Published in final edited form as:

Proteins. 2005 December 1; 61(4): 900–906.

Rational Genomics I: Antisense Open Reading Frames and Codon Bias In Short Chain Oxido Reductase Enzymes and the Evolution of the Genetic Code.

W.L. Duax¹, R. Huether¹, V. Pletnev², D. Langs¹, A. Addlagatta³, S. Connare¹, L. Habegger¹, and J. Gill¹

¹Hauptman-Woodward Medical Research Institute, Buffalo, NY 14203

²Institute of Bioorganic Chemistry RAS, Moscow, Russia

³Univ. Oregon, Eugene, OR

Abstract

The short chain oxidoreductase (SCOR) family of enzymes includes over 6000 members, extending from bacteria and archaea to humans. Nucleic acid sequence analysis reveals that significant numbers of these genes are remarkably free of stop codons in reading frames other than the coding frame, including those on the antisense strand. The genes from this subset also use almost entirely the GC-rich half of the 64 codons. Analysis of a million hypothetical genes having random nucleotide composition shows that the percentage of SCOR genes having multiple open reading frames exceeds random by a factor of as much as 1×10^6 . Nevertheless screening the content of the SWISS-PROT TrEMBL database reveals that 15% of all genes contain multiple open reading frames. The SCOR genes having multiple open reading frames and a GC-rich coding bias exhibit a similar GC bias in the nucleotide triple composition of their DNA. This bias is not correlated with the GC content of the species in which the SCOR genes are found. One possible explanation for the conservation of multiple open reading frames and extreme bias in nucleic acid composition in the family of Rossmann folds is that the *primordial* member of this family was encoded early using only very stable GC-rich DNA and that evolution proceeded with extremely limited introduction of any codons having two or more adenine or thymine nucleotides. These and other data suggest that the SCOR family of enzymes may even have diverged from a common ancestor before most of the AT-rich half of the genetic code was fully defined.

Introduction

As part of our effort to predict the fold, functions, cofactors, and substrate specificities of members of the short chain oxidoreductase (SCOR) family (1) we have been analyzing the amino acid and nucleic acid sequences of ~1600 members of the SCOR family. The SCOR family of enzymes is defined by an approximate length of 250-350 amino acid residues, a catalytic triad composed of a Ser residue, a Tyr and Lys in a YxxxK sequence, and a Gly-rich N-terminal $\beta 1\alpha 2$ -turn. Despite a common fold, cofactor binding, and oxidoreductase activity the only fully conserved residue in the family is the catalytic Tyr. In this paper we describe the nature of a codon and nucleotide bias detected in the SCOR family, and the possible implications of this bias with respect to the evolution of the genetic code.

Methods and Software

The Homology-Derived Secondary Structure of Proteins (HSSP) program of Sander and Schneider (2) was used to conduct homology searches and to identify and align all sequences in the SWISS-PROT and TrEMBL databases (3) that have at least 25% sequence identity with $3\alpha, 20\beta$ -hydroxysteroid dehydrogenase (PDB-Ac:1HDC) (4). The coding sequences from the

1593 SCOR proteins having 25% homology with 1HDC were retrieved and the codon usage and nucleotide composition of the genes examined in detail. Programs were written to examine these genes and test for the presence of “stop” codons in the five alternate ways in which each gene could be read. The five additional alternate reading frames are the two nucleotide frame shifts (+1 and +2) in the sense strand and the three possible reading frames (0,+1, +2) in the antisense strand. Programs were designed to tabulate the use of the 64 codons in each of the six reading frames and the total frequency of occurrence of the 64 possible ordered nucleotide triples in a gene. The program also analyzed codon use and nucleotide frequencies as a function of the 32 pairs of complementary codons or trinucleotides. Examples of the tables output by the program appear in deposited material. If a reading frame other than the coding frame was found to have no “stop” codons (TAA, TAG, and TGA), it was identified as an open reading frame (ORF). The majority of SCOR genes (1255 of 1593) do not have an alternate ORF (SORFs). However 254 genes are found to have a second ORF and we label these genes double open reading frames (DORFs). Furthermore 84 of the SCOR genes were found to have three ORFs. We labeled those 84 SCOR genes triple ORFs (TORFs). Graphs of the tabulated data were prepared using a Microsoft office Excel program.

