Regions of extreme synonymous codon selection in mammalian genes

Peter Schattner* and Mark Diekhans

Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA

Received January 31, 2006; Revised February 26, 2006; Accepted March 7, 2006

ABSTRACT

Recently there has been increasing evidence that purifying selection occurs among synonymous codons in mammalian genes. This selection appears to be a consequence of either cis-regulatory motifs, such as exonic splicing enhancers (ESEs), or mRNA secondary structures, being superimposed on the coding sequence of the gene. We have developed a program to identify regions likely to be enriched for such motifs by searching for extended regions of extreme codon conservation between homologous genes of related species. Here we present the results of applying this approach to five mammalian species (human, chimpanzee, mouse, rat and dog). Even with very conservative selection criteria, we find over 200 regions of extreme codon conservation, ranging in length from 60 to 178 codons. The regions are often found within genes involved in DNA-binding, RNA-binding or zinc-ion-binding. They are highly depleted for synonymous single nucleotide polymorphisms (SNPs) but not for non-synonymous SNPs, further indicating that the observed codon conservation is being driven by negative selection. Forty-three percent of the regions overlap conserved alternative transcript isoforms and are enriched for known ESEs. Other regions are enriched for TpA dinucleotides and may contain conserved motifs/structures relating to mRNA stability and/or degradation. We anticipate that this tool will be useful for detecting regions enriched in other classes of coding-sequence motifs and structures as well.

INTRODUCTION

Non-synonymous mutations, which by definition change the translated amino acid sequence, are generally under evolutionary constraint. In contrast, synonymous mutations (also known as ‘silent site’ mutations) have long been considered relatively exempt from evolutionary pressure. However, exceptions to the neutral evolution of synonymous mutations are known. For example, in bacteria and viruses, a single stretch of DNA may code for two different proteins, one on each strand (1). In such cases, codon selection is highly constrained and mutations are generally under negative evolutionary pressure unless they are synonymous on both strands. In contrast, in eukaryotes and especially in mammals, because of their relatively small population size, it has been much less clear that selection exists at synonymous sites. Moreover, it was believed that constraints on synonymous codon selection—if they existed at all—were driven by genome-wide constraints such as variations in relative abundances of specific tRNA species or in regional levels of G + C% (2,3).

However, this picture of largely neutral evolution of synonymous substitutions in mammals has been severely challenged recently (for a review see (4)). Several studies involving comparisons with pseudogene (5) and intergenic/intronic evolution (6) have suggested that as much as 39% of human silent sites might be under, at least weak, selection (6). In addition, evidence has accumulated that silent-site selection may arise not only from genome-wide constraints such as variations in G + C% or relative tRNA abundances, but also from local requirements imposed by mRNA processing. Alternative splicing regulation motifs such as exon splicing enhancers (ESEs) have been demonstrated to be under selective pressure at synonymous sites (7–10). mRNA stability requirements (11–13), RNA-editing sites (14) and coding-region miRNA targets (15) or other coding-region binding sites have also been suggested as contributing to local constraints on synonymous codon selection. As a result, new theoretical models of codon evolution have been developed that incorporate the possibility of varying rates of selective evolution at synonymous sites across genes and across the genome (16).

The realization that mRNA processing may impose constraints on synonymous codon selection has also led to
the idea that if these constraints are conserved between related species such conservation could be a signature for the presence of cis-regulatory signals involved in mRNA processing. This approach led to the detection of a splicing regulatory region in the BRCA1 gene (17,18) and an RNA-editing site in the KCNA1 gene (14).

Here we develop the idea that an extended region of codon conservation might serve as a signature for the presence of mRNA (or DNA) processing motif(s) into a computational scanning tool, by systematically searching for regions in homologous genes of related species with such extreme codon conservation as would be highly unlikely to occur by chance. We have applied this approach to a set of 11,786 homologous gene pairs between human and mouse, including 9127 also conserved in dog. Even when using very conservative and stringent criteria for synonymous codon purifying selection, we find over 200 extended regions of extreme codon conservation and estimate that there may well be over 1600 such regions throughout the mammalian genome. Our results provide additional support for the idea that synonymous codon selection is often a result of alternative splicing and the conservation of ESEs. However, we also find that many of these regions do not appear to be associated with alternative splicing at all, and hence may be enriched for conserved sequence and/or structural motifs important for other processes related to mRNA maturation and regulation.

MATERIALS AND METHODS

Data sources

RefSeq IDs for 11,786 pairs of human and mouse genes were obtained from the HomoloGene Database (build 39.1, February 2005) (19) and their corresponding sequences were extracted from the NCBI May 2004 (hg17) and March 2005 (mm6) assemblies of the human and mouse genomes, respectively. The sequences included all HomoloGene human–mouse gene pairs for which RefSeq mRNA accessions (prefix = NM_) existed in both species. Homologous genes from chim- panzee, rat, dog and chicken were also obtained from the HomoloGene Database with the corresponding sequences extracted from GenBank. For these genes, sequences with both RefSeq mRNA accessions and RefSeq computer-model mRNA accessions (prefix = XM_) were included. Single nucleotide polymorphism (SNP) locations (19,20), gene expression data (21), CpG island information, conserved alternate splice isoforms (22) and annotation data from InterPro (23) and the Gene Ontology (GO) Consortium (24) were extracted from the Genome Browser Database at UCSC (25). ESE and exon splicing silencer (ESS) motifs were taken from the literature (26–28).

