely to be underestimated. Indeed, if we use putatively neutral sequences as standard, we observe \( C_{\text{paired}} \) to be significantly higher than \( C_{\text{paired*}} \) (CIs of \( C_{\text{paired}} \) and \( C_{\text{paired*}} \) do not overlap for any category except for folds in UTR regions of ST genes). The values of \( C_{\text{paired}}(\text{dros}) \in (0.974, 0.991) \) and \( C_{\text{paired}}(\text{hom}) \in (0.895, 1.000) \) consistently exceed constraints at nonsynonymous sites in the two taxa and suggest that secondary structures are subject to strong evolutionary restrictions. This is also true for unpaired positions as already suggested by the reduced substitution rates. Although our choice to select for conserved regions might lead to increased values of \( C \), it is no exception to also observe such high values in regions that were not chosen according to strong conservation between taxa. For
Table 3. Selective Constraints (C) as well as Ratio of Observed to Expected Numbers of Nucleotide Substitutions for Paired (A,B) and Unpaired (C) Sites in folds of Different Sequence Types.

<table>
<thead>
<tr>
<th>Sequence Type</th>
<th>Constraint C (95% CI)</th>
<th>( \frac{N_{\text{obs}}}{N_{\text{exp}}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drosophilids</td>
<td>Hominids</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>intron_UTR_ST</td>
<td>0.989 (0.976, 0.998)</td>
<td>0.895 (0.682, 1.000)</td>
</tr>
<tr>
<td>intron_UTR_AS</td>
<td>0.989 (0.980, 0.997)</td>
<td>1.000 (1.000, 1.000)</td>
</tr>
<tr>
<td>intron_UTR_ST</td>
<td>0.991 (0.985, 0.997)</td>
<td>0.930 (0.861, 0.977)</td>
</tr>
<tr>
<td>intron_UTR_AS</td>
<td>0.986 (0.979, 0.992)</td>
<td>0.937 (0.860, 0.984)</td>
</tr>
<tr>
<td>exon_UTR_ST</td>
<td>0.936 (0.911, 0.960)</td>
<td>0.918 (0.881, 0.955)</td>
</tr>
<tr>
<td>exon_UTR_AS</td>
<td>0.974 (0.961, 0.986)</td>
<td>0.930 (0.876, 0.977)</td>
</tr>
<tr>
<td>exon_UTR_ST</td>
<td>0.981 (0.962, 0.996)</td>
<td>0.968 (0.921, 1.000)</td>
</tr>
<tr>
<td>exon_UTR_AS</td>
<td>0.974 (0.956, 0.988)</td>
<td>0.932 (0.849, 0.983)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>0.986 (0.983, 0.988)</td>
<td>0.937 (0.910, 0.963)</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Sequence Type</th>
<th>Constraint C (95% CI)</th>
<th>( \frac{N_{\text{obs}}}{N_{\text{exp}}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drosophilids</td>
<td>Hominids</td>
</tr>
<tr>
<td>intron_UTR_ST</td>
<td>0.910 (0.787, 0.984)</td>
<td>0.837 (0.231, 1.000)</td>
</tr>
<tr>
<td>intron_UTR_AS</td>
<td>0.897 (0.805, 0.967)</td>
<td>1.000 (1.000, 1.000)</td>
</tr>
<tr>
<td>intron_UTR_ST</td>
<td>0.912 (0.846, 0.969)</td>
<td>0.764 (0.488, 0.931)</td>
</tr>
<tr>
<td>intron_UTR_AS</td>
<td>0.882 (0.817, 0.935)</td>
<td>0.795 (0.464, 0.956)</td>
</tr>
<tr>
<td>exon_UTR_ST</td>
<td>0.681 (0.532, 0.807)</td>
<td>0.768 (0.619, 0.873)</td>
</tr>
<tr>
<td>exon_UTR_AS</td>
<td>0.783 (0.663, 0.888)</td>
<td>0.769 (0.553, 0.920)</td>
</tr>
<tr>
<td>exon_UTR_ST</td>
<td>0.876 (0.741, 0.975)</td>
<td>0.887 (0.647, 1.000)</td>
</tr>
<tr>
<td>exon_UTR_AS</td>
<td>0.845 (0.722, 0.933)</td>
<td>0.787 (0.437, 0.955)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>0.870 (0.842, 0.894)</td>
<td>0.836 (0.757, 0.905)</td>
</tr>
</tbody>
</table>