Results

Statistical significance of multiple ORFs

The examination of 1593 SCOR proteins having 25% amino acid sequence identity with 3α , 20β -hydroxysteroid dehydrogenase, a representative member of the SCOR family (1), revealed that 254 of them have sense/antisense double open reading frames (SAS-DORFs), and 84 have triple open reading frames (TORFs). In all cases a DORF consists of a pair of *in-frame* totally overlapping sense/antisense ORFs. The three ORFs of a TORF are always composed of the two sense/antisense ORFs and a third ORF in the “sense” strand related to the putative coding frame by a frame shift of + 2. To test the statistical significance of the presence of DORFs and TORFs in SCOR genes we generated a million hypothetical genes of random nucleotide composition and average GC content of 65%. The only constraints we placed upon the hypothetical genes was that one frame begin with a methionine codon and end with a stop codon, and that the gene be composed by 750 randomly generated nucleotides and contain no other in frame stop codes (TAG, TGA or TAA). There are 31 possible combinations of 2, 3, 4, 5, and 6 ORFs. Only double open reading frames (DORFs) were found in the randomly generated genes, regardless of GC content. A second ORF was generated in only 1.7% of the hypothetical genes. There are five possible types of DORF in which the coding frame is paired with each of the two translational frames (2 and 3) or three possible antisense frames (4, 5 and 6). We labeled the 5 possible DORFs according to the combination of frames that are open. Frame 1 is the coding frame, frame 2 is the frame related to frame 1 by a single +1 nucleotide frame shift. Frame 3 is related to frame 1 by a -1 nucleotide frame shift. Frame 4 is the frame that is antisense to the coding frame, frame 5 is antisense frame that is shifted +1 nucleotide from the coding frame and frame 6 is the antisense frame that is shifted -1 nucleotide from the coding frame. The observed distribution of the five possible types of DORFs (1,2, 1,3, 1,4, 1,5, 1,6) in a million hypothetical genes having a GC content of 65% is 0.14, 0.14, 0.14, 1.07, and 0.11%. Ironically the DORF that does exhibit a slightly higher frequency of occurrence in randomly generated genes is the 1,5 combination and not the most commonly observed sense/antisense 1,4 combination. Only one of these six possible DORFs is observed in the SCOR family, the sense/antisense overlapping DORF (1,4) which is present in 21% of the SCOR enzymes (254 DORFs and 84 TORFs). No genes having 3 or more open reading frames are found in one million randomly generated genes, whereas 84 of the SCOR genes (5%) have triple open reading frames. The occurrence of SCOR genes having double ORFs is 150 times greater than random and the occurrence of SCOR genes having triple ORFs is at least 6×10^4 times greater than random.

GC codon bias

The existence of sense/antisense ORFs indicates that six of the 64 codons are not present in either the sense (coding) or antisense (anticoding) strand of the DNA. The six codons that are not present are the three stop codons (TAG, TAA, and TGA) of the “standard” genetic code as well as their complements (CTA, TTA, and TCA). A closer examination of the nucleic acid sequences of DORFs and TORFs reveals that these are not the only codons in these genes that are never or rarely used. Remarkably 93% of the amino acids in the 84 SCOR proteins with three open reading frames are encoded by just 32 of the codons of the genetic code (Figure 1). The thirty two codons that are used include all eight of the GC only codons (green in Fig. 1), seventeen of the GC-rich codons (blue in Fig. 1), seven AT-rich codons and none of the AT-only codons. Furthermore over 80% of the amino acids in the 84 SCOR genes having TORFs (Figure 1a) and 71% of the amino acids in the 254 SCOR genes having DORFs are encoded by the **GC rich** half of the genetic code (blue and green symbols in Figure 1).