Algorithm description

An outline of the codonScan algorithm is shown in Figure 1. The coding portion of each HomoloGene gene pair was extracted, translated and aligned using protein BLAT (29) using default BLAT parameters. The aligned sequences were divided into 60 codon windows, and then scored for codon conservation. For comparisons with neutral-evolution models, only non-overlapping windows were included. In contrast, for the complete searches, overlapping windows were used, with start positions separated by either 1 or 20 codons. For these searches, only the highest scoring of multiple overlapping regions was retained.

The windowing and scoring procedure differed depending on whether the program was searching for regions of low relative synonymous codon substitution (i.e. relative to non-synonymous codon substitution) or low total synonymous codon substitution (i.e. independent of the local non-synonymous codon substitution rate). For the detection of low total synonymous codon substitution, the window length was set equal to (60 + n) codons where n is the number of non-synonymous substitutions in the window (n ≥ 0). For this screen, windows with either zero or one synonymous substitution were retained. These regions were then extended by the program to the maximum length possible, while maintaining the restriction of only a single synonymous substitution. For the detection of low relative synonymous codon substitution, we identified (fixed-length) 60 codon windows with at least three times as many non-synonymous as synonymous substitutions. These windows were further filtered to remove those with fewer than two or more than five synonymous substitutions (to remove regions found in both screens and those more likely caused by positive amino acid selection).

Probabilistic analysis—uniform and randomly varying synonymous codon evolution

We assessed the likelihood that the data could have occurred by chance by comparing them with several ‘null models’ involving spatially uniform or randomly varying codon substitution rates (30). In the simplest of these models, we assumed that the probability distributions of all aligned codon pairs (as either conserved, or synonymous or non-synonymous substitutions) are independent and identically distributed. This assumption is not strictly correct; substitution rates between the human and mouse do vary across the genome, generally as a function of regional G + C% (31). However, the observed
synonymous codon substitution rate only varies between 0.26 and 0.29 as a function of regional G + C% and has only a small influence on the location of regions with elevated codon conservation (Results).

We used the observed fractions of conserved, synonymous or non-synonymous codon-pairs in the entire set of alignments of homologous genes to estimate the corresponding probabilities: \( p_c \) for aligned codons conserved at the nucleotide level, \( p_s \) for codons with nucleotide substitution but coding for identical amino acids, and \( p_n \) for substitutions that also change the encoded amino acid. With this model, the probability that a window of \( N \) codon pairs has exactly \( x_c \) conserved, \( x_s \) synonymous and \( x_n \) non-synonymous codon pairs \((x_c + x_s + x_n = N)\) is given by the multinomial distribution as follows:

\[
P_M(x_c, x_s, x_n) = \frac{N!}{x_c!x_s!x_n!} p_c^{x_c} p_s^{x_s} p_n^{x_n}.
\]

For a window of \( M + x_n \) codon pairs with exactly \( x_n \) non-synonymous codon pairs \((x_c + x_s + M)\), the probability distribution for exactly \( x_c \) conserved and \( x_s \) synonymous codon pairs simplifies from the multinomial to the binomial distribution as follows:

\[
B(x_c | M, p_{cs}) = \binom{M}{x_c} p_{cs}^{x_c}(1 - p_{cs})^{M-x_c}.
\]

Here \( p_{cs} \) is the conditional Bernoulli probability that aligned codons are conserved given that they code for identical amino acids and is estimated from the counts of conserved and synonymous codon pairs:

\[
p_{cs} = \frac{x_c}{x_c + x_s},
\]

where the sum is over the entire set of gene alignments.

Hence, the probability of a window meeting the criterion for a ‘region of low total synonymous codon substitution’ (by our definition above, a window with either zero or one synonymous substitution) is as given below:

\[
p_s = B(0 | 60, p_{cs}) + B(1 | 60, p_{cs})
\]

For non-overlapping windows, the probability of finding \( 'k' \) such regions out of \( W \) windows can be approximated by a Poisson distribution:

\[
p(k | \lambda) = \frac{e^{-\lambda} \lambda^k}{k!},
\]

\( \lambda = W p_s \)

where \( W \) is the total number (68 069) of windows in the dataset and \( p \) is given by Equation 4. The Poisson model is valid as long as \( p << 1 \) and \( W >> 1 \). (For overlapping windows the assumption of independence among all window scores is no longer valid, and the Poisson distribution needs to be replaced by the ‘Clumped Poisson Distribution’ (32,33). By restricting our analysis to non-overlapping windows we avoid this complication).

More complex models of synonymous codon evolution were also compared with the data. These included a model in which \( p_{cs} \) has a different value for each of the 61 possible codons. For this model, we separately estimated \( p_{cs} \) for each codon from the observed synonymous changes among all the aligned pairs including that codon. In addition, a model [similar in spirit to those of Yang (34) and Pond and Muse (16)] was developed in which \( p_{cs} \) itself is a random variable with a value at each position in the genome randomly selected from a normal distribution. In both of these models, once \( p_{cs} \) has been determined at each position in the alignment, further calculations of the probabilities of finding a specific number of synonymous or non-synonymous substitutions within a 60 codon window are carried out in the same manner as for the basic model with constant \( p_{cs} \).

