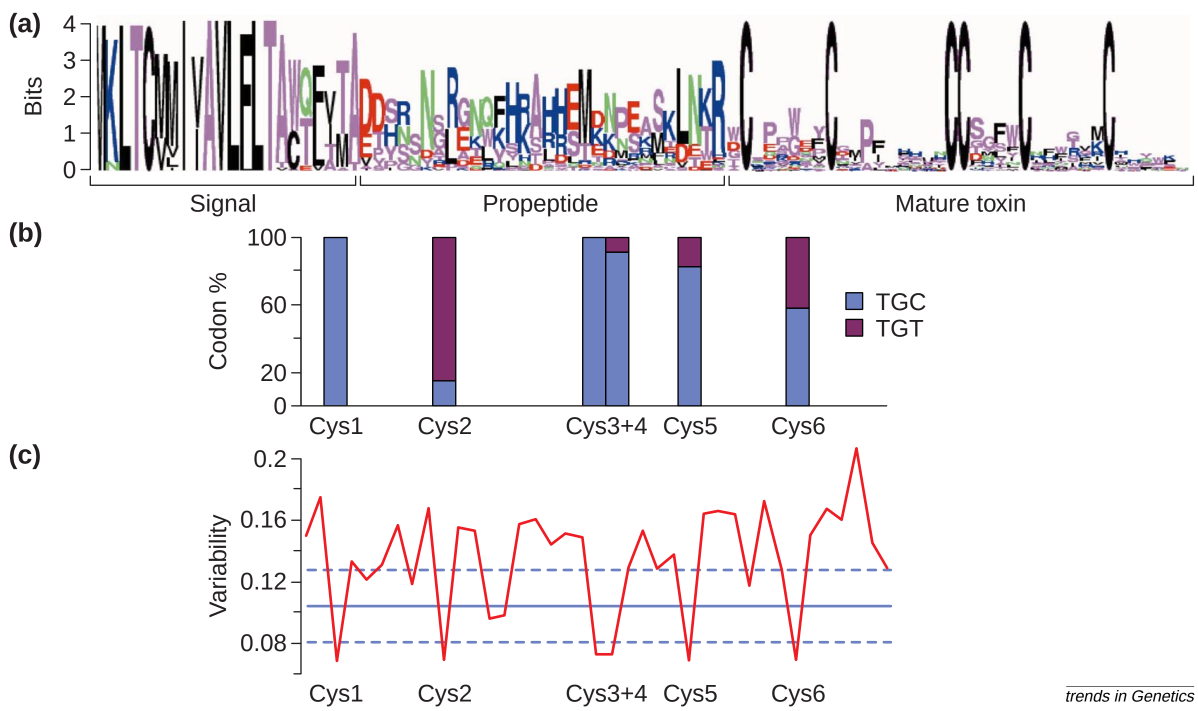
Position-specific codon conservation in hypervariable gene families

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[1]. 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[2]. 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[2], 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[12] 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[13]. 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.

Silvestro G. Conticello
silvoc@
wicc.weizmann.ac.il

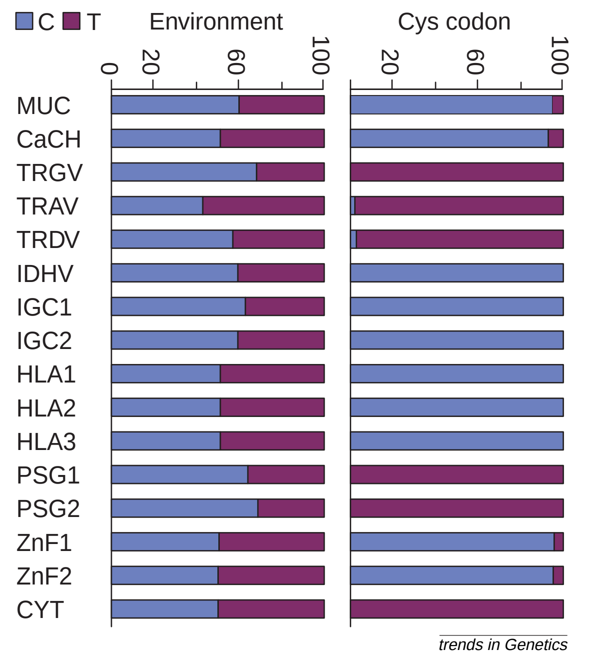
Yitzhak Pilpel*
tpilpel@
genetics.med.harvard.edu