In order to test whether the GC-rich codon bias observed in genes having DORFs and TORFs is a consequence of the GC content of the species in which the SCOR genes occur, we compared the GC content of the SCOR DORFs and TORFs with the GC content of the genus and/or species in which they are found. The 1593 SCOR enzymes are found in 280 species of bacteria (1008 genes), 13 species of archaea (62 genes), and 137 species of eukaryotes (523 genes). The 84 TORFs and 254 DORFs that exhibit a severe GC bias in nucleic acid composition are found in 311 genes from bacteria, 5-from archaea and 22-from eukaryotes. The GC content of the SCOR genes tends to be higher than the GC content of the species in which they are found. This is particularly true in species of lower GC content. SCOR genes having multiple open reading frames (MORFs) are found in 42 genes from 15 species (including three human genes) whose average GC content ranges between 46% and 59%. In all but one of these cases the GC content of the SCOR gene is from 0.1 to 18.0% higher than the average GC content of the species in which the SCOR gene is found (Table 1). A plot of the SCOR GC contents against the genomic CG contents is a scattergram with a linear R^2 value of 0.1 and a t-test probability of approximately 0.25. Clearly the biased GC rich codon use is correlated with the presence of a MORF in the SCOR genes and not the GC content of the species in which they are found.

Is there a bias in synonymous codon use?

We also examined the use of synonymous codons for all 20 amino acids in the SCOR TORFs, DORFs and SORFs. For each of the amino acids there is preferential use of the most GC-rich of the synonymous codons for that amino acid. A preference for GC-rich synonymous codons was found in SCOR genes having TORFs and DORFs from eukaryotes as well as prokaryotes. The preference is most extreme in genes having TORFs, and is only slightly less restricted in DORFs. When the use of synonymous codons in SCOR genes having single ORFs was examined the GC-bias was detected in SCOR genes from prokaryotes, little or no bias was observed in eukaryotes, and archaea had a bias favoring synonymous codons that were AT-rich. This further demonstrates that the GC codon bias is highly correlated with bacterial species MORFs and not simply the GC content of the species in which the gene is found.

Triple composition bias versus codon bias

The distribution of the 64 possible nucleotide triples is different in each of the six possible reading frames. A pattern of nucleotide triple bias in the coding frame does not mean that an identical or even similar pattern will be found in the other five frames.

For example given the sequence: -|GCA|TCG|CGA|AGC|CCT|AGG|CGT|AGG|CG-

The nucleotides in the codon frame are all GC-rich (two out of three nucleotides in each codon are G or C).

With a frame shift of +1 the nucleotide triple pattern is changed to: -G|CAT|CGC|GAA|GCC|CTA|GGC|GTA|GGC | G -

In this frame there are four AT-rich triples and 4 GC-only triples.

A further frame shift (+2) produces: -GC|ATC|GCG|AAG|CCC|TAG|GCG|TAG|GCG|-

which also has four AT-rich codons two of which are stop codons (TAG).

Analysis of the triple frequency in the SCOR genes reveals the presence in the DNA of a similar bias in which 32 GC-rich nucleotide triples account for 80% of the triples in the DNA of 84 TORFs (Figure 1b) and 73% of the triples in the DNA of 254 DORFs. The pattern of separation of occurrence of specific combinations of G,C, T and A becomes more pronounced when nucleotide triples are analyzed. Nucleotide triples composed of G and C only are more common than those composed of A and T only, and the GC-rich and AT-rich triples cluster in the upper and lower halves of the distribution, respectively. The separation of the four subpopulations of nucleotide triples (GC-only, GC-rich, AT-rich, and AT-only) that is illustrated in Figure 1b clearly reveals that the partitioning of the four subgroups of triples is more pronounced in the entire DNA than in the coding frame only. When we scramble the coding and amino acid sequence of each of the 284 SCOR DORFs by random exchange between pairs of codons a million times we find that new stops are introduced into the non-coding frames and the pattern of GC bias in triple frequency in the non coding frames is diminished. Examination of the triple content of the DNA rather than just the triple frequency in the coding frame demonstrates that the nucleotide triple bias is not restricted to the coding frame and is in fact a more fundamental property of the DNA of SCOR genes containing MORFs.