**RESULTS**

**Several hundred regions of extreme codon conservation exist between human and mouse**

We hypothesize that, if synonymous codons are under purifying selection because they overlap mRNA processing motifs, then there might exist extended gene regions—presumably containing multiple and/or extended motifs—that would show extreme silent-site conservation even when compared between relatively distant species, such as human and mouse. To test this hypothesis, we developed a program, codonScan, for the detection of unusual patterns of synonymous codon substitution. The codonScan algorithm is depicted schematically in Figure 1 and is described in detail in Materials and Methods. We applied codonScan to 11 786 homologous genes that had been aligned between human and mouse. We searched for windows with 60 conserved amino acids and a maximum of one synonymous substitution to ensure a low probability of finding conserved regions by chance. We initially detected 84 such regions out of 68 069 non-overlapping windows (0.12%). To find additional conserved regions, we re-screened all 1067 gene pairs that had regions with five or fewer synonymous substitutions, this time using all possible overlapping windows. This resulted in the identification of a total of 204 regions on 179 genes with no more than a single synonymous codon substitution. Of these, 84 regions have 60 or more consecutive codon pairs without a single synonymous substitution. Although our minimum length cutoff for these regions was 60 codons, many are significantly longer. The longest, found in the MATR3...
gene, consists of 178 codons. Four illustrative examples are shown in Figure 2.

As can be seen from the examples in Figure 2 (which are representative of the entire set, data not shown), the regions of extreme codon conservation are not simply parts of genes with uniformly low synonymous substitution rates. Rather these are regions which—for some reason—have been highly conserved at the synonymous codon level despite the fact that the remainder of the genes in which they are located do not show elevated conservation at silent sites.

Of these regions 55 (31%) are located within a single exon. In contrast, others (e.g. those in DMD and KPNA4) are spread over as many as six exons. The complete set is listed in Supplementary Table I and can be accessed as a ‘custom track’ on the UCSC Genome Browser (25) using the website at http://www.soe.ucsc.edu/research/compbio/codonScan/. In most cases the regions also have very low rates of non-synonymous substitutions, with an average number of non-synonymous substitutions/region = 1.55. However there are also regions with numerous non-synonymous substitutions; an extreme example is a region in the RAP80 gene which has 18 non-synonymous substitutions with only a single synonymous one.

CodonScan was also applied to search for regions of anomalously low synonymous substitution ratios, i.e. with at least three times as many non-synonymous substitutions as synonymous ones. Sixty codon windows with start-positions offset by twenty codons were screened and lower-scoring overlapping windows were removed. This resulted in the identification of 716 regions out of a total of 54 295 windows (1.3%). These regions were further filtered to remove those with less than two or more than five synonymous codon substitutions (Materials and Methods), resulting in 270 regions on 246 different genes. Seven of these (2.5%) are also among the 204 regions with low total synonymous codon substitutions. The complete set of regions with low synonymous substitution ratios is listed in Supplementary Table II and is also viewable on the UCSC Genome Browser via http://www.soe.ucsc.edu/research/compbio/codonScan/.

Regions with low synonymous substitution ratios are consistent with positive non-synonymous substitution as well as purifying synonymous codon selection. Consequently, purifying synonymous codon selection may be more difficult to identify in these regions than in those with low total synonymous codon substitutions. For this reason, we focus primarily on the regions with low total synonymous codon substitutions (i.e. those with no more than a single synonymous codon substitution), which we refer to simply as ‘conserved codon regions’ or ‘CCRs’.

CCRs are not expected under uniform and randomly varying models of synonymous codon substitution

For the entire set of human–mouse data, the mean conditional synonymous substitution rate, $p_{cs} = 0.28$. In a model with uniform substitution rate $p_{cs} = 0.28$, the expected length of a run with zero substitutions is equal to 9.2 codons; the expected length of a run with a single substitution = 19.4 codons. These values are much smaller than the runs of 60 codons or more that we observe. In fact, in this model, the probability of finding even one (or more) CCR (out of a total of 68 069 windows) is <0.005. The probability of finding five or more CCRs is <10^{-10}.

These conclusions are illustrated in Figure 3. Figure 3A shows the distribution of observed synonymous substitutions among all non-overlapping windows in the human–mouse

Figure 2. Counts of synonymous substitutions along sliding 60 codon windows of four genes (DMD, GRIK2, KCNC4 and PHC2). Each open circle represents the number of synonymous substitutions/window of 60 conserved amino acids. The horizontal red line indicates the expected number (16.8) of synonymous substitutions/window in a uniform substitution model. DMD has two regions of codon conservation (CCRs) while GRIK2, KCNC4 and PHC2 have a single CCR. KCNC4 is seen to have its conserved region at the extreme 5’ end of its coding region, in contrast to the other genes whose CCRs are in internal parts of their coding regions.
dataset (green points). Also shown in the figure (red line) is the predicted distribution of synonymous substitutions under the model where the number of synonymous substitutions in each window is binomially distributed with probability equal to the observed mean conditional synonymous substitution rate, $p_{cs} = 0.28$. This simple model clearly does not fit the observed data well. In particular, the observed number of highly conserved regions (the extreme left-hand side of the figure) is significantly greater than that predicted by the model.

