Disulfide bonds between conserved cysteine residues maintain the structural integrity of surface loops in large families of extracellular ligands and receptors, most strikingly in the variable loops of venom-derived toxins. Because cysteine can be encoded by two codons (TGT or TGC), T/C mutations in the third nucleotide of the triplet are expected to be silent mutations, that is, not selected for or against by the evolutionary forces acting on such peptides. Surprisingly, an analysis of cysteine codon usage in a number of hypervariable gene families reveals strict codon conservation in specific positions adjacent to or within the hypervariable regions of these genes. This phenomenon suggests the possible existence of specific positional codon-conservation mechanisms in certain genes and, furthermore, it can be used as a functional-genomics tool to identify critical residues in a particular protein family.

The conopeptides are a large family of venom-derived toxins, recently suggested to be undergoing accelerated evolution for hypervariability in the mature toxin domain. To estimate the possible range of variability of conopeptides, we examined their precursor cDNAs currently available in GenBank and, after eliminating redundant sequences, chose the largest available family (the so-called scaffold VI/VII grouping) for further study. The resulting multiple sequence alignment consisted of 53 conopeptide precursors from nine different species. As expected from previous studies of this family, the open reading frames revealed strong conservation in their N-terminal signal sequences, dropping somewhat in the pro-domain, and with almost no conservation in the mature toxin segment except for the invariant cysteine residues (Fig. 1a). Most strikingly, examination of the corresponding nucleotide alignments revealed that five out of the six cysteines in these peptides exhibit a pronounced position-specific codon conservation (Fig. 1b). This position-specific codon conservation is all the more remarkable because it appears in the most hypervariable region of the sequence (Fig. 1c). It is not a reflection of a global codon bias in these species.
because the codon usage of TGC/TGT in molluscs is close to 50%. Furthermore, the preferred codon for Cys1, Cys3, Cys4 and Cys5 in these conopeptides is TGC, whereas Cys2 is preferentially encoded by TGT, and Cys6 shows a less-biased ratio of TGC/TGT (Fig. 1b).

In order to find out if this observation is seen also in other variable gene families, we performed automated BLAST searches of GenBank to identify sets of similar sequence stretches of 50 residues terminating on a cysteine. The resulting data were then restricted to 260 alignments containing 50 or more members, and these sets were analysed for conservation of the cysteine codon versus the average T:C ratio in the nine nucleotide positions immediately following the cysteine codon. The probabilities of obtaining the observed cysteine codon bias were estimated from a binomial distribution assuming a priori probabilities of 42% TGC; 58% TGT, according to the human codon bias table; except for the Trypanosoma mucins for which the trypanosome bias of 66% TGC; 34% TGT was used. All biases shown are statistically significant (p values from 10^-13–10^-38). Gene families in the panel include human T-type calcium channels (CaCH), T-cell receptor γ (TRGV), T-cell receptor β (TRBV), Ig heavy chain (IGHV), Ig constant regions (IGC1, IGC2), HLA (HLA1, HLA2, HLA3), pregnancy-specific glycoproteins (PSG1, PSG2), zinc-finger domain proteins (ZnF1, ZnF2) and cytochrome P450 (CYT).

Although global codon biases in different phyla are well documented4–7, the only example for a region-restricted codon preference in a defined gene family that we are aware of is a preference for readily mutable codons in the hypervariable regions of immunoglobulins8. This tendency has been suggested to act as a possible facilitator of accelerated somatic mutation9,10. Our observations suggest that an even more cogent phenomenon might occur in gene families that reveal expanded variability, such as venom-derived toxins and various recognition molecules in the immune system. The stringent position-specific conservation of cysteine codons before or within hypervariable regions of these different gene families suggests that a specific mechanism might have arisen to ensure and maintain the observed codon conservation. Such a mechanism could have arisen in order to conserve the structurally crucial cysteines, thereby imposing the observed codon conservation as a byproduct of conserved DNA or RNA recognition and/or modification. Thus, one possibility is that specific ‘protecting’ molecules
bind to cysteine codons in hypervariable regions of selected genes, in order to protect them from the enhanced mutagenesis that might occur in close proximity. These protective molecules, whose role might be analogous to that of a lithographic mask, would thereby impose the observed codon conservation as a bioprotectant of the mechanism for conservation of the encoded amino acid residue. More intriguingly, some might speculate that complexes of such ‘hyperprotected’ codons could actually target mutability by providing recognition sites for a mutator complex, which could then operate on adjacent sequence stretches.

Regardless of the molecular or evolutionary mechanisms, the specific codon bias described above, it might be possible to take advantage of the phenomenon to identify structurally or functionally important residues in gene families. This would be generally useful, especially if such restricted codon biases could also be found for residues other than cysteine. To test this notion, we examined the olfactory receptor superfamilies in mammals, which is thought to be one of the largest gene families with defined hypervariable regions. Redundancy was eliminated from the data set by removing all sequences showing more than 80% identity to each other. Three of the cysteine codon positions in the following alignment of 71 unique DNA sequences were highly biased (80–90%), two in the hypervariable region that favor TGT, and one in a less variable region that favors TGC. Perhaps more interestingly, the first tyrosine residue in this segment was less biased. It is noteworthy in this context that Y of the MAYDRY consensus sequence is specific for olfactory receptors, while Y6 is generally conserved in all G-protein-coupled receptor families. Noteworthy in this context that Y3 of the MAYDRY consensus is less conserved in all G-protein-coupled receptor families.

To detect the origin of replication in Archaea, we have identified ten repeats (five direct and five reverse) of a 12 bp motif (AAACCTACCACC), which displayed some similarity with DnaA boxes of E. coli, surrounding an AT-rich central region (Fig. 1a). The salient singularity point (i.e. where the tetramers skew slopes are changing) between 155 060 and 162 813 bp was all the more likely to contain oriC as the two ribosomal operons (at about 190 kb and 1490 kb) were close to it and according to oriented. Moreover cumulative GC skew at third base codon position (Fig. 1a) is in agreement with this. The only large intergenic region of this stretch was located between genes TM0151 and TM0152, which encode hypothetical proteins. In this 359 bp region (136 960–137 518), we identified ten repeats (five direct and five reverse) of a 12 bp motif (AACCTACCACC), which displayed some similarity with DnaA boxes of Escherichia coli, surrounding an AT-rich central region (Fig. 1b). Thus, this region shows the typical features of bacterial