GC bias in distribution of complementary pairs of nucleotide triples

A graph of the distribution of occurrence of complementary pairs of nucleotide triples in the DNA of the 84 SCOR TORFs reveals an even more pronounced separation of the four classes of codons (GC-only, GC-rich, AT-rich, and AT-only) (Fig. 1c). Just as taking the average value of each of the four subgroups of triples (GC-only, GC-rich, AT-rich and AT-only) produces markedly different magnitudes of 8.5%, 3.9%, 1.2% and 0.2% respectively, combining the percentage use of complementary pairs of triples results in enhanced separation of the four subgroups. Of particular significance is the reduction or disappearance of overlap between the subgroups. The extreme bias in complementary triple frequency in the 84 TORFs is less pronounced in the 254 DORFs (supplemental Table 1) and further diminished in the 1255 SORFs.

Discussion

The presence of multiple open reading frames (MORFs) in so many genes of one protein family raises questions of how common the phenomenon is, whether vital information is retained on the antisense strand of the gene accounting for its conserved ORF and why so many genes having MORFs and a significant bias in nucleotide triple composition have been retained in one family of proteins through hundreds of millions of years of evolution.

How common are genes with multiple ORFs and a GC codon bias?

Upon screening a random collection of 1,089,310 genes in the SWISS-PROT and TrEMBL database, we found that 20% of them have MORFs. Furthermore, 13.5% of the human genes in the gene bank have MORFs. This demonstrates that the phenomenon is not restricted to species with uniformly high GC content. Although the genes of heat shock proteins (5), ribosomal proteins, transcription regulators, and oxido-reductases have significant numbers of MORFs (manuscript in preparation), most families of eukaryotic proteins have few if any

MORFs and do not exhibit a codon bias as severe as the GC bias found in SCOR proteins. Codon use and nucleotide triple frequency in a random collection of human genes (Fig. 2) exhibit no comparable pattern of relative use of GC-rich and AT-rich codons. Forsdyke detected a similar partitioning of the four classes of nucleotide triples (GC-only, GC-rich, AT-only, AT-rich) in the genome of the herpes simplex virus (6). However only four of the 77 genes in the herpes simplex virus have sense antisense ORFs and none have higher order MORFs. Merino reported a direct correlation between the GC content of a species and the frequency of occurrence of SAS-DORFs (7) but did not examine the genes for the presence of TORFs or codon bias.

Streptomyces coelicolor, an ancient bacteria in soil, is a source of many antibiotics containing unnatural amino acids that are synthesized nonribosomally by polyketide enzymes. It has one of the largest and most GC-rich genome thus far reported. Our analysis of the complete genome for *Streptomyces coelicolor* (7555 genes) revealed that in addition to 2819 DORFs the genome contains 2174 TORFs, 228 QORFs (quartet ORFs) and 56 PORFs (penta ORFs). All of the genes of *S. coelicolor* have a nucleic acid triple bias nearly identical to that found in the TORFs and DORFs of the SCOR family (manuscript in preparation). The wide spread existence of genes having MORFs and a correlated GC-rich trinucleotide composition that is orders of magnitude greater than random raises interesting possibilities about the evolution of families of proteins whose genes share these characteristics.

Do sense-antisense protein pairs have biological significance?

In 1994 LeJohn described the isolation and characterization of two proteins that were products of the sense and antisense strands of the same gene in *Achyla klebsiana*. Co-isolation of the two proteins using a monoclonal antibody suggested that the proteins, a 1063 amino acid eukaryotic glutamine dehydrogenase and a 600 amino acid heat shock protein (HSP-70) (8-10) formed a complex that was recognized by the antibody. Analysis of the genes responsible for production of the HSP-70 protein and the carboxy-terminal of the glutamine dehydrogenase protein containing the Rossman fold revealed a bias in the use of codons. Seventeen codons were entirely unused in the sense/antisense overlapping open reading frame and half of the 64 codons are responsible for the coding of 89% of the 1282 amino acids in the protein products of the overlapping strands. A search of the genebank found 36 other genes that encode HSP-70 proteins and have sense/antisense DORFs. BLAST searches using the amino acid sequences that are antisense to the 36 HSP-70 proteins having SASDORFs, detected homology with fragments of calcium binding proteins, calcium receptors, and glutamine dehydrogenases. These data suggest that primordial genes may have encoded calcium receptors, calcium binding proteins, and glutamine dehydrogenases on one DNA strand and HSP-70s on their antisense strand and that subsequent to gene duplication, the genes and the protein products evolved independently but retained some evidence of their origin on opposite strands of the same DNA. The structures isolated by LeJohn from *Achyla klebsiana* may be a rare case in which two proteins are still expressed by complementary antisense strands of DNA. Despite hundreds of millions of years of evolution the sense/antisense complementarity and severe codon bias has been retained. On the basis of similarity in sequence and folding detected between glutamate dehydrogenase and HSP-70, and aminoacyl tRNA synthetase from class I and class II respectively, has been suggested that the latter originated as SAS-DORFs (11). In contrast to the case of HSP-70 genes, there is no evidence of homology between hypothetical proteins antisense to the SCOR coding frames and any known protein in the gene bank.