We next investigated whether the occurrence of CCRs might be the result of unusual local transition-to-transversion substitution-ratios or G + C% values. However, we found that the transition-to-transversion ratio was almost identical between the CCRs (71.6%) and the entire dataset (72.4%). Similarly, we found that CCR G + C% levels were very similar to those of the entire dataset (detailed below).

We also checked whether the presence of CCRs might be due to variations in the available codon ‘wobble space’, i.e. in the number of alternative codons available for a given amino acid. As an extreme example, a (hypothetical) region of 60 methionines and tryptophans, which are each encoded by a single codon, would automatically appear as a ‘CCR’ even without any further constraints. We tested this hypothesis by means of a codon-substitution model in which the overall synonymous codon substitution rate, $p_{cs}$, is replaced by separate rates for each codon, which were determined from the empirical substitution rates for that codon over the entire dataset. However, with even with this modification, the model could not predict observed CCR frequency (data not shown).

Finally, we considered the possibility that the synonymous substitution rate might vary randomly across the genome. To test this hypothesis, we implemented a hierarchical model in which the synonymous substitution rate itself varies randomly according to a normal distribution. The mean of the normal distribution was set equal to the observed mean synonymous substitution rate and the SD was chosen to minimize the sum of the squared error residuals with the observed distribution. This empirical model is similar in spirit to the recent, more theoretical model of Pond and Muse (16), which also allows for the possibility of varying synonymous substitution rates at every codon location. In contrast to the other models we implemented, this last model has a much closer fit to the data (Figure 3, blue curve) confirming the observations of Pond and Muse (16) with a much larger dataset. However, even this model fails to predict the long stretches of purifying synonymous codon selection found in the CCRs as seen in the extreme tail of the distribution, as visualized in Figure 3B. In this model, the expected number of CCRs per 68 069 windows is 3.4 and the probability of finding eight or more CCRs is <0.024. Consequently, at this confidence level, the presence of nearly all of the CCRs is still not explained even in this class of models.

From these results, we conclude that simply including independent variation in the synonymous substitution rate at all synonymous sites is not sufficient to explain the presence of the extended regions of extreme conservation exhibited by the CCRs. Instead we find—perhaps, unsurprisingly from a biological perspective—that sites of synonymous purifying selection are highly clustered and, as a result, that a realistic model of mammalian synonymous substitutions will need to include some type of Markov process or other form of autocorrelation to reproduce these clusters.

Over 40% of CCRs overlap conserved alternative splice sites and are enriched for ESE motifs

It has become evident recently that splicing regulation (7,9) and, in particular, conservation of ESEs (8) are important factors in synonymous codon selection. Our data confirm these observations. Of 204 CCRs 88 (43%) overlap regions with experimentally supported, alternative splicing that is conserved between human and mouse (22). In contrast, only 11%
of the entire set of windows overlap conserved alternative splice sites (7652/68 069). This enrichment is highly statistically significant (t-test p-value < 10^{-10}). Interestingly, alternatively spliced regions with at least three supporting ESTs in human transcript libraries, but where alternative splicing is not conserved in mouse, are not enriched for CCRs. In fact, only 11.2% of the CCRs (23/204) overlap human-only alternative splice junctions, while 19.5% (13 295/68 069) of the windows in the entire dataset do.

The enrichment for conserved alternative splice regions suggests that CCRs may be enriched for conserved ESE and/or ESS motifs. To test this hypothesis, we screened the CCR sequences for the presence of ESE motifs as identified either by the RESCUE-ESE approach (26) or by the method of Zhang and Chasin (28). Similarly we searched for silencer motifs identified both by Zhang and Chasin’s method as well as by the splicing-reporter system of Wang et al. (27).

The 88 CCRs that overlap conserved alternative splicing regions have an average of 22 RESCUE-ESE motifs per 180 nt. Non-CCR, conserved, alternatively spliced regions have an average of 19 ESE motifs per 180 nt while regions without annotated conserved alternative splicing (whether or not they were CCRs) have an average of 17 ESE motifs per 180 nt. These differences, though small, are statistically significant (t-test p-value < 0.0002) because of the large number of sample regions available (88 alternatively spliced CCRs, 116 other CCRs, 7651 alternatively spliced non-CCRs and 61 653 other non-CCRs regions). ESEs identified by the method of Zhang and Chasin (28) also showed a consistent occurrence (though not statistically significant) in the CCRs with conserved alternative splicing. Neither dataset of ESS motifs showed statistically significant enrichment or depletion in the CCRs.

**CCRs are depleted for synonymous, but not non-synonymous, SNPs**

If the low synonymous substitution rate observed in the CCRs was the result of purifying selection, then we would expect a similar decrease in the rate of synonymous SNPs in these regions. For the specific case of synonymous codons overlapping ESEs, such a reduction in the synonymous SNP rate has been noted previously (36,37). To test whether the suppression of synonymous SNPs extends to all CCRs and not just to those that overlap ESEs, we compared the rate of human synonymous and non-synonymous SNP variations between the conserved regions and the entire set of non-overlapping windows. Regions from the entire data set were found to have 3.4 times as many synonymous SNPs as do CCRs (1.1 versus 0.33 SNPs/kb; t-test p < 10^{-15}). Interestingly the rate of non-synonymous SNPs in the CCRs is not significantly different from the overall genomic non-synonymous SNP rate (p = 0.70).