Gustavo Glusman*
bmgustav@
bioinfo.weizmann.ac.il

Mike Fainzilber
mike.fainzilber@
weizmann.ac.il

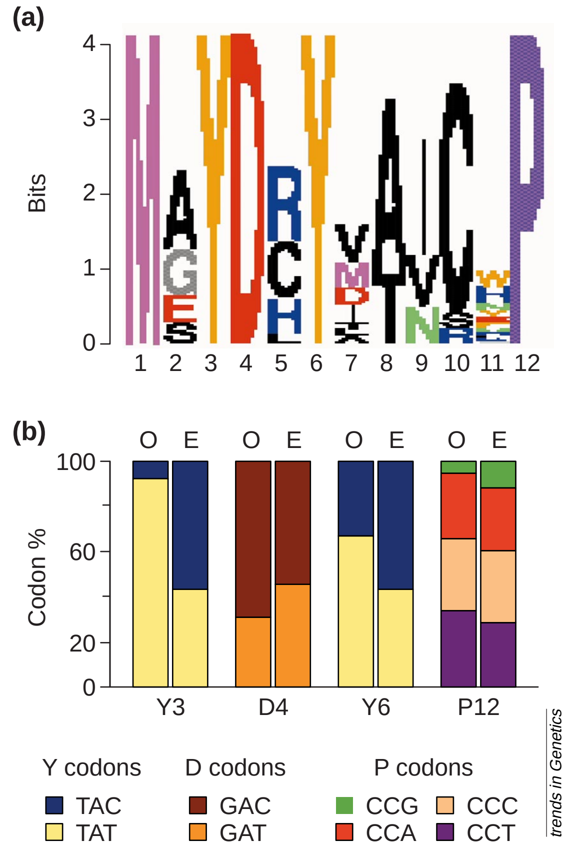
Laboratory of Molecular Neurobiology, Department of Biological Chemistry; and *The Crown Human Genome Center, Department of Molecular Genetics; Weizmann Institute of Science, 76100 Rehovot, Israel.

FIGURE 2. Cysteine codon conservation



Cys codon conservation at identified positions in a number of gene families, compared with average C:T ratios calculated over nine neighboring nucleotide positions 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: mucins (MUC); L-type calcium channels (CaCh); T-cell receptor γ CDR3 (TRGV); T-cell receptor α CDR3 (TRAV); T-cell receptor δ CDR3 (TRDV); Ig heavy chain CDR3 (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).

FIGURE 3. Tyrosine codon conservation



A position-specific tyrosine codon conservation in olfactory receptors. (a) Sequence logo in the region of the consensus Met, Ala, Tyr, Asp, Arg, Tyr (MAYDRY) derived from multiple alignment of 71 olfactory receptor DNA sequences. (b) O (observed) versus E (expected, as calculated from human codon bias table) codon usage for the conserved residues in the alignment. P value for Y_3 $p < 10^{-19}$ (calculated as above).

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[3] 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[4–7], 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 immunoglobulins[8]. This tendency has been suggested to act as a possible facilitator of accelerated somatic mutation[9,10]. 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 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 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 byproduct of the mechanism for conservation of the encoded amino acid residue. More intriguingly, one 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 underlying the position-specific codon bias described above, it might be possible to take advantage of the phenomenon to identify structurally or functionally important residues in variable 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 superfamily 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 resulting 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 from the transmembrane Met, Ala, Tyr, Asp, Arg, Tyr (MAYDRY) consensus sequence segment was found to be highly conserved (Fig. 3), whereas the codon for the second tyrosine residue in this segment was less biased. It is noteworthy in this context that Y_3 of the MAYDRY consensus is specific for olfactory receptors, while Y_6 is generally conserved in all G-protein-coupled receptor families[11]. Codons for other conserved residues in this region (D_4 and P_{12}) reveal codon usage that is close to the predicted value (Fig. 3). 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 gene families. Full data sets and supplementary information to this paper can be found at http://bioinformatics.weizmann.ac.il/papers/codon_pscc/.

Acknowledgements

This work was supported by funds from the Biotechnology Infrastructure Program of the Israeli Ministry of Science and the Crown Human Genome and Forchheimer Molecular Genetics Centers at the Weizmann Institute. We thank D. Lancet and E. Trifonov for stimulating discussions.

References

1 Norton, R.S. and Pallaghy, P.K. (1998) The cystine knot structure of ion channel toxins and related polypeptides. *Toxicon* 36, 1573–1583
2 Duda, T.F. and Palumbi, S.R. (1999) Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus. Proc. Natl. Acad. Sci. U. S. A.* 96, 6820–6823
3 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389–3402
4 Smith, J.M. and Smith, N.H. (1996) Site-specific codon bias in bacteria. *Genetics* 142, 1037–1043
5 Kliman, R.M. (1999) Recent selection on synonymous codon usage in *Drosophila. J. Mol. Evol.* 49, 343–351
6 Rodriguez-Trelles, F. *et al.* (1999) Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics* 153, 339–350
7 Duret, L. and Mouchiroud, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis. Proc. Natl. Acad. Sci. U. S. A.* 96, 4482–4487
8 Wagner, S.D. *et al.* (1995) Codon bias targets mutation. *Nature* 376, 732
9 Jolly, C.J. *et al.* (1996) The targeting of somatic hypermutation. *Semin. Immunol.* 8, 159–168
10 Kepler, T.B. (1997) Codon bias and plasticity in immunoglobulins. *Mol. Biol. Evol.* 14, 637–643
11 Ben Arie, N. *et al.* (1994) Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire. *Hum. Mol. Genet.* 3, 229–235
12 Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097–6100
13 Pilpel, Y. and Lancet, D. (1999) The variable and conserved interfaces of modeled olfactory receptor proteins. *Protein Sci.* 8, 969–977

Origin of replication of *Thermotoga maritima*

The complete genome of the hyperthermophilic bacterium *Thermotoga maritima* has been recently published[1]. Yet, its origin of replication (oriC) remains unknown, because classical approaches, such as G+C ratio, GC skew (G−C/G+C)[2] (see Fig. 1a) and asymmetric distribution of oligomers along the genome[3], have failed to find it[1].

To detect the origin of replication in Archaea[4], we have successfully used a slightly different method, based on tetramer skews, that is, the excess of a tetramer over its reverse complement, displayed in a cumulative way[5]. This method, when applied to the *T. maritima* genome, revealed a skewed distribution of the tetramer GAGT (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