C

<table>
<thead>
<tr>
<th>Sequence Type</th>
<th>Constraint C (95% CI)</th>
<th>( \frac{N_{\text{obs}}}{N_{\text{exp}}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drosophilids</td>
<td>Hominids</td>
</tr>
<tr>
<td>intron_UTR_ST</td>
<td>0.878 (0.849, 0.906)</td>
<td>0.369 (0.000, 0.816)</td>
</tr>
<tr>
<td>intron_UTR_AS</td>
<td>0.888 (0.864, 0.914)</td>
<td>0.399 (0.040, 0.703)</td>
</tr>
<tr>
<td>intron_UTR_ST</td>
<td>0.894 (0.875, 0.911)</td>
<td>0.698 (0.578, 0.814)</td>
</tr>
<tr>
<td>intron_UTR_AS</td>
<td>0.880 (0.861, 0.897)</td>
<td>0.700 (0.556, 0.826)</td>
</tr>
<tr>
<td>exon_UTR_ST</td>
<td>0.796 (0.758, 0.831)</td>
<td>0.648 (0.560, 0.734)</td>
</tr>
<tr>
<td>exon_UTR_AS</td>
<td>0.876 (0.850, 0.898)</td>
<td>0.699 (0.599, 0.790)</td>
</tr>
<tr>
<td>exon_UTR_ST</td>
<td>0.844 (0.799, 0.889)</td>
<td>0.724 (0.587, 0.860)</td>
</tr>
<tr>
<td>exon_UTR_AS</td>
<td>0.836 (0.799, 0.875)</td>
<td>0.693 (0.543, 0.836)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>0.885 (0.879, 0.891)</td>
<td>0.631 (0.568, 0.684)</td>
</tr>
</tbody>
</table>

Note.—Values of C and \( \frac{N_{\text{obs}}}{N_{\text{exp}}} \) were obtained from the comparisons of (A) paired versus neutral, (B) paired versus unpaired, and (C) unpaired versus neutral sites, respectively. Sequence types are as in table 1.

instance, we observe \( C_{\text{paired}}(\text{miRNA}[\text{microRNA}]) = 0.68 \) and \( C_{\text{unpaired}}(\text{miRNA}) = 0.41 \) as well as \( C_{\text{paired}}(\text{tRNA}) = 0.83 \) and \( C_{\text{unpaired}}(\text{tRNA}) = 0.76 \) for sets of orthologous miRNAs from the mirBase (Griffiths-Jones et al. 2008) and tRNAs from Rfam (Gardner et al. 2009) in human and rhesus macaque, respectively. These values fall into a similar range as our whole-genome estimates even though the phylogenetic range is much smaller. A comparison of constraints in folded molecules to constraints at first, second, and third codon positions that were selected according to the same criteria (supplementary table S4, Supplementary Material online) suggests that RNA molecules are generally subject to larger constraints than protein-coding regions.

We were not able to find significant differences in selective constraints between ncRNAs that belong to different sequence types of the drosophilid and hominid genomes. In drosophilids, we can exclude the low divergence between D. melanogaster and D. simulans as a potential reason because we did not find any differences when comparing D. melanogaster and its more distant relative D. pseudoobscura (results not shown). In hominids, the previously visible large variation of substitution rates (previous section) can also be observed when calculating selective constraints and results in large CIs. These observations suggest that evolution of conserved ncRNAs is not influenced by factors that are specific to a certain sequence type (related to a certain genomic context in terms of the location within the gene) but rather by variation in constraints along the genome or intrinsic factors related to the RNA molecule itself.

Previous studies have shown that secondary structures in introns of ST and AS genes are responsible for correct splicing (Eperon et al. 1988) and inclusion or exclusion of exons (Howe and Ares 1997). Furthermore, the importance of other functional elements that are required for alternative splicing was found to result in a larger number of constrained sites in genes with alternative splice forms (Eöry et al. 2010) and might suggest that folded regions in AS genes are under stronger constraints as well. We were, however, not able to observe this pattern which implies that the folded regions in our data set are either 1) not related to splicing, 2) free of this additional constraint, or 3) the constraint on structures related to splicing is as strong as constraints that preserve specific functions in other genomic locations. One notable exception are significantly reduced values of \( C_{\text{paired}} \) and \( C_{\text{unpaired}} \) in coding exons.
of drosophilid ST genes (nonoverlapping CIs for sequence types exon\textsubscript{N}UTR\_S\_T and exon\textsubscript{N}UTR\_AS in table 3).