**CCRs have increased frequencies of TpA dinucleotides**

We searched for correlations between codon conservation and dinucleotide bias. CpG dinucleotide frequencies are reduced in most genomic sequences because of higher mutation rates (38). TpA dinucleotide frequencies are also reduced, especially in coding sequences (38–40). We searched for dinucleotide dependencies in CCR occurrence, by determining overlaps between CCRs and annotated CpG islands as well as by counting local distributions of both CpG and TpA dinucleotides. No significant differences were found between CCRs and other gene regions in terms of the presence annotated CpG islands (t-test p = 0.73). Consistent with this result, we also found no significant difference between the absolute frequencies of CpG dinucleotides in the CCRs and in other gene regions.

In contrast, we observed a significant increase in the number TpA dinucleotides in the CCRs; the frequency of TpA dinucleotides is 26% higher in CCRs (t-test p < 10^{-7}) than in other coding regions. For junctional dinucleotides—those containing the third nucleotide of one codon with the first nucleotide from the subsequent codon and, hence, less constrained by amino acid selection (11,38), the TpA frequency is 49% higher in CCRs than for other gene regions (t-test p < 10^{-7}).

Although the literature includes conflicting explanations of the low frequencies of TpA dinucleotides in mRNA sequences (11,38,39,41), considerable evidence has been presented to indicate that these low levels are related to UpA being a ribonuclease cleavage site (11,38). Consequently finding a relative excess of TpA dinucleotides in the CCRs was unexpected. We hypothesize that the reason is that there may be other conserved motifs and/or secondary structures within these CCRs that compensate for the increased TpA incidence, in a manner analogous to the higher frequency of (conserved) ESEs in the neighborhood of weak alternative splice sites, as compared with strong, constitutive ones (26). In contrast, our data do not support an alternative explanation of TpA frequencies—that the observed relative TpA frequencies are simply artifacts resulting from variations in CpG frequency (39)—since the CpG rates of CCRs are approximately the same as those of non-conserved regions, while the TpA frequencies are quite different.

**Genes with CCRs do not show evidence of codon selection bias**

To test for unusual codon-selection bias among the CCRs, we both calculated directly the effective number of codons (ENC) (42) and looked for indirect signatures of codon-selection bias, such as higher expression rates among genes with CCRs. No significant ENC difference was observed between the CCRs and less conserved regions (t-test p-value = 0.26). In addition, using gene expression profiles as annotated in the GNF Human Gene Atlas2 (21), we observed essentially no difference (<5%) between the maximum transcript expression values of CCR genes and the entire gene set. Note that these results do not contradict the assertion that, in general, gene expression may be correlated with codon bias (43). Rather our data shows that any such effect does not explain the extended regions of purifying codon selection found in the CCRs.

Interestingly, we did observe variations in the distribution of tissues with the greatest transcript expression. Of 171 CCR genes 28 showed maximal expression in brain tissue (62% enrichment compared to entire gene set) while only two exhibited maximal expression in germline tissue (80% depletion). Although the idea that CCR genes have atypical tissue expression patterns is intriguing, applying Bonferroni corrections for multiple testing results in a p-value = 0.09 for the
Two-Sample Equality of Proportionality Test. This does not rule the null hypothesis at the 5% confidence level, and, consequently, we cannot yet exclude the possibility that this observation is the result of random variations in the expression data.

**Genes with CCRs are enriched for annotations associated with nucleic-acid-binding and transcriptional regulation**

To determine whether genes with CCRs are enriched for specific types of gene functions, we calculated the frequency of occurrence of all GO and InterPro annotations in CCR genes as well as in the entire gene set. We found several InterPro and GO annotations occur more frequently in genes with CCRs than in randomly selected genes. Figure 4 shows the frequencies of occurrence of the most enriched GO and InterPro annotations found in genes with CCRs.

Assessing the statistical significance of these results requires Bonferroni correction for multiple testing, which, in this case, is challenging because there are thousands of non-independent GO and InterPro annotations (see e.g. R. Gentleman, ‘Using GO for Statistical Analyses’, http://bioconductor.org/docs/papers/2003/Compendium/GOstats.pdf). Nevertheless, for the GO annotations of ‘DNA binding’, ‘Regulation of transcription’ and ‘Nuclear location’ the uncorrected t-test p-values \(< 2.6 \times 10^{-6}\) are small enough to imply that even with thousands of multiple tests, the observed enrichments are statistically significant. In the other cases, however, the differences observed between genes with CCRs and other genes, though suggestive, need to be viewed with caution.

**CCRs are two to four times as abundant on the X chromosome as on the autosomal chromosomes**

We observed a striking difference between the frequency of occurrence of CCRs on the X chromosome and on the autosomes. CCRs occur almost four times as frequently on the human X chromosome as on the autosomal chromosomes (0.41% on the X chromosome versus 0.11% on the autosomal chromosomes). Although these values depend on the specific CCR cutoff used, CCR density on the X chromosome is at least twice that of the autosomal chromosomes for any CCR cutoff value from one to six synonymous substitutions/window.