While studying an inflammatory vascular disease caused by *antineutrophil cytoplasmic* autoantibodies with specificity for proteinase 3 (PR3). Pendergraft, *et. al* found that patients harbored not only antibodies to the autoantigen (PR-3) but also antibodies to a peptide translated from the antisense DNA strand of PR-3 (12). The authors concluded that

autoimmunity can be initiated through an immune response against a peptide that is antisense or complementary to the autoantigen, which then induces anti-ideotypic antibodies that cross react with the autoantigen (12). In light of recent discoveries of a wide variety of functional RNAs (13) it is also possible that functional RNAs may be transcribed from alternate frames of the genes having multiple ORFs.

Why has a codon bias survived million of years of evolution?

One possible explanation for the conservation of a GC codon bias in the multiple ORFs in the SCOR family is that the primordial member of the family was encoded early in evolution in a GC-rich species. It has been repeatedly demonstrated that AT-rich DNA melts at a lower temperature than GC-rich DNA (14). This is consistent with the fact that GC base pairs are stabilized by three hydrogen bonds and AT base pairs by only two hydrogen bonds. The relative stability of 32 complementary pairs of nucleotide triples in DNA will depend upon the local environment and overall conformation of the DNA. Trifonov (15) presented a table of relative energies of complementary codon pairs based upon thermodynamic parameters derived from optical melting studies of 90 oligonucleotide duplexes containing only Watson-Crick base pairs (15,16). When the relative energies are plotted versus the order of complementary triple frequencies in the 84 SCOR-TORFs, (Figure 3) qualitative agreement is observed. It is conceivable that when the genetic code was evolving the more stable GC-only and GC-rich nucleotide triples gained and retained definitions with respect to coding specific amino acids more consistently and efficiently than less stable AT-only and AT-rich nucleotide triples. Fourteen codons of the standard or universal code are known to code for different amino acids in different species including bacteria and the mitochondria of eukaryotes (17). The majority of the codons having variable definitions are AT-only or AT-rich and are rarely found in the DNA of SCOR genes having multiple open reading frames. The codons having variable definition are identified in the figure legends. The fact that the majority of the codons that are used least in coding in the multiple open reading frames are AT-rich and include those that have multiple definitions in different species is consistent with the possibility that the earliest SCOR genes evolved before the AT-rich half of the coding system was fully defined and before some species separation. Although we have not attempted to predict an exact order of introduction of codons into the genetic code, our data supports, in part, the order of codon evolution proposed by Trifonov (15,18). However our pattern of separation of populations of the four classes of codons (GC-only, GC-rich, AT-rich, and AT-only) in SCOR MORFs and the *S. coelicolor* genome suggests that the evolution of triples used for coding might have been in the order, GC-only before GC-rich before AT-rich before AT-only, with triples corresponding to complementary antisense pairs in each of these subgroups entering the coding system at approximately the same time in the course of the evolution of the code. Additional data on MORF content, nucleotide triple occurrence, and codon usage in entire genomes and families of ancient proteins is needed to substantiate hypothetical schemes for the evolution of the coding system by single step mutations such as those proposed by Trifonov. The partitioning of nucleotide triples into subgroups that are GC-only, GC-rich, AT-rich and AT-only may be more pronounced in the entire gene than in the coding frame because the composition of the coding frame is influenced by evolving biological processes that will lead to greater frequency of use of codons corresponding to functionally important amino acids despite possible lower energetic stability of the nucleotide triples that code for them.

