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.

Figure 1. Position-specific cysteine codon conservation in conopeptides

(a) Sequence logo from alignment of 53 conopeptides from nine species. Note the high conservation in the signal peptide, lower in the pro region, and high variability in the inter-cysteine loops of the mature peptide. (b) Cysteine codon conservation from the conopeptide multiple alignment. The probabilities of obtaining the observed codon biases were estimated from a binomial distribution assuming a priori probabilities of 43.5% TGC versus 56.5% TGT, calculated from the codon bias tables for the five most sequenced molluscan species. p values for the cysteine codon biases are highly significant for Cys1 (p < 10^-19), Cys2 (p < 10^-5), Cys3 (p < 10^-19), Cys4 (p < 10^-14) and Cys5 (p < 10^-9), and on the borderline for Cys6 (p = 0.01). (c) Hypervariability in the immediate sequence environment of the conserved codons. Protein sequence variability at each alignment position was calculated as previously described. The solid horizontal line represents an ‘average variability’ taken from analysis of our 260 BLAST data sets (see text), and the two dashed lines represent the margins of two standard deviations in each direction. Note the extreme variability of the sequence environment, compared with the high conservation of cysteines and their codons.
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 after the cysteine codon (these nucleotides do not form part of the original BLAST query). After removal of the alignments showing an overall C or T bias in the nine neighboring nucleotide positions, a number of large gene families with specific cysteine codon conservation adjacent to an unbiased T/C environment were identified (Fig. 2). Most of the identified families are genes with hypervariable regions and, intriguingly for some of them (mucins, T-cell receptors, immunoglobulin heavy chain, and zinc-finger domain families), the conserved cysteine codon is that immediately preceding the hypervariable region. It is noteworthy that these families do not reveal different cysteine codon biases at different positions, as observed for the conotoxins.

Although global codon biases in different phyla are well documented, 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 immunoglobulins. This tendency has been suggested to act as a possible facilitator of accelerated somatic mutation. 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 conservation of the encoded amino acid residue. Although our observations do not shed light on the molecular nature of such a mechanism, they do indicate that it is likely to work at the level of specific DNA or RNA recognition and/or modification. Thus, one possibility is that specific ‘protecting’ molecules...
Gene conservation

The complete genome of the hyperthermophilic bacterium Thermotoga maritima has been recently published. Yet, its origin of replication (oriC) remains unknown, because classical approaches, such as G+C ratio, GC skew (G−C)/(G+C) (see Fig. 1a) and asymmetric distribution of oligomers along the genome, have failed to find it.

To detect the origin of replication in Archaea, we have successfully used a slightly different method, based on tetramer skew, that is, the excess of a tetramer over its reverse complement, displayed in a cumulative way. This method, when applied to the T. maritima genome, revealed a skewed distribution of the tetramer GACT (Fig. 1a). The salient singularity point (i.e., where the tetramer skew slopes are changing) between 155 060 and 162 813 bp was all the more likely to contain oriC as the main two ribosomal operons (at about 190 kb and 1490 kb) were close to it and accordingly 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 559 bp region (156 960–157 518), we identified ten repeats (five direct and five reverse) of a 12 bp motif (AAACCTACCACC), which displayed some similarity with DnaA boxes of other bacteria. Thus, the phenomenon of position-specific codon conservation might provide a functional-genomics approach to identify important residues for further structural and functional study. Moreover, this observation suggests that novel mechanisms could exist to conserve crucial residues in hypervariable regions or families.

Origin of replication of Thermotoga maritima

[References and acknowledgements]

The complete genome of the hyperthermophilic bacterium Thermotoga maritima has been recently published. Yet, its origin of replication (oriC) remains unknown, because classical approaches, such as G+C ratio, GC skew (G−C)/(G+C) (see Fig. 1a) and asymmetric distribution of oligomers along the genome, have failed to find it.

To detect the origin of replication in Archaea, we have successfully used a slightly different method, based on tetramer skew, that is, the excess of a tetramer over its reverse complement, displayed in a cumulative way. This method, when applied to the T. maritima genome, revealed a skewed distribution of the tetramer GACT (Fig. 1a). The salient singularity point (i.e., where the tetramer skew slopes are changing) between 155 060 and 162 813 bp was all the more likely to contain oriC as the main two ribosomal operons (at about 190 kb and 1490 kb) were close to it and accordingly 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 559 bp region (156 960–157 518), we identified ten repeats (five direct and five reverse) of a 12 bp motif (AAACCTACCACC), 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