This result may in part be because purifying selection against recessive, deleterious mutations needs to be stronger on the X chromosome (44) or because of increased mutation rates in the generation of sperm cells (45). In fact we find that the conditional synonymous substitution rate, \(p_{cs}\), is lower throughout the X chromosome \((p_{cs} = 0.24)\) than on the autosomal chromosomes where \(p_{cs}\) ranges from 0.26 (on chromosome 2) to 0.32 (chromosome 19), with mean \(p_{cs} = 0.28\) on all the autosomal chromosomes. Nevertheless, it is doubtful whether this modest 14% decrease in overall substitution rate can account for the 2- to 4-fold increase in the occurrence of CCRs. Consequently, we suspect that, in addition, there may be constraints specific to at least some of the X chromosome CCRs that also contribute to the high incidence of CCRs on the X chromosome.

**CCRs are more abundant in regions of low G + C%**

We investigated the dependence of codon conservation on regional G + C% (as measured over 100 kb regions) and third codon position G + C% (GC3%). We found that windows in the highest quartile of regional G + C% (G + C > 50%) have a higher mean conditional synonymous substitution rate \((p_{cs} = 0.29)\) than those in the lowest quartile (G + C < 40%) where \(p_{cs} = 0.26\). Similarly, regions in the highest GC3 quartile (GC3 > 75%) have a higher mean synonymous substitution rate (0.29) than those in the lowest quartile (GC3 < 43%) where \(p_{cs} = 0.27\).

The set of windows with low G + C% and low GC3% also have a somewhat increased incidence of CCRs (Figure 5). Interestingly, this variation is more pronounced among those CCRs that overlap conserved, alternatively spliced regions than among those that do not (Figure 5). Nevertheless, as can be seen from the figure, the correlation between G + C% and GC3% with CCR abundance is limited; 34% of the CCRs have greater than average regional G + C% and 31% have greater than average GC3%.

**CCRs are conserved among five mammalian species**

To determine the extent CCRs exist beyond the human and mouse genomes, we aligned 50 386 non-overlapping windows from 9127 Canis familiaris Homologene homologs to mouse genes for which we also had human-mouse alignments. In addition, we extracted dog, rat and chimpanzee HomoloGene homologs of all of the CCR genes and aligned them to their mouse homologs as well. For nearly all human–mouse CCRs, the pattern of extreme codon conservation was preserved among all five mammalian species. The median number of synonymous substitutions/window between mouse and their rat, chimpanzee and dog homologs were one, one and two, respectively. The results for rat and chimpanzee are hardly surprising considering their close relationships to mouse and human. More striking is that two thirds (88/131) of the mouse–human CCRs, for which a dog homolog was available, have two or fewer synonymous substitutions in dog and that 92% (121/131) have five or fewer. Similarly, 96% (54/56) of the detected mouse–dog CCRs are also conserved in human. In other words, in nearly all cases, the very same gene regions exhibiting extreme codon conservation between mouse and
human are also conserved between mouse and dog, indicating that not only are CCRs under purifying selection, but that the mechanisms underlying this selection are conserved among mammals.

Interestingly there are a few exceptions to mouse–dog codon conservation in homologs of mouse–human CCRs. Four dog CCR homologs (3%) have 10 or more synonymous substitutions, 2 of which (SCN8A:codons1265–1328 and RP42/tes3:140–212) have 33 and 38 substitutions, respectively. One possible reason is that the proper homolog has not yet been identified from the C.familiaris genome. Alternatively, the data may indicate that, for these genes, the constraint underlying the codon conservation no longer exists in dog.

Many CCRs are also conserved in chicken

We tested whether extended regions of extreme codon conservation could be found in other vertebrates by extracting 6055 HomoloGene mouse–human–chicken gene homologs and analyzing the resulting alignments of 29019 non-overlapping windows with codonScan.

Using the stringent CCR criterion of only a single synonymous substitution/60 conserved amino acids, only two mouse–chicken CCRs were detected in the entire mouse–human–chicken dataset. In contrast, 45 mouse–human CCRs were detected in the same dataset. However, this comparison is misleading because it ignores the higher overall synonymous substitution rate between chicken and mouse.

Specifically, the observed mouse–chicken conditional synonymous substitution rate, \( p_{cs} \), is 0.46; for mouse–human alignments \( p_{cs} = 0.28 \). Consequently, the (binomial) probability for a 60 codon human–mouse alignment with no more than a single synonymous substitution is approximately equal to that for chicken–mouse alignment with eight or less synonymous substitutions (in both cases the binomial probability \( \sim 10^{-7} \)).

If we therefore relax our definition of a chicken–mouse ‘conserved-codon region’ to one with no more than eight synonymous substitutions/window, we find 83 such regions between chicken and mouse. What is more interesting is that 60 of them (72%) also have less than five synonymous substitutions between human and mouse. The point of all this is that while total codon conservation is lower between chicken and mouse, many of the patterns of codon conservation—i.e. the gene locations of the most conserved regions—are conserved between chicken and mammals.

DISCUSSION

Our data demonstrate the existence of hundreds of extended (180 nt or longer) regions, in mammalian genes, exhibiting codon conservation at ‘silent’ sites. As such, these data provide further support for the hypotheses that there is extensive purifying selection at synonymous sites (4) and that efforts to measure neutral nucleotide substitution rates should focus on ancestral-repeat and other intergenic regions rather than on codon third-base ‘wobble’ positions (31).