Conclusions

An unexpected pattern of multiple open reading frames (MORFs) has been found in genes of the short chain oxidoreductase (SCOR) family in which 21% of the genes have antisense open reading frames (ORF) overlapping the entire sense strand and (25%) of these genes (5% of the total) have a third ORF related to the coding frame by a frame shift of +2. The SCOR genes

having MORFs also exhibit a severe bias in nucleotide triple distribution in which 85% of amino acid residues of those gene products are encoded by the energetically more stable GC-rich codons. Although hypothetical proteins encoded by the antisense sequence of the SCOR DORFs show less than 25% homology with known proteins, the antisense strand may encode functional RNAs or small peptides. The data suggest that SCOR enzymes originated in a GC-rich species before the entire genetic code was fully defined and that the ORFs and nucleotide triple bias has been retained for hundreds of millions of years because mutation provides little or no enhanced stability to the DNA or the protein fold and function of the SCOR enzymes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Grant sponsor: The National Institutes of Health Grant Number: DK026546

One of the authors (WLD) would like to dedicate this paper to the memory of Herbert LeJohn whose work ignited his imagination.

The technical assistance of Robert Collins, Patricia Yu, Melda Tugac, and Jean Gallmeyer is gratefully acknowledged. We thank Dr. Jane Griffin for careful reading of the manuscript and critically important advice and discussion. Lastly, we also thank one of the reviewers whose care and constructive criticism made a significant contribution to the final form.

References

1. Duax WL, Pletnev V, Addlagatta A, Bruenn J, Weeks CM. Rational Proteomics I. Fingerprint Identification and Cofactor Specificity in the Short-Chain Oxidoreductase (SCOR) Enzyme Family, PROTEINS – Structure. Function and Genetics. 2003in press
2. Sander C, Schneider R. Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment. *Proteins* 1991;9:56–68. [PubMed: 2017436]
3. Bairoch A, Apweiler R. The SWISS-PROT Protein Sequence Database and Its Supplement TrEMBL in 2000. *Nucleic Acids Res* 2000;28:45–48. [PubMed: 10592178]
4. Ghosh D, Wawrzak Z, Weeks CM, Duax WL, Erman M. The Refined Three-Dimensional Structure of 3 α ,20 β -Hydroxysteroid Dehydrogenase and Possible Roles of the Residues Conserved in “Short-Chain” Dehydrogenases. *Structure* 1994;2:629–640. [PubMed: 7922040]
5. Konstantopoulou I, Ouzounis CA, Drosopoulou E, Yiangou M, Sideeras P, Sander C, Scouras ZG. A *Drosophila hsp 70* Gene Contains Long, Antiparallel, Coupled Open Reading Frames (LAC ORFs) Conserved in Homologous Loci. *J. Mol. Evol* 1995;41:414–420. [PubMed: 7563128]
6. Forsdyke DR. Relative Roles of Primary Sequence and (G + C)% In Determining the Hierarchy of Frequencies of Complementary Trinucleotide Pairs in DNAs of Different Species. *J. Mol. Evol* 1995;41(5):573–581. [PubMed: 7490771]
7. Merino E, Balbas P, Puente JL, Bolivar F. Antisense Overlapping Open Reading Frames in Genes From Bacteria to Humans. *Nucleic Acids Res* 1994;22:1903–1908. [PubMed: 8208617]
8. LeJohn HB, Cameron LE, Yang B, Reenie SL. Molecular Characterization of an NAD-Specific Glutamate Dehydrogenase Gene Inducible by L-Glutamine. *J. Biol. Chem* 1994;269:4523–4531. [PubMed: 8308022]
9. LeJohn HB, Cameron LE, Yang B, MacBeath G, Barker DS, Williams SA. Cloning and Analysis of a Constitutive Heat Shock (Cognate) Protein 70 Gene Inducible by L-Glutamine. *J. Biol. Chem* 1994;269:4513–4522. [PubMed: 8308021]
10. Yang B, LeJohn HB. NADP⁺-Activatable, NAD⁺-Specific Glutamate Dehydrogenase. *J. Biol. Chem* 1994;269:4506–4512. [PubMed: 8308020]
11. Carter CW, Duax WL. Did tRNA Synthetase Classes Arise on Opposite Strands of the Same Gene. *Mol. Cell* 2002;10:705–708. [PubMed: 12419215]