Even though no significant differences in the strength of selection between sequence regions and between ST and AS genes were detected, there are clear overall differences in selective constraints between drosophilid and hominid folds. As suggested by the theory of nearly neutral evolution (Ohta and Gillespie 1996), there exists a range of selective constraints \((\frac{1}{2N_e(}\text{dros}) < s < \frac{1}{2N_e(}\text{hom})\) that are nearly neutral for a species of small \(N_e\) (e.g., hominids) but deleterious for one of large \(N_e\) (e.g., drosophilids). Assuming different effective population sizes in the most recent common ancestor (MRCA) of hominids \((N_e \approx 52,000–96,000;\) Chen and Li 2001) and drosophilids \((N_e \approx 200,000 D. melanogaster\) worldwide sample; Schug et al. 1998), this fact is then reflected in low constraints in the former and strong constraints in the latter species. Indeed, all estimates for \(C\) (except for paired positions in UTR introns of AS genes) follow this theoretical prediction and exhibit lower values for hominids than for drosophilids. When grouping all estimates from different sequence types (table 3), a significant difference between drosophilid and hominid constraints can be observed at paired and unpaired sites (Wilcoxon one-sided \(W = 69, P = 0.0053\) (\(C_{\text{paired}}\); \(W = 81, P = 2.057 \times 10^{-5}\) (\(C_{\text{unpaired}}\); fig. 2). It is important to note that the difference for \(C_{\text{unpaired}}\) between the two genera is much larger than for \(C_{\text{paired}}\) and \(C_{\text{paired\_ST}}\). This observation can be explained by a longer fixation time (\(T\)) of a double mutant compared with \(T\) of a mutation at a single locus for the same \(N_{es}\) (Kimura 1980, 1985). With growing \(N_{es}\), \(T\) (and thus also \(\frac{N_{es}}{N_{eh}}\)) increases more slowly at independently evolving sites than for nucleotide pairs (supplementary fig. S1, Supplementary Material online). This in turn results in greater differences of \(C\) at independent sites than at paired sites for the same difference in \(N_e\) (assuming that \(s\) is the same in both genera) and leads to the observed effect in figure 2.

Selection Coefficients Obtained from the Bidirectional Model of Sequence Evolution

Based on the ratio of the expected to observed numbers of substitutions \((\frac{N_{ex}}{N_{eh}})\) in paired positions of folded RNA molecules, it is possible to estimate the scaled coefficient of selection \((\theta)\) for the scaled mutation rate in drosophilids according to previous studies in \(D. simulans\) and \(D. melanogaster\) (Kliman et al. 2000; Andolfatto 2001; Hutter et al. 2007) and report further results for the two boundaries of this interval. For humans, \(\theta = 0.001\) was chosen according to Nachman and Crowell (2000). Scaled selection coefficients \(N_{es}\) were obtained from \(\frac{N_{ex}}{N_{eh}}\) using the bidirectional model of sequence evolution (Innan and Stephan 2001) for the comparisons of 1) paired with unpaired and 2) paired with putatively neutral sites (fig. 3). For the comparison of paired with unpaired sites, \(\frac{N_{ex}}{N_{eh}}\) was taken from table 3 as a value from the interval \((3.131, 11.368)\) and \((4.243, 8.811)\) for drosophilids and hominids, respectively. The same ratio reaches substantially higher values in the range of \((15.601, 114.900)\) for drosophilids and \((9.516, 31.619)\) for hominids when relating paired to putatively neutral sites. These ratios were translated into \(N_{es}\) using the corresponding values for \(\theta\) in drosophilids and hominids (fig. 3a and b and supplementary table S5, Supplementary Material online). From the comparison of paired with unpaired sites, we obtain values for \(N_{es}\) that fall into the ranges \((0.811, 1.063)\) for hominids, and \((0.735, 1.184)\) and \((0.824, 1.351)\) for drosophilids assuming \(\theta = 0.003\) and \(\theta = 0.03\), respectively (fig. 3a). These estimates for \(N_{es}\) are only slightly higher than a previously obtained result of \(N_{es} \approx 0.6–0.7\) from the Drosophila bicoid 3’UTR (Innan and Stephan 2001) and suggest that most mutations in paired regions of ncRNAs in hominids and drosophilids are only slightly deleterious \((s < \frac{1}{2N_e}\)). Interestingly, for small \(\theta\), relatively low values of \(N_{es}\) are already sufficient to obtain large \(\frac{N_{ex}}{N_{eh}}\) ratios due to the nonlinear (nearly exponential) nature of the relationship between these two parameters. Hence, when relating evolution at paired sites to the neutrally evolving standard in hominids (fig. 3b), we obtain \(N_{es}\) in the range of \((1.088, 1.469)\) and, thus, only slightly higher values for the selection coefficient than from the paired versus unpaired comparison (even though the difference of \(\frac{N_{ex}}{N_{eh}}\) between the comparison of paired versus unpaired and paired versus neutral is 2.3- to 3.6-fold). The same is true for estimates in drosophilids: \(N_{es} \in (1.288, 1.949)\) when \(\theta\) is assumed to be small.
Selective Constraints in Folded RNAs