Moreover, as illustrated in Figure 2, we have found that CCRs are typically localized within mRNAs with overall synonymous substitution rates not significantly different from other mRNAs. This suggests that these extended regions of codon conservation are the result of localized constraints rather than gene-wide or genomic constraints such as variations in regional \( G + C\% \) or in relative abundances of tRNA species. This conclusion is supported by our failure to find correlations between codon-usage bias or overall gene expression between genes with CCRs and other genes and the very limited correlation between the presence of CCRs and the levels of GC3% or regional \( G + C\% \).

Instead, our data confirm, using a larger dataset, recent work indicating that codon conservation is often associated with splicing regulation and the presence of ESEs (7–9). In fact 43% of the detected CCRs overlap known regions of conserved alternative splicing and are enriched for known ESEs. Moreover the data confirm that multiple splicing regulation motifs may often be clustered over extended regions.

However, >50% of the conserved regions do not overlap regions of known conserved alternative splicing nor are they enriched for known ESEs. To be sure, some of these regions may include conserved alternative splicing isoforms that have not yet been observed. However, for several reasons, we believe it is unlikely that all the observed codon conservation is simply the result of conserved alternative splicing. First, this would imply that more than half of the conserved alternative splicing in human and mouse has not yet been observed. In addition, this would beg the question as to why codon conservation is not elevated in thousands of regions that nevertheless exhibit conserved alternative splicing isoforms.

Finally, we have seen that the CCR regions without conserved,
alternative-splicing annotations are not enriched ESE motifs, again suggesting that they are indeed not involved in alternative splicing regulation.

Instead, we consider it more likely that these regions are constrained to conserve motifs or secondary structures critical for other post-transcriptional processes, such as RNA-editing (14) or mRNA stabilization and degradation (11–13). Some may also result from constraints from overlapping reading frames, though the examples of overlapping reading frames found so far in mammals do not appear to be conserved between human and mouse (46). The large percentage of CCRs on the X chromosome suggests X-specific mechanism(s), such as dosage compensation, or its evasion, may also be involved in some cases (47).

Although these possibilities have all been proposed before, few actual examples of any of them are currently known. We believe that, perhaps, the most important result of the present work is the identification of hundreds of extended regions of extreme synonymous codon purifying selection that should be ideal places to look for examples of these phenomena. For example, CCRs that overlap conserved alternatively spliced regions but that are not enriched for known ESEs might contain ESE motifs that have not yet been identified. CCRs with elevated TpA levels could be examined for additional secondary structure or stability motifs that might explain why mRNAs with these regions are not subject to UpA ribonuclease degradation. CCRs could be screened for enrichment for putative coding-region miRNA target motifs (15). Yet another example, which we have initiated in collaboration with others, is to search the eight CCRs and six ‘approximate CCRs’ (180 nt regions with only two or three synonymous substitutions between human and mouse) that are located on glutamate receptor and potassium channel genes—i.e. that belong to gene families that include the few known mammalian-coding-sequence RNA-editing sites—for additional RNA-editing sites.

Comparison with previous work describing localized codon conservation

The only prior systematic screen for regions of codon conservation of which we are aware is the Cosmo program (48). However, there are major differences between Cosmo and the present work. Cosmo searches for short (two or three codon) regions. Since such short stretches of conserved codons occur often by chance, Cosmo is limited to searching for genes for which aligned homologs are available in a large number of species. Moreover, the codon conservation must exist in all, or nearly all, of these species. Consequently, Cosmo was only applied to a small number of genes (48). In contrast, the codonScan approach searches for conserved regions of at least 60 codons, which are far less likely to be conserved by chance.

CodonScan also shares some features with the screen for ultra conserved elements (‘ultras’) of Bejerano et al. (49). However, one major difference is that ultras need not be in coding sequence; in fact the majority of ultras are not (49). Also, ultras generally do not have any non-conserved elements; this is in contrast to CCRs that may include non-synonymous substitutions and/or span multiple exons separated by weakly conserved introns (Figure 6). As a result, only 22% of the genes that we identified as having CCRs and 4% of the regions with high relative codon conservation also include ultra-conserved elements (for a complete list of such genes see the Supplementary Tables).

How many extended regions of purifying synonymous codon selection are there?

The 204 CCRs and 270 regions of high relative codon conservation do not represent all extended regions of synonymous codon purifying selection in human and mouse. Our stringent limit of one synonymous substitution per 60 conserved amino acids was chosen primarily to minimize the likelihood of finding conserved regions by chance. In fact, biological features that underlie codon conservation, such as ESEs and binding protein target sites, are generally much shorter than 60 codons in length. Similarly, our cutoff for regions of low relative synonymous codon substitution as having three times as many non-synonymous as synonymous substitutions was arbitrary and stringent; criteria for positive selection of amino acids typically use non-synonymous to synonymous substitution cutoff-ratios of one-to-one (50) rather than three-to-one as in the present work.

Figure 6. CCR of BAT2 from UCSC Human Genome Browser. Tracks show chromosomal position, presence of overlapping genes, conservation among mammals of exons and introns and positions of SNPs (non-synonymous SNPs are red). In contrast to many genes that show increased conservation not only in their exons but also in their intronic regions proximal to their splice sites, the conserved region of BAT2 shows extreme (exonic) codon conservation, but very little sequence conservation in its intervening introns.
The stringency of the selected cutoff-values is demonstrated by the fact that the three regions of unusual codon conservation that motivated the present work—those in KCNA1, BRCA1 and DRD2—were not detected in the present screens. The highest scoring region of KCNA1 has two synonymous substitutions and zero non-synonymous ones, BRCA1 has seven synonymous substitutions (and 23 non-synonymous ones), and DRD2 has eight synonymous and two non-synonymous ones. Nevertheless, the present work shows that even with stringent cutoffs, many regions of extreme codon conservation are detected.