12. Pendergraft WF III, Preston GA, Shah RR, Tropsha A, Carter CW Jr, Jennette JC, Falk RJ. Autoimmunity is Triggered by cPR-3(105-201), a Protein Complementary to Human Autoantigen Proteinase-3. *Nature Medicine* 2004;10:72–79.
13. Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, Nemzer S, Pinner E, Walach S, Bernstein J, Savitsky K, Rotmman G. Widespread Occurrence of Antisense Transcription in the Human Genome, *Nature. Biotechnology* 2003;21:379.
14. Marmur J, Doty P. Determination of the Base Composition of Deoxy tribonucleic Acid From Its Thermal Denaturation Temperature. *J. Mol. Biol* 1962;5:109–118. [PubMed: 14470099]
15. Trifonov EN. Consensus Temporal Order of Amino Acids and Evolution of the Triplet Code. *Gene* 2000;26:139–151. [PubMed: 11164045]
16. Xia T, Santa Lucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs. *Biochemistry* 1998;37:14719–14735. [PubMed: 9778347]
17. Maeshiro T, Kimura M. The Role of Robustness and Changeability on the Origin and Evolution of Genetic Codes. *Proc. Natl. Acad. Sci. USA* 1998;95:5088–5093. [PubMed: 9560233]
18. Trifonov EN, Kirzhner A, Kirzhner M, Berezovsky N. Distinct Stages of Protein Evolution as Suggested by Protein Sequence Analysis. *J. Mol. Evol* 2001;53:394–401. [PubMed: 11675599]

Figure 1.

(a) Codon use (b) nucleotide triple frequency and (c) complementary pairs of nucleotide triples in 84 SCOR genes having three open reading frames. Codons that have different definitions in different species (TTG,CTT,CTC,CTA,CTG,ATA,ATG,GTG,TAA,TAG,AAA,TGA,AGA, and AGG) are identified by X.

Figure 2.

(a) Codon use and (b) frequency of occurrence of complementary pairs of nucleotide triples in a random collection of 91 human genes (color code as in Figs. 1). For purposes of comparison 2a corresponds to 1a and 2b corresponds to 1b.

Figure 3.
The relative thermostability of the complementary codon pairs in the order of frequency of occurrence in the 84 SCOR TORFs.

Table 1

Comparison of the GC content of the SCOR genes having multiple open reading frames in species with GC content between 46% to 59%. The codon bias is a property of the SCOR genes and not the GC content of the species in which they are found. Δ - SCOR average GC content minus species GC content.

Species	# SCOR Genes	SCOR GC content		SCOR Average GC	Species GC content	Δ
		Range				
Agrobacterium tumefaciens	14	59.2 - 64.8		62.3	59.8	2.5
Brucella melitensis	7	54.5 – 63.2		59.4	58.3	1.1
Burkholderia cepacia	3	64.0 - 65.0		64.7	64.3	0.4
Drosophila melanogaster	3	59.0 - 63.0		61.2	54.0	7.2
Gallus gallus	1	70.3		70.3	51.3	19.0
Glomerella lagenarium	1	62.8		62.8	56.5	6.3
Homo sapiens	3	59.3 - 64.62		62.8	52.5	10.3
Mus musculus	1	48.1		48.1	52.3	-4.2
Neisseria meningitidis	2	57.4		57.4	53.3	4.1
Oryctolagus cuniculus	1	61.0		61.0	54.7	6.3
Thermotoga maritima	1	51.1		51.1	46.4	4.7
Xylella fastidiosa	2	65.0 - 67.1		66.6	53.7	12.9
Zea mays	1	72.4		72.4	54.6	17.8
Oryza sativa	1	73.6		73.6	55.6	18.0
Magnaporthe grisea	1	63.8		63.8	56.5	7.3