Fig. 3. Scaled selection coefficients at paired sites of folded RNA molecules in hominids ($\theta = 0.001$) and drosophilids ($\theta = 0.003$ and $\theta = 0.03$). Figures 3(a) and (c) show values obtained from the comparison of paired versus unpaired sites. Figures 3(b) and (d) show values calculated from the comparison of paired versus putatively neutral sites. Shaded areas under the curves in (a) and (b) display the range of selection coefficients for the given data. Box plots in (c) and (d) describe the distribution of selection coefficients in 1) hominids, 2) drosophilids with $\theta = 0.003$, and 3) drosophilids with $\theta = 0.03$. Brackets indicate significant differences in selection coefficients between drosophilids and hominids. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

(i.e., 0.003). If, however, larger values for $\theta$ are taken (i.e., 0.03), the estimates for $N_s$ in drosophilids fall into a broader range (1.504, 3.994) due to a slower increase of $N_{obs}/N_{exp}$ with growing $N_s$. Although the calculation of $C$ for drosophilids and hominids (previous section) only showed moderate differences at paired positions between the two genera (Fig. 2), the comparison of $N_s$ allows to observe significant differences of the scaled selection coefficient at paired positions (Fig. 3c and d). Again, this difference is due to the different effective population sizes for the MRCA of hominids and drosophilids and leads to nearly neutral effects of mutations in the former but deleterious effects in the latter.

Although compensatory evolution in human mt-tRNAs was shown to cross deep valleys of reduced fitness ($s \approx 10^{-3}$ to $10^{-2}$; Meer et al. 2010), our study suggests that evolution in ncRNAs that are positioned in the nuclear genome proceeds through valleys of smaller depth ($s \approx 10^{-6}$). This difference may be attributed to a higher mutation rate in the mitochondrial genome that requires stronger selective pressures that purge deleterious mutations from the sequence.

The estimation of $N_s$ might be affected by factors that were not accounted for by our analysis. For instance $N_s$ might be underestimated due to gene conversion that can increase $N_{obs}$ and thus bias the estimates for $N_s$ downwardly. However, gene conversion alone cannot explain the difference in $s$ by one order of magnitude between folded RNAs encoded on the mitochondrial and nuclear genomes. Furthermore, we evaluated the ratio $N_{obs}/N_{exp}$ of diploid organisms in a haploid framework. It was shown before (Ichinose et al. 2008) that compensatory evolution is accelerated in diploid populations and leads to higher $N_{obs}$. This is especially true when mutation rate is not low and deleterious effects of mutants are recessive. Again, this suggests that our values of $N_s$ obtained from the bidirectional haploid model could be underestimated. However, when mutation rate is low ($\theta < 0.01$), which is true in our case, compensatory evolution does not depend strongly on the haploid or diploid selection scheme but the fixation time is rather limited by low mutation rates. Therefore, our use of a haploid model should affect the estimates of $N_s$ only marginally.

It is important to note that the strength of selection is usually not of a constant magnitude but subject to variation. The values presented here depict the average over all fragments that were used in the analysis—regardless of the underlying structure of the folded molecules. However, we have shown in previous studies that heterogeneity in rates of compensatory evolution is caused by several factors that are related to the two-dimensional structure of the RNA molecule. These include the distance in sequence between pairing nucleotides, the length of pairing regions and the position of the base pair within the pairing region (Parsch et al. 2000; Piskol and Stephan 2008). The influence of these factors may be attributed to the detrimental effect of recombination on double mutations (Stephan and Kirby 1993; Stephan 1996; Chen et al. 1999), the increased tolerance for base pair disrupting mutations, and their influence on stability and structure of the molecule (Mimouni et al. 2009), respectively. This effect of structural variation is not only reflected in different rates of compensatory evolution but can also be seen in variation of selective coefficients (supplementary fig. S2, Supplementary Material online).
Conclusions

The results of our study show the strong restrictions imposed on the evolution of ncRNA molecules. Often the mechanisms these structures are involved in are unknown and hence also the direct source for their constrained evolution. However, we were able to show that their restricted evolution is mostly driven by the basic need of the structure to maintain pairings between nucleotides and just to a smaller extent by the specific region the RNA molecule is located in. It is important to note that the estimation of selective constraints is strongly influenced by the choice of the neutral standard. Our comparison of estimates that were obtained using 1) unpaired regions of folded molecules and 2) repetitive regions and intronic regions shows large differences in sequence constraint C and moderate to strong differences in Ns depending on the choice of θ. When comparing constraints between drosophilids and hominids, we were able to confirm previous theoretical predictions that species of larger Ne are subject to stronger evolutionary restrictions and that differences in Ne have a greater effect on s at independently evolving sites than at sites that evolve in pairs.

Supplementary Material

Supplementary tables S1–S5 and figures S1–S3 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank Olivier Tenaillon and one anonymous reviewer for very valuable comments on the manuscript. This work was supported by grant Deutsche Forschungsgemeinschaft Ste 325/8 to W.S.

References


1528