A cutoff-independent estimate of the number of CCRs in human and mouse can be made from the difference between the observed number of regions and the prediction of the randomly varying substitution model, as shown in Figure 3B. This approach yields an estimate of 652 regions or a little <1% of the 68,069 non-overlapping windows. However, this is an underestimate since only non-overlapping windows are included. For example, with a cutoff of a single synonymous substitution per window, using non-overlapping windows, 84 CCRs were found. In contrast, by searching all (overlapping) windows, we found 204 such regions, indicating a detection sensitivity of 41% when only non-overlapping windows are screened. Applying this sensitivity correction, leads to an estimate of ~1600 extended regions of extreme synonymous codon purification in the set of 11,786 gene pairs.

Future directions

We realize that, as currently implemented, our methods are only a first step in characterizing CCRs and in identifying the cis-regulatory motifs and secondary structures that may underlie them, and we are currently investigating several approaches for further characterizing them. One such approach is to classify and partition the CCRs into clusters (CCRs exhibiting conserved alternative splicing, regions enriched for TpA dinucleotides, CCRs contained within a single exon, multiple-exon CCRs in which the intervening introns are also highly conserved and so on) and applying motif-finding algorithms such as MEME (51) or Gibbs-Sampling (52) to search for motifs enriched in each of these clusters. We are also investigating building more rigorous, probabilistic models of codon-conservation in which CCR parameters, such as region length or ratios of synonymous to non-synonymous substitutions would be determined from the data by maximum likelihood or Bayesian methods [e.g. see chapter 7 of Ref. (53)]. Such a more formal, probabilistic framework should also facilitate the simultaneous incorporation of sequence data and phylogenetic data from multiple species, which is not possible within the present model.

In conclusion, we believe that the identification of regions exhibiting purifying synonymous codon selection will become an important tool in the detection of novel mRNA-processing motifs. Moreover, although in the present work we have only applied this method to genes from mammalian species, the approach should be equally applicable to the detection of coding-sequence, cis-regulatory motifs in homologous genes from other groups of species. The only requirements are that the species are close enough that their amino acid sequences can be confidently aligned and that the cis-regulatory motifs being sought are conserved among the species, yet distant enough that silent sites which are not under common selective pressure will have had enough time to diverge.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are indebted to Charles Sugnet for providing the set of alternatively spliced regions conserved between human and mouse and to Katherine Pollard for advice regarding the statistical analysis. We thank Gill Bejerano, Sean Eddy, Katherine Pollard and Marijke J. van Baren for critical reading of an earlier version of the manuscript and Hiram Clawson and Todd Lowe for helpful discussions. M.D. was supported by the National Cancer Institute, National Institutes of Health (N01-C0-12400). Funding to pay the Open Access publication charges for this article was provided by the National Cancer Institute, National Institutes of Health (N01-C0-12400).

Conflict of interest statement. None declared.

REFERENCES

1. Rogozin, I.B., Spiridonov, A.N., Sorokin, A.V., Wolf, Y.I., Jordan, I.K., Tatusov, R.L. and Koonin, E.V. (2002) Purifying and directional selection in overlapping prokaryotic genes. Trends Genet., 18, 228–232.
2. Duret, L. (2002) Evolution of synonymous codon usage in metazoa. Curr. Opin. Genet. Dev., 12, 640–649.
3. Eyre-Walker, A. and Hurst, L.D. (2001) The evolution of isochores. Nature Rev. Genet., 2, 549–555.
4. Chamary, J.V., Parmley, J.L., and Hurst, L.D. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Rev. Genet., 7, 98–108.
5. Bustamante, C.D., Nielsen, R. and Hartl, D.L. (2002) A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. Mol. Biol. Evol., 19, 110–117.
6. Hellmann, I., Zoller, S., Enard, W., Ebersberger, I., Nickl, B. and Paabo, S. (2003) Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res., 13, 831–837.
7. Baek, D. and Green, P. (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proc. Natl Acad. Sci. USA, 102, 12813–12818.
8. Parmley, J.J., Chamary, J.V. and Hurst, L.D. (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol. Biol. Evol., 23, 301–309.
9. Xing, Y. and Lee, C. (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequeces. Proc. Natl Acad. Sci. USA, 102, 13526–13531.
10. Pagani, F., Raponi, M. and Baralle, F.E. (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc. Natl Acad. Sci. USA, 102, 6368–6372.
11. Duan, J. and Antezana, M.A. (2003) Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J. Mol. Evol., 57, 694–701.
12. Duan, J., Wainwright, M.S., Comeron, J.M., Saitou, N., Sanders, A.R., Gelernter, J. and Geiman, P.V. (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum. Mol. Genet., 12, 205–216.
13. Chamary, J.V. and Hurst, L.D. (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol., 6, R75.
14. Hoopengardner, B., Bhailla, T., Staber, C. and Reenan, R. (2003) Nervous system targets of RNA editing identified by comparative genomics. Science, 301, 832–836.
